Г. Х. Гуд,
Р. Э. Макол

СИСТЕМОТЕХНИКА

ВВЕДЕНИЕ В ПРОЕКТИРОВАНИЕ БОЛЬШИХ СИСТЕМ

„Советское радио"

SYSTEM ENGINEERING

an introduction to the design of large-scale systems

HARRY H. GOOD

ROBERT E. MACHOL

MCGRAW-HILL BOOK COMPANY, INC.

New York Toronto London

1957

Г. Х. ГУД, Р. Э. МАКОЛ

СИСТЕМОТЕХНИКА

ВВЕДЕНИЕ В ПРОЕКТИРОВАНИЕ БОЛЬШИХ СИСТЕМ

ПЕРЕВОД С АНГЛИЙСКОГО
К. Н. ТРОФИМОВА, С. Е. ЖОРНО, И. В. СОЛОВЬЕВА
ПОД РЕДАКЦИЕЙ
Г. Н. ПОВАРОВА

ИЗДАТЕЛЬСТВО „СОВЕТСКОЕ РАДИО"
МОСКВА — 1962

В книге излагается общая методика проектирования сложных технических систем. Авторы рассматривают сложность как особое явление и обсуждают некоторые общие пути преодоления трудностей, связанных со сложностью проектируемых систем.

Книга носит прикладной характер. Материал излагается в доступной форме. Не требуется больших знаний по высшей математике, необходимый математический аппарат излагается попутно.

Книга может представить интерес для инженеров различных профилей, планово-экономических и научных работников, занимающихся сложными техническими проблемами.

О СИСТЕМОТЕХНИКЕ И О КНИГЕ ГУДА И МАКОЛА

ОТ РЕДАКТОРА ПЕРЕВОДА

I

Предлагаемая вниманию читателей книга двух американских ученых, сотрудников Мичиганского университета Г. Х. Гуда и Р. Э. Макола* посвящена вопросам проектирования современных крупных, сложных, высокоавтоматизированных технических систем — «систем большого масштаба», как их называют авторы книги. Комплекс этих вопросов, получивший в США наименование «системотехника» (system engineering), авторы стараются упорядочить, по возможности теоретически осмыслить и представить в целостном, последовательном изложении, хотя они и считают широкие обобщения преждевременными.

В том виде, как она разрабатывается в США и излагается в настоящей книге, «системотехника» затрагивает многие стороны того важнейшего направления техники, которое мы в СССР обычно обозначаем словами «комплексная автоматизация». Комплексная автоматизация достигла больших успехов в нашей стране, непрерывно развиваясь вширь и вглубь и стремительно увеличивая гигантскую индустриальную мощь советской державы. Для анализа наших достижений в столь важной области и подготовки новых технических побед необходимо глубокое критическое ознакомление со всем существующим мировым опытом. С этой точки зрения и следует подходить к книге Гуда и Макола о системотехнике.

Советский читатель уже знаком с переведенными на русский язык работами зарубежных ученых по двум другим новым направлениям техники — по кибернетике и по исследованию операций. Системотехника весьма близка к этим направлениям и во многом перекликается с ними.

Критическое обсуждение системотехники, касающейся разнообразных технических вопросов, использующей сложный математический аппарат и во многом еще находящейся в неустойчивом, переходном состоянии, требует, конечно, как и проблемы кибернетики или проблемы исследования операций, внимания и участия многих специалистов разных профилей. Здесь мы изложим лишь некоторые общие, предварительные соображения.

II

Системы «большого масштаба» — детище комплексной автоматизации и развития электронных вычислительных машин и в этом отношении представляют собой характерное и знаменательное явление современной техники. Они отличаются от прежних, «малых» технических систем не только количественно — обилием частей и органов, но и качественно — иным, более высоким уровнем организации, иными, более сложными функциональными взаимосвязями этих частей и органов. Правда, граница между большими и малыми системами не очень резка, и существует ряд переходных форм.

В качестве примеров систем большого масштаба или приближающихся к ним систем авторы книги называют большие транспортные системы (системы уличного движения в городах, системы железнодорожных, морских и других сообщений), системы слепой посадки самолетов (систему «Телеран» и др.),

---

* Гарри Х. Гуд — профессор электротехники и организации производства Мичиганского университета, бывший директор исследовательского центра «Уиллоу Ран» при Мичиганском университете. Роберт Э. Макол — доктор философии, научный сотрудник Инженерного исследовательского института при Мичиганском университете.

автоматические телефонные системы (в частности, американскую координатную систему № 5), системы автоматического производства электронных компонентов (систему «Тинкертой» и др.), автоматические системы обработки больших массивов коммерческих и научных данных (системы учета в крупных организациях, системы для обработки данных с радиозондов и др.) и такие системы военного назначения, как система предупреждения гражданской обороны (МПВО), система зенитных управляемых реактивных снарядов «Ника» с соответствующим наземным оборудованием, континентальная система ПВО США.

Авторы указывают следующие характерные черты систем большого масштаба:

1. Определенная целостность, или единство, системы — наличие у всей системы какой-то общей цели, общего назначения. Система должна отвечать определенными сериями оптимальных ответов («выходов») на данные множества возмущений («входов») применительно к определенному критерию эффективности.

2. Большие размеры системы. Система большого масштаба является большой и по числу частей, и по числу выполняемых функций, и по числу входов, и по своей абсолютной стоимости.

3. Сложность поведения системы — такие сложные, переплетающиеся и перекрывающиеся взаимосвязи между переменными, встречающимися в системе, что изменение одной переменной влечет изменения многих других переменных. Эта сложность проявляется также в сложных и переплетающихся петлях обратной связи в системе.

4. Высокая степень автоматизации системы. В частности, для таких больших систем весьма характерно широкое применение новейших автоматических вычислительных машин в целях гибкого управления системами и механизации умственного труда человека, работающего с системой.

5. Нерегулярное, статистически распределенное во времени поступление внешних возмущений («входов») — невозможность точного предсказания нагрузки.

6. Наличие (в большинстве случаев) состязательных, конкурирующих сторон.

Конечно, этот перечень определяющих особенностей систем большого масштаба нельзя считать каноническим; весьма возможно, что-то здесь нужно добавить или убавить. Можно было бы также подобрать иные и, пожалуй, более яркие примеры систем большого масштаба. Однако, как бы ни колебалась теоретическая граница между большими и малыми техническими системами, реальное существование больших систем и их глубокое отличие от малых не подлежат сомнению.

III

Все расширяющееся развитие систем большого масштаба — закономерный этап в истории мировой техники, результат количественного и качественного роста техники, ее глубоких, поистине революционных перестроек в XX в. В эпоху комплексной автоматизации, в которую мировая техника вступила в годы после II мировой войны, автоматизируются уже не отдельные операции и процессы, не работа отдельных устройств и машин, а целые производства, работа целых цехов и предприятий, сначала в основных функциях, а в перспективе — полностью. Этим путем и создаются большие автоматизированные системы, управляющие сложными и многочисленными вещественными, энергетическими и информационными потоками в разнородных их переплетениях и комбинациях.

Такие большие системы — системы большого масштаба — содержат обычно огромное количество разнородного рабочего, силового и измерительного оборудования и центральной и местной управляющей аппаратуры, соединенных друг с другом разветвленными, многосторонними связями для автоматического — или в основном автоматического — выполнения данного комплекса функций или целей (скажем, производственного цикла данного предприятия) в условиях сложной окружающей среды, при наличии помех и противоборствующих факторов. Во многих случаях эти системы большого масштаба настолько сложны, что различные их функциональные подсистемы сами являются системами большого масштаба.

Конечно, развитие техники обусловливается иными факторами, чем развитие живой природы, но если ограничиться внешней, образной аналогией, то, по-видимому, можно было бы отважиться на следующее сравнение. Животное царство в своем развитии прошло через целый ряд последовательных уровней возрастающей организации животных на пути от одноклеточных простейших до таких сложнейше организованных, высокодифференцированных многоклеточных, как высшие позвоночные; например, различают уровни организации: протоплазматический, клеточный, тканевый, органный и, наконец, самый высший — системный *.

* См. К. Виланд. «Биология». Изд-во иностранной литературы, 1959, стр. 201.

Техника также прошла в своем развитии через ряд последовательных уровней сложности и организации: природный камень, брошенный каким-нибудь австралопитеком в преследуемую дичь; шёлльское кремневое орудие; составное орудие из многих деталей (античность, средние века); простейшие машины античности и XVII—XVIII вв.; сложные, удивительные машины XIX в. и первых десятилетий XX в.; малые системы машин первой половины XX в. (автоматические линии и т. п.); наконец, системы большого масштаба, о которых мы здесь толкуем *. И если система большого масштаба отличается от шелльского рубила не меньше, чем собака (системный уровень организации) от амебы (протоплазматический уровень организации), то и от обычной машины, пусть даже весьма совершенной и хитроумной, система большого масштаба отличается не меньше, чем собака от какого-нибудь плоского червя (органный уровень организации).

Скажем еще раз: это только метафора. Но она показывает, о каком скачке идет речь.

IV

Разнообразие оборудования и сложность функций, свойственные системам большого масштаба, требуют особого подхода к изучению и проектированию таких систем. Приходится учитывать качественно новый характер подобных систем. Отсюда и возникает комплекс особых теоретических и практических вопросов, объединяемых в США — удачно или нет — под названием «системотехника».

Такой особый подход к системам большого масштаба еще не нашел законченного, отчетливого теоретического выражения; многое в методике проектирования больших технических систем пока открывается и осознается чисто эмпирически, интуитивно, многое только-только начинает вырисовываться. В частности, термин «системотехника» хотя и пользуется большой популярностью в американской технической литературе и в обиходе американских фирм, но не имеет четкого, общепризнанного определения и нередко наполняется весьма различным смыслом. Как пишут авторы книги, «редко удается найти хотя бы двух слушателей, согласных в понимании предмета».

У нас, в СССР, такие вопросы разрабатываются обычно под рубрикой «проблемы комплексной автоматизации», без выделения в особую область, хотя, по-видимому, такое выделение имеет определенный смысл.

Таким образом, указанные вопросы еще требуют значительного развития и обсуждения, и, в частности, встает вопрос о том, насколько целесообразно использовать для обозначения данной области существующий термин «системотехника». Не предрешая окончательных ответов, можно, однако, отметить следующее.

Как показывает практика, проектирование систем большого масштаба распадается на две достаточно ярко выраженные самостоятельные стадии: выбор и организацию функций и структуры системы в целом и выбор и проектирование физических единиц оборудования, т. е. компонентов системы. Это разделение настолько значительно, что каждая стадия фактически требует проектировщиков существенно другого профиля, существенно другой квалификации.

Первая стадия, функционально-структурная (ее, пожалуй, можно было бы назвать «макропроектированием» системы), требует главным образом специалистов нового, широкого профиля, обладающих более или менее обширными общими знаниями и широким кругозором в технике и понимающих специфику систем большого масштаба и трудности проектирования их, так что они могут охватить систему в целом и судить о ней в целом и в случае необходимости могут правильно подобрать специалистов других, более узких профилей для более глубокой проработки различных частных сторон и функций системы. Эти специалисты нового профиля, именуемые в книге «инженерами-системотехниками» (system engineers), и суть главные носители указанного общего подхода к системам большого масштаба, главные фигуранты в «системотехнике».

Вторая стадия, проектирование компонентов (так сказать, «микропроектирование» системы), требует прежде всего классических инженеров: электриков, механиков, связистов и т. д. Чтобы подчеркнуть особое положение инженеров-системотехников и их отличие от классических инженеров, авторы даже называют инженеров-системотехников не «специалистами», т. е. людьми, устремленными на частное, специальное, а «универсалистами» (generalists), т. е. людьми, устремленными на общее.

Если проектирование компонентов ведется с помощью традиционных, опирающихся на физику и химию технических наук (теория

---

* См., например, Г. В. Осипов. «Техника и общественный прогресс». Изд-во АН СССР, 1959; С. Лилли. «Автоматизация и социальный прогресс». Изд-во иностранной литературы, 1958 г. и другие книги об истории и новейшем развитии техники.

электрических машин, теория электрических аппаратов, теория проводной связи, радиотехника, техническая механика, техническая оптика и т. д.) или их новейших вариантов (атомная техника), то функционально-структурное проектирование систем большого масштаба требует, очевидно, новых технических наук и прежде всего некой общей теории систем большого масштаба, «теоретической системотехники», которая уже не будет ни частью теоретической электротехники, ни частью технической механики, а будет чем-то совсем иным, над ними и вне их.

Что касается таких более поздних технических наук функционального характера, как теория автоматического регулирования, теория дискретных автоматов, теория релейных схем, то они занимают промежуточное положение между традиционными физико-химическими теориями компонентов и общей теорией систем большого масштаба, поскольку они занимаются уже не физикой, а функциями, но не столько систем, сколько их органов. В общем они нужны в той или иной мере и проектировщикам системы *. Функционально-структурное проектирование систем большого масштаба, по-видимому, потребует, кроме «общей теоретической системотехники», также ряда дисциплин подобного среднего уровня, имеющих дело с какими-то частными аспектами больших систем, и, разумеется, будет нуждаться в соответствующем математическом аппарате, в различных и, несомненно, достаточно сложных математических методах. Такие дисциплины и методы нужно подыскать среди существующих или создать новые.

Итак, по-видимому, существует достаточно оправданная потребность в создании общей технической науки о системах большого масштаба и в создании и подыскании еще целого ряда смежных дисциплин. Как называть общую теорию таких систем — «системотехникой» или иначе, — вопрос особый, но дело, конечно, не в выборе названия, а в правильных обобщениях существующего опыта создания больших систем и в развернутых теоретических работах.

В целом же можно думать, что теория комплексной автоматизации требует гораздо более дифференцированной разработки и разбивки на большее число отделов, чем обычно до сих пор считалось.

---

* О различии между физическими и функциональными теориями в технике см. мою статью «Логика на службе автоматизации и технического прогресса» в «Вопросах философии», 1959, № 10.

V

Авторы позволяют себе лишь простейшие обобщения в области «теоретический системотехники». Главное внимание они уделяют собиранию и упорядочению имеющегося практического опыта проектирования больших систем и энциклопедическому изложению тех общих технических и математических дисциплин, которые имеют дело с анализом и разработкой важнейших частных, главным образом функциональных, аспектов систем и которые должны составлять обязательный теоретический багаж уже сегодняшнего проектировщика больших систем, нередко вынуждаемого практикой к самодеятельному, эмпирическому вторжению в теоретические terrae incognitae — «неведомые земли». В качестве таких дисциплин в книге фигурируют теория вероятностей, математическая статистика, теория игр, теория вычислительных машин и т. д.

По каждой из этих дисциплин авторы дают определенный, более или менее обширный объем основных сведений, общих положений и главных методов, который должен позволить инженеру-системотехнику ориентироваться в соответствующих сторонах и этапах проектирования систем большого масштаба и использовать в случае необходимости различную более узкую и специальную литературу. Общее описание процесса проектирования систем большого масштаба и рудиментарные общие положения системотехники, излагаемые в книге, дают читателю представление о связях и взаимодействиях всех этих сторон и изучающих их дисциплин и концентрируют внимание читателя на проблеме целостного, общего подхода к технике больших систем.

Авторы приводят также многочисленные указания и соображения по поводу организации проектных работ исходя из опыта американских проектных организаций.

Таким образом, в целом книга представляет собой своеобразное учебное пособие для проектировщиков систем большого масштаба и их руководителей: первым оно должно облегчать решение системных задач, вторым — понимание специфики этих задач. Авторы полагают — и, по-видимому, с этим можно согласиться, — что для успешного проектирования важно не только одно, но и другое.

VI

Имеет смысл остановиться несколько подробнее на перечне упомянутых выше общих дисциплин, близкое знакомство с которыми

авторы считают необходимым для системотехника: теория вероятностей, математическая статистика, теория вычислительных машин, системная логика (теория алгорифмов), теория массового обслуживания, теория игр, линейное программирование, групповая динамика, кибернетика, теория моделирования, теория информации, теория автоматического регулирования (теория следящих систем), техническая психология.

Большинство дисциплин из этого списка, конечно, не вызывает никаких возражений. Почти все они возникли или окончательно сложились уже в 40-е и 50-е годы синхронно с триумфальным шествием комплексной автоматизации. Поэтому они молоды. Исключение составляют теория вероятностей и математическая статистика, но и они получили окончательный, строгий вид чуть ли не в 30-х годах, многие важные их разделы (теория марковских процессов, теория планирования экспериментов, теория статистических решений и др.) сравнительно молоды, и в технике они стали применяться в основном тоже в 40-е и 50-е годы.

Заметим, что в создании и развитии этих научных дисциплин большую роль сыграли ученые России. Это следует иметь в виду при чтении книги, так как авторы большей частью указывают лишь имена западных ученых.

Особое место в перечне занимает кибернетика, или, как ее определил автор этого слова Н. Винер, «наука о связи и управлении в машине и животном». Авторы книги строго ограничивают себя рамками технических систем и оставляют открытым вопрос об аналогиях между техническими системами большого масштаба и большими системами в биологии, экономике и т. п. Однако они рассматривают кибернетику как ценный источник направляющих идей для системотехники. Как известно, кибернетика долгое время была предметом ожесточенных научных и идеологических споров, но в настоящее время это направление в целом оценивается нашими видными учеными положительно и пользуется большой популярностью в нашей печати.

Наконец, в списке фигурирует так называемая «групповая динамика» — направление, возникшее в недрах американской социологии, которое стремится изучать «групповое поведение» людей математическими, а именно топологическими, методами. Такое направление, естественно, вызывает серьезные возражения ввиду своих исходных идеологических установок *. В то же время представителями групповой динамики был предложен ряд любопытных математических (топологических) методов анализа сложных сетей, образуемых различными организационными структурами, и авторы настоящей книги, по-видимому, более интересуются именно этими аспектами групповой динамики, в применении к организационным схемам проектных коллективов. Математическое существо этих аналитических методов и применимость их к системотехнике, вероятно, заслуживают какого-то специального обсуждения, подобно тому как не очень давно у нас обсуждались математические методы анализа экономических явлений, использованные в так называемой «эконометрике» **. В таком критическом плане и следует читать приводимые в книге материалы о групповой динамике.

Естественно, что не все дисциплины, включаемые в теоретический багаж системотехников, имеют в нем одинаковый вес. Главными из перечисленных выше авторы считают теорию вероятностей и теорию вычислительных машин. Первая, по их мнению,— ключ ко внешнему проектированию систем большого масштаба, т. е. к правильной постановке «системной задачи», формулировке того, чего мы хотим от системы; вторая — ключ к внутреннему проектированию, т. е. к правильному решению системы, к выбору того, как заставить систему делать то, что мы хотим. Проектировщик системы должен особенно хорошо владеть этими двумя дисциплинами.

Можно спросить, нет ли еще каких-либо претендентов на место в списке «орудий» системотехники. Это конечно, сложный вопрос, и ответ на него требует более глубокого рассмотрения системотехники. Однако на некоторые дисциплины можно указать сразу. Мы имеем в виду некоторые теории, которые стали развиваться более или менее интенсивно лишь в последнее время, уже после написания книги: теорию дискретных автоматов, теорию логических машин (моделирующих такие процессы мышления, как дедукция, индукция, абстрагирование и т. д.), теорию самоорганизующихся систем и — что весьма важно — теорию надежности, которая получает сейчас одну из ведущих ролей в технических науках вообще.

---

* См. Г. Беккер и А. Босков. «Современная социологическая теория». Изд-во иностранной литературы, 1961 г. и послесловие к этой книге «Исторический материализм и современная буржуазная социология».

** См. «О применении математики в экономических исследованиях и об отношении к эконометрике» (Материалы совещания, созванного редакцией журнала «Вестник статистики»), Госстатиздат, 1959; «Применение математики в экономических исследованиях», под ред. В. С. Немчинова, Соцэкгиз, 1959 и другие публикации.

VII

Разработка и испытание систем большого масштаба требуют больших затрат денег, рабочей силы и времени. Здесь речь идет не о человеко-часах, а о человеко-годах. Пять—десять лет — вполне обычный в системотехнике срок. Сложность и разнообразие проблем, встающих при проектировании системы большого масштаба, и средств, используемых для их решения, требуют слаженной, планомерной работы специалистов многих профилей, подключающихся к проектированию на разных его этапах и выполняющих разные функции. Проектирование систем большого масштаба есть настоящая «индустрия открытий», где созданием новой системы занимаются уже не отдельные инженеры и изобретатели, как в случае малых систем, а целые проектные «мануфактуры» или даже «фабрики», если учесть широкое применение цифровых вычислительных машин и другого сложного оборудования в современных проектных организациях.

Поэтому понятно огромное значение, которое авторы придают в системотехнике коллективным, бригадным методам работы. Главным действующим лицом в книге является не отдельный проектировщик системы, а бригада проектировщиков-системотехников, состоящая из специально подготовленных инженеров-системотехников и некоторых других специалистов, полезных при «макропроектировании», и взаимодействующая с группами проектирования оборудования, группами эксплуатационных испытаний и т. д. Авторы уделяют много места обсуждению методов организации работы системотехнической бригады и проектной «фабрики» в целом, обсуждению требований к подготовке инженеров-системотехников, условий, обеспечивающих максимальную эффективность коллективной работы, и т. п. Все это, конечно, заслуживает самого пристального внимания. Во всяком случае, техника систем большого масштаба требует не только общей теории систем большого масштаба, но и особой науки о рациональной организации труда проектировщиков. Такая наука, конечно, будет нуждаться в специальном теоретическом и математическом аппарате. Групповая динамика, о которой мы уже говорили выше, и должна, по мнению авторов, служить частью этого аппарата.

VIII

Даже без особого подчеркивания авторами роли, принадлежащей теории вероятностей и теории вычислительных машин, приведенный список теоретических орудий системотехники достаточно ясно показывает, что в отличие от традиционных технических дисциплин, где ведущую роль играло *детерминированное* и *непрерывное*, в системотехнике, как и во многих более новых технических дисциплинах, огромную роль играет *стохастическое* (случайное) и *дискретное*. Теория массового обслуживания, теория игр, теория информации — все это основано на рассмотрении и расчете различных вероятностей. Цифровые вычислительные машины, системная логика, линейное программирование — здесь вещи по своей природе дискретны. Непрерывное и детерминированное преобладает еще лишь в аналоговых вычислительных машинах, теории моделирования, теории автоматического регулирования и технической психологии, да и то со значительными уступками дискретному и случайному: возьмем хотя бы современные вероятностные методы в теории автоматического регулирования.

Замена строгого, однозначного детерминизма более свободной и многозначной стохастической, вероятностной картиной связи между событиями есть также общая особенность более новых физических наук (термодинамика, квантовая физика и т. д.) по сравнению с более старыми (ньютоновская механика, классическая электродинамика и т. д.). Об этом глубоком перевороте в физике уже много писали, и причина здесь, как и в технике, ясна: от изучения простых систем и единичных явлений мы переходим к изучению сложных систем и массовых явлений, когда нам уже важен не результат отдельного события, а общий эффект основной массы событий.

Роль дискретного также выросла в современной науке, и это опять-таки является результатом существенного усложнения изучаемых естественных и создаваемых нами искусственных систем. Сложные системы состоят из разнообразных, разнородных элементов и потому обычно обладают ярко выраженной, дискретной (прерывной) структурой, сложность которой требует специальных средств изучения и специальной математики, каковы математическая логика, теория алгорифмов, современная комбинаторика, диофантов анализ и т. д. Таким же дискретным, прерывным строением должно обладать и поведение сложных систем во времени, запутанные и нередко грандиозные процессы, разыгрывающиеся в них.

Конечно, между непрерывным и дискретным и между стохастическим и детерминированным нет непроницаемой стены, они диа-

лектически переходят друг в друга и вырастают друг из друга. Тем не менее различия между ними оказывают огромное влияние на науку и технику.

Еще одна особенность теоретического аппарата системотехники заслуживает быть отмеченной. В системотехнике уделяется большое внимание возможностям человека обслуживать сложную технику. Развитие техники, облегчая в целом труд человека и увеличивая его производительность, в то же время предъявляет к человеку, теперь оператору автоматизированных установок, гораздо более высокие требования в отношении скорости реакций, точности движений, быстроты решений, собранности внимания, а следовательно, и общего морального и физического состояния. Счет нередко идет на секунды и доли секунд, ошибки или опоздания стоят весьма дорого или даже совсем фатальны (скажем, в военных или аварийных системах). Это делает необходимым тщательное изучение физиологии и психологии человека в условиях систем «человек — машина» (техническая психология).

С другой стороны, огромный размах действий организаций, применяющих системы большого масштаба на практике, требует разработки специальных научных методов, облегчающих принятие ответственных решений руководящим персоналом. Это второе направление получило название «исследование операций». В целом исследование операций в настоящее время рассматривается как самостоятельная научно-техническая область, родственная системотехнике, но отличная от нее. Исследование операций соприкасается с системотехникой при внешнем проектировании систем, когда определяются функции системы и изучаются нужды и желания потребителей. Теоретический аппарат исследования операций во многом схож с теоретическим аппаратом системотехники и включает теорию вероятностей, математическую статистику, теорию игр, теорию массового обслуживания, линейное программирование и т. д.

Углубленное изучение функций и действий человека (оператора, наблюдателя и т. д.) опять-таки весьма характерно для современной физики и в целом опять-таки есть симптом усложнения изучаемых и создаваемых систем.

IX

Сложность изучаемых и создаваемых современным человеком систем — одна из больших проблем, стоящих ныне перед наукой и техникой. Авторы книги посвящают первую главу как раз показу растущей сложности задач и решений в технике, и многие другие книги и статьи были посвящены этому вопросу.

В частности, У. Р. Эшби видел одну из заслуг кибернетики именно в том, что она предлагает некий общий подход к анализу и познанию этого феномена сложности. В своей книге о кибернетике * он уделил много места обсуждению больших и «очень больших» систем, и многое сказанное там перекликается с содержанием настоящей книги, посвященной большим техническим системам.

Конечно, даже самым большим и сложным системам далеко по своей сложности до центральной нервной системы человека с ее таким удивительным органом, как головной мозг. Однако, если даже технические системы и не достигнут уровня сложности головного мозга человека, потолок их сложности наступит нескоро, и существующая сегодня сложность систем большого масштаба, несомненно, еще значительно возрастет.

В этих сверхсистемах будущего, насколько можно сейчас судить, органы управления будут необычайно гибкими, чуткими и могущественными. Таким орудиям системотехники, как теория логических машин и теория самоорганизующихся систем, о которых мы упомянули раньше как о новейших направлениях, не охваченных в книге, суждена, несомненно, большая и важная роль в этом развитии. Несомненно также, что техника очень сложных и очень больших систем вызовет к жизни еще много новых научных дисциплин — и прикладных, и абстрактных. Научный и технический прогресс в этих областях идет сейчас с громадной и все возрастающей скоростью. Не исключена, скажем, возможность, что фантастическая биполярная математика, рисуемая профессором И. А. Ефремовым в его известном романе о будущем, с какими-нибудь странными для нас репагулярными и кохлеарными исчислениями, станет чем-то реальным в не столь отдаленном будущем.

X

Общая проблемная формулировка основных вопросов системотехники и упорядоченное энциклопедическое изложение основных ее теоретических орудий составляют главное

---

* У. Р. Эшби. «Введение в кибернетику». Изд-во иностранной литературы, 1959.

достоинство книги. Авторы суммируют американский опыт конца 50-х годов. Книга содержит описание многих американских больших систем из мирной и военной техники и опирается на опыт американских проектных организаций. Сложные вопросы целого ряда разнородных наук излагаются умело и довольно ясно. Можно думать, что ознакомление с книгой поможет нашим специалистам по разработке сложных технических систем в их творческой деятельности и в повышении культуры проектирования и будет способствовать обсуждению и разработке проблем комплексной автоматизации, теории систем большого масштаба и смежных дисциплин. Кроме того, книга может быть интересна также другим читателям, которые интересуются вопросами кибернетики, вычислительной техники, автоматизации и т. п. и нуждаются в литературе более сложной, чем популярные книжки, и более простой, чем специальная литература.

Однако, как явствует из сказанного выше, при чтении книги необходимо самостоятельное, критическое отношение ко взглядам и утверждениям авторов. Не все их советы и рекомендации одинаково приемлемы, и не со всеми их положениями можно согласиться. Предмет системотехники слишком нов и сложен, чтобы формулировка ее положений была окончательной и исчерпывающей. Здесь предстоит еще основательная работа мысли.

Далее, хотя перед нами книга по технике, она в разных местах затрагивает те или иные социально-экономические проблемы, упоминает те или иные черты американской жизни. Естественно, советский читатель будет здесь смотреть на многое совсем иначе, чем стоящие на других идеологических позициях американские авторы; взять, например, оптимистическую трактовку авторами отношений между рабочими и предпринимателями. Нельзя забывать также, что условия работы проектных организаций при социалистическом хозяйстве совсем другие, чем при капиталистическом; это относится, в частности, к экономической стороне проектирования.

Наконец, следует иметь в виду, что некоторые вопросы специальных дисциплин излагаются авторами неточно (например, в области радиолокации) или решаются большинством наших ученых иначе (субъективная концепция вероятностей); некоторые вопросы спорны, например ряд вопросов, связанных с групповой динамикой, кибернетикой и технической психологией.

XI

Перевод и редактирование такой энциклопедической книги представляли, конечно, значительную трудность. Приходилось иметь дело с крайне широким диапазоном специальных терминов, понятий и выражений. Кроме того, русская терминология во многих случаях, особенно в молодых дисциплинах, установилась еще не полностью или еще только устанавливается, используется много омонимов и синонимов, и наш выбор, возможно, был не всегда наилучшим.

У авторов книги своеобразный литературный стиль, более легкий и разговорный, чем мы привыкли ожидать от аналогичных технических книг. Такой стиль, несколько эмоциональный, с какой-то дозой юмора, довольно распространен в современной научно-технической литературе США. Русский читатель уже имел достаточно яркий его образчик в книге Дж. Д. Вильямса «Совершенный стратег» («Советское радио», 1960). Настоящая книга написана, конечно, менее гротескно, но некоторые читатели все же могут счесть те или иные места слишком экстравагантными. Здесь приходится учитывать разницу в традициях подачи материала. Впрочем, легкость формы не всегда говорит о несерьезном содержании, и, наоборот, торжественный, высокоученый «язык жрецов» может скрывать отсутствие мысли.

Мы снабдили книгу рядом примечаний, указывающих читателю на иное понимание вопроса или объясняющих отдельные подробности, термины и т. д. Кроме того, добавлен список некоторой дополнительной литературы, связанной с содержанием книги и примечаниями к ней или дающей позднейшие разработки.

28 декабря 1961 г.

Г. Н. ПОВАРОВ

ПРЕДИСЛОВИЕ АВТОРОВ

Вот уже свыше десяти лет, как инженеры и руководители предприятий стали свидетелями возникновения все более широкого подхода к проблеме проектирования технического оборудования. Это явление было плохо понято и описывалось неточно. Его называли *системотехникой* (system engineering) *, *проектированием систем, анализом систем* и часто *системным методом*. Но редко кто, произнося эти слова, имел в виду те же понятия, которые они вызывали в умах слушателей, и вряд ли найдутся хоть двое слушателей, согласных между собой по этим вопросам.

Такое несоответствие понятий нельзя считать чем-то необычным. Проектирование систем — сравнительно новое дело. Хотя широкий подход к проблемам был характерной особенностью мышления многих людей на протяжении всей истории, однако необходимость в таком подходе к проектированию комплексного оборудования возникла лишь в последнее время. Кроме того, проектирование систем вызывает к жизни много нового: новый научный аппарат, новую классификацию частей, особую организованность (хотя на первый взгляд она может показаться беспорядком) и коллективные, бригадные методы работы. Сейчас настало время соединить все это в единое целое, и в этом состоит первая честолюбивая цель настоящей книги.

Вторая цель, столь же честолюбивая, заключается в том, чтобы дать инженеру, входящему или собирающемуся войти в какую-либо бригаду проектирования системы, достаточный запас общих технических знаний, которые могли бы облегчить ему работу. Наша книга излагает методы проектирования систем; указывает надлежащее место для каждой из новых наук, ставших служанками системотехники; описывает главные проблемы, назначение и языки этих наук; дает ряд практических сведений о работе бригады проектирования системы.

Руководитель предприятия и административный работник, даже пропустив те разделы книги, которые требуют знания высшей математики, получат из остальных представление о сложных задачах, стоящих перед проектировщиками систем. Благодаря этому руководство получит возможность достигать своих целей легче и быстрее путем лучшего удовлетворения технических нужд проектировщиков систем.

Хотя в книге имеется теоретический материал, она не является сугубо теоретической. По существу все ее содержание вполне доступно для усвоения при наличии элементарных знаний по высшей математике. В тех немногих местах, где используется более сложный аппарат, читатель, не имеющий математической подготовки, должен принять на веру только один-единственный шаг в построении. Исключение составляет гл. 29, где мы были вынуждены применить дифференциальные уравнения и преобразование Лапласа, и § 7.3, где используется преобразование Фурье.

Предлагаемая вниманию читателя книга является практической еще и с другой точки зрения. Мы имели возможность обмениваться мнениями по проектированию систем со многими инженерами и учеными. Нам кажется, что в этой области абстрагирование началось слишком рано. Сейчас нередко предлагаются общие теории проектирования систем, слишком легко декларирующие аналогию и общность. В противоположность этому настоящая книга не дает никакой общей теории. Она излагает опыт — части и фрагменты и связи между ними; лишь изредка, в небольших

* Американский термин «system engineering», который мы передаем словом «системотехника», переводился у нас иногда словами «комплексная техника».— *Прим. ред.*

областях делается попытка соединить несколько фрагментов в целое с общим выводом. Однако даже в этих случаях, где можно было опираться на наличный опыт, формулировка таких общих положений очень трудна, и, высказывая их, мы старались проявить максимальную осторожность. С другой стороны, мы избегали компилирования очевидностей.

Специалистам, чьи области мы здесь обсуждаем, мы трижды приносим извинение. Первое извинение — за степень охвата областей. Дать полное введение в каждую область мы не считали ни целесообразным, ни возможным. Имелся в виду лишь достаточный запас сведений для ознакомления читателя с языком и аппаратом данной области и предоставления ему возможности прийти к разумному заключению, нужно ли ему для решения его задачи обращаться к соответствующему специалисту или же он должен сам покопаться в этой области поглубже. При этом указывается литература, необходимая для выполнения последней задачи. Наш выбор направлялся нашим опытом ученых-универсалистов, и мы вряд ли можем надеяться, что результат совпадет с суждением специалиста.

Второе извинение — за уровень строгости. К примеру, излагая теорию информации, мы не сделали оговорки о стационарности и эргодичности процессов, управляющих рассматриваемыми источниками. Однако такие тонкости отмечаются лишь специалистами для специалистов или потенциальных специалистов. По нашему мнению, это привело бы к ненужному отвлечению от главной цели книги. Если же читатель заинтересуется той или иной областью, то приводимая литература поможет ему поближе познакомиться с предметом.

Третье извинение — за библиографию. Мы не пытались перечислять всю литературу даже в более узких областях. Основные ссылки даны на книги; ссылки на периодику делались большей частью с целью выразить признательность авторам соответствующих статей. Мы старались указать в каждой области те немногие книги и статьи, которые смогут принести наибольшую пользу читателю, желающему углубиться в предмет. Эти перечни снабжены критическими аннотациями, излагающими наше мнение о ценности каждой работы. Более подробные перечни литературы можно найти в упомянутых нами книгах и статьях.

В конце некоторых глав вниманию читателей предлагаются задачи, ответы на которые в большинстве случаев приводятся в конце книги. Эти задачи были выбраны для иллюстративных целей. При использовании книги в качестве учебного пособия преподаватель должен подобрать дополнительные задачи; их можно найти в рекомендуемой нами литературе.

Остается выразить многим нашу признательность. Простой взгляд на оглавление показывает неосуществимость такой задачи. Но мы можем показать хотя бы часть кредитовой стороны нашего баланса в местах наибольшего долга. Мы особенно благодарны Джину Фелкеру и его коллегам из Белловских телефонных лабораторий за чтение всей рукописи. Следующие лица читали одну или несколько глав в пределах своей компетенции или делали замечания по большим разделам рукописи: Дж. А. Бойд, Уильям Браун, А. Брюс Кларк, Томас Коннорс, Лун Кутрона, Джон Де Тэрк и др. Некоторые не имели времени для подробного чтения; другие предложили изменения, которых мы не сделали. Возможные недостатки этой книги возникли вопреки их критике, но не по их недосмотру. Среди многих лиц, помогавших нам в механике переписки на машинке, копирования и т. п., мы особенно хотели бы выразить благодарность Виолетте Гейл и Перл Ламкин.

ЧАСТЬ I

ВВЕДЕНИЕ

ГЛАВА 1

СЛОЖНОСТЬ КАК ПРОБЛЕМА

1.1. Увеличение сложности

Сложность человеческой жизни увеличивается со все возрастающей скоростью. Если брать субъективную сторону дела, то еще двести лет назад это увеличение сложности было бы совсем незаметно за время от рождения до смерти одного человека. В наши же дни высказывания представителей старшего поколения обнаруживают их изумление перед тем, какие изменения в темпе и сложности жизни произошли с поры их детства. Даже наблюдательный молодой человек лет двадцати пяти сумеет заметить разницу в темпах жизни и числе обязанностей, которые приходится выполнять цивилизованным людям, и в частности американцам.

Если брать объективную сторону дела, то изменение темпа жизни можно грубо измерить с помощью разных статистических данных. Нас стало больше, мы взаимодействуем друг с другом более часто, мы передвигаемся с большей скоростью. Мы создаем больше вещей, больше разновидностей вещей и более сложные вещи. Некоторые из наиболее очевидных количественных характеристик этой сложности изображены на рис. 1.1—1.4. Население Земного шара (рис. 1.1) возросло с 800 млн. человек в 1750 г. до 1200 млн. в 1850 г. и 2400 млн. в 1950 г. Рост населения каждого из континентов обнаруживает такую же экспоненциальную тенденцию. Эти числа говорят, конечно, об увеличении плотности населения. Так, при численности населения Земного шара, существовавшей в 1750 г., плотность населения составляла 17 человек на одну кв. милю; в 1850 г. плотность населения достигла 24 человек на кв. милю, а в настоящее время составляет 50 человек на кв. милю *.

Плотность населения в отдельных местностях, особенно в Соединенных Штатах, гораздо выше. Это сразу видно, если посмотреть на увеличение городского населения Соединенных Штатов. В 1850 г. 17% населения Соединенных Штатов жило в городских

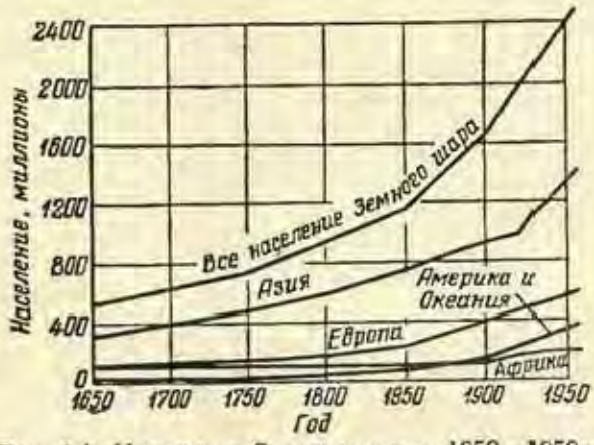


Рис. 1.1. Население Земного шара, 1650—1950 гг.

центрах; в 1875 г. эта цифра возросла до 27%, в 1900 г.— до 40%, в 1925 г.— до 50%, а в 1950 г. она составила 63%. Это относительное увеличение городского населения еще не остановилось и происходит на фоне непрерывного абсолютного роста сельского населения. Примером исключительно высокой концентрации людей может служить Рокфеллеровский центр в Нью-Йорке, где на 12,5 акра работает 32 000 служащих (не считая посетителей), что составляет более 1 500 000 человек на кв. милю **.

Но мы не только видим вокруг себя больше людей — мы ездим дальше и быстрее, чтобы встретиться с другими людьми (рис. 1.2). Максимальная скорость передвижения составляла в 1850 г. около 40 *миль/час*,

* 1 кв. миля=2,59 *км*². — *Прим. ред.*

** 1 акр=1/640 кв. мили≈0,405 *га*. — *Прим. ред.*

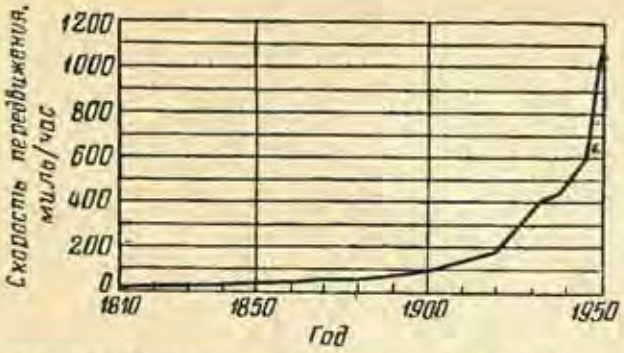


Рис. 1.2. Скорость передвижения, 1810—1950 гг.

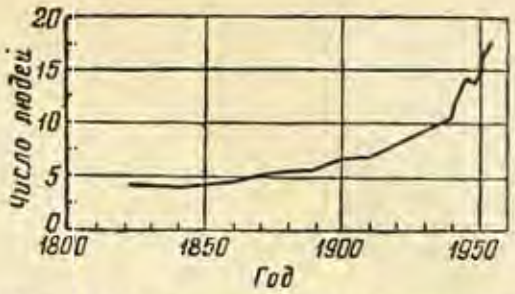


Рис. 1.4. Число людей, содержимых одним сельскохозяйственным рабочим, 1820—1950 гг.

а в 1900 — около 100 *миль/час* *. В 1950 г. экспериментальные полеты производились со скоростью 1000 *миль/час*, а скорость в 350 *миль/час* считалась обычной коммерческой скоростью; в наши дни ни у кого не возникает сомнения, что в самом ближайшем будущем столь же обычны будут скорости порядка 1000 *миль/час* **.

Если же мы не можем встретиться с нужными нам людьми лично, то посылаем им сообщения со все более возрастающей скоростью и частотой. Почта, телефон, телеграф, телетайп, радио, телевидение, фототелеграф и массовая печать обеспечивают исключительно большие возможности связи. Так, например, количество телефонных аппаратов в США (рис. 1.3) возросло с 1 350 000 в 1900 г. до 55 000 000 в 1955 г., т. е. увеличилось во много больше, чем население страны. При

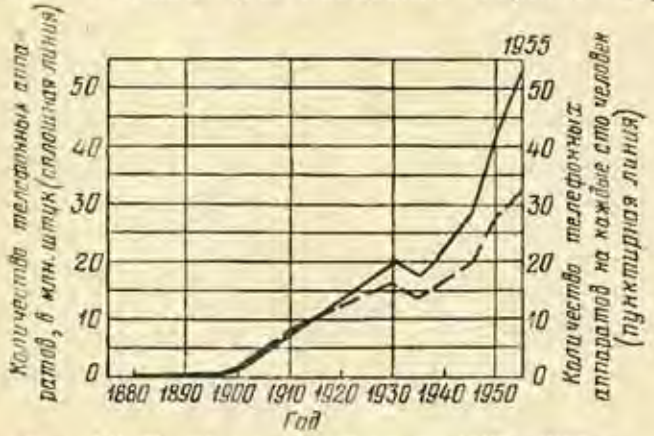


Рис. 1.3. Телефонные сообщения, 1880—1955 гг.
(указаны все действующие телефонные аппараты в США, включая добавочные телефоны к главным аппаратам).

этом плотность размещения телефонных аппаратов возросла с 1,7 до 33 аппаратов на каждые сто человек населения.

Производительность труда рабочего также увеличивается по экспоненциальному закону. Для промышленного рабочего эта закономерность хорошо известна, а данные о сельскохозяйственном рабочем приведены на рис. 1.4. В общем случае человек с машиной может производить настолько больше по сравнению с человеком без машины, что по выплате процентов и амортизации на машину и выделении разумной прибыли тому, кто вложил в машину деньги, остается еще достаточный избыток, чтобы значительно увеличить заработную плату рабочему. Непрерывный рост автоматизации и производительности труда ведет к увеличению свободного времени, бо́льшим возможностям учебы и, в конечном счете, к росту изобретений, ведущих в свою очередь к дальнейшему усложнению жизни ***.

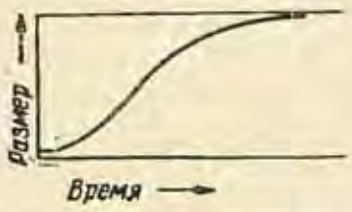


Рис. 1.5. Типичная логистическая кривая—рост, видоизменяемый ограничивающим фактором.

Все эти характеристики обнаруживают непрерывное и экспоненциальное возрастание, все эти кривые представляют собой «кривые роста». Но, конечно, в свое время вступят в действие ограничивающие факторы, а потому каждая из этих кривых должна быть частью логистической кривой **** такого вида, как на рис. 1.5. По форме кривых на

---

* 1 миля=1,609 *км*. — *Прим. ред.*

** Космические ракеты развивают скорость свыше 40 000 *км/час*=25 000 *миль/час* (вторая космическая скорость). — *Прим. ред.*

*** Вопросы о заработной плате и рабочем времени, разумеется, отнюдь не столь просты и тесно связаны с классовой борьбой. — *Прим. ред.*

**** Логистической кривой называется кривая роста, выражаемая уравнением

$$y = \frac{\lambda}{1 + \gamma e^{-\chi x}},$$

где λ, γ и χ — константы. В зарубежной научно-технической литературе логистическая кривая часто используется для описания роста различных совокупностей (клеток, людей, телефонных вызовов и т. д.), развития деловых связей между различными странами, воспитания в людях различных физических навыков, культуры и т. д., хотя применение ее основано не столько на теоретическом анализе соответствующих процессов, сколько на ее наглядности [Д. 13]. — *Прим. ред.*

рис. 1.1—1.4 нельзя с уверенностью сказать, что мы уже достигли точки перегиба. Следовательно, можно ожидать, что сложность нашей жизни будет продолжать увеличиваться с огромной скоростью.

1.2. Усилия человека справиться со сложностью

Увеличивающаяся частота встреч человека с человеком с годами сопровождалась все более и более сложным решением жизненных проблем. Например, в области транспорта последовательно появляются лошадь и телега, велосипед, железнодорожный поезд, автомобиль, самолет и вскоре, возможно, ракета *. Транспортные средства обнаруживают растущую сложность конструкции, сопровождаемую соответствующим ростом трудностей проектирования. Усложняются и средства управления движением транспорта по транспортным сетям. С увеличением скорости и усложнением конструкций безопасность сообщений все более и более зависит от этих систем управления движением. Железнодорожная сеть, автобусная сеть, перевозки на грузовиках, океанские и морские линии, управление авиатранспортом на воздушных трассах и в районе аэропортов — все это предъявляет все более сложные требования и нуждается во все более сложных решениях.

Кроме транспорта, существуют и другие большие области, в которых можно обнаружить аналогичный рост сложности решений: связь, торговля, промышленность, война, наука.

Техника телефонии за 75 лет перешла от соединения двух аппаратов непосредственно между собой к ручному коммутатору, к лабиринту таких коммутаторов и затем к панельной системе АТС; сейчас внедряются новые коммутационные системы, основанные на усовершенствованных электромеханических конструкциях, и в то же время разрабатываются электронные АТС.

Торговля перешла от карандаша и бумаги через эпоху кассовых аппаратов и арифмометров к перфорационным счетным машинам и адресографам. Сейчас создаются большие деловые системы, в которых это оборудование соединяется с новейшими средствами связи и быстродействующими электронными вычислительными машинами.

Слово «промышленность» некогда означало занятие одной семьи. Но после этапа потогонного и детского труда промышленность пришла через ручные поточные линии к механическим поточным линиям. Недавно мы были свидетелями рождения завода-автомата **, поглощающего сейчас творческие усилия многих конструкторов.

Более новое и более сложное явление представляет собой крупная промышленная организация, рассматривающая все свои заводы, склады и конторы как части одной большой системы.

Военная наука была на передней линии развития систем большого масштаба, и здесь больше чем в какой-либо другой области приходилось думать категориями целого, не ограничиваясь рассмотрением работы частей. Так, например, военным привычно иметь дело с системами противовоздушной обороны, которые охватывают всю страну, подготавливать объединенные комбинированные удары сухопутных и морских сил и координировать совместные действия тысяч бомбардировщиков.

Чтобы справиться со всем этим, наука — источник всех решений — должна была создать свои собственные системы большого масштаба, среди которых получили теперь широкую известность электронные цифровые вычислительные системы.

1.3. Характеристики системы

Такие системы большого масштаба и являются главными объектами, изучаемыми в этой книге. Однако мы не будем предпринимать никаких попыток точного определения границ, очерчивающих рассматриваемые системы. Как обычно бывает в любой области, эти границы проходят по широким неопределенным территориям и поиски их точного положения вызвали бы большие, но бесплодные споры. Гл. 2 посвящена приблизительному описанию типичных систем большого масштаба, и после чтения этой главы характер рассматриваемых нами систем должен стать ясным. Аппарат и методы настоящей книги часто оказываются полезными в случае систем, которые не являются настоящими системами большого масштаба, настоящими

* Ныне ракетный транспорт уже стал действительностью, как показывают замечательные полеты наших отечественных космических кораблей. — *Прим. ред.*

** Это иногда называют автоматизацией. Автоматизация, по-видимому, должна означать вторую промышленную революцию — замену человека машиной в функциях управления, так же как в результате первой промышленной революции машина заменила человека в качестве двигателя. Однако автоматизация означает также и много других вещей, например более целостный, комплексный подход к производству, распределению и продаже [55]. Термин «автоматизация» приобрел много нетехнических значений и поэтому нигде в этой книге не применяется — *Прим. авт.*

большими системами; однако такие применения представляют собой побочный выход по сравнению с нашей главной целью.

Отсутствие формальных определений не помешает нам отметить характеристики, которые часто бывают свойственны системам большого масштаба. Во-первых, мы имеем дело с системами оборудования, системами аппаратуры. Социальные, биологические, экологические и тому подобные системы имеют много интересных свойств, часть которых присуща и нашим системам; однако первая обязанность инженера, проектирующего системы, заключается в принятии решений по выбору оборудования, и цель нашей книги — объяснить, как делать такой выбор.

Каждая система большого масштаба отличается определенной целостностью. Хотя система может и не иметь жесткого управления из одного центрального пункта, однако во всяком случае все части системы служат некоторой общей цели; в каком-то смысле все они способствуют выработке определенного множества оптимальных выходов из данного множества входов, причем оптимальность оценивается согласно некоторому критерию эффективности.

Как мы увидим в гл. 9, это назначение системы не всегда очевидно, и во многих случаях одной из первых задач проектировщика системы должно быть четкое определение такого назначения; тем не менее единство цели существует всегда. В этом смысле современный город не является системой, хотя и может заключать в себе много систем. Будущий же город, по-видимому, может стать системой значительно большего масштаба, чем любая известная нам сейчас система (но жить в нем может оказаться менее приятным, чем в современных городах).

Такие системы являются большими — большими по числу своих частей, по числу различных типов частей, по числу выполняемых функций, по числу входов и — что может оказаться самым важным — по абсолютной стоимости. Как следствие этой большой стоимости, могут оказаться оправданными значительные затраты средств и труда на поиски лучшей конструкции системы и на повышение эффективности ее работы.

Такие системы обычно бывают сложными. Под *сложностью* мы понимаем, что изменение одной переменной будет влиять на многие другие переменные системы, и лишь редко линейным образом. Эти влияния могут быть косвенными, окольными. В гл. 10 и гл. 21 станет ясно, что до тех пор, пока эти взаимосвязи не выявлены, система изучена не полностью и что продумывание их является одной из основных задач проектировщика системы.

Другим признаком сложности служат множественные петли обратной связи и петли обратной связи в петлях обратной связи. Инженер по следящим системам хорошо знает, что система, содержащая много петель обратной связи, поддается анализу с чрезвычайным трудом. У такого инженера понятие обратной связи не совпадает с нашим: он делает упор на обратные связи по передаче энергии, тогда как мы делаем упор на обратной связи по информации. Однако, как мы увидим в § 25.3 и гл. 28, эти понятия не столь различны, как может показаться с первого взгляда; как в том, так и в другом случае мы могли бы пользоваться словом «сигнал».

Рассматриваемые в этой книге системы являются автоматическими. Степень автоматизма, конечно, различна (для различных систем), но никогда не достигает того низшего предельного значения, когда все функции управления осуществляются людьми, и, по-видимому, никогда не достигнет второго предельного значения, при котором не потребуется никакого участия людей. Поэтому в рассматриваемых системах мы всегда найдем вычислительные устройства, и их изучение составит важный отдел системотехники. С другой стороны, не менее важной задачей является определение взаимосвязей между машиной и оператором-человеком, и мы обязаны уделить некоторое время (гл. 30) технической психологии.

Входы* системы почти всегда будут распределены так, что их можно описать только статистически. Входы могут быть распределены во времени (телефонные вызовы). Это приводит к невозможности предсказать точные нагрузки и к альтернативным методам определения пиковых нагрузок. Кроме того, входы бывают множественными и обычно не-

---

* Термин «вход» (input) обозначает в этой книге не только и не столько точку воздействия на систему извне (то, что обычно понимается под «входом» в русской технической литературе), сколько само это внешнее воздействие (то, что обычно называют у нас «входным фактором», «входным возмущением», «входным сигналом»). Мы сохранили при переводе это обобщенное значение термина «вход», чтобы лучше передать строй мысли авторов. Подобно этому, термин «выход» (output) означает не только и не столько точку воздействия системы вовне (то, что обычно понимают под «выходом» в русской технической литературе), сколько само это воздействие системы на внешний мир (то, что обычно называют у нас «выходным фактором», «выходным действием», «выходным сигналом»). — *Прим. ред.*

скольких различных типов; операции, которым они подвергаются в системе, также являются множественными, осуществляясь при этом и последовательно, и параллельно.

Наконец, в наиболее сложных системах имеются также и состязающиеся, соперничающие стороны. Иными словами, некоторая разумная сила (враг) пытается уничтожить или уменьшить эффективность системы.

Большие системы отличаются от простых систем как в количественном, так и качественном отношениях. Изобретенный Беллом телефон был сложным устройством, и последовательные усовершенствования сделали его еще более сложным. Но лишь многократное повторение, необходимость соединения друг с другом 55 000 000 телефонных аппаратов, привело к созданию большой системы. Непосредственное попарное соединение частей становится неосуществимым даже при небольшом числе абонентов. В соответствии с тем обстоятельством, что одновременно необходимо будет осуществлять только небольшую часть всех возможных соединений, требуется какая-то переключающая, коммутационная система. Это приводит к потрясающе сложной системе, обладающей в то же время гигантскими размерами,— что и является системой большого масштаба.

Каковы компоненты такой большой системы? Тот, кто проектирует катушку индуктивности, считает компонентом кусок провода; инженер, проектирующий реле, считает компонентом катушку; инженер, конструирующий переключатель,— реле; проектировщик телефонного линейного искателя — переключатель; конструктор автоматической телефонной станции — линейный искатель; инженер, проектирующий телефонную систему,— телефонную станцию. Совершенно очевидно, что понятие «система» определяется точкой зрения человека, использующего это слово, и проектировщик больших систем должен обладать широким кругозором.

1.4. Развитие бригадного метода

Рост сложности наших задач и создаваемых для их решения систем приводит к появлению все большего числа научных работников как ввиду разнообразия их интересов, так и в силу общественных потребностей. С другой стороны, увеличение научного персонала способствует дальнейшему росту сложности посредством изобретений и организации.

В старину ученый имел всеобъемлющие интересы. Однако по мере роста своих знаний он изучал детальнее уже меньшие области. Еще в эпоху Возрождения такой человек, как Леонардо да Винчи, мог стать знаменитым и в науке, и в искусстве. Позже подобная широта интересов стала трудной и ученый стал посвящать все свое время изучению своей собственной специальности — науки. Еще позже естественные науки (отделившиеся от общественных наук) превратились сначала в занятие, требующее всего времени ученого, а затем в занятие, требующее еще больше времени, что привело к таким специальностям, как физические и биологические науки. Но и эти науки испытали дальнейшую специализацию.

Примерно около 1800 г. появились химик и физик как особые существа. Математика стала специальностью около 1700 г., а инженерное дело (мы говорим о гражданских инженерах в отличие от военных) — около 1850 г. Вскоре инженерное дело распалось на ряд частных специальностей (строительная техника, механика, электротехника), как распалась химия (на органическую, физическую и аналитическую) и физика (на оптику, механику, электричество).

И, наконец, XX столетие стало свидетелем расцвета еще далее зашедшей специализации: появилась инфракрасная спектроскопия, химия крови, техника следящих систем. Но эти «осколки» отнюдь не являются занятием лишь какого-нибудь одного специалиста. Они составляют целые области исследования. Имеются тысячи инженеров по следящим системам, и они могут специализироваться по электромеханическим следящим системам, или по нелинейным электромеханическим следящим системам, или по устройствам измерения ошибок в нелинейных электромеханических следящих системах.

Поскольку в разных областях применяется различный аппарат, различные методы и различный язык, эти специалисты при общении между собой очень часто испытывают большие трудности. И вот вырастают гибридные науки. Физическая химия была одной из первых; теория информации, перекинувшая мост между теорией вероятности и техникой связи, является одной из последних. Но эти гибридные науки решили только небольшую часть задачи. Связывая науки между собой лишь попарно, они не дают координированного подхода к задачам, затрагивающим многие науки. Для создания такой сложной системы, как телефонная, требуются специалисты десятков профилей: прежде всего инженеры всякого рода — электрики, механики, строители, химики; далее — математи-

ки, геологи, экономисты, психологи, архитекторы.

Раньше проектирование больших систем было повторением в расширенном масштабе проектирования малых систем. Это, естественно, приводило к детальной спецификации заказчиком компонентов большой системы. Формулировка требований к компонентам при незнании тех конструктивных трудностей, с которыми проектировщик компонентов мог столкнуться при попытке выполнить спецификации, приводила к непредвиденным осложнениям, когда доставленные компоненты начинали соединять друг с другом. Из-за обнаруженных при этом несоответствий приходилось либо переконструировать компоненты, либо вводить новые, соединительные компоненты. Чем больше и сложнее система, тем больше и сложнее (и относительно, и абсолютно) становилось такое перепроектирование или введение новых соединительных компонентов. Соединительные компоненты бывали больше тех компонентов, которые они должны были соединять.

Возникла необходимость в новой идее и новом методе. С инженерной точки зрения эта идея заключается в формировании ученого-техниковеда, т. е. научно образованного универсалиста с широким кругозором*. Метод же заключается в коллективной, бригадной работе. Над проблемами больших систем работают целые бригады ученых и инженеров как универсалистов, так и специалистов, и совместными усилиями они стараются найти решение и осуществить его физически.

Таким образом, мы пришли к идее бригады проектирования системы — небольшой группы инженеров и ученых, которые должны вести большой проект и организовать работу по созданию системы. Такие люди назывались по-разному: *техниковедами, инженерами-системотехниками, системными аналитиками, проектировщиками систем большого масштаба.* Описанный метод называли *системным методом* или *методом бригадного проектирования.*

---

* Словами «ученый-техниковед» мы передаем американский термин «engineering scientist». К сожалению, у нас не сложилось общепринятого краткого наименования для представителей технических наук, хотя надобность в этом, как можно думать, существует. Слово «техник» прочно означает специалиста со средним техническим образованием, слово «технолог» — специалиста по процессам обработки или переработки (что соответствует американскому термину «process engineer»). Словом «техниковед», в отличие от слова «технолог», мы и хотим обозначить представителя технических наук в широком смысле, хотя это, может быть, и не самый удачный выбор. — *Прим. ред.*

Этому проектировщику и его товарищам по бригаде и адресуется наша книга. Поскольку ясно, что он не может приобрести новые знания во всех нужных областях, чтобы стать специалистом в каждой из них, то мы даем материал лишь в таком объеме, чтобы он мог понимать язык и проблемы специалиста. Такой универсалист — новая величина в техническом мире, и следует заняться вопросом его подготовки.

1.5. Построение книги

Существуют четыре различные основы для построения книги по системотехнике. Первая из них — хронологические *фазы*, через которые проходит работа по проектированию системы, каковы, например, организационная фаза и предварительный анализ. Вторая — логические *этапы* проектирования, каковы, например, анализ единичной нити (операции над одним входом) и большой нагрузки (методы обслуживания множественных входов). Третья — *части* системы, каковы, например, устройства связи и устройства индикации. Четвертая — *орудия* проектирования систем, каковы, например, теория информации и теория массового обслуживания.

Гл. 3 построена по фазам проектирования; однако, как мы убедимся при чтении этой главы, каждая фаза повторяет предыдущую, имея целью осуществить те же этапы проектирования, но на более разработанной базе. Таким образом, хронология не является подходящей основой для построения книги, и принятое нами построение базируется на трех других основах.

Так как эти три основы перекрываются в некоторой мере, хотя и не полностью, то может показаться, что материал излагается непоследовательно. Например, такое важное и мощное орудие, как теория массового обслуживания, требует себе целой главы независимо от того, рассматривать ли ее как орудие или как этап (анализ большой нагрузки); однако это орудие и этот этап применяются ко многим частям системы. С другой стороны, столь же важное орудие — теория вероятностей — применяется главным образом к одной части системы (устройствам связи), но не используется при рассмотрении этапов проектирования.

Наконец, существуют определенные этапы проектирования, такие, как построение математической модели, которые не связаны ни с одной частью системы и к которым не применяется никакого специального научного

орудия*; существуют определенные части, такие, как средства передвижения или подачи материала, для которых нельзя указать специальных этапов проектирования или специальных научных орудий среди этапов и орудий, рассматриваемых в этой книге; существуют специальные орудия, такие, как кибернетика и линейное программирование, которые нельзя в настоящее время связать с каким-либо этапом проектирования или какой-либо частью системы, хотя мы уверены в том, что не пройдет и десяти лет, как это станет возможно. Таким образом, три нити — этапы, орудия и части — тянутся через всю книгу, и мы заранее извиняемся перед читателем, если ему покажется, что при описании важных вопросов системотехники мы перебрасываемся с одной из них на другую.

В гл. 2 даются краткие описания нескольких систем большого масштаба. Этим преследуется двоякая цель: пояснить на примерах, какого рода объекты рассматриваются в книге, и дать материал, необходимый для понимания приведенных по всей книге конкретных иллюстративных примеров. В гл. 3 говорится о комплексном подходе к процессу проектирования системы и рассматривается взаимосвязь между частями системы, этапами проектирования, научными орудиями и фазами.

Гл. 4—8 посвящены изложению основного аппарата теории вероятностей; при этом мы подходим к индуктивной теории математической статистики, но не углубляемся в нее. В гл. 9—13 обсуждается существо проблемы проектирования системы и методика исследования входов системы. Основными дисциплинами, изучаемыми в этой части книги, являются исследование операций и математическая статистика. Гл. 14—20 вводят еще одно основное орудие — теорию вычислений, рассматриваемую с широкой точки зрения, включая системы счисления, аналоговые и цифровые машины и входные и выходные устройства.

В гл. 21 дается дальнейшее развитие вопросов взаимосвязей между частями системы, этапами проектирования и научными орудиями; эта глава является также вступительной к последующим главам, в которых рассматривается начало решения проблемы системного проектирования. В гл. 22—24 на первый план выдвигается принцип разделения логики системы, большой нагрузки и состязательных аспектов системы. В гл. 25—30 рассматриваются этапы проектирования, связанные с выбором оборудования. В каждой из глав 22—30 рассматривается одно из орудий проектирования систем (за исключением гл. 27, в которой рассматриваются части системы).

Наконец, гл. 31 завершает картину работы проектировочной бригады и касается кратко вопросов экономики и вопросов испытания и оценки созданной системы. На этой стадии работы проектировщик системы обычно завершает свои исследования пересмотром задачи с целью изучить возможности дальнейших улучшений.

ЛИТЕРАТУРА

По этим вопросам нет хороших книг. «Кибернетика» Винера [53], подробнее обсуждаемая в гл. 25, очень важна, но его взгляды существенно отличаются от точки зрения нашей книги. Несколько интересных статей опубликовано в периодической печати. Хант [54] дал хорошее популярное обсуждение вопросов системотехники, с конкретными ссылками на достижения одной группы, хотя он и слишком ее восхваляет. Драккер [55] рассматривает проблемы автоматизации с позиций, очень близких к точке зрения настоящей книги, несмотря на то, что ни разу не пользуется термином «проектирование систем» или «системотехника».

---

* Можно утверждать, что исследование операций представляет собой специальное научное орудие для построения математических моделей. Однако мы не считаем исследование операций вполне определенным орудием в том же смысле, в каком является орудием теория массового обслуживания. В самом деле, исследователь операций может рассматривать теорию массового обслуживания как вспомогательную часть исследования операций. — *Прим. авт.*

ГЛАВА 2

ПРИМЕРЫ СИСТЕМ БОЛЬШОГО МАСШТАБА

Цель этой главы состоит в пояснении основного содержания книги с помощью конкретных примеров. Эти примеры вместе с характеристиками, перечисленными в предыдущей главе, должны служить определением того, что мы понимаем под системами большого масштаба. Некоторые из рассматриваемых в этой главе систем нельзя полностью отнести к системам большого масштаба; другие достаточно велики, чтобы поразить даже

самое большое воображение. В ряде случаев мы просто указываем на существование проблемы, в других случаях предлагаются решения или описывается их осуществление.

Нам хочется, чтобы читатель обратил внимание на огромное количество и разнообразие проблем, связанных с уже существующими системами. Чтобы не увеличивать чрезмерно объема книги, мы не могли большинство из этих проблем рассмотреть достаточно подробно, поэтому описанию каждой из них мы можем посвятить только от нескольких строк до нескольких страниц. В результате при попытке вникнуть в детали могут возникнуть неясности. Однако наша задача состоит в том, чтобы дать читателю представление о широте и актуальности проблемы и о научных орудиях, которыми можно воспользоваться.

2.1. Транспорт

Наиболее известными проблемами большого масштаба в области железнодорожного транспорта являются диспетчеризация, маневровая работа и сигнализация; эти проблемы взаимосвязаны, и для их решения уже применяется в широких масштабах автоматическое оборудование.

Аналогичные проблемы возникают в системах метрополитена, которые в таких крупных городах, как Нью-Йорк, становятся весьма сложными. Значительный интерес представляет также рассмотрение оборота грузовых железнодорожных вагонов; они смешиваются и рассеиваются по всей стране, как это видно, например, по любому товарному поезду, и эффективное использование их и возможно быстрый возврат в надлежащие пункты дают большой денежный выигрыш. Проблемы морских перевозок будут рассмотрены кратко в § 2.6. Сейчас же мы остановимся на проблемах авиационного и автомобильного сообщения.

Проблематика автомобильных сообщений включает такие частные задачи, как регулирование и диспетчеризация потоков грузовиков и автобусов, а также общие вопросы внутригородских и междугородных сообщений. Мы опишем один подход к проблеме внутригородского движения, связанный лишь с одной ее стороной, а именно с управлением сигналами светофоров, опуская значение таких других вопросов, как выбор мест стоянки автомашин, введение одностороннего движения, ограничение переходов через улицы, запрещение доставки грузов в часы пик, или таких радикальных шагов, как реконструкция магистральных улиц.

Системы слепой посадки. Существуют два основных метода управления посадкой самолетов в условиях плохой видимости: 1) наземное управление посадкой; 2) система посадки по приборам. При наземном управлении посадкой [136] самолет наблюдается одной или несколькими наземными радиолокационными станциями (при этом одна из станций может быть двухкоординатной и использоваться для определения положения самолетов в горизонтальной плоскости, а вторая станция может применяться для определения высоты; кроме того, каждой из этих основных станций может придаваться станция для наблюдения за сравнительно удаленными самолетами и одна или более станций для сопровождения самолетов в ближней зоне). Операторы наземных станций следят за положением самолета, заходящего на посадку, и, если он отклонился от требуемого курса, по радиотелефону передают на самолет корректирующую команду.

В системе посадки по приборам [137] применяется радиомаяк с двумя веерообразными диаграммами излучения, направленными вдоль линии снижения самолета. Одна из этих диаграмм расположена в вертикальной плоскости; если летчик удерживает самолет в пределах этой диаграммы, то самолет заходит на посадку правильным курсом, т. е. идет точно по направлению взлетно-посадочной дорожки. Второй радиолуч направлен перпендикулярно первому лучу и имеет небольшой угол наклона относительно горизонтальной плоскости. Удерживая самолет в зоне этого луча, летчик обеспечивает снижение самолета по глиссаде, т. е. снижается с нужной скоростью. На приборной доске в кабине самолета расположен двухстрелочный прибор; положение каждой из стрелок прибора указывает летчику положение самолета относительно соответствующего радиолуча наземной станции. Таким образом, чтобы вывести самолет требуемым курсом к началу взлетно-посадочной дорожки, летчик в момент снижения самолета должен вести самолет курсом, при котором обе стрелки будут удерживаться в центральном положении.

По многим соображениям представляется целесообразным разработать автоматическую систему слепой посадки самолетов. Основной вопрос, на который необходимо при этом ответить, заключается в том, создавать ли автоматическую систему наземного управления посадкой или автоматическую систему посадок по приборам. Иными словами, необходимо решить, где, на земле или в воздухе, должна осуществляться основная функция управ-

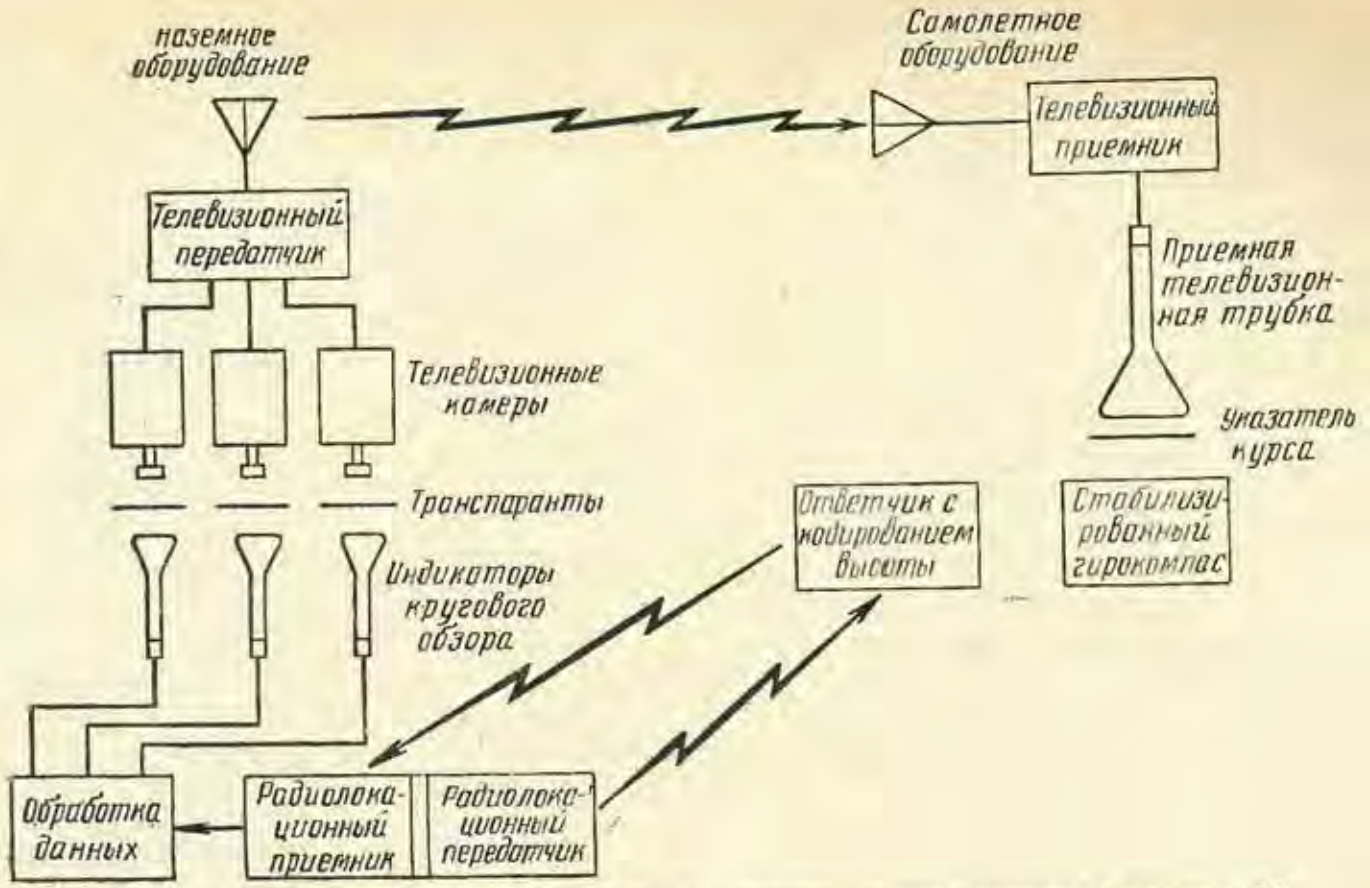


Рис. 2.1. Основные элементы системы ближней телевизионно-радиолокационной навигации [4].

ления — сравнение положения самолета с положением луча радиомаяка или радиолокационной станции и выбор корректирующих действий.

Так как этой аппаратурой можно оборудовать тысячи аэропортов и сотни тысяч самолетов, стоимость аппаратуры может дойти до миллиардов долларов, что заставляет чрезвычайно серьезно отнестись к решению. Так как оборудование всех аэропортов и самолетов должно быть сходным и обеспечивать совместную работу и так как необходимость в такой системе чувствуется крайне остро, решение должно быть принято в кратчайший срок. Совершенно очевидно, что при проектировании такой сложной системы возникает много других проблем; однако мы привели только одну проблему в качестве типичного примера такого решения по системе большого масштаба, которое нельзя легко достичь путем компромисса.

Система ближней телевизионно-радиолокационной навигации [4]. Система ближней телевизионно-радиолокационной навигации («Телеран»*) (рис. 2.1) была предложена в 1946 г. фирмой «Радио корпорейшен оф Америка» как средство управления полетами самолетов. Оборудование этой системы состоит из наземной и самолетной аппаратуры; при ее проектировании учитывалось, что в воздухе всегда могут находиться частные самолеты, не оборудованные аппаратурой этой системы, но что все самолеты гражданской коммерческой авиации будут оснащены такой аппаратурой.

Бортовая аппаратура этой системы состоит из радиолокационного ответчика (маяка) и телевизионного приемника, а также специальной индикаторной аппаратуры. Наземная аппаратура включает радиолокационную станцию, телевизионный передатчик и соответствующую аппаратуру обработки данных. Радиолокационная станция должна обеспечивать наблюдение за всеми самолетами, находящимися в зоне ее действия, получая мощные ответные сигналы от самолетов с ответчиками и обычные отраженные сигналы от всех других самолетов.

Сигналы самолетных ответчиков должны кодироваться, с тем чтобы наряду с другими данными получать также данные о высоте полета самолета. Вычислительное устройство наземного оборудования должно рассортировывать ответные сигналы в соответствии с высотой полета самолета, от которого получен

* Название «Телеран» (teleran) — сокращение английских слов «television radar navigation». — *Прим. ред.*

ответ, и посылать каждую изовысотную группу сигналов в соответствующее индикаторное устройство. Экраны этих индикаторов должны фотографироваться вместе с надлежащими транспарантами, и получаемая картина — передаваться по телевизионному каналу на все самолеты, летящие на соответствующей высоте. Таким образом, летчик каждого самолета мог бы наблюдать картину, содержащую положения всех других самолетов, летящих на той же или на близкой высоте, но не самолетов, летящих на других высотах.

Положение самолета, принимающего это изображение, отмечалось бы на этой картине специальным знаком. Применение транспарантов имеет целью: передачу топографической информации низколетящим самолетам; указание местоположения аэродромов и радиомаяков всем самолетам; передачу метеорологических данных тем самолетам, которые запросят их; передачу других данных, например частот связи, установок высотомера по высоте аэродрома и т. п., смотря по необходимости. Кроме того, были приняты меры для обеспечения «автоматического полета» при помощи двух расположенных на приборной доске приборов, указывающих азимут и дальность до той станции телевизионно-радиолокационной навигации, на частоту которой настроена аппаратура самолета.

Наземные станции предполагалось иметь нескольких типов. Станция одного типа (для управления движением по авиатрассам) должна иметь дальность действия 50 миль. В районах страны с интенсивным воздушным движением (между Бостоном и Вашингтоном) эти станции должны обеспечивать сплошное перекрытие территории, в других районах страны они могли бы обеспечивать перекрытие только авиационных трасс.

Станция второго типа (для управления воздушным движением в зоне аэропорта) должна иметь дальность действия около 20 миль. С этой станции должны передаваться телевизионным методом диаграммы маршрутов выхода на аэродром и специальные инструкции по посадке.

Станцию третьего типа (для управления посадкой самолетов) предполагалось устанавливать на некоторых определенных аэродромах. Она подобна системе посадки по приборам: на экране самолетного телевизионного приемника предполагалось показывать две стрелки, которые при правильном снижении самолета должны были совпадать с отметкой, изображающей приземляющийся самолет.

В 1947 г. была закончена разработкой и демонстрировалась модель единичной нити этой системы. В ее состав входили радиолокационные станции дальнего обнаружения, аэродромного обзора и захода на посадку; в состав самолетного оборудования входил радиолокационный ответчик. Аппаратура обработки данных обеспечивала сортировку принимаемой информации и распределение ее в соответствии с высотами полетов самолетов на три группы (предполагалось, что аппаратура в окончательном виде должна обеспечивать восемь или более градаций высоты). В состав этой опытной аппаратуры входили также наземные телевизионные передатчики и самолетные телевизионные приемники. Некоторые предполагаемые функции системы, например автоматическое управление посадкой самолетов, этим демонстрационным образцом не выполнялись.

Системы резервации мест. Проблема резервации мест на самолеты и железнодорожные поезда является весьма серьезной. Пассажиры делают много предварительных заказов на билеты и затем изменяют их весьма существенно, причем много заказов поступает или изменяется в самые последние минуты. Однако, несмотря на такое изменение заказов, должен вестись строгий учет проданных билетов и свободных мест. С одной стороны, разница между 90 и 100% проданных мест может стать разницей между прибылью и убытком для всей системы; с другой стороны, 101% проданных мест приведет к большому недовольству пассажиров, в результате чего будут потери клиентуры, а правительство по требованию пассажиров, возможно, окажет определенное давление. Наиболее известными действующими автоматическими системами для регистрации заказанных и купленных билетов являются системы, используемые на Пенсильванской железной дороге и авиакомпанией «Американ эйрлайнс».

У компании «Американ эйрлайнс» с 1945 г. (в Бостоне) работает полуавтоматическая система резервации. В 1952 г. эта компания ввела в действие в Нью-Йорке автоматическую систему, получившую название «Резервизор» [3] (рис. 2.2). Основой этой системы является небольшая специализированная цифровая вычислительная машина, в которой в качестве запоминающего устройства используется магнитный барабан. На магнитном барабане записываются данные о свободных местах (в количестве до 127) для каждого из 10 000 различных участков рейсов в течение ближайших 10 дней плюс любых двух дальнейших дней (такими днями могут быть, например, праздничные дни, когда количество пассажиров резко возрастает). Каждый

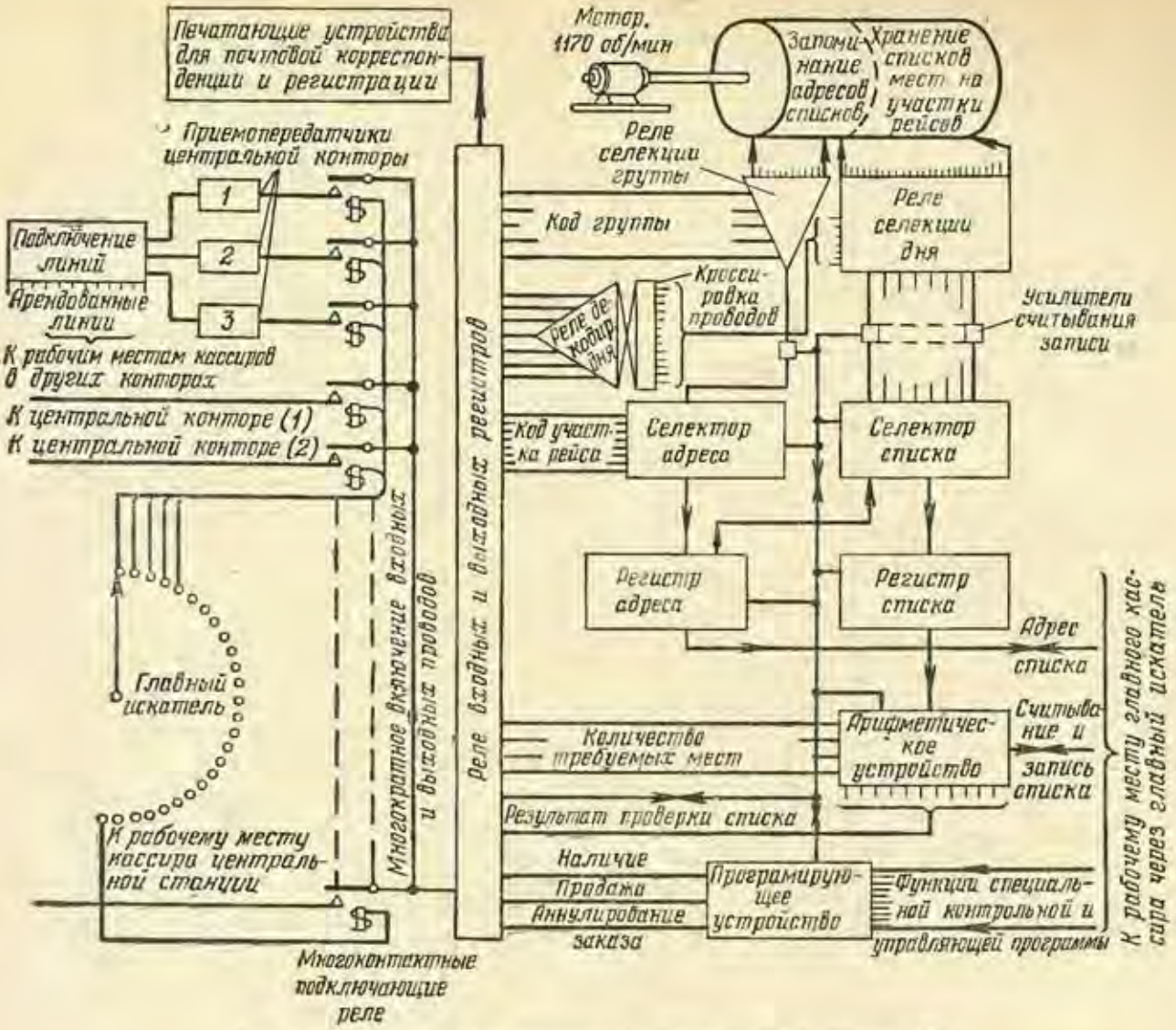


Рис. 2.2. Система „Резервизор"—блок-схема оборудования [3].

рейс делится на участки (от взлета до посадки самолета), и данные о местах для каждого участка одного и того же рейса записываются независимо.

Этот магнитный барабан может автоматически запрашиваться любым из 200 кассиров, находящихся в Нью-Йорке и других пунктах страны. Запрос осуществляется с помощью клавишного устройства, в которое кассир вводит свои данные, и *схем полетов*, на которых указано расписание полетов на участках авиатрасс, входящих в соответствующую группу. Затем клавишное устройство кодирует введенную информацию, преобразуя ее в сигналы, которые могут передаваться автоматически.

Система «Резервизор» может запрашиваться через одну-две секунды, разными кассирами. Система обладает способностью выполнять некоторые логические операции, такие, как извещение служащего агентства об отказе от билета на рейс, все места на который распроданы.

Цикл работы этой системы взят равным 24 *час*. Центральная вычислительная машина дублирована, так что текущий ремонт или выход из строя аппаратуры не прерывает работы системы. В течение 1953 г. эта система работала 99,8% запланированного времени [3].

Система регулирования уличного движения в Денвере. 5 мая 1952 г. в г. Денвере, штат Колорадо, была введена в действие центральная система регулирования уличного движения, работа которой основана на автоматической регулировке сигналов светофоров в соответствии с непрерывно изменяющимися условиями движения. Наиболее интенсивное движение транспорта в этом городе происходит между деловым центром и северным и северо-восточным районами города: в утренние «часы пик» — в направлении к центру и в вечерние часы — в направлении

от центра. Однако в городе все время существует значительное поперечное движение, и обычный пик движения может возникать в необычное время или даже изменить свое направление. Так, например, если во второй части дня идет дождь, то около 5 часов вечера иногда наблюдается интенсивный поток машин в деловой центр города, связанный, по-видимому, с тем, что жены направляются на автомашинах за своими мужьями.

В обычных условиях светофоры работают поочередно, в так называемом «прогрессивном» режиме, при котором автомашина, едущая с надлежащей скоростью, синхронизируется с переключениями светофоров и встречает на своем пути только зеленые огни*. Такая система характеризуется несколькими переменными. Очень трудно спланировать переключение светофоров таким образом, чтобы оно было поочередным в обоих направлениях, особенно когда длина цикла и скорость прогрессии (см. ниже) изменяются. Следовательно, приходится благоприятствовать либо потоку в центр, либо потоку из центра. *Скоростью прогрессии* (вторая переменная) называется скорость переключения светофоров, при которой автомашины подходят к светофорам синхронно с включением на них зеленого света, т. е. та «групповая скорость», с которой волна изменения сигналов светофоров движется вдоль улицы. Под термином «длина цикла» (третья переменная) понимают продолжительность всего отрезка времени между предыдущим и последующим включением зеленого света светофора.

Дополнительными переменными могут быть отношение времен горения зеленого и красного света на протяжении всей магистрали и возможное дополнительное время, отводимое на переход перекрестка пешеходами, когда сигналы светофора должны быть красными для автомашин во всех направлениях.

Когда движение не очень интенсивное, скорость прогрессии должна быть высокой, а длина цикла — небольшой, чтобы обеспечивалось быстрое движение автомашин вдоль главных магистралей и в то же самое время минимальная задержка перед светофорами автомашин, движущихся в перпендикулярных направлениях. Когда движение становится гуще и потому медленнее, оказывается предпочтение потоку в центр либо потоку из центра, смотря по обстоятельствам; в результате скорость прогрессии снижается, а длина цикла возрастает.

Так как продолжительность горения красного света на главной улице (и зеленого для поперечного движения) поддерживается постоянной, то удлинение цикла приводит к увеличению числа автомашин, которые могут проехать по главной улице; снижение скорости прогрессии удерживает скорость движения на уровне, при котором автомашины еще редко будут останавливаться на перекрестках из-за красных сигналов светофора.

Это регулирование движения осуществляется при помощи датчиков давления (педалей), установленных в надлежащих точках города. Три таких датчика измеряют поток в центр, и три — поток из центра. Когда над одним из датчиков проезжает автомашина, он вырабатывает электрический импульс, который посылается в центральный пост. Эти импульсы подводятся к двум конденсаторам: одному для потока в центр и другому для потока из центра. Каждые 6 *мин* измеряется заряд конденсаторов, после чего они разряжаются.

Измерительное устройство имеет шесть градаций уровней заряда. В соответствии с измеренным уровнем устанавливается одна из шести предварительно рассчитанных скоростей прогрессии и соответственно одна из шести возможных длин цикла. Некоторое дополнительное усложнение работы системы вызвано применением на некоторых перекрестках световых табличек «стойте — идите» для пешеходов.

Оборудование всей системы (включая датчики, простое вычислительное устройство и линии связи) стоит около 130 000 долларов, или несколько более 1000 долларов на один перекресток. Эта система оказывается очень полезной, так как сокращает продолжительность «часа пик» на 30 *мин* и значительно облегчает условия движения автомашин в этот период времени. Аналогичная, но более сложная система была затем введена в действие в г. Балтиморе, штат Мэриленд.

К интересным проблемам, связанным с такой системой, можно было бы отнести: выбор оптимального количества и местоположения датчиков, возможность прогрессивного переключения светофоров в обоих направлениях и (или) в направлении некоторых перпендикулярных улиц одновременно, преимущества от применения световых сигналов для пешеходов (уменьшение замедления правых поворотов) по сравнению с отрицательными свойствами таких сигналов при заданной длине цикла (меньшая продолжительность зеленого света), учет влияния различной густоты потоков по разным направлениям и соответ-

---

* У нас такая организация работы светофоров известна под названием «зеленой волны». — *Прим. ред.*

ствующий выбор логики системы и, наконец, выбор оптимального темпа, с которым система регулирования должна делать замеры движения.

Выбор критерия эффективности для такой системы является весьма серьезной задачей. Конечно, число минут, на которые сокращается «час пик», производит впечатление, но такой критерий не содержит в себе характеристик, необходимых для оценки работы системы. Хотя этот критерий является единственным, который можно найти в опубликованных отчетах, инженер-системотехник, по-видимому, больше интересовался бы числом минут, за которые автомашины проходят путь от делового центра города до пригородов, скоростью потока автомашин в милях в час или количеством автомашин, пересекающих перекресток в течение часа.

2.2. Связь

Современная телефонная система является, какое бы определение ни взять, системой большого масштаба; действительно, она состоит из совокупности различных подсистем, активно взаимодействующих между собой, причем каждую из этих подсистем можно назвать системой большого масштаба. Ниже мы очень коротко ознакомим читателя с пятью такими подсистемами: с одной из основных коммутационных систем центральной телефонной станции, с системой автоматического учета для выписки счетов за телефонные разговоры, с автоматической системой ремонта, с методом защиты системы от воздействия необычных пиковых нагрузок и с автоматической системой дальнего набора номера.

Координатная система № 5. В настоящее время Белловская телефонная система* эксплуатирует несколько различных ручных систем и четыре автоматические системы для центральных телефонных станций: шаговую, панельную, координатную № 1 и координатную № 5. Кроме того, разработана чисто электронная система, однако пока еще не решено, когда она будет установлена в эксплуатацию. Координатная система № 5 должна успешно работать со всеми другими системами; любой телефонный аппарат должен иметь возможность соединения с любым другим из 55 000 000 телефонных аппаратов.

Центральная телефонная станция, использующая координатную систему № 5 обслуживает [6], обычно 10 000 главных абонентских установок (каждая главная абонентская установка состоит из абонентов, находящихся на одной коллективной линии). Таким образом, к станции подходят и оканчиваются в *стативах шнуров абонентских линий*** 10 000 абонентских линий (абонентская линия всегда образуется парой проводов). Система должна обеспечивать возможность соединения в любой момент времени любой из этих линий с любым возможным оборудованием другого рода. Основным коммутационным устройством, применяемым для этой цели, служит координатный переключатель***, представляющий собой матрицу (координатную сетку) из вертикальных и горизонтальных проводов. В одном типе применяемых в системе координатных переключателей к вертикальным проводникам матрицы подключено 10 абонентских линий, а к горизонтальным проводникам — 10 «соединителей»****.

В каждой точке пересечения («точке коммутации») вертикальных и горизонтальных проводников матрицы установлено реле (с несколькими контактными точками для надежности), которое возбуждается при помощи электромагнитов и сохраняет контакт между проводниками до тех пор, пока оно не примет сигнала о размыкании цепи. Благодаря каскадному соединению таких координатных переключателей типовой статив шнуров абонентских линий может соединить любую из 500 абонентских линий с любым из 100 соединителей (приводимые в этом параграфе числа относятся к средней, или типовой, станции с координатной системой № 5; для отдельных АТС эти числа могут изменяться в широких пределах).

Соединители оканчиваются в 10 *стативах шнуров соединительных линий,* которые так-

---

* Белловская телефонная система («Белл Телефон Систем») — один из крупнейших концернов в США, образованный Американской телефонно-телеграфной компанией и ее дочерними предприятиями, контролирует большую часть телефонной сети США.

Кроме Американской телефонно-телеграфной компании, в Белловскую систему входят компания «Вестерн Электрик», Белловские телефонные лаборатории и 21 эксплуатационная компания. — *Прим. ред.*

** В координатных системах Белловской телефонной системы соединительные линии между переключателями одного статива принято называть «шнурами» (links), а соединительные линии между отдельными стативами — «соединителями» (junctors) [Д. 27]. — *Прим. ред.*

*** В русской литературе координатный переключатель довольно часто именовался также на английский лад переключателем «Кроссбар» (от его английского названия crossbar switch), соответственно координатная телефонная система именуется также системой «Кроссбар». — *Прим. ред.*

**** Эти «соединители» (junctors) называют также в русской литературе о координатной системе № 5 «промежуточными линиями». — *Прим. ред.*

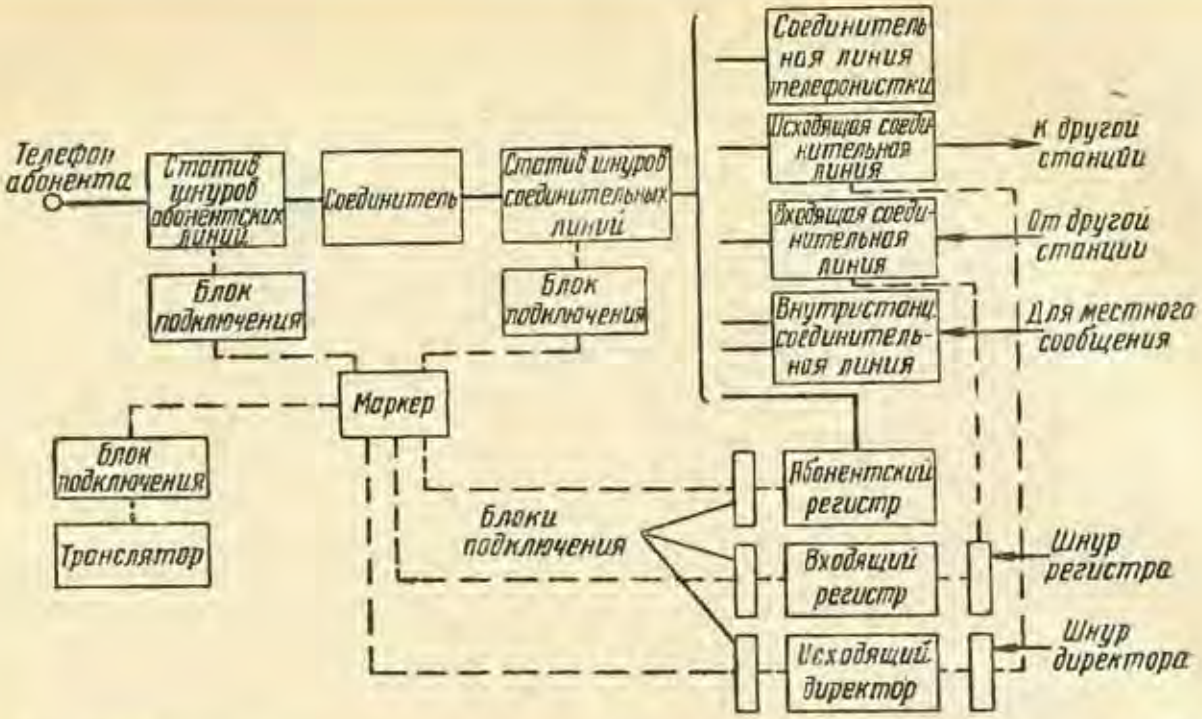


Рис. 2.3. Основные элементы коммутационного оборудования координатной системы № 5 [63].

же состоят из координатных переключателей. Через эти стативы шнуров соединительных линий каждый соединитель может подключаться либо к одной из внутристанционных соединительных линий, либо к одной из входящих соединительных линий, либо к одной из исходящих соединительных линий, либо к одной из соединительных линий телефонисток *, либо к одному из абонентских регистров (originating registers), либо к одному из входящих регистров, либо, наконец, к одному из «директоров». *Регистрами* называются устройства, принимающие импульсы набора номера от абонента и хранящие их для последующего использования при осуществлении вызова; *директор* (sender) — устройство, которое может передавать к другой станции последовательность импульсов, аналогичных импульсам, поступающим от номеронабирателя телефонного аппарата; другие термины не требуют объяснений.

Управление всей этой коммутацией осуществляют примерно семь сложных устройств, называемых *маркерами* и являющихся в действительности специализированными цифровыми вычислительными машинами, хотя они и не называются так. Маркер используется только для установления и нарушения соединений и не участвует в удержании созданного соединения. В этом отношении он подобен телефонистке на ручной телефонной станции, которая вставляет шнуры в гнезда, чтобы осуществить соединение, а затем, во время разговора абонентов, может обслуживать другие соединения. В противоположность этому в более ранних автоматических телефонных системах сложные устройства, необходимые для коммутации цепи, должны были применяться в больших количествах, так как в тех системах они оставались занятыми в течение всего времени, пока цепь замкнута.

Работа координатной системы № 5 при типичном вызове показана на рис. 2.3. Когда абонент снимает микротелефонную трубку с рычага своего аппарата, на телефонной станции срабатывает реле, которое заставляет соответствующий статив шнуров абонентских линий через блок подключения ** маркеров к шнурам абонентских линий занять один из свободных маркеров и сообщить этому маркеру, что требуется установить соединение для набора номера. Маркер определяет местоположение («адрес») вызывающей абонентской линии и одновременно выбирает один из свободных абонентских регистров.

Через свой блок подключения шнуров соединительных линий маркер подключается к надлежащему стативу шнуров соединительных линий и передает выбранному абонент-

* Телефонистки обслуживают некоторые классы вызовов и в случае необходимости оказывают помощь абонентам при других вызовах — *Прим. ред.*

** Блок подключения (connector) состоит из специальных многоконтактных реле, несущих до 60 нормально разомкнутых контактов. В целом блок подключения создает 150—250 независимых соединений. — *Прим. ред.*

скому регистру информацию о местоположении вызывающей абонентской линии (эта информация не нужна абонентскому регистру для приема импульсов набора номера, но, как мы увидим, она ему потребуется позже).

Затем маркер выбирает свободный соединительный путь от линии вызывающего абонента через переключатели стативов шнуров абонентских линий и стативов шнуров соединительных линий к этому выбранному абонентскому регистру, причем сначала производится опробование на занятость вводов шнуров абонентских линий, вводов соединителей и вводов шнуров соединительных линий на возможных путях, а потом выбранный путь фиксируется срабатыванием магнитов в координатных переключателях. После установления соединительного пути маркер, блок подключения шнуров соединительных линий, блок подключения шнуров абонентских линий *, блок подключения маркера к шнурам абонентских линий освобождаются и абонентский регистр посылает абоненту сигнал готовности станции. Все описанные операции совершаются обычно в пределах одной секунды, после чего все эти устройства, и в частности маркер, могут обслуживать другие соединения в течение всего времени (в среднем 12 *сек*), пока набирается номер.

После того как абонент окончил набор номера, абонентский регистр занимает свободный маркер и сообщает ему адрес линии вызывающего абонента и номер телефона вызываемого абонента. *Номер телефона* и *адрес линии* — это не одно и то же; по многим соображениям оказывается желательным предоставить телефонной компании свободу изменять местоположение той или иной абонентской линии на стативе шнуров абонентских линий (и тем самым адрес линии) без изменения номера телефона, и обратно.

Следовательно, один из этапов осуществления вызова заключается в том, что маркер через соответсвующий блок подключения занимает особое устройство — *транслятор*, который переводит номер телефона в адрес линии и определяет сигнал вызова (вызываемый абонент может находиться на коллективной линии и иметь особый звонок). Однако до того, как это может быть сделано, маркер должен определить, относится ли данный вызов к местному сообщению станции или к сообщению с другой станцией; мы будем предполагать, что вызов относится к местному сообщению.

Затем маркер пробует, не занята ли линия вызываемого абонента. Если она оказывается занятой, то маркер подключает линию вызывающего абонента к специальной внутристанционной соединительной линии для посылки сигнала занятости, после чего маркер освобождается. Если же линия не занята, то маркер осуществляет соединение двух телефонов и подключает к линии вызываемого абонента устройство посылки сигнала вызова. Затем все другие устройства, в том числе абонентский регистр, маркер и различные блоки подключения, освобождаются, а координатные переключатели удерживаются в нужном положении с помощью реле.

Изложенное выше, как бы сложно оно ни выглядело, представляет собой лишь неполное описание одного класса вызова. Но, разумеется, существует много разных классов вызовов. В гл. 13 мы еще раз рассмотрим работу этой системы.

Автоматический учет разговоров. За 1948 г. Белловская телефонная система осуществила 50 миллиардов вызовов для своих абонентов. Более 1600 миллионов из этих разговоров были междугородными разговорами с разовой оплатой (из них примерно 600 миллионов стоили только от 5 до 10 центов за разговор, но, несмотря на это, на них приходилось выписывать разовые счета, что в ряде случаев вызывалось правительственными законами). Кроме того, для значительной части других разговоров, хотя они и оплачивались суммарно, необходимо было определять их продолжительность и стоимость в соответствии с разными тарифами; так, например, оплата могла составлять 10 центов за первые 4 *мин* и 5 центов за каждые последующие 2 *мин* для одного разговора, а для следующего разговора могла быть совсем иной. Вполне понятно, что эти расчеты обходятся весьма дорого.

Описываемая ниже система автоматического учета разговоров [7] подключается к координатной системе № 5. При этом система № 5 должна выполнять некоторые дополнительные функции; так, например, абонентский регистр должен получать и хранить сведения о классе обслуживания вызывающего абонента, а маркер, получая от регистра эти сведения вместе с номерами вызывающего и вызываемого абонентов, должен определить, подлежит ли этот вызов оплате и если да, то по какому тарифу. Маркер должен также осуществить соединение с определенной исхо-

* Блок подключения шнуров абонентских линий служит для подключения к маркеру статива шнуров абонентских линий при опробывании вводов шнуров. — *Прим. ред.*

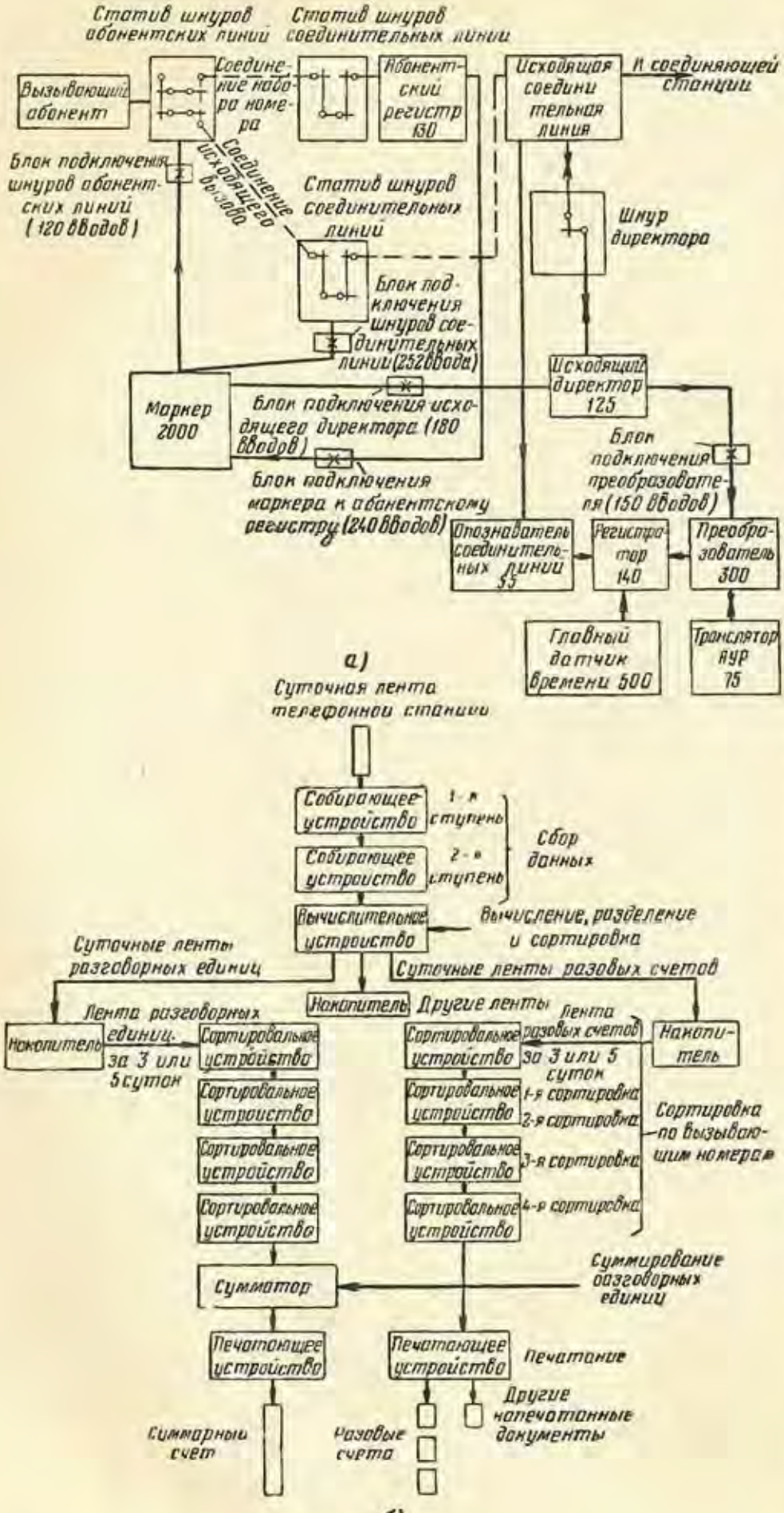


Рис. 2.4. Блок-схема системы автоматического учета разговоров — АУР [7]:

а — центральная телефонная станция; б — учетный центр.

дящей соединительной линией (при исходящем вызове) и подключить директор (который посылает необходимую информацию к другой телефонной станции) к *преобразователю*.

Преобразователь сначала подключается к *транслятору*, чтобы определить номер телефона вызывающего абонента, затем преобразует всю относящуюся к этому вызову информацию в код, который можно набить на перфоленту. Если соединение с другой станцией осуществлено с положительным результатом, то, прежде чем зазвонит телефон вызываемого абонента, преобразователь автоматически подключится к *регистратору*, который пробивает все данные на бумажной перфоленте. Каждый регистратор обслуживает группу в 100 исходящих соединительных линий, и одним из элементов записываемой информации является двузначное число (от 00 до 99), обозначающее соединительную линию; этот двузначный *код разговора* является ключом для работы системы автоматического учета разговоров.

Если вызываемый абонент занят или не отвечает, то никакой дальнейшей информации на перфоленте не записывается и, как мы увидим в дальнейшем, никакой оплаты не начисляется. Если вызываемый абонент ответил, то на перфоленте делается одна запись: фиксируется время начала разговора и двузначный код разговора. Наконец, когда разговор заканчивается и вызывающий абонент от-

ключается, на перфоленте делается другая запись, указывающая время окончания разговора и код разговора.

В конце 24-часового периода (в 3.00 утра, т. е. в наименее загруженный час) лента обрезается и передается в учетный центр (рис. 2.4,*б*), производящий расчеты для 250 000 абонентов. Рулон ленты может иметь диаметр 45,72 *см* и содержать данные о 25 000 разговорах; однако данные об этих разговорах перемешаны: между записью о начале одного разговора и записью о его окончании может находиться много записей о других разговорах.

Три записи, относящиеся к одному разговору, определяются по одинаковому коду разговора. Они располагаются в определенном порядке: первая запись указывает номера телефонов, тарифную сетку и тому подобное, вторая — время, когда ответил телефон вызванного абонента, и третья — время окончания разговора. Каждая из этих отметок времени представляет собой трехзначное число (минуты и десятые доли минуты, так как час пробивается на перфоленте каждый раз, когда начинается новый час).

Сначала лента вводится в *собирающее устройство,* которое считывает записанные на ленте данные и одновременно перфорирует 10 новых перфолент. На первую из этих лент наносятся все данные о разговорах, код которых оканчивается на 0; на вторую — разговоры, код которых оканчивается на 1, и т. д. Когда лента окажется полностью переписана таким образом, десять выходных лент скрепляются своими концами друг с другом и вновь вводятся в собирающее устройство, которое переводится при этом в режим сортировки данных по первой цифре кода разговора. Десять новых выходных лент снова скрепляются своими концами друг с другом, в результате чего создается итоговая лента, на которой три элемента информации о каждом состоявшемся разговоре оказываются собранными вместе. Эта лента содержит также информацию о несостоявшихся разговорах и другую дополнительную информацию, как, например, часовые отметки времени.

Затем эта лента вводится в *вычислительное устройство,* которое определяет продолжительность каждого состоявшегося разговора, исключает данные о несостоявшихся разговорах, подсчитывает стоимость разговоров, перфорирует на отдельных выходных лентах данные о разговорах с суммарной оплатой и о разговорах с разовой оплатой, а также выполняет различные другие операции. Так, например, оно набивает специальную ленту для *проскальзывающих разговоров,* т. е. разговоров, которые происходили в 3 часа утра, когда снималась лента. Оно также набивает на специальной ленте данные о каждом разговоре любого абонента, который просит сделать для него такую выборку или уточнить его счет.

Полученные ленты каждого типа соединяются между собой и вводятся в *сортировальное устройство,* которое сортирует данные по номерам телефонов, подобно тому, как собирающее устройство производило сортировку по кодам разговоров. Тем самым сводятся вместе все данные, необходимые для выписки счета за телефон. Лента, содержащая эти данные, вводится в *сумматор,* который складывает стоимости разговоров, подлежащих суммарной оплате, и перфорирует новую ленту. Эта лента вводится в *печатающее устройство,* которое в необходимых случаях преобразует цифровой код в буквы (например, цифру 12, обозначающую декабрь, — в DEC; цифру 236, обозначающую Adams 6, — в AD6) и печатает автоматически ярлыки суммарных и разовых счетов за междугородные телефонные разговоры. В настоящее время эти счета соединяются ручным способом с другими, вручную подготовленными счетами, однако итоговый счет абонента можно печатать и с помощью автоматически действующих устройств.

При одном типовом испытании этой системы было обслужено 38 198 разговоров. При этом не было ни одного неправильного подсчета (и, следовательно, ни одного крайне нежелательного превышения правильной суммы счета); три разговора, или 0,0078% всех разговоров, оказались неучтенными (и, следовательно, неоплаченными): один разговор был записан на дефектной ленте, а в остальных двух случаях было неисправно оборудование.

Автоматический ремонт [6]. Для обнаружения важнейших неисправностей в координатной системе № 5 предусмотрены акустические и оптические сигналы. Существует, однако, несколько типов неисправностей, не носящих аварийного характера, так что их устранение не является срочным. Например, контакты в какой-либо точке коммутации могут загрязняться и создавать неисправность перемежающегося характера.

В случаях, когда маркер не может осуществить данный вызов из-за неисправности оборудования, он подключается к *регистратору неисправностей* и передает ему всю информацию, относящуюся к несостоявшемуся вызову, включая соединительный путь, который он пытался установить. Пока регистратор неисправностей переносит на перфокарту эту

информацию (а также время в кодированной форме), маркер предпринимает вторую попытку осуществить вызов, используя другой путь. Если вторая попытка также оказывается неудачной, пробивается вторая перфокарта и абоненту посылается сигнал о неисправности оборудования. Кроме того, имеется *автоматическое контрольное устройство*, проверяющее автоматически выборочным способом импульсные характеристики регистров и директоров во время их текущей работы. Время от времени дежурные техники собирают карты от регистратора неисправностей и выполняют необходимые ремонтные работы.

Регулирование линейной нагрузки. Телефонная система построена на вполне разумном предположении, что лишь небольшой процент абонентов пожелает использовать свои телефоны в одно и то же время; строить ее в расчете на 100%-ное использование было бы слишком дорого. Точный процент абонентов, которые могли бы одновременно использовать свои телефоны, определить очень трудно; он зависит от используемой системы (скажем, от того, используется ли система № 1 или № 5), района расположения телефонной станции, вида переговоров и метода обслуживания возникающих вызовов.

Как мы уже знаем, в координатной системе № 5 соединителей хватает на одновременное обслуживание примерно лишь 20% абонентов, но эти соединители делятся на разные классы, и если все абоненты пожелают сделать вызовы одного определенного класса (скажем, с абонентами той же самой станции), то система сможет обслужить только очень небольшой процент абонентов. Наконец, абонентские регистры устанавливаются из расчета на одновременное обслуживание только 2% абонентов. Таким образом, если одновременно снимет трубки большое число абонентов, то лишь небольшая часть их услышит сигналы готовности станции.

Однако в действительности положение еще хуже. Если большое число абонентов, сняв трубки и не услышав сигнала готовности станции, будут продолжать держать трубки, ожидая появления этого сигнала, то при этом абоненты будут занимать маркеры, заставляя их искать незанятый регистр. А затем, когда абонент, которому был послан сигнал готовности станции, закончит набор номера, соответствующий регистр окажется не в состоянии найти свободный маркер для осуществления следующего этапа вызова. Приходящие тем временем другие вызовы также не будут обслуживаться, что увеличит число ожидающих маркеров и выключит из работы другое оборудование, например соединители и соединительные линии. Приходящие вызовы, ожидая своей очереди обслуживания, будут распространять цепную реакцию задержки на другие телефонные станции до тех пор, пока не будет парализована вся телефонная система.

Этот «кошмар» инженеров-телефонистов не является каким-то досужим вымыслом. Когда по радио было передано сообщение о смерти Франклина Рузвельта, очень большое число людей почти одновременно пожелало позвонить об этом по телефону своим знакомым. В то время координатная система № 5 еще не была введена в эксплуатацию, но координатная система № 1 была полностью парализована почти таким же образом, как описывалось выше, и прошло немало часов, прежде чем восстановился нормальный режим ее работы. Медленно возрастающая нагрузка не вызывает таких неприятностей, как внезапное увеличение нагрузки, однако если она оказывается достаточно высокой, то может вызвать такую же цепную реакцию; это случилось в одном из больших городов, когда, внезапно изменив направление, к городу стал приближаться ураган.

Поэтому перед инженерами Белловской телефонной системы возникла серьезная задача по предотвращению паралича телефонной сети в случае возникновения широко распространяющейся новости, вроде сообщения об угрозе воздушного нападения.

Найденное решение просто и изящно. Прием входящих вызовов, т. е. требований обслуживания со стороны вызывающих абонентов, должен приостанавливаться, но посылка исходящих вызовов, т. е. звонка абонентам, являющегося последним этапом осуществления соединения, не должна приостанавливаться. Приостановка исходящих вызовов не только противоречила бы назначению всей системы, но и способствовала бы дальнейшему росту цепной реакции.

Следовательно, система регулирования линейной нагрузки, позволяющая решить эту проблему, должна просто лишать абонентские линии возможности занимать абонентские регистры. Во всех же других отношениях система должна продолжать работать нормально. Такая приостановка работы производится по группам линий, вплоть до 80% всех линий станции, в зависимости от характера события. Остающиеся 20% являются наиболее важными и не должны отключаться от регистров. Но телефонная коммутационная система может выдержать почти любую нагрузку, создаваемую этими оставшимися телефонами.

Поскольку система регулирования линей-

ной нагрузки введена в действие, она может служить и в нормальных условиях перегрузки, которые нельзя квалифицировать как критические. Когда нагрузка на телефонной станции или в какой-либо части телефонной станции приближается к опасному уровню, загораются лампочки предупреждения и частично срабатывает система регулирования, отключая небольшую группу абонентов. Абоненты об этом, конечно, не знают, так как их телефоны принимают приходящие вызовы, а когда они снимают телефонные трубки, то замечают только, что станция дает сигнал готовности к набору номера с некоторой небольшой задержкой. При обычно встречающихся перегрузках применение такого регулирования на несколько секунд почти всегда приводит к желаемым результатам.

Прямой набор номера при дальней связи. Прямой набор номера при дальней телефонной связи был одной из целей инженеров Белловской телефонной системы с 1933 г. и учитывался при проектировании таких систем, как координатная система № 5. В случае менее сложного коммутационного оборудования абонент должен был бы при наборе номера сообщать полную информацию о всем пути прохождения вызова.

В системе прямого дальнего набора номера (или системе *вызова абонента в «чужой зоне»*), которая впервые была введена в коммерческую эксплуатацию в 1951 г., для вызова любого телефона, находящегося в пределах страны, необходимо набрать только три цифры (плюс местный номер вызываемого абонента). Это оказалось возможным благодаря способности маркера декодировать информацию, опробовать различные возможные пути прохождения вызова и, наконец, выполнить следующий этап соединения и пропустить всю необходимую информацию. При этом, конечно, потребовалась специальная дополнительная коммутационная аппаратура.

2.3. Промышленность

Хотя первые системы большого масштаба в основном использовались для обработки и передачи информации, вполне возможно, что наиболее эффективное применение в будущем они найдут при управлении процессами переработки материалов. В настоящее время уже создано несколько типов заводов-автоматов, и с течением времени их количество и разнообразие будут значительно возрастать. Мы рассмотрим только одну типичную задачу — автоматическое производство электронных схем.

Существует два основных метода решения

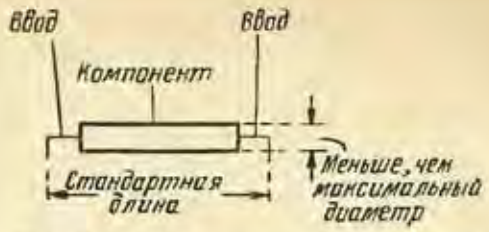


Рис. 2.5. Компонент для автоматической сборки.

этой задачи. Один метод известен под названием *автоматической сборки компонентов* [8]. В этой системе используются более или менее стандартные детали и задача заключается в разработке специального оборудования, которое при сборке придавало бы необходимое положение детали и удерживало ее в этом положении с высокой точностью до окончания соответствующего этапа сборки *.

Другой метод известен под названием *модульного конструирования и механизированного производства электроники* [9]. В этой системе применяются компоненты специальной конструкции, облегчающей сборку **. Обе системы имеют целью избежать больших затрат труда, неизбежных при сборке бесчисленного количества мелких деталей методом ручной пайки. Обе системы основаны на применении *печатных схем и пайки погружением:* сначала с помощью раствора, содержащего серебро, проводники, которые должны соединять детали, наносятся в виде необходимых линий на пластмассовую пластинку, а затем эта пластинка опускается в ванну с расплавленным припоем, который пристает только к серебру.

В системе автоматической сборки компонентов почти все компоненты (сопротивления, конденсаторы и индуктивности) изготовляются лишь в небольшом количестве классов со стандартными размерами (рис. 2.5). После того как схема разработана и вычерчена, она разбивается на подсборки (каков, например, узел высокой частоты), каждая из которых будет собираться на отдельной плате. Затем схемы подсборок стандартизируются путем придания каждому компоненту стандартной длины и расположения его либо в вертикальном, либо в горизонтальном положении, либо под углом к этим двум положениям, который может иметь значения, кратные 10°.

---

* Этот метод кратко называется «автосборкой» (Auto-Sembley); его разработала техническая лаборатория Корпуса связи США. — *Прим. ред.*

** Этот метод называется также «проектом Тинкертой» (Tinkertoy project) и был разработан в 1950 г. Национальным бюро стандартов США. — *Прим. ред.*

После этого с точностью до тысячной дюйма измеряются положения концов этих компонентов и результаты измерений записываются с помощью специального кода. Так, запись 35721853 означает, что один конец компонента располагается в 5,572 дюйма от левого края платы и в 1,853 дюйма от ее нижней кромки. Другие цифры указывают ориентировку компонента, и еще другие — его тип, размер и электрическое значение. Окончательно все эти данные кодируются в одно число, состоящее из 50 двоичных разрядов.

Затем числа, найденные таким образом для каждого компонента подсборки, сводятся в таблицу, а вся эта группа чисел переносится на *перфокарту памяти*, которая вводится в вычислительное устройство. Вычислительное устройство за каждый цикл работы берет из соответствующего хранилища нужный компонент и вставляет его в устанавливающую головку, которая доставляет этот элемент в определенное место и поворачивает его на заданный угол.

Одновременно с этим механизм подачи устанавливает плату в нужное положение. В соответствующих точках платы заранее просверливаются отверстия, а на ее обратной стороне заранее печатается схема, что обеспечивает соединение этих отверстий друг с другом соответствующими проводниками. Затем компонент вставляется в плату и вычислительное устройство освобождается для установки следующего компонента. После того как все элементы будут вставлены, плата отодвигается и производится пайка погружением и автоматическое испытание подсборки (компоненты проходят автоматическую проверку до сборки).

В системе модульного конструирования основным узлом является «модуль», который по своим размерам несколько меньше рассмотренной выше подсборки и делится на небольшие узлы, называемые *галетами*; каждая из этих галет может нести по два или три компонента. Галета имеет размеры приблизительно почтовой марки. На поверхность галеты наносят компоненты и печатные схемы; провода от них выводятся в одну из 12 выемок, расположенных по кромке галеты. От четырех до шести таких галет монтируется одна над другой, образуя модуль, причем электрические соединения между галетами создаются с помощью 12 вертикальных проводов, проходящих через выемки.

После пайки проводов в выемки и, если нужно, обрезки некоторых проводов весь модуль (размеры которого лишь немногим больше кубического дюйма) покрывается пластиком. На верхней части модуля может монтироваться ламповая панель; 12 проводов, выступающих из основания модуля, образуют стандартный соединитель для установки и подключения модуля к шасси. Сравнительно небольшое число соединений между модулями осуществляется с помощью нанесенной на шасси печатной схемы. На шасси может также укрепляться какой-либо нестандартный элемент схемы, который не удается встроить в модуль.

В этой системе компоненты изготовляются во время сборочных операций. Сопротивления получаются путем напыления угля на асбестовую ленту, покрываемую затем пластиком, и последующей нарезки этой ленты на куски нужной длины. Конденсаторы получаются путем нанесения проводящей (серебряной) краски на керамический материал и горячей сушки. Катушки индуктивности печатаются на маленьких цилиндрах. Все компоненты и все сборки проходят автоматическую проверку, и неисправные компоненты и сборки автоматически выбраковываются.

Система модульного конструирования претендует на большее, чем система автоматической сборки компонентов, так как она пытается сочетать изготовление компонентов со сборкой. Поэтому она принципиально сложнее, особенно в таких аспектах, как возможные допуски на компоненты, но, по-видимому, при длительной работе будет лучше системы автоматической сборки, если ее можно заставить работать правильно.

В системе автоматической сборки компонентов серьезные трудности вызывает необходимость высокой точности в работе механизма подачи и устанавливающей головки. К преимуществам аппаратуры с модульной конструкцией следует отнести возможность ее ремонта путем замены целого модуля. Такой ремонт при больших масштабах производства и массовой эксплуатации аппаратуры обходится дешевле, чем попытки найти вышедшую из строя деталь. Кроме того, такой ремонт можно производить путем смены вставных штепсельных блоков, не прибегая к паяльнику, что является значительным преимуществом.

Наконец, следует заметить, что обе системы экономически оправдывают себя лишь в случае массового выпуска аппаратуры при достаточно постоянном темпе производства. Однако система автоматической сборки компонентов обладает большей гибкостью: как только изготовлены перфокарты памяти, машина может изготовить и одну, и много подсборок без какого-либо изменения настройки. Система модульного конструирования оказы-

вается менее эффективной, если она не может делать достаточно большое количество модулей одного типа.

2.4. Коммерция

К коммерческим системам мы причисляем также системы сбора и обработки данных в правительственных учреждениях. О проблемах сбора и обработки данных в правительственных учреждениях мы сможем сказать немногое, и не с целью описать какое-нибудь решение, а лишь для того, чтобы показать чудовищную сложность проблем.

Министерство финансов США каждый день получает миллион погашенных правительственных чеков, каждый из которых должен проверяться; в этом случае в системе имеются соперничающие стороны, так как некоторое небольшое количество людей делает все, что в их силах, чтобы обмануть правительство. Администрация социального обеспечения сталкивается с аналогичной проблемой: работа по обработке данных в этой организации настолько велика, что для выполнения счетных работ каждый день приходится пробивать около 700 000 новых перфокарт.

Бюро переписей должно учитывать не только рост страны, но также и непрерывно растущие требования о дополнительных данных и специальных обследованиях. Без применения счетно-перфорационных машин оно оказалось бы теперь не в состоянии обработать материал производимой каждое десятилетие переписи даже в течение 10 лет, но без применения новейших электронных вычислительных машин Бюро переписей вскоре вновь попало бы в то же положение. Обработка информации в настоящее время стала настолько эффективной, что ее стоимость составляет лишь небольшую долю от стоимости сбора информации, подготовки ее для обработки и публикации результатов обработки.

ВВС США стоят перед нерешенной задачей военно-тыловой работы в своем Управлении материально-технического обеспечения. Этому управлению приходится иметь дело с 1 200 000 различных предметов снабжения при складском учете, оно должно иметь в запасе достаточное количество предметов каждого наименования, хранить их в требуемом месте и быстро выдавать по требованиям.

Говоря о частных организациях, следует указать на массовые журналы, которые рассылаются миллионам подписчиков; все эти подписчики должны быть занесены в списки, они могут менять свои адреса, и их приходится просить о продолжении подписки. Большие компании по страхованию жизни насчитывают миллионы держателей страховых полисов. Эти держатели имеют различный возраст и пол, обладают различным правом отказа от продолжения страховки, включают в договор различные условия страхования и живут в 48 различных штатах *, в каждом из которых действуют свои особые законы. Не говоря даже о сложности записи и хранения всех этих данных, следует указать на сложность проблемы составления полиса, который оказывался бы законным с юридической точки зрения в том или ином штате и имел бы соответствующую страховую премию. Для решения этой задачи несколько страховых компаний используют большие вычислительные машины.

Канадская почтовая система [60]. В 1955 г. в Оттаве для сортировки и распределения почты была введена в эксплуатацию автоматически действующая система. Наиболее интересной задачей, которую надо было решить при создании системы, была задача стандартизации входов. Если бы все письма были одинакового размера и формы и имели определенный механизированный код (подобный применяемому для перфокарт), то для механизации процессов обработки почты можно было бы применить много различных методов.

Метод решения, которым в действительности воспользовались, состоит в том, что специальный код печатается большими черными знаками в нижнем правом углу передней стороны конверта на определенном, фиксированном расстоянии от правого края и низа конверта. Сначала письма сортируются и вращаются механически до тех пор, пока они не примут правильного положения; в качестве исходного ориентира машина использует положение почтовой марки (предполагается, что она наклеена в правом верхнем углу на лицевой стороне конверта). Если машина совершает ошибку, что иногда случается, то такая ошибка устраняется вручную. Письма необычного размера или формы или имеющие другие отклонения от общепринятого типа отбраковываются в небольшую пачку, которая также обрабатывается вручную.

Затем отсортированные письма, имея правильное положение, большими партиями поступают к операторам специальных пишущих машинок, которые наносят на них нужный код. Каждое письмо при этом наблюдается оператором через окошко, в которое виден написанный на письме адрес, и оператор набирает на клавиатуре код места назначения (эти

* Сейчас в США 50 штатов. — *Прим. ред.*

коды он скоро запоминает наизусть), затем оператор нажимает особую кнопку и на письме автоматически печатается адрес, набранный кодом, письмо передвигается в сторону, а на его место устанавливается новое.

Операторы работают настолько быстро (в больших городах до 90 писем в минуту), что механическое движение писем снижает скорость их работы; поэтому каждый оператор работает за двумя окошками, попеременно наблюдая за каждым из них. Если оператору необходимо посмотреть код, работа машины приостанавливается. Если оператор не может прочитать адрес или не может нанести код по каким-либо другим причинам, он нажимает вспомогательную кнопку и такое письмо направляется к группе писем, обрабатываемых вручную.

После того как письма закодированы, чтение кода и последовательная сортировка и раскладка писем выполняются уже автоматическими устройствами. Управление этими операциями осуществляется цифровой вычислительной машиной. Она размещается на нескольких небольших стойках, в то время как аппаратура автоматической обработки писем занимает, конечно, значительно больший объем.

Луисвилльская система фирмы «Дженерал Электрик». Компания «Дженерал Электрик» построила недалеко от г. Луисвилля, штат Кентукки, большой центр по производству электрических бытовых приборов. Этот центр состоит из пяти связанных единым производственным циклом фабрик, на которых работают 12 000 человек. В 1954 г., когда строительство этого комбината было завершено только частично и на нем работало около 5 000 человек, компания «Дженерал Электрик» установила там [10] систему большого масштаба для обработки данных, основой которой явилась быстродействующая электронная вычислительная машина. Использование этой машины (стоимость ее эксплуатации составляет полмиллиона долларов в год) большей частью основывалось на прежних методах, однако сама идея использовать ее для каждодневных коммерческих операций была новшеством.

Это новшество оказалось настолько удачным, что оно, несомненно, предвещает введение большого числа подобных систем на других предприятиях, и притом в гораздо меньших фирмах, чем «Дженерал Электрик». Компания же «Дженерал Электрик» рассматривает вопрос об установке в Луисвилле пяти таких вычислительных машин, по одной на каждой фабрике.

Так как этот эксперимент был «первым», работа системы осуществлялась в несколько неблагоприятных условиях; так, например, до тех пор пока люди не стали полностью доверять вычислительной машине, приходилось продолжать изготовлять в печатной форме некоторые старые ведомости отчетности, хотя никакой необходимости в них уже не было. При подсчете стоимости вычислений не учитывались такие реальные, хотя и неуловимые преимущества, как более быстрое и более точное получение данных, необходимых руководству для принятия решений.

Применение электронной вычислительной машины началось с решения на ней четырех перечисленных ниже основных задач, жизненно важных для предприятия; это при частично построенном комбинате загрузило машину работой только на 10 часов в неделю. Однако даже при таких неблагоприятных обстоятельствах отношение к вычислительной машине быстро изменилось. При предварительном анализе подсчитали, что после введения в действие всего комбината вычислительная машина снизит расходы на 500 000 долларов в год, даже если она будет использоваться только для решения четырех указанных основных задач, что загрузит ее только на 20 часов в неделю; в действительности же, конечно, она используется параллельно и для решения многих других дополнительных задач, принося этим самым дополнительный доход.

Четыре основные задачи таковы:

— составление платежных ведомостей, включая печатание чеков, отчетность по рабочей силе и учет элементов себестоимости;

— планирование материального снабжения и складской учет, включая ежедневную запись итогов и анализ предполагаемых изменений планов производства;

— учет заказов и составление счетов, включая выписку накладных;

— ведение общей отчетности и калькуляция себестоимости.

Обсуждается также ряд других строго коммерческих применений, включая подготовку годовых бюджетов и квартальных прогнозов, исследование существующего порядка учета сбыта и материального учета, а также подготовку графика загрузки фабричного оборудования. Дальнейшие планы предполагают составление долгосрочных прогнозов сбыта и комплексный «анализ динамики распределения». Кроме того, вычислительная машина используется для выполнения различных инженерных расчетов, каково, например, решение задачи по линейному программированию, описанной вслед за последним примером в § 25.1.

2.5. Наука

Цифровые вычислительные машины, включая входные и выходные устройства, являются системами большого масштаба; некоторые специалисты склонны их считать даже типичными по отношению к любым системам большого масштаба, однако такую точку зрения следует считать слишком узкой. Вычислительная машина сама, конечно, является компонентом некоторых других систем большого масштаба, рассматриваемых здесь. Ученые используют вычислительные машины в разных системах, применяемых при исследованиях. Особенно интересным примером служит обработка данных, полученных при испытаниях в аэродинамической трубе.

Аэродинамика сверхзвуковых скоростей настолько сложна, что значительная часть исследований должна осуществляться при помощи испытаний в аэродинамических трубах. Результаты таких испытаний также весьма сложны, и раньше для их обработки обычно затрачивалось несколько недель. Это приводило к тому, что экспериментатор проводил работы фактически вслепую. Он вынужден был планировать серию экспериментов, рассчитанных на те значения параметров, которые, по его мнению, представляли интерес; а затем он должен был проводить эти опыты и ждать, пока не будут обработаны данные, чтобы найти смысл результатов. При этом нередко получалось, что некоторые критические значения параметров оказывались пропущенными и экспериментатор должен был проводить новую серию подобных испытаний и снова ждать.

В настоящее время по крайней мере одна аэродинамическая труба оборудована системой автоматической обработки данных. В состав этой системы входит большая вычислительная машина, получающая данные непосредственно от измерительных приборов аэродинамической трубы и выдающая результат через несколько секунд или минут, так что экспериментатор может принимать разумные решения о значениях интересующего его параметра для следующего испытания.

Радиозонд [138]. Радиозонд представляет собой устройство для получения данных о метеорологических условиях в верхних слоях атмосферы. Запуск радиозондов полезен не только для определения специфических данных, необходимых для полетов на таких высотах, но и для получения общей картины состояния погоды, на основании которой можно составлять прогноз погоды.

Простейший радиозонд состоит из шара, наполненного легким газом и несущего метеорологические измерительные приборы и радиопередатчик; обслуживающее его наземное оборудование состоит из радиоприемника и устройства слежения для определения координат радиозонда, которое может быть либо оптическим (теодолит), либо радиоэлектронным (радиолокационная станция). Бортовые метеоприборы измеряют такие величины, как температура и давление; направление же и скорость ветра в верхних слоях атмосферы определяются непосредственным слежением за полетом свободно поднимающегося радиозонда.

Определение скорости ветра на основании показаний наземных устройств требует проведения вычислений, занимающих определенное время. Однако значительно более сложным является установление корреляции в показаниях различных радиозондов, выпущенных в свободный полет из разных пунктов, составление соответствующих изоплет и предсказание дальнейшего поведения воздушных масс.

Для того чтобы сделать бортовую аппаратуру радиозонда (которая часто теряется) как можно более дешевой, метеоприборы, так же как и радиолиния связи, делаются сравнительно простыми. Это приводит к необходимости применять сложные методы телеизмерения, в связи с чем наземная аппаратура принимает результаты измерений в виде кодовых многократных сигналов, что еще больше усложняет задачу обработки данных. Наконец, в целях корреляции результатов измерений данные, получаемые от нескольких радиозондов, выпущенных из различных пунктов, должны привязываться к одной точке.

Совершенно очевидно, что проектировщик системы при рассмотрении этой задачи должен выбрать тип вычислительного устройства, тип линии связи между бортовым передатчиком и наземным приемником, а также количество наземных пунктов, которые должны быть связаны в общую сеть. Он может также рассмотреть такие проблемы, как выбор правильного расположения наземных станций, вид радиосигналов телеизмерительной линии и целесообразность применения таких же кодовых сигналов при связи между наземными станциями.

Проектировщику системы могут также встретиться такие вопросы, как степень точности используемых метеоприборов и сравнение этой точности (которая не должна значительно ухудшаться в аппаратуре связи, аппаратуре обработки данных и вычислительной аппаратуре системы) со стабильностью

атмосферы (представляется нецелесообразным измерять что-либо ежечасно с такой высокой точностью, что данные будут существенно изменяться через 5 *мин*) и с точностью теории, используемой для предсказания погоды. Аналогично этому он может заняться сравнением относительных преимуществ использования большего числа наземных станций или их более тесного расположения, с одной стороны, и, например, более частого запуска радиозондов каждой станцией, с другой стороны.

2.6. Военная техника

Каждый вид вооруженных сил сталкивается с рядом проблем, решение которых связано с применением очень больших и очень сложных систем. С проблемой материально-технического обеспечения в ВВС мы уже познакомились. Основная проблема для сухопутных войск заключается в определении дислокации и действий противника. Это приводит к определенным системам большого масштаба, какой является, например, система радиолокационного наблюдения. Эта система составляет часть общей системы наблюдения за полем боя, которая, в свою очередь, является частью общей системы разведки. Военно-морской флот создал интересные системы для обнаружения и уничтожения вражеских подводных лодок. Он также столкнулся с огромной проблемой использования в военное время нашего большого торгового флота, включая решение таких вопросов, как целесообразность или нецелесообразность применения конвоев, выбор маршрутов, целесообразность стандартизации габаритов твердых грузов, оптимальное планирование и управление всей операцией.

Система предупреждения службы гражданской обороны [61]. Одной из нерешенных проблем Федеральной администрации гражданской обороны является проблема оповещения населения в случае воздушного нападения. При современном состоянии военной техники промежуток времени между моментом оповещения и началом воздушного нападения может составлять всего несколько часов; через несколько лет могут быть созданы межконтинентальные баллистические ракеты, и этот разрыв может сократиться до нескольких минут *. Количество потерь при таком воздушном нападении можно значительно сократить, обеспечив своевременное оповещение и укрытие населения в убежищах. Применение сирен является лишь частичным решением этой задачи.

В США около 99% домов в районах, представляющих цели для воздушного нападения (и почти такой же процент домов в сельских районах), имеют радиоприемники и телевизоры, которые можно было бы использовать для оповещения, если бы они всегда были включены и настроены на соответствующую частоту; однако, хотя рассматривался вопрос о специальных радиоприемниках, представляется, что применить для этой цели немодифицированные обычные радиоприемники нельзя.

Около 99% семей в районах целей и примерно такой же процент семей в сельских районах живут в домах, подключенных к коммунальной системе электроснабжения. Можно предложить несколько методов посылки специального сигнала по всем системам электроснабжения (несколько увеличивая или уменьшая частоту, применяя модуляцию различной частотой, посылая импульс напряжения или временно выключая систему). Можно также сконструировать очень простое устройство для каждой семьи, которое бы реагировало на такие изменения питающего напряжения, создавая соответствующий сигнал тревоги. Однако, к несчастью, система энергоснабжения обладает такими свойствами, что практическое осуществление любого из этих предложений является чрезвычайно сложной, а может быть, и вообще неосуществимой задачей.

Наконец, примерно 80% семей в районах целей и почти такое же количество семей в сельских районах имеют телефоны. Однако, как указывалось в §2.2, одновременно включить не только все, но даже значительную часть телефонов нельзя.

Таким образом, рассматриваемая задача создания системы связи, на первый взгляд допускающая несколько легко осуществимых решений, является в действительности чрезвычайно трудной. Уместно заметить, что оптимально сконструированные системы большого масштаба довольно трудно бывает эффективно использовать для других целей. Так, например, телефонная система, сконструированная для соединения лишь абонента с абонентом, отнюдь не пригодна для вещания. Систему энергоснабжения, рассчитанную на то, чтобы напряжение и частота оставались неизменными при больших и внезапных изменениях нагрузки, трудно приспособить для передачи заметного изменения частоты, напряжения или других параметров. Далее, обычные наши вещательные приемники, сконструированные в расчете на применение только в определен-

* Как известно, в настоящее время такие межконтинентальные баллистические ракеты уже существуют. — *Прим. ред.*

ные периоды времени при большом потреблении мощности, нельзя приспособить для непрерывного применения при небольшом потреблении мощности. Кроме того, они сконструированы для одновременного приема только одной станции и в них приняты специальные меры для подавления сигналов других станций, в связи с чем они не могут принимать сигнал, имеющий частоту, отличную от частоты станции, на которую они настроены.

Система зенитных управляемых реактивных снарядов «Ника». «Ника» * — зенитный управляемый реактивный снаряд (ЗУРС) ближнего действия, разработанный по заказу армии США для объектовой ПВО, т. е. для обороны одиночных объектов небольших размеров, таких, как города или основные военные сооружения. Ее основными элементами являются: радиолокационная станция сопровождения цели; собственная система обнаружения (которая может отсутствовать при подключении системы «Ника» к другой системе обнаружения), обеспечивающая обнаружение цели и передачу ее радиолокационной станции сопровождения цели; радиолокационная станция сопровождения снаряда; вычислительное устройство; сам реактивный снаряд и пусковые устройства. В системе используется наведение по командным сигналам, при котором траектория наведения выбирается наземными устройствами. Для разгона снаряда до крейсерской скорости, которая значительно превышает скорость звука, используются ракетные двигатели.

Рассмотрим кратко некоторые вопросы, которые надо решать при проектировании такой системы ЗУРС.

Должна ли траектория наведения выбираться на земле (наведение по командным сигналам), в воздухе (самонаведение) или отчасти на земле и отчасти в воздухе (наведение по радиолучу)? Или траектория наведения должна выбираться в начале полета на земле, а затем — в воздухе (наведение по командным сигналам на маршевом участке полета и самонаведение на последнем участке)?

Следует ли применить систему ближнего действия (что позволит создать более дешевые ЗУРС и, следовательно, закупить больше ЗУРС) или систему дальнего действия (площадь обороняемого района увеличивается пропорционально квадрату дальности действия)?

Должен ли снаряд сопровождаться радиолокационной станцией с момента запуска или он должен захватываться после того, как пролетит некоторое расстояние (в любом из этих случаев возникают специальные радиолокационные проблемы)?

Если мы остановимся на последнем, то рискнем ли мы применить радиолокационных станций сопровождения снарядов меньше, чем число снарядов, которые по нашим планам будут одновременно находиться в воздухе?

Будет ли на снаряде устанавливаться радиолокационный ответчик (при котором легче осуществлять сопровождение, а следовательно, представляется возможность уменьшить размеры и снизить стоимость радиолокационных станций сопровождения) или нет (т. е. этот отведенный нам объем и вес мы используем для каких-либо других целей)?

Если мы примем решение не применять ответчик, то для каких целей следует использовать освободившийся объем и вес: для увеличения топлива (что увеличит дальность действия ЗУРС), улучшения аппаратуры наведения (что позволит уменьшить ошибки сближения) или увеличения количества взрывчатого вещества (что повысит вероятность поражения цели при данной величине промаха)?

Должна ли радиолокационная станция сопровождения цели иметь узкий луч (при котором обеспечивается лучшая разрешающая способность и точность) или широкий (что облегчает захват цели)?

Должен ли двигатель реактивного снаряда быть ракетным (обеспечивает большую скорость, но мало экономичен), прямоточным воздушно-реактивным (позволяет получить меньшую скорость, но зато более экономичен) или турбореактивным (скорость еще меньше, но зато двигатель еще экономичнее)?

В дополнение к этим и многим другим аналогичным вопросам встают обычные технические вопросы (такие, как вопрос о рабочей частоте радиолокационных станций или о носовом или хвостовом расположении рулей управления), оказывающие большое влияние на проектирование системы, а также вопросы более общего характера (применять ли аналоговое или цифровое вычислительное устройство, использовать ли централизованное или децентрализованное управление и т. д.).

Система ПВО континента. Система «Ника» является частью системы объектовой ПВО армейского командования, которая имеет в своем составе различные виды оружия (систему «Ника», другие ЗУРС, неуправляемые реактивные снаряды, зенитные орудия), средства

* Греческое мифологическое название ракет «Ника» (Nike) транскрибируется в переводной литературе по-разному: пишут, например, «Найк», и т. п. Собственно американское произношение — «Найки». — *Прим. ред.*

наблюдения, аппаратуру передачи данных, а также аппаратуру для оценки обстановки и принятия решения о том, оружие какого типа наиболее целесообразно применить при данной конкретной ситуации. Эта система объектовой ПВО (или, вернее, целый ряд таких систем, по одной на каждом обороняемом объекте в Соединенных Штатах) является частью системы ПВО континента, которая включает средства зональной (районной) ПВО ВВС и находится в ведении командования ВВС.

При выборе характеристик каждой такой частной системы, входящей в общую систему ПВО, необходимо производить их оценку не только с точки зрения этой частной системы, но и с точки зрения общей системы. Так, например, решения относительно выбора рабочих частот радиолокационных станций и использования бортовых радиолокационных ответчиков будут обусловлены допустимым распределением частот. Кроме того, некоторые решения относятся ко всей общей системе: соотношение между ЗУРС дальнего и ближнего действия (конечно, в общем арсенале средств ПВО должны быть и те и другие); надежность работы радиолокационных станций, средств связи и другого оборудования, уязвимого для действий противника; решение о централизованном и децентрализованном управлении общей системой и т. п. (при этом центральные элементы частных систем являются периферийными элементами общей системы).

Наконец, система ПВО континента должна быть приспособлена для выполнения многих сложных функций, не возлагаемых на частные системы, входящие в нее. Она должна обеспечивать объединение информации, получаемой от многих разнородных источников, каковы телефонные донесения от корпуса наземных наблюдателей, сообщения радиопеленгаторных пунктов, а также донесения радиолокационных станций нескольких различных типов. Она должна также обеспечивать принятие решения о распределении целей между различными боевыми системами: ясно, что пилотируемый истребитель нельзя наводить на самолет, обстреливаемый управляемыми реактивными снарядами или зенитными орудиями. При принятии таких решений приходится учитывать много различных факторов. Так, например, зенитные орудия оказываются неэффективными против высоколетящих целей, в то время как некоторые ЗУРС, наводимые на цели при помощи радиолокационных средств, неэффективны против низколетящих целей, так как отражения от наземных предметов ухудшают качество работы аппаратуры наведения.

Совершенно ясно, что система ПВО является настолько большой и сложной, что мы могли бы лишь очень кратко рассмотреть ее проблемы, если бы даже такой материал и не содержал секретных сведений. Чтобы не возникало вопросов, насколько эта система велика, можно сказать, что на ее оснащение ежегодно расходуется несколько миллиардов долларов. И решения по системе должна принимать, опираясь на общий план работ, небольшая группа военных и гражданских инженеров-системотехников.

ГЛАВА 3

КОМПЛЕКСНЫЙ ПОДХОД К ПРОЕКТИРОВАНИЮ СИСТЕМЫ

Момент, когда проектировщик системы вступает в процесс проектирования, зависит от характера задачи. Вообще говоря, существует четыре момента вступления: формулировка задачи, определение входов, определение выходов и собственно разработка системы. Когда при проектировании системы основным требованием является снижение стоимости, первый вопрос гласит: что мы пытаемся получить за наши деньги? Другими словами, это значит, что в качестве измерителя эффективности, который очень тесно связан с формулировкой задачи, выбирается число долларов на единицу достигнутого результата. Это в свою очередь говорит о необходимости четкого определения тех результатов, которые мы хотим получить, проектируя систему.

Так, например, первое, что побудило компанию «Дженерал Электрик» заняться созданием системы управления для комбината в Луисвилле, была стоимость продукции, поэтому разработки начались с исследования вопросов стоимости. В других случаях первым объектом изучения являются входные возмущения.

Так, например, при проектировании системы ПВО мы должны в первую очередь исследовать типы вражеских самолетов, против которых нужно организовать оборону, определить их вероятные характеристики, тактику, возможные маршруты и т. д. За этим должен следовать выбор измерителя эффективности (в рассматриваемом случае таким измерителем, по-видимому, будет количество долларов

на одно ожидаемое уничтожение цели). При проектировании завода-автомата первая фаза исследования будет заключаться в изучении выходов, т. е. продукции, после чего необходимо будет перейти к определению входов системы и выбору измерителя эффективности. Наконец, в том случае, когда система уже существует, может оказаться целесообразным начать работу с изучения собственно системы и лишь затем перейти к изучению входов, выходов и измерителя эффективности; на сегодня примером такого подхода является внедрение дальнего набора номера в Белловскую телефонную систему.

В любом из рассмотренных случаев кто-то — руководитель предприятия, заказчик, потребитель или сам проектировщик системы — решает, что наступило время действовать, и организует работу по проектированию системы. Этот процесс лучше всего описать, предположив, что проектирование проходит через определенные фазы в хронологическом порядке; но при этом надо иметь в виду, что нередко фазу невозможно опознать, пока она не окончилась.

3.1. Первая фаза. Начало работы

Цель этой фазы, занимающей от одного дня до одного месяца, в том, чтобы положить начало работе. Бригада системного проектирования на этой стадии состоит из одного-трех человек. Руководитель бригады говорит с каждым имеющим отношение к исследуемой задаче, кто только захочет слушать и отвечать. Мысленно процесс проектирования уже начался; его можно разделить на две части: внешнее проектирование и внутреннее проектирование (рис. 3.1).

Трудно себе представить, чтобы в первой фазе проектирования системы сразу было найдено решение стоящей перед проектировщиками задачи, однако нельзя отрицать и того, что острый и глубокий ум может в некоторых случаях найти решение уже на этом этапе работы. Первые мысли почти всегда ошибочны, но они используются как хороший нож, позволяющий заострить наши вопросы, а также как своеобразный импульс, который позволяет добиться более полных ответов от молчаливых информаторов.

Некоторые проектировщики систем избегают принятия столь быстрого решения, исходя из того, что существует нелюбовь к априорным идеям и что лица, не участвующие в проектировании, могут сделать вывод о том, что заключения делаются слишком поспешно. Однако этого не следует бояться. Во-первых, каждый выдающийся исследователь имеет ту или иную априорную идею; опасность заключена не в этой априорной идее, а в нежелании отказаться от нее перед лицом фактов. Во-вторых, нет никакой необходимости оповещать всех о принятом решении. Это решение является вспомогательным орудием, первой гипотезой, и оно должно использоваться именно так.

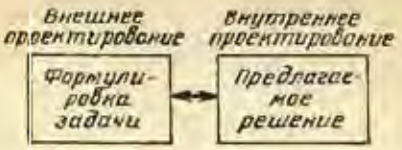


Рис. 3.1. Этапы проектирования системы — начальная фаза.

Выражения *внешнее проектирование* и *внутреннее проектирование* обозначают здесь зачатки двух основных разделов, на которые делится проектирование системы: раздела, занимающегося требованиями к системе и ее окружению, т. е. к вещам вне системы, и раздела, занимающегося конструктивными решениями, относящимися к оборудованию, способам работы и людям, т. е. к самой системе. В дальнейшем мы увидим, как эти разделы будут расти при переходе к каждой новой фазе работы.

Отчетность. Продукция бригады проектирования системы — это всегда документ. Удовольствие показать законченную вещь в «металле» должно оставаться привилегией инженера по оборудованию, проектировщика компонентов. Проектировщик-системник может сослаться только на свою долю в окончательной большой картине. Таким образом, результатом начальной фазы, равно как и других фаз, является документ. Этот документ должен содержать:

1) формулировку проблемы (в основном результат обсуждений, не подтвержденный экспериментом или детальным анализом);

2) предполагаемые варианты решений (с указанием преимуществ каждого из них);

3) предложения по необходимому составу специалистов и их количеству в бригаде проектирования системы, а также по срокам их подключения к работе (эта бригада будет в последующем вести проект);

4) грубую оценку времени и денежных средств, необходимых для завершения проекта.

Объем этого документа редко превышает несколько страниц.

3.2. Вторая фаза. Организация работы

После того как разработка проекта началась, работа в период организационной фазы имеет три цели: 1) развернуть бригаду проектирования системы до полного состава; 2) составить план всех работ над проектом; 3) приступить к выбору наилучшего решения. Эта фаза длится от двух недель до трех месяцев. Заканчивается она, когда полностью укомплектованная и хорошо функционирующая группа проектировщиков начинает разрабатывать более детально систему, первые слабые контуры которой стали обрисовываться уже раньше.

Бригада проектирования системы в общем случае должна состоять по крайней мере из пяти человек, что обеспечит достаточное количество точек зрения, необходимое для выбора правильных решений и исключения возможности ошибок. Эффективно работающая бригада должна состоять максимум из 12 человек; более крупные бригады имеют тенденцию разделения на фракции, отстаивающие разные идеи. Действительно, система допускает разные решения, и, чтобы работа могла продвигаться, необходимо выбирать из числа решений, кажущихся наиболее подходящими, какое-то оптимальное решение. Некоторая разумная часть сил должна быть направлена на альтернативные решения; однако, когда наступает время выбора, любое решение лучше, чем отсутствие решения.

Большинство или все члены бригады должны быть учеными-универсалистами того типа, о котором идет речь в нашей книге. Кроме того, каждый член бригады обычно является еще и специалистом в какой-нибудь узкой области, причем состав специалистов бригады должен подбираться очень тщательно. Обычно по крайней мере один член бригады должен быть специалистом по электронике, один — математиком, один — специалистом в той области, к которой относится рассматриваемая задача: аэронавигационным инженером при проектировании системы самолетного оборудования, инженером-связистом при проектировании телефонной системы или экспертом в области учета при проектировании деловой информационной системы. Другие члены добавляются смотря по наличию и потребности.

Рис. 3.2. Этапы проектирования системы—организационная фаза.

Опыт работы говорит о том, что в течение всего проектирования желательно обеспечить постоянство состава бригады.

Планы выполнения всего проекта системы должны содержать график распределения времени и кривую роста денежных средств, персонала и материалов, необходимых для завершения проекта, хотя следует иметь в виду, что на этой стадии работы такие оценки могут содержать значительные ошибки. Как правило, ошибки делаются в сторону недооценки потребности во времени, средствах и людях, если только бригада проектирования системы не состоит из очень опытных специалистов. Планы работ должны включать также достаточно подробные заявки на персонал для эксплуатационных испытаний, на исследовательские группы и группы по оборудованию, причем группы по оборудованию должны быть подразделены на группы, входящие в основную организацию, и группы, работающие по субдоговорам. Кроме того, в планах обычно указывается, какие консультации специалистов желательно организовать.

Приступая к решению проблемы, бригада осуществляет ряд этапов проектирования, подобных показанным на рис. 3.2. Каждый из нарисованных прямоугольников обозначает предмет одной из дальнейших глав. Формулировка задачи включает в себя выбор измерителя (критерия) эффективности и определение проектных критериев (гл. 9).

Многократно пересматриваемая формулировка задачи отражает в ее последовательных вариантах состояние знаний о требуемой системе. Математическая модель служит средством, помогающим лучше разобраться в окружении (окружающей среде) системы и предлагаемых решениях; она рассматривается подробно в гл. 10. Планирование экспериментов описывается в гл. 11, но так как бригада не является полной и так как у нее пока нет людей для обслуживания аппаратуры и для исследований, то с опытами приходится подождать.

Здесь мы вынуждены познакомить читателя с одним положением, к которому будем возвращаться на протяжении всей книги. Система должна отвечать на входы так, чтобы она: 1) замечала, что с входами каждого ти-

па происходят нужные вещи для получения требуемого выхода; 2) делала это в нужное время при множественных входах, часто случайно распределенных во времени; 3) если входы стремятся разрушать систему (война или конфликт любого другого вида), то специальное внимание должно быть обращено на выбор ответных действий, которые в этом случае могут быть множественными и даже случайными, т. е. различными при тождественных входах и тождественных обстоятельствах. Опыт говорит о том, что эти стороны проектирования системы в практической работе допускают разделение. Работу, связанную с первой стороной, мы называем *проектированием единичной нити* (гл. 22); работу, связанную со второй стороной, — *проектированием большой нагрузки* (гл. 23); и, наконец, работу, связанную с третьей стороной, мы называем *состязательным проектированием* (гл. 24).

При проектировании системы эти этапы работы должны осуществляться одновременно в течение всего процесса проектирования. Соединительные линии на рис. 3.2 указывают логические, но не хронологические соотношения.

Отчетность. В течение этой фазы идеи изменяются очень быстро, и без надлежащих отчетов будет трудно держать всех, кто связан с проблемой, в курсах новых разработок. Ходом работы интересуется руководство и заказчик, а также и начальник отдела оборудования, который может ожидать, что на него в ближайшее время будет возложена определенная часть работы; сами члены бригады проектирования системы также будут получать информацию из этих отчетов. Для составления отчетов на этой стадии наиболее удобны периоды приблизительно в две недели.

Документ, подводящий итог организационной фазе, кратко описывает в начале состав бригады проектирования системы, а в конце — планы работы под проектом. Основное содержание отчета должна составить весьма ясная формулировка проблемы, насыщенная гораздо большим содержанием, чем это было возможно раньше, и несколько разумных решений проблемы, из которых одно обычно выбирается как наиболее подходящее.

Этапы проектирования системы. Цикл выполнения этапов проектирования системы начинается на организационной фазе. Эти же самые и еще немногие дополнительные этапы будут неоднократно повторяться на дальнейших фазах проектирования, каждый раз с дополнительным усовершенствованием. Этапами внешнего проектирования системы являются формулировка проблемы, создание математической модели, планирование экспериментов и, наконец, проведение экспериментов.

Первичными этапами внутреннего проектирования системы являются проектирование единичной нити, проектирование большой нагрузки и, наконец, состязательное проектирование. Дальнейшие этапы включают выявление подсистем и их проектирование. Эти последние этапы содержат большое число меньших этапов, часть которых будет упомянута в последующих главах книги; но в силу того, что эти этапы изменяются от системы к системе, мы не перечисляем их здесь.

3.3. Третья фаза. Предварительное проектирование

Цель предварительного проектирования заключается в разработке первого варианта того, что действительно может быть названо системой. Эта фаза может продолжаться от двух месяцев до двух лет и считается законченной, когда будет разработано подробное техническое (функциональное) задание системы. Этапы работы, показанные на рис. 3.2, расширяются, как показано на рис. 3.3.

Для выполнения экспериментов, планирование которых было начато раньше, требуется достаточное число обслуживающих групп (под *обслуживающими группами* понимаются группы, обслуживающие работу основной бригады проектирования системы). Эти эксперименты доставляют данные, позволяющие создать лучшую математическую модель, и одновременно ту необходимую интуицию, которую можно приобрести путем испытания в реальных условиях эксплуатации.

В то же самое время группы оборудования начинают всматриваться в проблему проектирования с теми небольшими долями инженерного таланта, которые отмерены исследователям частных задач. Предложенная система никогда не может быть продуктом творчества только универсалистов. Системники без помощи инженеров-аппаратчиков будут, конечно, бесплодной группой. Задачи, возникающие при предварительном проектировании и затрагивающие группы по оборудованию, приводят либо к анализу и исследованию, либо к экспериментам, которые должны разрешить спорные вопросы о работе аппаратуры.

При проектировании любой сложной системы, даже когда делается попытка оставаться в рамках достаточно хорошо известной аппаратуры или оборудования, всегда возникает потребность в проверке той или иной идеи, которая в большей или меньшей

Рис. 3.3. Этапы проектирования системы—фаза предварительного проектирования.

степени представляет собой экстраполяцию существующей техники в новые условия. Такие скачки в области аппаратуры требуют, чтобы идея получила доказательство, и эти доказательства даются ссылками на критические опыты. Крайне важно, чтобы результаты новых теоретических исследований и экспериментов находили применение в системе.

Фаза предварительного проектирования является такой стадией работы, когда решаются вопросы, допускающие лишь многозначные ответы. Многие из этих вопросов возникают неоднократно при проектировании различных систем. Будет ли система (и в первую очередь вычислительное устройство, но также и другие части) аналоговой или цифровой? Должно ли управление системой осуществляться централизованно, децентрализованно или представлять собой комбинацию этих двух методов? Имеет ли смысл стандартизировать каждый вход или следует допустить использование входов различных видов? Или, что еще труднее, где следует осуществлять эту стандартизацию? Заменять ли ту или иную функцию человека автоматической функцией? Какую лучше применить индикацию — оптическую или акустическую? Не лучше ли тормозить входы применительно ко всей системе в целом или к какому-либо из ее компонентов и обрабатывать их в порядке очереди или следует применить параллельные каналы? Или, что в технике связи то же самое, сделать ли канал широкополосным или узкополосным, но с задержкой? Следует ли делать данную установку в военной системе стационарной или подвижной? Должна ли вся информация доводиться до сведения всех, чтобы на основе ее каждый из операторов сам определял свои действия, выбирая, что ему в этой информации нужно (циркулярное управление), или следует адресовать передаваемую информацию, указывая, кому из операторов она предназначена (коллективная линия), или, наконец, для передачи информации каждому из операторов следует выделить отдельный канал связи (индивидуальная линия)? Какой срок службы должны иметь компоненты? Как велик должен быть наименьший сменяемый блок?

Эти вопросы составляют существо проектирования системы, и системотехника необходима именно потому, что ответы на эти вопросы отнюдь не являются делом простого выбора. Всегда имеются ограничения, недостаток информации, непредвиденные будущие события — короче говоря, ответы всегда неоднозначны.

Отчетность. На протяжении фазы предварительного проектирования отчеты должны делаться примерно один раз в месяц. Каждый отчет сообщает о разных мелких изменениях в идеях и о некоторых подробных исследованиях и экспериментах. Однако основные усилия должны быть направлены на подготовку документа, который содержал бы в себе первый вариант технического (функционального) задания системы. Когда система понята

достаточно хорошо, разработка такого задания будет относительно несложным делом.

Первая часть такого документа должна последовательно знакомить читателя с обоснованиями соответствующих решений, принятых проектировщиками, в то время как в последней части должны содержаться предложения относительно необходимого числа людей, необходимого времени и материалов, влияния окружающих факторов и т. п. Документ содержит следующий материал:

1) достаточно подробное описание работы всей системы в целом;

2) четкое описание подсистем;

3) для каждой подсистемы:

а) полное описание формы, числа и времени появления её входа и выхода;

б) ясное и полное описание её функционирования, т. е. операций над входом, необходимых для производства выхода;

в) перечень предельно допустимых общих габаритов, весов и т. п.;

г) по крайней мере один метод физической реализации предложенного способа функционирования в пределах указанных ограничений плюс любая дополнительная информация об исследованиях, посвященных другим методам.

Документ должен содержать последовательные — от этапа к этапу — рассуждения, чтобы рассеять возможные сомнения относительно правильности проектирования и предвосхитить всегда возникающие вопросы относительно полноты исследования системы. Конечно, если этот документ обнаружит плохие мысли или отсутствие мыслей по поводу какого-либо пункта, возбуждающего сомнение у проектировщика аппаратуры, то проектировщик аппаратуры захочет и будет должен провести дополнительные исследования. Это потребует дополнительного времени и может даже оказаться катастрофическим в случае серьезного просчета, что еще раз указывает на особую важность надлежащего предварительного проектирования.

Описание метода, каким достигается физическая реализация предложенного способа функционирования, необходимо для того, чтобы инженеры-аппаратчики имели уверенность в практической осуществимости системы. С другой стороны, проектировщику аппаратуры не следует говорить, как он должен решать свою задачу. Проектировщик-системник является универсалистом и обычно недостаточно хорошо подготовлен для того, чтобы сделать окончательный выбор компонентов. Конечно, функциональное задание не должно нарушаться. Оно отображает системный подход, требование к системе в целом. Оно не может быть создано по принципу «как попало».

Части системы. Любой компонент любой системы может быть отнесен к одному из следующих шести типов: вход, связь, логическое управление, рефлексивное управление, подача и выход.

Входное оборудование предназначается непосредственно для приема входов. Таким оборудованием могут быть чувствительные устройства, например радиолокационные станции, номеронабиратель телефонного аппарата или входное устройство вычислительной машины. Выходное оборудование подразделяется на два резко различных типа: на аппаратуру индикации и исполнительные органы. К устройствам подачи мы относим транспортные средства; в самолетных системах эти средства могут оказаться критическими компонентами.

Различие между двумя типами управляющих устройств является принципиальным; термины для их обозначения взяты нами по аналогии с человеком, который совершает определенные рефлекторные действия (например, отдергивание руки от горячей печки) мгновенно и просто, но совершает более сложные действия (например, решение задач) более медленно и с участием умственной деятельности. Линия разграничения между средствами управления этих двух типов будет, конечно, весьма условной; однако устройства одного типа удобно рассматривать под рубрикой системной логики, а устройства другого — под рубрикой автоматического регулирования.

Часть системы — это отнюдь не то же самое, что подсистема. Хотя часто и бывает удобно думать о логическом управлении как о подсистеме, тем не менее редко оказывается удобным думать о связи как о подсистеме. В самом деле, некоторые подсистемы могут содержать в себе все вышеуказанные части.

3.4. Четвертая фаза. Основное проектирование

Цель основного проектирования заключается в уточнении функционального задания. Эта фаза может продолжаться от одного года до десяти лет и заканчивается, когда получено *проектное задание*. Это задание является в основном функциональным и во многом похоже на то задание, которое было получено в конце предыдущей фазы. Вместе с тем оно имеет два важных отличия: оно, во-первых, детализировано в значительно

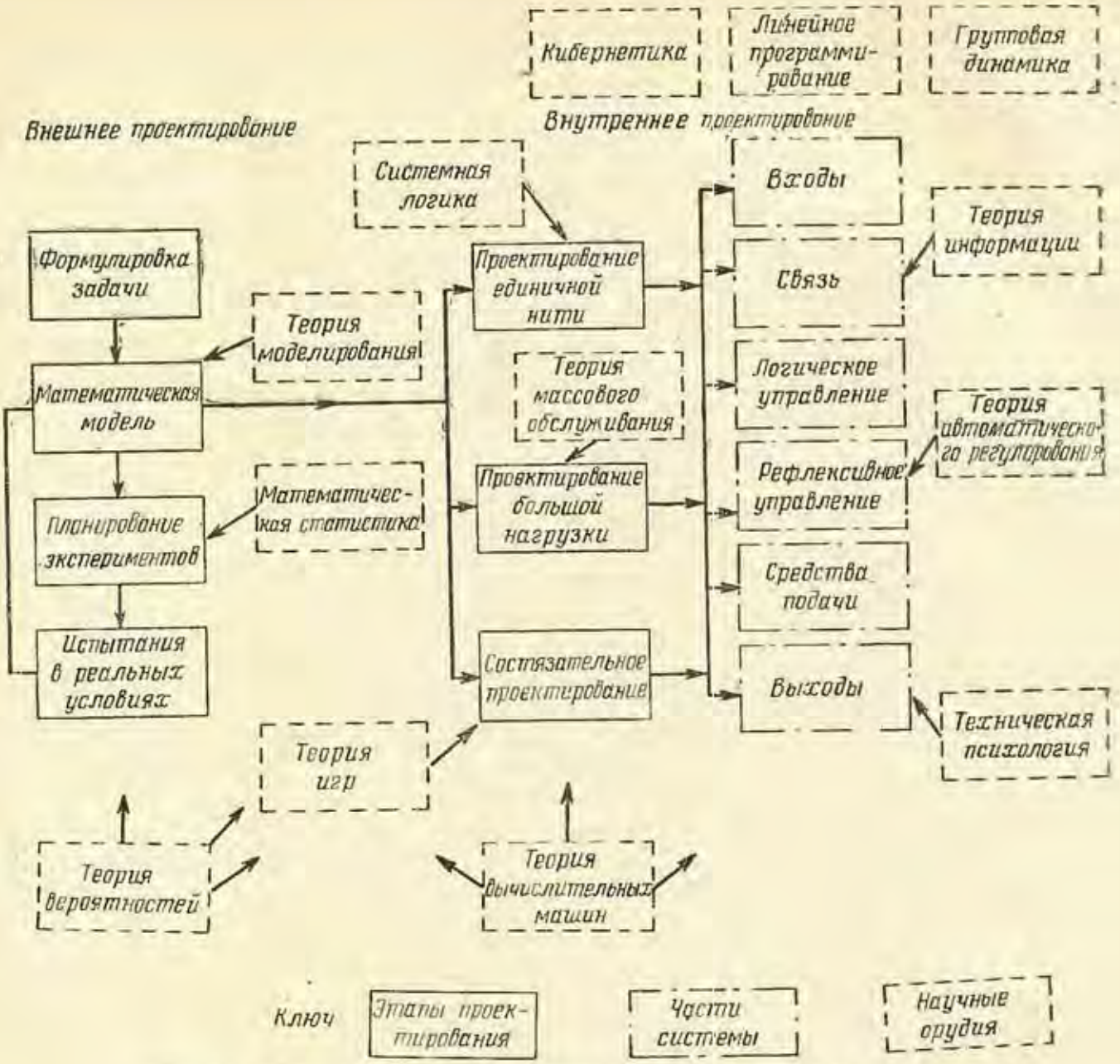


Рис. 3.4. Этапы проектирования системы—фаза основного проектирования.

большей степени и, во-вторых, является «замороженным»; т. е. бригада проектирования системы согласилась, что разработчики оборудования могут идти дальше и конструировать образец без каких-либо новых изменений задания, кроме тех, которые будет нужно ввести по требованию самих разработчиков оборудования.

Основное проектирование является своего рода отделочной работой. Как гончарный круг используется для обработки и придания необходимой формы куску глины, предварительно сформованному вручную, так и процесс проектирования системы по мере его осуществления обрабатывает и формует окончательный вариант системы. И как гончар работает то над одной частью, то над другой, так же поступает и разработчик системы.

Во время этой фазы проект разрастается до размеров, которые требуют полностью укомплектованных групп оборудования, либо входящих в проектную организацию непосредственно, либо работающих по субподрядным договорам. Общее число людей, занятых проектированием системы, в этот период работы может достигать десятков, сотен и тысяч. В процесс проектирования должен включаться персонал для испытаний системы в реальных условиях эксплуатации. Сборка монтажа на столе, создание отдельных образцов аппаратуры, проверка и испытание определенных групп оборудования, проектирование и перепроектирование — все это является

элементами процесса проектирования системы.

Во время этой фазы работы используются самые разнообразные научные орудия. Каждой части системы и каждому этапу процесса проектирования необходимо уделять большое и постоянное внимание. Зависимости между этапами схематически показаны на рис. 3.4.

Отчетность. Во время основного проектирования системные бригады должны докладывать о результатах своей работы примерно каждые три месяца. Группы оборудования должны докладывать о своей работе приблизительно столь же часто, но это зависит от характера проблемы. Продуктом основного проектирования является проектное задание, тщательно сформулированное таким образом, чтобы оно удовлетворяло требованиям заказчика, дирекции, руководства проектными работами, проектировщика системы и, что самое главное, изготовителя оборудования. Проектное задание должно быть вполне готовым для конструирования опытного образца системы, хотя это конструирование и могло быть начато несколько раньше.

Орудия проектирования систем. В принципе нет такой научной дисциплины, которая не могла бы потребоваться при проектировании какой-либо системы большого масштаба. В нашей книге мы уделяем особое внимание тринадцати следующим орудиям: теории вероятностей, математической статистике, теории вычислительных машин, системной логике, теории массового обслуживания, теории игр, линейному программированию, кибернетике, групповой динамике, теории моделирования, теории информации, теории автоматического регулирования, технической психологии.

Эти научные орудия были выбраны потому, что они имеют самое непосредственное отношение к процессу проектирования систем, очень широко используются для самых различных систем и в то же время сравнительно новы (их развитие началось в основном с 1940 г.) и поэтому мало известны. Последний довод не относится, конечно, к теории вероятностей; однако эта дисциплина настолько необходима, что каждый проектировщик системы должен быть очень хорошо знаком с ней. Математическая статистика занимает промежуточное положение: большинство рассматриваемых нами вопросов было разработано в течение XX столетия, но преимущественно до 1940 г.

На рис. 3.4 схематически показаны взаимосвязи восьми из этих орудий с различными этапами работы и частями системы. Три из пяти оставшихся орудий еще нельзя отнести к какому-либо этапу работы или части системы, и они показаны на рисунке изолированно. Остальные два орудия, а именно теория вероятностей и теория вычислительных машин, настолько широко применяются, что они применимы к подавляющему большинству этапов работы и частей системы, а также тесно связаны с другими орудиями.

Так, теория вероятностей является основой математической статистики, теории информации и теории массового обслуживания, и ее понятия необходимы в теории моделирования и в теории игр; теория вычислительных машин лежит в основе системной логики и моделирования и используется для решения задач теории игр, линейного программирования и теории массового обслуживания. Этим двум фундаментальным научным орудиям будут посвящены две из шести частей книги.

Между упомянутыми нами научными орудиями существуют и другие взаимосвязи. Теория массового обслуживания тесно связана с теорией моделирования, и каждая из этих дисциплин обязательно необходима для правильного понимания другой. Аналогично этому кибернетика связана с теорией автоматического регулирования, с одной стороны, и технической психологией, с другой. Все эти взаимосвязи станут более понятны по мере знакомства читателя с материалом книги.

3.5. Пятая фаза. Конструирование опытного образца

Цель этой фазы состоит в создании лабораторного опытного образца системы. Фаза длится не более шести месяцев для малых и двух лет для больших систем. Заканчивается она, когда создается опытный образец, готовый для проведения испытаний. Термин *опытный* используется нами для наименования широко различающихся между собой типов конструкций, охватывающих весь диапазон между простыми макетами на столе и заводскими образцами, а иногда даже и перекрывающих его. Мы будем только проводить различие между лабораторными и заводскими образцами.

Лабораторный образец, предназначенный и подготовленный для проведения испытаний, изготовляется обычно вручную. Он содержит одну полную единичную нить системы, снабженную обычно устройствами для моделирования дополнительных входов, а также наиболее ответственные (критические) фрагменты из аппаратуры большой нагрузки системы.

Так, например, вычислительная машина с разделением по времени должна испыты-

ваться полностью вместе с входными и выходными устройствами (которые, однако, могут иметь в своем составе по одному экземпляру различных повторяющихся компонентов); испытывать единичную нить с помощью более медленной вычислительной машины, пригодной только для одного входа, было бы недостаточно. Образец должен также иметь в своем составе контрольную аппаратуру. В хорошо исполненном образце контрольную аппаратуру должны составлять не лабораторные приборы, а опытные образцы той контрольной аппаратуры, которая фактически будет встроена в производственные модели системы.

Количество персонала, входящего в бригаду проектирования системы на этой фазе работы, не следует изменять, несмотря на то, что при проектировании систем большого масштаба в работе могут быть заняты тысячи людей. Если проектирование выполнено качественно, продолжительность этой фазы работы будет составлять лишь сравнительно небольшую часть общего времени. Однако если проектирование выполнено недостаточно хорошо, что всегда случается при работе «методом проб и ошибок», это конструирование может занять большую часть общего времени разработки. Очень часто потребность иметь систему уже «в металле» слишком велика, и, желая скорее закончить разработку, слишком рано принимают решения о начале этой фазы работы, что приводит к большим издержкам времени и средств.

Отчетность. В период конструирования образца отчеты должны выдаваться через каждые три месяца. Кроме того, уже на начальном периоде этой фазы необходимо начать разработку технических руководств по эксплуатации и уходу за самим образцом. Если эти руководства не будут созданы к моменту поступления образца на испытания, можно бесцельно потерять много месяцев или, что еще хуже, испытания могут быть организованы неправильно и по разработанной системе будет составлен не отвечающий действительности неблагоприятный отчет.

3.6. Шестая фаза. Испытание, отладка и оценка

Мы уже говорили, что существует значительное количество различных образцов. Не меньше и различных видов испытаний: *лабораторные испытания, испытания у заказчика* и т. п.

Мы будем различать здесь только испытания лабораторного образца, изготовляемого вручную и обслуживаемого непосредственно создавшими его инженерами (или под их наблюдением), и «оценку» производственного образца, который изготовляется (или может быть изготовлен) обычными заводскими методами и обслуживается теми, кто будет эксплуатировать его в реальных условиях. Обучение таких операторов следует начинать уже на этапе конструирования образца, и этих операторов нужно по возможности использовать при проведении испытаний.

Всегда возможно несколько решений любой проблемы, связанной с созданием системы, и почти каждое из этих решений можно реализовать практически, затратив на это определенное время, усилия и средства. Основная цель испытаний заключается в том, чтобы получить подтверждение, что аппаратура работает так, как это предусматривалось при ее проектировании, и исключить неизбежные технические дефекты. Но нет никаких сомнений в том, что система будет в конечном счете работать и что она не будет отлично работать на начальной стадии испытаний (если она вообще будет работать на этой стадии).

Если проектирование системы осуществлено правильно, то в течение нескольких месяцев или года система будет *проверена* и можно будет приступать к изготовлению заводского образца (в действительности к его изготовлению часто приступают еще тогда, когда испытания лабораторного образца только начинаются). Если в процессе испытаний будут выявлены серьезные недостатки в проектировании системы или если фаза конструирования образца будет неоправданно сокращена из-за желания как можно раньше начать испытания, то испытания могут растянуться на годы; в общем случае такие действия приводят обычно к бесполезной потере времени и средств.

Оценка заводского образца оказывается еще более предрешенной, чем суждение о результатах испытаний. Цель оценки состоит в том, чтобы определить, соответствует ли конструкция своему назначению. Однако практически это никогда нельзя сделать для системы большого масштаба, развернутой в реальных условиях ее эксплуатации. Если проектирование системы проведено достаточно хорошо, то в случае, когда система работает (в том значении, которое придают этому понятию при испытаниях), она отвечает своему назначению (в том смысле, которое должно быть придано этому утверждению при оценке системы).

Оценку системы большого масштаба можно осуществить только до начала ее строительства. Совершенно очевидно, что в течение времени, прошедшего от начала работы до оценки системы и составляющего от двух до пятнадцати лет, произойдет много различных изменений: может измениться сама проблема; произойти непредвиденные, но весьма важные изменения в технологии производства; во время проектирования систем возникнуть много новых хороших идей. Все это может привести к целесообразности повторения процесса проектирования, который снова может продолжаться от двух до пятнадцати лет. В результате оценка, осуществляемая в конце процесса проектирования системы, частично перекрывается с оценкой, производимой в начале этого процесса.

ЧАСТЬ 2

ТЕОРИЯ ВЕРОЯТНОСТЕЙ — ОСНОВНОЕ ОРУДИЕ ВНЕШНЕГО ПРОЕКТИРОВАНИЯ СИСТЕМ

ГЛАВА 4

ОСНОВНЫЕ ПОЛОЖЕНИЯ

В классической логике предложения являются либо истинными, либо ложными, без какой-либо возможности сомнения. Мы говорим: «Если будет дождь, я пойду в кино, в ином случае — нет». Фраза свидетельствует о неуверенности говорящего в своих будущих действиях, однако мера этой неуверенности отсутствует. Я не решил окончательно, что буду делать; мои действия зависят от события (будущего состояния погоды), относительно которого я нахожусь в настоящий момент в неведении. Мое незнание может быть почти полным (если дождь будет в день визита Джо, когда бы это ни случилось) или же частичным (если дождь будет через час от текущего момента, в чем можно быть почти полностью уверенным, так как об этом говорит вид неба, барометр и имеющиеся данные о положении грозового фронта и его движении на запад). Мы хотели бы уметь выразить степень вероятности, которую следует придать безоговорочному утверждению «я пойду в кино», так как исход «зависит от случая».

Чтобы определить, что такое случай, случайность, представим себе некоторый эксперимент, реальный или мысленный, который может дать нам два или больше исходов в условиях, являющихся фактически или по наилучшему нашему знанию вполне тождественными. При таких условиях говорят, что исход является случайным. Следует, конечно, обратить внимание, насколько субъективно это определение, зависящее от объема знаний. Вероятность есть мера случайности.*

Рабочее определение числовой вероятности события приводится ниже, однако его нельзя считать вполне удовлетворительным,

---

* Авторы исходят из субъективной концепции вероятности, когда вероятность понимается как мера незнания или мера сомнения; теория вероятности сближается при этом с многозначной («модальной») логикой. Существует, однако, объективная концепция вероятности, когда вероятность понимается как мера объективной случайной связи событий. Исследование субъективных состояний «вероятности» может, конечно, иметь определенный интерес и не обязательно должно связываться с идеалистической философией, однако для применений теории вероятностей в технике и естествознании наиболее адекватной является объективная концепция вероятности.

Субъективная концепция вероятности была весьма распространена в XIX в., однако в наше время в отечественной математике прочно утвердилась объективная концепция. Так, Б. В. Гнеденко пишет в работе [Д. 9]: «Каждый исследователь, имеющий дело с применениями теории вероятностей к физике, биологии, артиллерийской стрельбе, экономической статистике или любой другой конкретной науке, в своей работе исходит по существу из убеждения, *что вероятностные суждения выражают собой некоторые объективные свойства изучаемых явлений*. Утверждать, что при некотором комплексе условий появление события A имеет вероятность p, это значит утверждать наличие между комплексом условий $\mathfrak{S}$ и событием A некоторой вполне определенной, хотя и своеобразной, но от этого не менее объективной, существующей независимо от познающего субъекта связи».

В рассматриваемом примере будущий дождь есть событие A, а тот или иной, но каждый раз один вполне определенный набор характеристик текущего состояния погоды (давление, облачность, движение грозового фронта и т. п.) есть комплекс $\mathfrak{S}$. Событие A *случайно* относительно *данного* комплекса $\mathfrak{S}$, если при осуществлении комплекса $\mathfrak{S}$ оно может произойти, а может и не произойти; при этом по отношению к другим комплексам характеристик погоды $\mathfrak{S}'$ событие A может оказаться уже не случайным, а необходимым или, наоборот, невозможным. Таким образом, вместо нашего знания и незнания мы рассматриваем здесь полноту или неполноту связи между комплексом $\mathfrak{S}$ и событием A. *Вероятность как мера случайности есть мера этой объективной полноты связи.*

С этой точки зрения при изучении случайности в эксперименте со многими исходами нас должен интересовать не столько объем наших знаний об условиях эксперимента, сколько точное выявление нужного нам исходного комплекса $\mathfrak{S}$ в этих условиях. — *Прим. ред.*

так как в нем употребляется слово «возможно», которое, конечно, означает «вероятно». Определить понятие вероятности так же трудно, как и другие наиболее фундаментальные научные понятия *. Такие единицы, как единицы времени и длины, определить очень трудно; однако, несмотря на это, они постоянно используются инженерами с большой пользой. Для наших целей достаточно указать, что возникающая в этом случае трудность вызвана применением математических определений к физическим явлениям **.

Для инженера часто бывает вполне достаточно знать только, что теория дает приемлемое предсказание физических явлений и что ничего лучшего пока еще нет. Ему важно лишь, что теория вероятностей «работает». Инженерное оправдание геометрии Евклида заключается не в ее истинности, а в том факте, что мосты, спроектированные с применением законов этой геометрии, не рушатся. Справедливость теории вероятностей устанавливается на той же самой утилитарной основе ***.

Однако неясности по части философских основ теории вероятностей не должны вести к беззаботности по отношению к тем исходным допущениям, которые делаются в этой теории. Значительная часть трудностей, возникающих при применении теории вероятностей, вызвана пренебрежением к основным допущениям. Кажущийся парадокс в задаче 8.1 перестает быть парадоксом, если читатель ясно представляет себе допущения, принимаемые при определении случайной переменной.

4.1. Определение числовой вероятности

Если эксперимент (аналогичный тому, о котором шла речь при определении вероятности) может иметь n различных исходов (или результатов), причем эти исходы являются несовместимыми, исчерпывающими и равновозможными, и если *благоприятными* являются m исходов, то мы можем сказать, что вероятность благоприятного исхода равна m/n, и написать

$$P = \frac{m}{n} \text{ или } P(Б.И.) = \frac{m}{n}. \qquad (4.1)$$

Каждое из вышеуказанных ограничений является существенно важным. Выражение *«несовместимые»* говорит о том, что можно получить лишь не более одного исхода; выражение *«исчерпывающие»* указывает, что по меньшей мере один исход наступит и что этот исход будет одним из n перечисленных выше исходов; следовательно, эксперимент должен иметь в точности один исход. Слово *«благоприятный»* в теории вероятностей часто используется не в обычном значении этого слова и обозначает немного больше, чем «имеющий отношение к...». Более точно следовало бы сказать: «Если m исходов из n суть собы-

---

* По логическим основаниям теории вероятностей существует обширная литература. Этот философский вопрос рассматривается с различных точек зрения. Одни авторы, к которым относятся Райхенбах [18] и фон Мизес [19], пытаются определить вероятность на частотной основе, т. е. если количество экспериментов стремится к бесконечности, то вероятность благоприятных результатов можно определить как предел доли экспериментов с благоприятным исходом. Вторая группа авторов, представленная Карнапом [20], Джеффрисом [21] и Кейнсом [22], определяет вероятность как логическое отношение, аналогичное логической импликации, но допускающее градации. Третья группа авторов, в частности Купмен [23] и Колмогоров [24], пытаются определить вероятность на аксиоматической основе. Они считают, что вероятность есть игра, разыгрываемая по определенным правилам, разработанным на строгой математической основе. Наше определение вероятности больше всего походит на последнее определение. — *Прим. авт.*

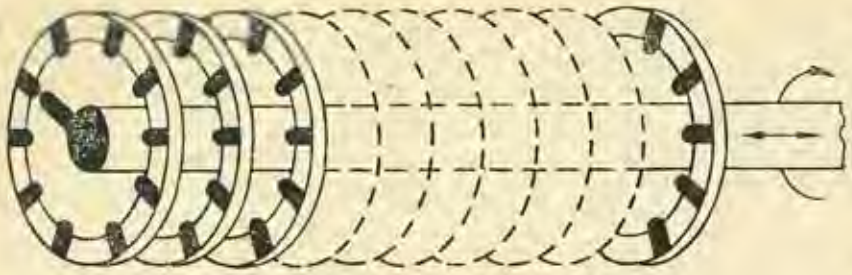
** Аксиоматическое построение теории вероятностей, предпринятое русским математиком акад. А. Н. Колмогоровым и другими, следует общим принципам аксиоматического построения математических наук и в этом отношении является необходимым этапом развития теории вероятностей, подытоживающим всю ее предыдущую историю.

Долгое время теория вероятностей представляла собой не вполне сложившуюся в математическом смысле науку, без четко выделенных исходных понятий и положений, что часто приводило к парадоксам в вероятностных выводах и рассуждениях. Развитие приложений привело к необходимости строгой математической формулировки теории.

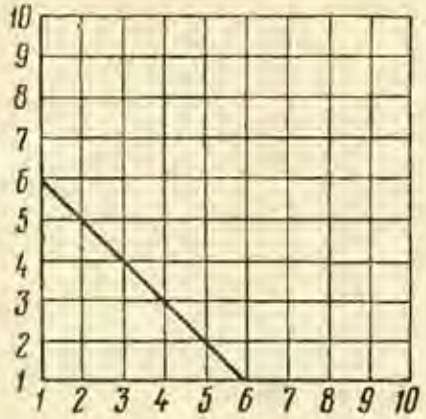
Сравнение аксиоматического метода с «игрой», конечно, не более, чем метафора: ведь аксиомы выбираются не произвольно, а так, чтобы они обобщали весь имеющийся опыт изучения вероятностных явлений и связей.

Классическое определение вероятности с помощью понятия равновозможности (см. следующий параграф), частотное определение и другие специальные определения, в том числе «субъективные», суть лишь частные интерпретации и попытки интерпретации аксиоматического определения. В этом отношении их нельзя ставить на одну доску с аксиоматическим определением [Д. 9, Д. 10]. — *Прим. ред.*

*** Теория тем и доказывает свою истинность, что она успешно применяется на практике (критерий практики в теории познания). Однако, как известно, необходимо различать истину абсолютную и истину относительную. Например, геометрия Евклида относительно истинна в определенных пределах (в обычных земных условиях) и относительно ложна в более широких пределах (в космических масштабах). Так же относительно истинны и классическая и современная (аксиоматическая) концепции вероятности. — *Прим. ред.*



а)



б)

Рис. 4.1. Шаговый переключатель:
а—схематическое изображение; *б*—эквивалентная решетка.

тие A, то вероятность этого события A равна m/n».

Для раскрытия смысла этого определения рассмотрим работу шагового переключателя, применяемого в качестве линейного искателя в телефонной системе [33]. Переключатель подключен к 100 абонентским линиям и состоит из 10 рядов контактов, по 10 в каждом ряду, как показано на рис. 4.1,*а*. Подвижный контакт (щетка) движется от одного ряда к другому, затрачивая 0,1 *сек* на подход к первому ряду и 0,1 *сек* на переход к каждому последующему ряду. Когда подвижный контакт достигает требуемого ряда, он совершает круговое вращение, затрачивая 0,1 *сек* на каждый переход с одного неподвижного контакта на соседний. Перемещение подвижного контакта происходит до тех пор, пока он не попадет на неподвижный контакт, соединенный с требуемой абонентской линией.

Задача заключается в том, чтобы рассмотреть первый телефонный вызов, который состоится после пяти часов, и определить вероятность того, что для подключения (соединения) требуемой линии этот линейный искатель затратит менее 0,85 *сек*.

Для решения задачи мы сначала произведем топологическое преобразование шагового переключателя в эквивалентную решетку, изображенную на рис. 4.1,*б*. Левый нижний угол решетки соответствует первому контакту в первом ряду и т. д. Каждой точке пересечения рядов в этой решетке мы можем поставить в соответствие период времени, в течение которого подвижный контакт достигает соответствующей абонентской линии. Например, изображенная на рисунке наклонная линия соединяет все точки, для достижения которых требуется 0,6 *сек*. Таким образом, мы можем подсчитать все точки, соответствующие каждому данному времени соединения с вызываемым абонентом. Результаты такого подсчета приведены в табл. 4.1.

Таблица 4.1

Число точек, соответствующих разным временам соединения

Время соединения, *сек*	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1,0
Число точек	1	2	3	4	5	6	7	8	9	10
Время соединения, *сек*	1,1	1,2	1,3	1,4	1,5	1,6	1,7	1,8	1,9	—
Число точек	9	8	7	6	5	4	3	2	1	—

Общая сумма этих чисел точек равна 100, как и должно быть.

Теперь мы можем подсчитать число всех точек, для соединения с которыми затрачивается меньше 0,85 *сек*; это число равно 36. Следовательно, согласно нашему определению вероятность соединения равна 36/100, или 0,36.

Это решение будет правильным только в том случае, если выполняются все сделанные нами допущения. Требование, что исходы должны быть *несовместимыми*, означает, что подвижный контакт не может остановиться одновременно на двух неподвижных контактах и подключить сразу две линии; указание, что исход должен быть *исчерпывающим*, предполагает, что вызов наверняка произойдет после пяти часов и будет относиться к абоненту, к которому идет одна из 100 рассматриваемых нами линий.

Условие *равновозможности* проверить наиболее трудно. Если одним из абонентов является доктор, звонки к нему более вероятны, чем к любому другому абоненту; с другой стороны, если некоторые абоненты уехали в отпуск, им вряд ли будут звонить; и т. п. Для рассматриваемой здесь задачи мы можем допустить, что ни один из этих случаев не имеет места. В практической же ситуации такого рода мы могли бы сделать сначала столь же простое допущение, произвести вычисления и затем сравнить теоретический

результат с экспериментальными данными. Если сравнение укажет на существенное расхождение, то следовало бы сделать более сложные допущения.

Какова вероятность того, что соединение произойдет меньше чем за 0,05 *сек*? Совершенно очевидно, что $m=0$, и ответ напрашивается сам собой: вероятность равна нулю. В этом случае, как и во всех других дискретных случаях, вероятность, равная 0, является синонимом невозможности. Однако следует здесь же заметить, что это утверждение становится ошибочным в случае непрерывного процесса.

Какова вероятность, что шаговый переключатель затратит на установление соединения меньше 2 *сек*? Так как $m=100$, то вероятность равна 1. В случае дискретного процесса вероятность, равная единице, является синонимом достоверности.

4.2. Предварительные формулы

В этом разделе мы напомним читателю несколько формул, необходимых при выводе соотношений теории вероятностей. Некоторые из них, например формулы для перестановок и сочетаний, взяты из алгебры; другие, подобно формуле гамма-функции, взяты из математического анализа.

Сочетания и перестановки. Число сочетаний n различимых друг от друга элементов, взятых по m одновременно, вычисляется по формуле

$$C_m^n = \frac{n!}{(n-m)!\,m!} = C_{n-m}^n . \qquad (4.2)$$

Число перестановок n различимых друг от друга элементов, взятых по m одновременно, вычисляется по формуле

$$P_m^n = \frac{n!}{(n-m)!} . \qquad (4.3)$$

Вспомним, что

$$n! \equiv \prod_{i=1}^{n} i \quad \text{и} \quad 0! \equiv 1.$$

Формулы (4.2) и (4.3) поясняются табл. 4.2, которая показывает 20 различных перестановок пяти различных элементов (чисел от 1 до 5), взятых одновременно по два. В каждом ряду приводится полный перечень десяти различных сочетаний, так как две записи в каждом столбце изображают разные перестановки, но одно и то же сочетание.

Нам необходима только одна более сложная комбинаторная формула, позволяющая

Таблица 4.2

Сочетания и перестановки пяти элементов, взятых одновременно по два

1—2	1—3	1—4	1—5	2—3	2—4	2—5	3—4	3—5	4—5
2—1	3—1	4—1	5—1	3—2	4—2	5—2	4—3	5—3	5—4

определять число перестановок n элементов, взятых по n одновременно, для случая, когда некоторые, но не все элементы различимы.

Предположим, мы имеем n элементов, k_1 из которых отличимы от всех остальных, но не отличаются друг от друга, и так далее вплоть до k_m, где $\sum_{i=1}^{m} k_i = n$. Например, мы можем иметь k_1 голубых шаров, k_2 оранжевых и т. д. Нам надо определить общее число возможных различимых перестановок. Для решения этой задачи мы сначала пометим каждый шар так, чтобы можно было бы отличить каждый из них от других шаров. Из этой группы мы получим $n!$ перестановок. Затем мы определим, сколько из этих перестановок станут одинаковыми, если стереть с шаров номера. Существует $k_1!$ перестановок, получаемых простым обменом различных голубых шаров друг с другом, и так как все эти шары неразличимы, мы должны разделить $n!$ на $k_1!$, однако для каждой из них существует $k_2!$ способов расположения оранжевых шаров и т. д. Следовательно, общее число различимых перестановок равно

$$P_{k_1,k_2,\ldots,k_m}^n = \frac{n!}{\prod_{i=1}^{m} k_i!} . \qquad (4.4)$$

Эта формула применима также к сочетаниям n элементов, взятых по $k_1, k_2, \ldots, k_m$ одновременно, например для определения количества различных способов, которыми n элементов можно распределить в m групп с условием, чтобы в первой группе было точно k_1 элементов, во второй — k_2 элементов и т. д.

Гамма-функция. Гамма-функция, обозначаемая символом $\Gamma(n)$, определяется следующей формулой:

$$\Gamma(n) = \int_0^\infty x^{n-1} e^{-x}\, dx. \qquad (4.5)$$

Интересная особенность этой функции заключается в ее связи с $n!$. Для определения этой связи проинтегрируем $\Gamma(n+1)$, воспользовавшись известной формулой

$$\int v\,du = uv - \int u\,dv$$

и приняв, что

$$v = x^n \text{ и } du = e^{-x}\,dx,$$

получим

$$\Gamma(n+1) = [-x^n e^{-x}]_0^\infty - \int_0^\infty -e^{-x} n x^{n-1}\,dx.$$

При $x=0$ отношение x^n/e^x равно нулю, и при $x=\infty$ оно также равно нулю (так как e^x возрастает значительно быстрее, чем x^n при конечных значениях n, что легко показать путем дифференцирования числителя и знаменателя n раз). Следовательно,

$$\Gamma(n+1) = n\int_0^\infty e^{-x} x^{n-1}\,dx =$$

$$= n\Gamma(n) = n(n-1)\Gamma(n-1) = \ldots$$

Но

$$\Gamma(1) = \int_0^\infty x^0 e^{-x} dx = \int_0^\infty e^{-x}\,dx = 1 = 0!.$$

Следовательно, при всех целых значениях

$$\Gamma(n+1) = n!. \qquad (4.6)$$

Гамма-функция определена для всех действительных значений n, тогда как факториал имеет сам по себе смысл только для целых положительных значений. Например,

$$\Gamma(1/2) = \sqrt{\pi}. \qquad (4.7)$$

Так как $\Gamma(1) = 1 = 0 \times \Gamma(0)$, то $\Gamma(0)$ должно равняться бесконечности, так же как и $\Gamma(n)$ при всех отрицательных целых числах. График гамма-функции показан на рис. 4.2.

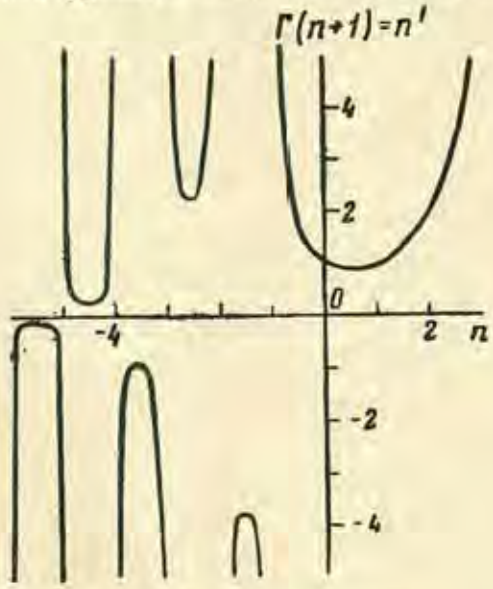


Рис. 4.2. Гамма-функция.

Приближенные формулы. Несколькими методами, слишком сложными, чтобы их приводить здесь, можно показать, что

$$\Gamma(n+1) = n! = \sqrt{2\pi n}\cdot n^n e^{-n}\left(1 + \frac{1}{12n} + \right.$$

$$\left. + \frac{1}{288n^2} - \frac{139}{51{,}840n^3} + \ldots\right).$$

Эта формула является точной для положительных n, если учитываются все члены бесконечного ряда. Практическая польза этой формулы обусловлена тем, что она становится очень простой, когда берется только первый член ряда. Тогда формула принимает вид

$$n! \approx \sqrt{2\pi n}\cdot n^n e^{-n}. \qquad (4.8)$$

Эта формула, известная под названием *приближения Стирлинга*, сравнительно проста, и ее использование при практической работе не вызывает трудностей. Ее относительная ошибка, равная примерно $1/12n$, при больших значениях n становится очень небольшой. Такую ошибку следует принимать во внимание только в случаях, когда вычисляется отношение двух больших чисел; но и в этих достаточно часто встречающихся случаях использование приближенной формулы Стирлинга оказывается полезным.

Большие числа часто даются в экспоненциальной форме — тогда их легко вычислять с помощью логарифмов. Так, например, желая найти число 2^n, мы сразу же можем записать, что логарифм этого числа по основанию 2 равен n. Затем этот логарифм можно преобразовать в обычный десятичный, воспользовавшись формулой, которая непосредственно следует из самого определения логарифма:

$$\log_{10} n = \log_2 n \log_{10} 2. \qquad (4.9)$$

Так как десятичный логарифм от 2 почти точно равен 0,3, эта формула приводит к полезному практическому правилу $2^n \approx 10^{0.3n}$. Так, например, 2^{20} приблизительно равно 10^6, или одному миллиону (действительное значение 2^{20} составляет 1 048 576).

4.3. Полная, сложная и условная вероятности

Под *полной вероятностью* понимается вероятность наступления хотя бы одного из нескольких несовместимых событий. Она равна, как мы убедимся в дальнейшем, хотя интуитивно это совершенно ясно без всякого доказательства, сумме вероятностей отдельных событий.

Рассмотрим эксперимент с несколькими возможными различными результатами, которые мы обозначим через A, B и так далее до K. Из n исходов, которые все равновозможны, n_1 исходов приводит к результату A, n—к B и так далее, причем $\Sigma n_i = n$. Тогда

$$P(A)+P(B)=\frac{n_1}{n}+\frac{n_2}{n}=\frac{n_1+n_2}{n}=P(A+B). \quad (4.10)$$

Исходя из этого, «вероятность наступления A или B» записывают обычно как $P(A+B)$. Отсюда следует, что сумма вероятностей всех несовместимых исходов равна 1, как и следовало ожидать. Так, возможности того, что линейный искатель из предыдущего примера затратит на соединение 0,1, 0,2 *сек* и так далее, являются несовместимыми; вероятность соединения с абонентом менее чем за 0,85 *сек* равна сумме вероятностей соединения за каждый из возможных интервалов от 0,1 до 0,8 *сек*.

Это положение символически изображается *диаграммой Эйлера* (другое название — диаграмма Венна) на рис. 4.3. Площадь круга изображает совокупность всех возможных исходов, а часть площади круга, обозначенная буквой A, изображает ту часть исходов, которые приводят к результату A. Чтобы показать, что полная вероятность результатов A и B равна $P(A)+P(B)$, нам не нужно знать точное распределение вероятностей других исходов, что иллюстрируется на диаграмме обозначением одной части площади словами «от D до K» или «другие исходы».

Сложная вероятность понимается как вероятность совместного появления двух определенных исходов в двух экспериментах. Эти два эксперимента могут быть одинаковыми или различными. Исход одного может зависеть от исхода другого.

Предположим, например, что эксперименты состоят в прибытии к перекрестку движущихся последовательно в северном направлении автомашин, а интересующие нас исходы — в том, повернут или нет эти автомашины влево. При этом часто случается, что водитель намеревается повернуть влево вообще в какой-нибудь не вполне определенной точке магистрали и осуществляет свое намерение при первой благоприятной возможности. На данном перекрестке появление такой возможности особенно вероятно, если едущий впереди автомобиль тоже поворачивает налево. Следовательно, последовательные эксперименты в этом случае не являются независимыми один от другого.

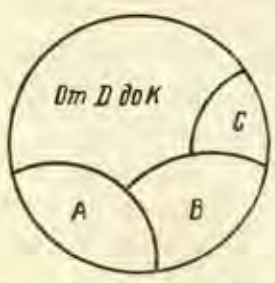


Рис. 4.3. Диаграмма для несовместимых исходов

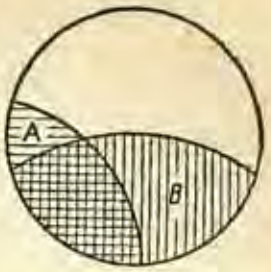


Рис. 4.4 Диаграмма для совместимых исходов.

Условная вероятность n-й автомашине повернуть влево, если $(n-1)$-я повернула влево, больше, чем условная вероятность n-й автомашине повернуть влево, если $(n-1)$-я машина не повернула в этом направлении. Безусловная вероятность n-й автомашине повернуть влево, при отсутствии сведений о поведении предыдущей машины, лежит где-то между этими двумя вероятностями. Но мы можем рассмотреть не только условную вероятность поворота второй машины после поворота первой, но и условную вероятность поворота первой машины в случае, когда нам известно, что вторая машина также повернула. Так как эти две условные вероятности могут отличаться одна от другой, мы должны рассматривать эксперименты последовательно.

Для графического изображения обобщенного понятия условной вероятности предположим, что нас интересует совместное наступление исхода A в первом из двух последовательных экспериментов и исхода B во втором эксперименте. Обозначим через $P(A)$ безусловную вероятность наступления исхода A в первом эксперименте; через $P_A(B)$ — условную вероятность наступления исхода B во втором эксперименте, если нам известно, что первый эксперимент имел исход A; и через $P(A,B)$ — вероятность совместного наступления интересующих нас исходов.

Эти условия изображены на рис. 4.4 с помощью диаграммы Эйлера. Отношение заштрихованной горизонтальными линиями площади круга (включая площадь, заштрихованную перекрещивающимися линиями) к площади всего круга есть дробь, равная вероятности наступления исхода A в первом эксперименте. Аналогично этому вертикально заштрихованная площадь (также включая участок, заштрихованный перекрещивающимися линиями) изображает наступление исхода B во втором эксперименте. Общий участок этих двух областей, заштрихованный перекрещивающимися линиями, изображает интересующее нас совместное наступление двух исходов. Условная вероятность $P_A(B)$ представ-

ляет собой дробь, равную отношению заштрихованной перекрещивающимися линиями площади к площади, заштрихованной горизонтальными линиями. Как следует из рис. 4.4, эта вероятность больше, чем безусловная вероятность $P(B)$. Однако она может быть равна безусловной вероятности или быть меньше нее.

Если через n обозначить общее число совместно наступающих исходов, через m — число исходов, благоприятных для A, а через m_1 — число исходов, благоприятных для A и B, то согласно нашему определению вероятности вероятность совместного наступления событий A и B равна

$$P(A,B)=\frac{m_1}{n}=\frac{m_1}{n}\frac{m}{m}=\frac{m}{n}\frac{m_1}{m}=P(A)\,P_A(B). \tag{4.11а}$$

Из этого несложного преобразования, как и из диаграммы Эйлера, легко показать, что

$$P(A,B)=P(B)\,P_B(A). \tag{4.11б}$$

Однако $P_B(A)$, вообще говоря, не равна $P_A(B)$; они не равны и на чертеже рис. 4.4.

Введем теперь для обозначения „не-A“ символ $\bar{A}$. Тогда $P(\bar{A})$ обозначает вероятность ненаступления события A. Разумеется, $P(A)+P(\bar{A})=1$. Фраза „B независимо от A“ обозначает условие, что

$$P_A(B)=P_{\bar{A}}(B)=P(B).$$

В этом случае

$$P(A,B)=P(A)\,P_A(B)=P(A)\,P(B). \tag{4.12}$$

Эта формула справедлива только в том случае, если события A и B независимы. Так, например, для случая, изображенного на рис. 4.4, $P(A,B)>P(A)\,P(B)$. Если площадь, обозначенную буквой A, изменять таким образом, чтобы общий участок уменьшался, то в конечном счете можно было бы получить диаграмму, изображающую независимость событий. Затем, при дальнейшем уменьшении площади совместного участка, $P(B)$ снова стала бы зависимой от $P(A)$, причем

$$P(A,B)<P(A)\,P(B).$$

Пример. Проиллюстрируем теперь все эти понятия тривиальным примером. Предположим, что имеется группа автомашин, из которых мы должны выбрать одну. Первый эксперимент заключается в определении того, будет ли выбранная нами автомашина грузовой (T) или негрузовой ($\bar{T}$). Второй эксперимент состоит в определении того, хорошие ли тормоза у выбранной нами машины (G) или плохие ($\bar{G}$). В табл. 4.3 и 4.4 показаны соответственно случаи независимых и зависимых событий. Из табл. 4.3 следует, что вероятность грузовой машине иметь хорошие тормоза равна вероятности негрузовой машине иметь хорошие тормоза, а следовательно, равна (безусловной) вероятности машине иметь хорошие тормоза, а следовательно, равна 0,75.

Таблица 4.3

Пример независимых исходов

	T	$\bar{T}$	Всего
G	0,30	0,45	0,75
$\bar{G}$	0,10	0,15	0,25
Всего	0,40	0,60	1,00

Таблица 4.4

Пример зависимых событий

	T	$\bar{T}$	Всего
G	0,30	0,50	0,80
$\bar{G}$	0,10	0,10	0,20
Всего	0,40	0,60	1,00

Из табл. 4.4 следует, что условная вероятность машине, о которой уже известно, что она грузовая, иметь хорошие тормоза, составляет по-прежнему 0,75, однако условная вероятность машине, о которой известно, что она не грузовая, иметь хорошие тормоза, составляет 0,83, безусловная же вероятность машине иметь хорошие тормоза равна 0,80.

Эти таблицы можно анализировать многими другими способами. Будет полезно проверить все формулы, приведенные в этой главе, по этим таблицам и объяснить полученные результаты словами.

Мы уже показали, что полная вероятность несовместимых исходов одного эксперимента (рис. 4.3) равна

$$P(A+B)=P(A)+P(B).$$

Распространим теперь эту формулу на общий случай, когда исходы не являются несовместимыми.

Заметим сначала, что

$$P(A)=P(A,B)+P(A,\bar{B}),$$

откуда

$$P(A)+P(B)=P(A,B)+$$
$$+[P(A,\bar{B})+P(A,B)+P(\bar{A},B)].$$

Но выражение в квадратных скобках определяет не что иное, как вероятность наступ-

ления A или B, так как оно является суммой вероятностей, что A наступит без B, или B без A, или оба вместе. Следовательно,

$$P(A)+P(B)=P(A,B)+P(A+B),$$

откуда получаем

$$P(A+B)=P(A)+P(B)-P(A,B). \quad (4.13)$$

В случае, когда события A и B несовместимы, вероятность $P(A, B)$ равна нулю, и это выражение сводится к рассмотренному ранее. В других случаях (рис. 4.4) вероятность $P(A,B)$ определяется площадью, которая учитывается дважды, когда $P(A)$ складывается с $P(B)$, и поэтому должна вычитаться. Распространение формулы (4.13) на три исхода дает формулу

$$P(A+B+C)=P(A)+P(B)+P(C)-$$
$$-P(A,B)-P(A,C)-P(B,C)+P(A,B,C). \quad (4.14)$$

В случае двух экспериментов применимы нижеследующие тождества, в которых исходы $i=A, B, \ldots$ наступают в одном эксперименте, а исходы $j=A, B, \ldots$ — в другом:

$$\sum_j P(i,j)=P(i),\ \sum_i P(i,j)=P(j); \quad (4.15)$$

$$\sum_i P(i)=\sum_j P(j)=1,\ \sum_i\sum_j P(i,j)=$$
$$=\sum_j\sum_i P(i,j)=1; \quad (4.16)$$

$$\sum_j P(i)\,P_i(j)=P(i),\ \sum_i P(i)\,P_i(j)=P(j); \quad (4.17)$$

$$\sum_j P_i(j)=1. \quad (4.18)$$

Равенство (4.17) вытекает из того обстоятельства, что величина, стоящая под знаком суммы, в точности совпадает с величиной, стоящей под знаком суммы в равенстве (4.15).

Равенство (4.18) вытекает из того обстоятельства, что если наступает определенный исход i (например, A), должен наступить какой-то исход j (например, B или $\bar{B}$); если рассмотреть и просуммировать все эти вероятности, то и получим $P=1$.

4.4 Марковские цепи

Последовательность событий, в которой вероятность данного исхода n-го события полностью определяется исходом $(n-1)$-го события, называется *марковским процессом* или *марковской цепью* *. Рассмотренный нами пример с левым поворотом автомашины на перекрестке (§ 4.3) мы можем рассматривать как пример марковской цепи. Предположим, что вероятность поворота автомашины влево равна 0,2, если предыдущая автомашина повернула влево, и равна 0,1, если предшествующая автомашина не сделала такого поворота. Какова вероятность n-й автомашине сделать левый поворот?

Нам известно, что вероятность $P(k)$ поворота k-ой автомашины плюс вероятность того, что эта автомашина не повернет, равна 1. Следовательно, мы можем выразить $P(k+1)$ через $P(k)$:

$$P(k+1)=0{,}2P(k)+0{,}1\,[1-P(k)]=$$
$$=0{,}1+0{,}1P(k). \quad (4.19)$$

Пусть вероятность поворота нулевой автомашины, равная $P(0)$, нам неизвестна. Тогда

$$P(1)=0{,}1+0{,}1P(0),$$
$$P(2)=0{,}1+0{,}01+0{,}01\,P(0)=0{,}11+0{,}01P(0),$$
$$P(3)=0{,}1+0{,}011+0{,}001P(0)=0{,}111+$$
$$+0{,}001\,P(0),$$
$$\ldots$$
$$\ldots$$
$$\ldots$$

$P(\infty)=1/9$ независимо от значения $P(0)$.

ЛИТЕРАТУРА И ЗАДАЧИ

См. гл. 8.

---

* Названы по имени известного русского математика акад. А. А. Маркова (1856—1922); впервые систематически изучившего свойства этих случайных процессов. — *Прим. ред.*

ГЛАВА 5

РАСПРЕДЕЛЕНИЯ ДИСКРЕТНЫХ ПЕРЕМЕННЫХ

5.1. Биномиальное распределение

В § 4.1 мы показали, что вероятность гипотетическому линейному искателю затратить менее 0,85 *сек* на соединение равна 0,36. Теперь мы поставим перед собой другую задачу: определим вероятность того, что из десяти по-

следовательных вызовов два вызова потребуют менее 0,85 *сек*, а остальные восемь — больше. Предположим при этом, что при каждом вызове подвижный контакт переключателя начинает движение из одного и того же начального положения и что время, затраченное на один вызов, не зависит от времени, затраченного на другой вызов.

Для решения задачи найдем сначала вероятность того, что каждый из первых двух вызывов потребует менее 0,85 *сек*, а каждый из оставшихся восьми требует больше. Так как времена соединения для разных вызовов не зависят друг от друга, то вероятность первым двум вызовам потребовать менее 0,85 *сек* равна $(0{,}36)^2$, вероятность каждому последующему вызову потребовать более 0,85 *сек* равна 0,64 и вероятность наступления восьми таких событий подряд равна $(0{,}64)^8$. Таким образом, вероятность возникновения всей рассматриваемой последовательности равна $(0{,}36)^2 \cdot (0{,}64)^8$. Но вероятность появления точно двух вызовов со временем соединения менее 0,85 *сек* в любой другой паре опытов, например в первом и третьем опыте или во втором и девятом будет та же. Число таких сочетаний равно, конечно, точно C_2^{10}. Таким образом, ответом к нашей задаче служит число $\frac{10\times 9}{1\times 2}\ (0{,}36)^2(0{,}64)^8 = 0{,}164$.

Этот метод решения задачи применим не только для двух вызовов. Вероятность того, что k вызовов из 10 потребуют каждый менее 0,85 *сек*, равна $C_k^{10}(0{,}36)^k(0{,}64)^{10-k}$.

Если это выражение преобразовать к еще более общему виду, обозначив буквой n полное число вызывов и буквой p вероятность появления соответствующего (*благоприятного*, или успешного) исхода при каждом вызове, то мы получим формулу

$$P(k) = C_k^n p^k (1-p)^{n-k}.$$

Эта формула оказывается весьма общей. Она определяет вероятность получения в точности k успешных исходов при проведении n независимых испытаний в случае, когда вероятность успешного исхода при любом испытании равна p. Обычно эта формула записывается в следующем виде:

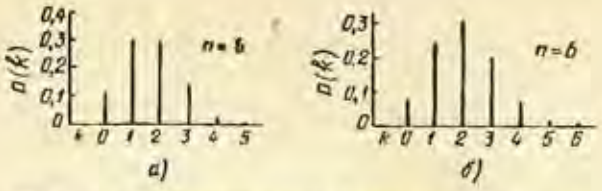


Рис. 5.1. Биномиальное распределение для $p = 1/3$.

$$P(k) = C_k^n p^k q^{n-k}, \qquad (5.1)$$

где $q = 1 - p$ обозначает вероятность неудачи.

Используя выражение (4.2) для C_k^n, получаем

$$P(k) = \frac{n!}{k!\,(n-k)!}\, p^k q^{n-k}. \qquad (5.2)$$

Это выражение для $P(k)$ называется *распределением вероятностей*, так как оно показывает, как вероятность распределяется по различным возможным исходам. Формула (5.2) принадлежит к группе выражений, которые могут служить распределениями вероятностей, и известна под названием *биномиального распределения* или *распределения Бернулли*.

Вспомним, что формула бинома имеет вид

$$(p+q)^n = \sum_{k=0}^{n} C_k^n p^k q^{n-k}.$$

Таким образом, вероятность появления в точности k удач равна $(k+1)$-му члену разложения бинома. Так как $p+q=1$, то и $(p+q)^n = 1$ при любых значениях n, вследствие чего сумма вероятностей, определяемых формулой (5.2), по всем значениям k должна равняться единице.

Таблица 5.1

Биномиальное распределение для $p = 1/3$

	$n=5$						$n=6$						
k	0	1	2	3	4	5	0	1	2	3	4	5	6
p^k	1	1/3	1/9	1/27	1/81	1/243	1	1/3	1/9	1/27	1/81	1/243	1/729
q^{n-k}	32/243	16/81	8/27	4/9	2/3	1	64/729	32/243	16/81	8/27	4/9	2/3	1
$p^k q^{n-k}$	32/243	16/243	8/243	4/243	2/243	1/243	64/729	32/729	16/729	8/729	4/729	2/729	1/729
C	1	5	10	10	5	1	1	6	15	20	15	6	1
$P(k)$	32/243	80/243	80/243	40/243	10/243	1/243	64 729	192/729	240/729	160/729	60/729	16/729	1/729

Чтобы читатель мог лучше представить себе биномиальное распределение, вероятность $P(k)$ для случая $p=1/3$ приведена в табл. 5.1 и изображена графически на рис. 5.1 для всех значений k от 0 до n, применительно к $n=5$ и $n=6$.

5.2. Ожидаемые значения. Математическое ожидание и дисперсия

Многие важные величины, характеризующие распределения, в частности математическое ожидание, определяющее положение распределения, и дисперсия, указывающая на его рассеяние, строятся с помощью оператора, называемого *ожидаемым значением*.

Мы определяем ожидаемое значение от произвольной функции $f(x)$ дискретной переменной выражением

$$E[f(x)] \equiv \sum_{x=-\infty}^{\infty} f(x)\, p(x). \qquad (5.3)$$

Ожидаемое значение переменной поэтому равно просто

$$E(k) = \sum_{k=-\infty}^{\infty} k p(k). \qquad (5.4)$$

Заметим, что ожидаемое значение является *линейным оператором*, т. е.

$$E[af(k) + bg(k)] = aE[f(k)] + bE[g(k)]. \qquad (5.5)$$

Это следует из того, что знак суммы является линейным оператором:

$$E[af(k) + bg(k)] = \Sigma[af(k) + bg(k)]\, p(k) =$$
$$= a\Sigma f(k)\, p(k) + b\Sigma g(k)\, p(k) = aE[f(k)] +$$
$$+ bE[g(k)].$$

Ожидаемое значение не обязательно является наиболее вероятным значением, но оно является разновидностью *взвешенного среднего значения*.

Мы определяем *математическое ожидание произвольного распределения* как ожидаемое значение переменной.

$$\mu \equiv E(k). \qquad (5.6)$$

Математическое ожидание определяет положение распределения (т. е. его центра) * и является наиболее удобным видом «среднего значения». Популярного термина «среднее значение» лучше избегать, если только мы умышленно не хотим применить туманную формулировку, так как этот термин не вполне определенный. Под ним часто понимают математическое ожидание, однако он может также относиться и к медиане, и к моде (см. § 6.3) или даже к какой-нибудь другой величине.

Для биномиального распределения математическое ожидание определяется выражением

$$\mu = \sum_{k=0}^{n} k C_k^n p^k q^{n-k}. \qquad (5.7)$$

Мы вычислим μ с помощью *образующей функции* $(q+pt)^n$, содержащей *вспомогательную переменную* t. Сначала мы введем эту переменную, а затем приравняем ее единице, чтобы она исчезла из полученного выражения. Начнем с применения к образующей функции формулы разложения бинома:

$$(q+pt)^n = \sum_{k=0}^{n} C_k^n (pt)^k q^{n-k}.$$

Дифференцируя по t, получаем

$$n(q+pt)^{n-1} p = \sum_{k=0}^{n} C_k^n k p^k t^{k-1} q^{n-k}. \qquad (5.8)$$

При $t=1$

$$n(q+p)^{n-1} p = \sum_{k=0}^{n} k C_k^n p^k q^{n-k}.$$

Но левая часть этого выражения равна np, а правая часть согласно формуле (5.7) равна μ.

Таким образом, математическое ожидание биномиального распределения дается формулой

$$\mu = np.$$

Дисперсия распределения, обозначаемая символом σ^2, является мерой его рассеяния или разброса, т. е. характеризует насколько значения переменной группируются вокруг математического ожидания.

Дисперсия определяется формулой

$$\sigma^2 = E[k - E(k)]^2. \qquad (5.9)$$

Квадратный корень из дисперсии называется *стандартным отклонением* (или *стандартом*) и обозначается буквой σ. Стандартное отклонение является среднеквадратиче-

* Будучи функцией вида $P(k)$, где k — случайная дискретная переменная, а P — вероятность данного значения этой переменной, распределение является кривой в плоскости (P, k) и потому характеризуется определенным положением в этой плоскости. — *Прим. ред.*

ским отклонением значений отдельных событий от ожидаемого значения. При работе удобно применять как дисперсию, так и стандартное отклонение: первую из-за ее математических свойств, а второе потому, что оно выражается в тех же единицах, что и математическое ожидание.

Стандартное отклонение является наиболее распространенной мерой рассеяния распределения, хотя для этой цели иногда применяются и другие меры, такие, как размах* и среднее отклонение от математического ожидания.

Если в выражение (5.9) вместо $E(k)$ подставить μ и раскрыть скобку, получим

$$\sigma^2 = E(k-\mu)^2 = E(k^2 - 2k\mu + \mu^2).$$

Используя формулу (5.5), преобразуем это выражение к следующему виду:

$$\sigma^2 = E(k^2) - 2\mu E(k) + \mu^2 = E(k^2) - 2\mu^2 + \mu^2 =$$

$$= E(k^2) - \mu^2 = E(k^2) - [E(k)]^2. \qquad (5.10)$$

Для биномиального распределения дисперсия равна

$$\sigma^2 = \sum_{k=0}^{n} k^2 C_k^n p^k q^{n-k} - \mu^2.$$

Используем вновь образующуюся функцию. Дифференцируя выражение (5.8) второй раз, получаем

$$n(n-1)(q+pt)^{n-2} p^2 =$$

$$= \sum_{k=0}^{n} C_k^n k(k-1) p^k t^{k-2} q^{n-k}.$$

При $t=1$

$$n(n-1)(q+p)^{n-2} p^2 =$$

$$= \sum_{k=0}^{n} k(k-1) C_k^n p^k q^{n-k},$$

так что

$$n^2p^2 - np^2 = \sum_{k=0}^{n} k^2 C_k^n p^k q^{n-k} -$$

$$- \sum_{k=0}^{n} k C_k^n p^k q^{n-k}.$$

Но

$$n^2p^2 = \mu^2 \quad \text{и} \quad \sum_{k=0}^{n} k C_k^n p^k q^{n-k} = np.$$

Следовательно,

$$\mu^2 - np^2 = \sum_{k=0}^{n} k^2 C_k^n p^k q^{n-k} - np.$$

Произведя перестановку членов, прийдем к

$$\sum_{k=0}^{n} k^2 C_k^n p^k q^{n-k} - \mu^2 = np - np^2 = np(1-p). \qquad (5.11)$$

Но левая часть выражения (5.11) равна σ^2. Следовательно,

$$\sigma^2 = npq. \qquad (5.12)$$

Пример. Какова вероятность того, что при серии в 10 вызовов линейный искатель затратит более 1,25 *сек* в двух или более из них?

Решение. Чтобы определить вероятность того, что число благоприятных исходов при n испытаниях будет больше или равно m (целое число), используем формулу

$$P(k \geqslant m) = \sum_{k=m}^{n} C_k^n p^k q^{n-k}. \qquad (5.13)$$

В рассматриваемом случае из табл. 4.1 находим, что $p = 0{,}28$. Суммирование от 2 до 10 займет много времени, поэтому мы произведем сложение от 0 до 1 и вычтем полученный результат из суммы $\sum_{0}^{n}$, которая, как мы знаем, равна 1. Итак,

$$P(k \geqslant 2) = 1 - P(k \leqslant 1) = 1 - P(0) - P(1) =$$

$$= 1 - C_0^{10}(0{,}28)^0 \cdot (0{,}72)^{10} - C_1^{10}(0{,}28)^1 (0{,}72)^9 =$$

$$= 1 - (0{,}72)^{10} - 10 \times 0{,}28 \times (0{,}72)^9 = 0{,}817.$$

При решении задач подобного типа важно не смешивать благоприятный исход (удачу) с неблагоприятным (неудачей), к чему легко может привести обычный смысл этих слов. В следующей задаче под благоприятным исходом понимается несчастный случай со смертным исходом.

Пример. Какова вероятность гибели одного пассажира, совершающего 100 000 полетов на самолете, если вероятность гибели за один полет на самолете равна 0,00001?

Решение. В данном случае у нас биномиальное распределение с $n = 10^5$ и $p = 10^{-5}$. Однако следует избегать двух опасностей впасть в ошибку, на которые мы обратим внимание перед изложением решения.

Первая опасность—подсчитать $np = 1$ и назвать это ответом. Ясно, что ни об одном пассажире, решившем совершить любое конечное число полетов, нельзя с пол-

---

* Размах $W(k)$ равен разности между наибольшим и наименьшим возможным значением случайной переменной k. — *Прим. ред.*

ной уверенностью сказать, что он погибнет. Отсюда следует, что вероятность должна быть меньше 1. Конечно, величина np имеет определенный смысл: это — ожидаемое значение. Ожидаемое число полетов, заканчивающихся гибелью пассажира, составляет 1 на 100 000 полетов.

Однако отсюда не следует, что точно на каждые 100 000 полетов необходимо ожидать точно одного несчастливого полета. Эта пропорция говорит только о том, что если совершить большое количество групп полетов, по 100 000 полетов в каждой группе, то в среднем мы можем ожидать такой процент несчастливых полетов. Таким образом, на 100 000 000 полетов мы должны были бы ожидать, что число несчастливых полетов приближается к 1 000. Это обстоятельство будет рассмотрено нами подробнее в § 8.3.

Вторая опасность — подсчитать вероятность гибели пассажира исходя из предположения, что пассажир может погибнуть только один раз. Конечно, в действительности человек не может быть убитым более одного раза. Однако математический процесс вычисления вероятности гибели пассажира приводит к правильному ответу только в том случае, если предположить, что каждый пассажир может погибнуть в результате аварии один или более раз, в то время как предположение, что каждый пассажир может быть убит только один раз, приводит к ошибочному результату. Справедливость этого становится очевидной, если учесть, что в действительности мы хотим найти не что иное, как величину, равную единице минус вероятность гибели пассажира 0 раз.

Этот кажущийся парадокс можно объяснить, рассмотрев допущения, которые мы невольно ввели, когда применили биномиальное распределение. Если мы предполагаем, что $n=100\,000$, то мы должны гарантировать, что все 100 000 опытов будут осуществлены; однако если человек погибает в каком-либо полете, он не может «участвовать» во всех последующих испытаниях. Поэтому если считать, что каждый пассажир может погибнуть только один раз, то для непосредственного определения вероятности гибели пассажира мы должны вычислить вероятность его гибели во время первого полета, затем вероятность того, что пассажир «пережил» первый полет и погиб во втором полете, и еще 99 998 аналогичных величин. Математически этот ряд имеет вид

$$P = p + qp + q^2p + \ldots = p(1 + q + q^2 + \ldots) =$$

$$= p\sum_{m=0}^{n-1} q^m = p\frac{1-q^n}{1-q} = 1 - q^n. \qquad (5.14)$$

Вместо этого мы хотим применить биномиальное распределение, гарантировав проведение 100 000 опытов, и вычислить все случаи возможной гибели пассажира один или большее число раз. Это, конечно, приводит к тому же самому ответу:

$$P(k \geqslant 1) = 1 - P(k = 0) =$$

$$= 1 - C_0^n p^0 q^n = 1 - q^n = 1 - (1-p)^n. \qquad (5.14')$$

В нашем случае $P=1-(0{,}99999)^{100\,000}$. Для оценки этой величины с точностью до второго знака необходимо применить семизначные таблицы логарифмов. Логарифм от 0,99999 равен —0,0000043; умножая на 100 000, мы получим —0,43. Антилогарифм от —0,43 равен 0,37. Таким образом, мы получили ответ $P=0{,}63$.

5.3. Биномиальное распределение при больших n

Ясно, что в некоторых случаях не представляется возможности применить методы сокращенного вычисления для уменьшения расчетных работ. Так, например, при необходимости определить вероятность не менее 20 удач в 100 испытаниях потребуется вычислить каждый член суммы от 0 до 20 (пример в конце § 6.4). Однако в большинстве случаев для вычисления таких вероятностей имеются приближенные формулы. В общем случае приближенные формулы получаются заменой дискретных функций непрерывными и справедливы, лишь когда число испытаний весьма велико. Так, например, читатель с математической подготовкой может легко заметить, что выражение $(0{,}99999)^{100\,000}$ из предыдущей задачи почти точно равно $1/e$ (это будет показано в общем виде в § 5.6; ошибка такого приближения настолько ничтожна, что для подсчета с помощью биномиального распределения более точного ответа, чем по этой формуле, потребовались бы одиннадцатизначные таблицы логарифмов).

Для вывода приближенного выражения для биномиального распределения предположим, что n неограниченно возрастает. Это можно сделать при разных системах ограничений. Так, например, можно ввести условие, что математическое ожидание должно сохраняться постоянным и равным np, в связи с чем при возрастании n вероятность p будет уменьшаться. Это приводит к распределению Пуассона (§ 5.6). При другой системе ограничений строго неизменными должны быть математическое ожидание и дисперсия, что приводит к нормальному распределению (§ 6.4). Здесь мы рассмотрим две другие системы ограничений.

Если вычертить чертежи, подобные тем, которые были приведены на рис. 5.1, но для больших значений n (рис. 5.2), то из них следует, что по мере возрастания n все ординаты уменьшаются, а их положение смещается вправо, в то время как сумма всех ординат остается равной единице. Таким образом, математическое ожидание и дисперсия будут возрастать пропорционально n. Мы можем вычислить вероятность получения в точности значения математического ожидания (предположив, что оно является целым числом), производя подстановку $k=np$, в результате чего получаем

$$P(np) = \frac{n!}{(np)!\,(n-np)!}\, p^{np}\,(1-p)^{n-np}.$$

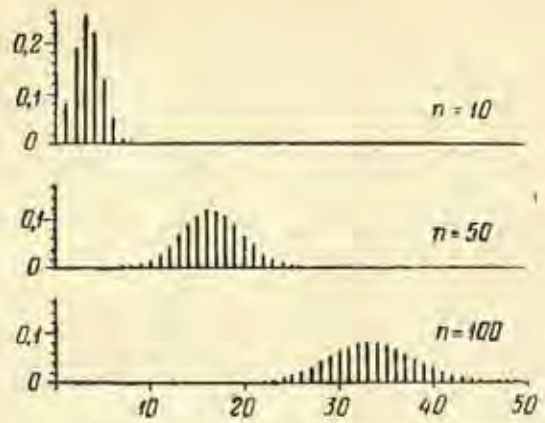


Рис. 5.2. Биномиальное распределение для больших $n\,(p = 1/3)$ (по Фрею [33]).

Для вычисления факториала применим приближенную формулу Стирлинга (§ 4.8):

$$P(np) = \frac{\sqrt{2\pi n}\, n^n\, e^{-n}}{\sqrt{2\pi np}\,(np)^{np}\, e^{-np}\sqrt{2\pi(n-np)}\,(n-np)^{n-np} e^{-n+np}} \times \times p^{np}(1-p)^{n-np}.$$

После сокращений выражение приводится к виду

$$P(\mu) = P(np) = \frac{1}{\sqrt{2\pi\, np\,(1-p)}} = \frac{1}{\sqrt{2\pi npq}} = \frac{1}{\sqrt{2\pi}\,\sigma}. \qquad (5.15)$$

Отсюда легко видеть, что вероятность получения наиболее вероятного результата при увеличении n уменьшается пропорционально квадратному корню из n. Однако вероятность получения результата, лежащего относительно близко к наиболее вероятному результату, при возрастании n увеличивается.

В этом нетрудно убедиться, рассматривая рост n при различных ограничениях. Именно, пусть n неограниченно возрастает, и вычертим кривые вероятностей получения не k, а отношения k/n (рис. 5.3) для нескольких последовательно возрастающих значений n. Математическое ожидание этого распределения при $n \to \infty$ остается постоянным и равным p. Дисперсия равна pq/n и при $n \to \infty$ уменьшается, стремясь к 0.

5.4. Применение биномиального распределения

Биномиальное распределение находит широкое практическое применение, так как при наличии соответствующих условий позволяет с хорошей степенью точности предсказывать

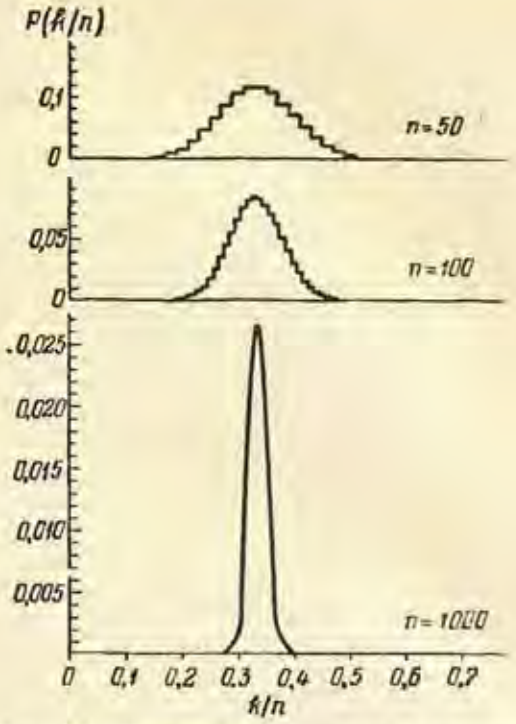


Рис. 5.3. Распределение доли благоприятных исходов при биномиальном распределении для больших $n\,(p = 1/3)$ (по Фрею [33]).

результаты реальных событий. Однако важно помнить предположения, сделанные при выводе формул, а именно, что все последовательные испытания имеют одну и ту же вероятность благоприятного исхода и что все последовательные испытания независимы друг от друга.

При практических применениях чаще всего определяют вероятность появления одного или более благоприятных исходов (5.14) либо же m или более благоприятных исходов (5.13). Выражение такого вида, как (5.13), рассматриваемое как функция от m, называется *кумулятивной функцией распределения*.

Биномиальное распределение описывает результат эксперимента, протекающего по принципу «вышло — не вышло», имеющего два возможных исхода, т. е. любую ситуацию, настолько поляризованную, что исход будет положительным (ответ «да») либо отрицательным (ответ «нет»), приемлемо хорошим либо неприемлемо плохим. Так, например, при контроле качества продукции нас может интересовать число снимаемых с конвейера бракованных изделий. Конечно, если нас заинтересует степень качества изготовленной продукции, тогда следует применить распределение непрерывной переменной. Оба эти случая достаточно подробно будут рассмотрены в § 12.3. Аналогично при артиллерийской

стрельбе, если нас интересуют только промах или попадание, мы можем применить биномиальное распределение; однако если нас интересует расстояние точек попадания снаряда от цели, то обычно применяется нормальное распределение (§ 6.4).

Биномиальное распределение может применяться для определения степени, с которой появление на практике определенных событий будет соответствовать данным, полученным с помощью формул и законов биномиального распределения.

Если мы проводим реальный эксперимент из пяти испытаний, причем вероятность удачи в каждом из них равна $^1/_3$, то из табл. 5.1 следует, что вероятность точно двух удач равна 80/243. В действительности мы либо получим, либо не получим точно два благоприятных исхода в наших пяти испытаниях. Однако если мы проведем 10 серий по пять испытаний в каждой, то как часто мы будем получать два благоприятных исхода в серии?

Ответ на этот вопрос можно получить из формул биномиального распределения, взяв $p=80/243$ и $n=10$; и, конечно, наборы таких серий также могли бы быть подвергнуты аналогичному анализу с помощью биномиального распределения.

С помощью биномиального распределения могут рассматриваться также проблемы надежности. Предположим, что вероятность безотказной работы мотора самолета во время определенного полета равна p. Тогда вероятность благополучного полета одномоторного самолета (исключая другие возможные причины аварии) также равна p, в то время как вероятность безаварийного полета двухмоторного самолета, который может совершать полет и при одном работающем моторе, равна $1-(1-p)^2$, т. е. значительно больше, чем p. С другой стороны, если полет двухмоторного самолета может осуществляться только при двух работающих моторах, вероятность безаварийной работы снижается до p^2, что существенно меньше p.

На рис. 5.4 показаны кривые для выражения (5.14′), полезные при решении проблем надежности.

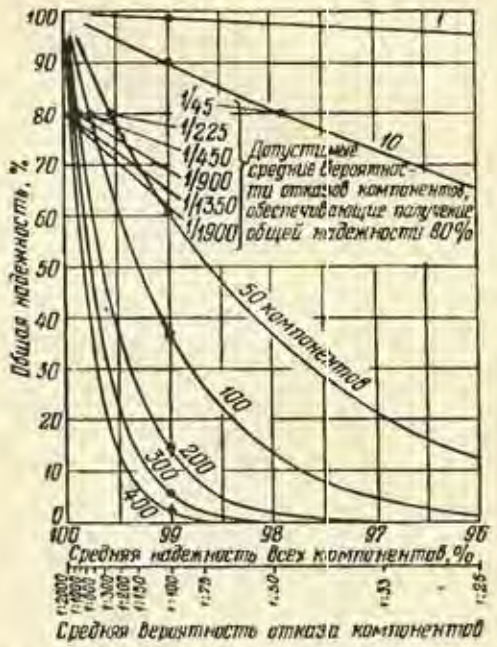


Рис. 5.4. Общая надежность устройств как функция их сложности и средней надежности их компонентов (согласно [28]).

5.5. Полиномиальное распределение *

При биномиальном распределении имелось два возможных исхода с вероятностями соответственно p и q. Существуют определенные ситуации, при которых может быть несколько исходов, и для них мы хотим знать вероятность получения k_1 исходов первого типа, имеющих вероятность p_1, и так далее вплоть до вероятности получения k_m исходов m-го типа, имеющих вероятность p_m.

Используя эти обозначения, можно записать полученную нами формулу для биномиального распределения в виде

$$P(k_1, k_2)=Cp_1^{k_1}p_2^{k_2},$$

где C — число способов, которыми n событий может быть разделено на m групп (в рассматриваемом случае на две группы). Для полиномиального распределения то же самое рассуждение дает формулу

$$P(k_1, k_2, \ldots, k_m)=Cp_1^{k_1}p_2^{k_2}\ldots p_m^{k_m},$$

где $\sum p_i=1$ и $\sum k_i=n$. Соответствующее значение C в этом случае должно определяться по формуле (4.4). Следовательно, можно написать

$$P(k_1, k_2, \ldots, k_m)=$$
$$=\frac{n!}{k_1!\,k_2!\ldots k_m!}\,p_1^{k_1}p_2^{k_2}\ldots p_m^{k_m}. \qquad (5.16)$$

Правая часть в формуле (5.16) является общим членом в разложении полинома

$$(p_1+\ldots+p_m)^n.$$

5.6. Распределение Пуассона

Как уже указывалось раньше, понятие о распределении вероятностей между различ-

* Полиномиальное распределение называется также мультиномиальным. — *Прим. ред.*

ными возможными исходами случайного процесса * является совершенно общим. Биномиальное распределение — только один пример. Любое математическое выражение, представляющее $P(k)$ как функцию от k, есть распределение вероятностей, если только оно удовлетворяет двум очевидным условиям:

$$p(k) \geqslant 0 \text{ для всех } k, \qquad (5.17а)$$

$$\sum_{k=-\infty}^{\infty} P(k) = 1. \qquad (5.17б)$$

Выражение (5.17а) просто указывает на то, что ни одно событие не может иметь отрицательной вероятности, так как это было бы бессмыслицей. Выражение (5.17б) указывает, что сумма вероятностей всех возможных исходов должна равняться 1.

Некоторые из функций, удовлетворяющих этим двум условиям, встречаются реально в инженерной практике, значительное же большинство — нет. Оказывается, что подавляющее большинство случайных процессов, наблюдаемых в природе и при проектировании систем, могут быть представлены с достаточно хорошим приближением одним из четырех рассмотренных здесь и в следующей главе распределений, а именно: биномиальным, Пуассона, равномерным и нормальным, или гауссовым.

Как уже говорилось раньше, распределение Пуассона является аппроксимацией (приближением) биномиального распределения, когда число испытаний весьма велико; эта аппроксимация справедлива при выполнении нескольких определенных ограничений. Мы получим распределение Пуассона двумя методами. Первый из них легче, но требует больших ограничений и потому менее полезен. При этом методе мы предполагаем неограниченное возрастание n и заставляем вероятность благоприятного исхода непрерывно уменьшаться, так что произведение np остается постоянным и равным μ. Число благоприятных исходов k остается небольшим. Таким образом, результаты представляют интерес в случаях (к которым относится, например, рассмотренная нами задача о вероятности аварий самолетов при 100 000 полетов), когда k равно нулю или очень мало, p очень мало, а n очень велико.

---

* Другой специальный термин, употребляемый в том же смысле, — *«стохастический процесс»*. Некоторые авторы [11] ограничивают термин *«случайный процесс»* понятием, которое вкладывается нами в термин *равномерное распределение* (§ 6.1). — *Прим. авт.*

Начнем с формулы (5.2) для биномиального распределения и подставим μ/n вместо p и приближение Стирлинга вместо $n!$ и $(n-k)!$, но не вместо $k!$.

Тогда

$$P(k) = \frac{n!}{(n-k)!\,k!}\left(\frac{\mu}{n}\right)^k\left(1-\frac{\mu}{n}\right)^{n-k},$$

или

$$P(k) \approx \frac{\sqrt{2\pi n}}{\sqrt{2\pi(n-k)}}\,\frac{n^n}{(n-k)^{n-k}}\,\frac{e^{-n}}{e^{-n+k}} \times \times \frac{1}{k!}\left(\frac{\mu}{n}\right)^k\left(1-\frac{\mu}{n}\right)^{n-k}. \qquad (5.18)$$

Первая дробь равна $n^{1/2}/(n-k)^{1/2}$. Группируя одинаковые члены, получаем

$$P(k) \approx \frac{n^{n+1/2}e^{-n}}{(n-k)^{n-k+1/2}e^{-n+k}n^k}\,\frac{\mu^k}{k!}\left(1-\frac{\mu}{n}\right)^{n-k} = = \frac{n^{n-k+1/2}}{(n-k)^{n-k+1/2}e^k}\,\frac{\mu^k}{k!}\left(1-\frac{\mu}{n}\right)^{n-k} = = \frac{1}{(1-k/n)^{n-k+1/2}e^k}\,\frac{\mu^k}{k!}\left(1-\frac{\mu}{n}\right)^{n-k}. \qquad (5.19)$$

Допустим теперь, что n неограниченно возрастает, в то время как k остается небольшим. Тогда $n-k \approx n$ и $n-k+1/2 \approx n$, так что

$$P(k) \to \frac{(1-\mu/n)^n}{(1-k/n)^n}\,\frac{1}{e^k}\,\frac{\mu^k}{k!}.$$

Из курса дифференциального исчисления вспомним, что

$$\lim_{n\to\infty}(1+x/n)^n = e^x.$$

Следовательно,

$$P(k) \to \frac{1}{e^{-k}e^k}\,\frac{\mu^k}{k!}e^{-\mu}$$

и в пределе

$$P(k) = \frac{\mu^k}{k!}e^{-\mu}. \qquad (5.20)$$

Это и есть формула распределения Пуассона.

Распределение Пуассона является дискретным; это значит, что переменная принимает только целые значения (включая 0). Однако математическое ожидание может иметь любое положительное значение.

В табл. 5.2 и на рис. 5.5 приведены примеры распределений Пуассона для математических ожиданий 0,1; 1 и 10.

Таблица 5.2

Распределение Пуассона

$\mu = 0{,}1$

k	0	1	2	3	4
$P(k)$	0,90484	0,09048	0,00452	0,00015	0,00000

$\mu = 1$

k	0	1	2	3	4	5	6	7
$P(k)$	0,3679	0,3679	0,1839	0,0613	0,0153	0,0031	0,0005	0,0000

$\mu = 10$

k	0	1	2	3	4	5	6	7
$P(k)$	0,00004	0,0004	0,0023	0,0076	0,0189	0,0378	0,0631	0,0901
k	8	9	10	11	12	13	14	15
$P(k)$	0,1126	0,1251	0,1251	0,1137	0,0948	0,0729	0,0521	0,0347
k	16	17	18	19	20	21	22	23
$P(k)$	0,0217	0,0128	0,0071	0,0037	0,0018	0,0009	0,0004	0,0002
k	24	25	—	—	—	—	—	—
$P(k)$	0,0001	0,0000	—	—	—	—	—	—

Выражение для распределения Пуассона (5.20) удовлетворяет нашим требованиям к распределению, (5.17). Так как каждый из трех множителей $e^{-\mu}$, μ^k и $k!$ всегда положителен, то $P(k)$ также должна всегда быть положительна. Сумма всех значений выражения $P(k)$ равна, конечно, 1, так как $\sum_{k=0}^{\infty} (\mu^k/k!)$ является разложением для e^{μ} и $e^{\mu} e^{-\mu} = 1$.

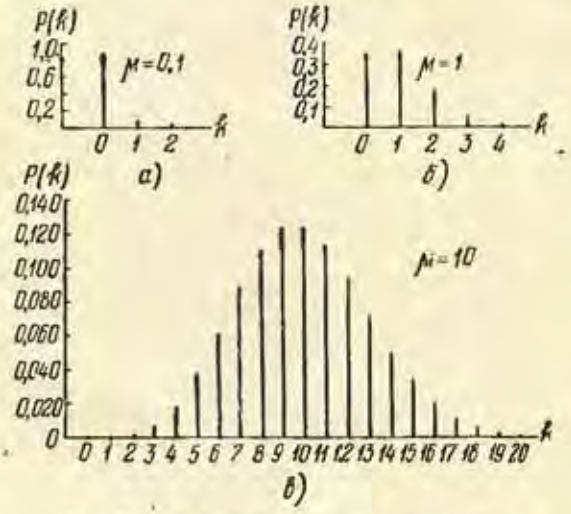


Рис. 5.5. Распределение Пуассона.

Математическое ожидание и дисперсия распределения Пуассона. Мы называли μ математическим ожиданием. Чтобы показать, что эта величина действительно является математическим ожиданием, и определить дисперсию, мы снова введем *вспомогательную переменную* t, которую затем приравняем 1 и исключим из полученного выражения.

Разложение $e^{\mu t}$ в ряд имеет вид

$$e^{\mu t} = \sum_{k=0}^{\infty} \frac{(\mu t)^k}{k!}.$$

Умножая на $e^{-\mu}$, получаем

$$e^{-\mu} e^{\mu} = e^{-\mu} \sum_{k=0}^{\infty} \frac{\mu^k t^k}{k!}.$$

Дифференцируя по t, получаем

$$e^{-\mu}\mu e^{\mu t} = e^{-\mu}\sum_{k=0}^{\infty}\frac{k\mu^k t^{k-1}}{k!}. \qquad (5.21)$$

Повторное дифференцирование (что будет использовано нами для определения дисперсии) дает

$$e^{-\mu}\mu^2 e^{\mu t} = e^{-\mu}\sum_{k=0}^{\infty}\frac{k(k-1)\mu^k t^{k-2}}{k!}. \qquad (5.22)$$

Положив $t=1$ в (5.21) и (5.22), соответственно получим:

$$\mu = e^{-\mu}\sum_{k=0}^{\infty}\frac{k\mu^k}{k!} = \sum_{k=0}^{\infty}kP(k) \qquad (5.23)$$

и

$$\mu^2 = e^{-\mu}\sum_{k=0}^{\infty}\frac{(k^2-k)\mu^k}{k!}. \qquad (5.24)$$

Но правая часть выражения (5.23) есть по определению $E(k)$. Следовательно, мы доказали, что математическое ожидание распределения Пуассона равно μ. Из выражения (5.24), производя перестановку, получаем

$$e^{-\mu}\sum_{k=0}^{\infty}\frac{k\mu^k}{k!} = \sum_{k=0}^{\infty}\frac{k^2 e^{-\mu}\mu^k}{k!} - \mu^2 =$$

$$= \sum_{k=0}^{\infty}k^2P(k) - \mu^2. \qquad (5.25)$$

Мы только что показали, что левая часть выражения (5.25) равна μ; следовательно, правая часть по определению есть дисперсия. Следовательно, дисперсия распределения Пуассона дается формулой

$$\sigma^2 = \mu. \qquad (5.26)$$

Пример. Если вероятность гибели самолета во время одного полета равна 10^{-5}, то какова вероятность гибели пассажира, решившего совершить 10^5 полетов?

Решение. Это та же самая задача, которую мы решали раньше при помощи биномиального распределения. Мы имеем распределение Пуассона с математическим ожиданием 1; k в данном случае есть число роковых полетов среди 100 000 полетов. Мы хотим найти решение для $k=0$. Итак, $P(0)=e^{-1}$ 10/0/$=e^{-1}=0{,}3679$. Ответ на задачу мы получим, если из 1 вычтем эту величину, что даст 0,6321.

Как указывалось раньше, распределение Пуассона и этом случае является очень хорошим приближением к биномиальному распределению; это объясняется тем, что ограничения, о которых мы говорили в начале вывода формул, в этой задаче очень хорошо выполняются.

Второй вывод распределения Пуассона. Существует много случаев, когда эти ограничения не выполняются, а распределение Пуассона тем не менее хорошо соответствует реальному закону распределения событий. Так, например, можно ожидать, что число автомашин, проходящих через некоторую точку (удаленную от любого пункта, где движение нарушается, например от светофора) в определенный интервал времени, подчиняется распределению Пуассона. Смысл математического ожидания для такого распределения достаточно ясен, однако определить подходящие n и p таким образом, чтобы наш предыдущий вывод формулы оставался в силе, затруднительно.

Распределение Пуассона применимо также ко многим другим случайным величинам, как-то: число потребителей электроэнергии, повернувших выключатели в течение определенного интервала времени; число автомашин, прибывающих в течение определенного интервала времени к заставе для сбора пошлин *; число телефонных вызовов или число телефонных вызовов какого-нибудь определенного класса в течение определенного интервала времени; число бомб, упавших на определенную площадь; и т. д.

Рассмотрим обобщенный случай, при котором некоторое событие происходит случайно во времени, и возьмем единичный интервал времени. Очень важно всегда выяснить весьма тщательно точный смысл слова «случайный», т. е. точно найти, какой именно механизм событий объявляется случайным. В рассматриваемом случае мы будем предполагать, что единица времени разделена на n равных подынтервалов (продолжительностью $1/n$) и что существует вероятность p наступления одного или более событий в течение каждого из подынтервалов (т. е. что существует вероятность $1-p$, что никакое событие не наступит в одном рассматриваемом подынтервале).

То обстоятельство, что вероятность одинакова для каждого из подынтервалов, обладающих одинаковой продолжительностью, определяет случайный механизм событий, если только мы введем еще одно ограничение, а именно, что наступление событий в течение одного подынтервала совершенно не зависит от событий в течение всех других подынтервалов. Это ограничение является допущением, и притом весьма важным. Можно было бы

* В США многие автомобильные магистрали являются частной собственностью и за пользование ими взимается плата при въезде на дорогу. — *Прим. ред.*

сделать разные допущения. Например, мы могли бы предположить, что строго определенное число событий распределено случайно в пределах единичного интервала. Это было бы совсем другим ограничением, так как очевидно, что наступление или ненаступление события в том или ином подынтервале будет тогда изменять вероятность событий в последующих подынтервалах.

В нашем случае мы получаем биномиальное распределение, в котором благоприятным исходом является наступление одного или более событий в течение одного подынтервала, а число испытаний равно числу подынтервалов n. Ожидаемое значение числа благоприятных исходов (т. е. числа содержащих события подынтервалов в единичном интервале времени) равно, следовательно, np.

Ожидаемого значения общего числа событий мы пока не знаем, так как нам не известно, как часто подынтервалы будут содержать больше одного события. Однако, если заставить n расти неограниченно, то длина dt каждого подынтервала становится бесконечно малой и вероятность наступления в течение подынтервала более чем одного события становится по сравнению с dt бесконечно малой высшего порядка. Тогда мы можем утверждать, что вероятность p есть вероятность наступления в течение подынтервала точно одного события. Это можно доказать непосредственно из допущения о независимости событий, однако для того чтобы подчеркнуть всю значимость этого соотношения при применении распределения Пуассона к физическим процессам, мы будем считать его нашим вторым допущением. Таким образом, число подынтервалов, содержащих события, становится равным общему числу событий.

Когда n возрастает неограниченно, произведение np стремится к своему пределу, который мы обозначим через μ. Но тем самым мы пришли точно к тем же предположениям, которыми пользовались раньше при выводе распределения Пуассона, а именно, получили биномиальное распределение, в котором произведение np остается неизменным при возрастании n.

Другими словами, если:

а) наступления некоторого определенного события рассеяны случайно во времени или пространстве;

б) ожидаемое число наступлений события в единице времени или пространства равно μ;

в) наступление события в определенной точке времени или пространства не зависит от наступления или ненаступления события в любой другой точке;

г) наступление двух событий точно в одной и той же точке фактически невозможно—

то вероятность наступления точно k событий в единице времени или пространства определяется выражением.

$$P(k)=\frac{e^{-\mu}\mu^k}{k!}. \qquad (5.20)$$

Таким образом, распределение Пуассона следует рассматривать не только как предельный случай биномиального распределения, но и как непосредственный аналог определенных физических явлений.

Пример. а) Если автомашины проходят через данный пункт в соответствии с распределением Пуассона с математическим ожиданием 1 автомашина в 1 *сек*, то чему равно стандартное отклонение? б) Если автомашины проходят с математическим ожиданием 60 автомашин в 1 *мин*, то чему равно стандартное отклонение? в) Если они проходят с математическим ожиданием 3600 автомашин в час?

Р е ш е н и е. Так как дисперсия распределения Пуассона численно равна его математическому ожиданию, а стандартное отклонение равно корню квадратному из дисперсии, то ответы соответственно равны: а) 1 автомашина; б) 7,74 автомашины; в) 60 автомашин. Однако распределения в этих трех случаях тождественны. Как мы увидим в § 6.4, стандартное отклонение во многих распределениях (включая распределения Пуассона, в которых математическое ожидание достаточно велико) представляет собой отклонение, значение которого превышается примерно в одном случае из каждых трех. Хотя это эмпирическое правило не выполняется достаточно хорошо для распределения Пуассона с математическим ожиданием, равным единице или меньше ее, все же стандартное отклонение дает грубую оценку порядка среднего отклонения от математического ожидания.

Таким образом, полученные нами ответы говорят о следующем. Если мы наблюдаем за дорогой в течение одной секунды, то число машин, которое мы ожидаем увидеть, равно 1, но мы не должны удивляться, увидев вместо этого две или не увидев ни одной машины; если наше наблюдение продолжается 1 *мин*, то шансы будут приблизительно 2 к 1, что мы увидим от 52 до 68 автомашин; если наше наблюдение продолжается один час, то шансы будут приблизительно 2 к 1, что мы увидим от 3540 до 3660 автомашин.

Наблюдаемая здесь тенденция уменьшения в процентном отношении отклонения числа наблюдаемых автомашин от математического ожидания и увеличения этого отклонения в абсолютных единицах является весьма общей; о ней говорит также характер кривых на рис. 5.3. С количественной точки зрения эта тенденция будет рассмотрена нами в § 8.3.

Это обстоятельство можно было бы учесть при выводе распределения Пуассона, взяв интервал длины t вместо единичного интервала. Тогда, если μ — число событий за единицу времени, распределение Пуассона принимает следующий вид:

$$P(k)=\frac{e^{-\mu t}(\mu t)^k}{k!}. \qquad (5.27)$$

Математическое ожидание этого распределения равно μt и дисперсия также равна μt.

В рассмотренной выше задаче временные интервалы между событиями остаются без изменения, рассматриваем ли мы период, равный 1 *сек*, 1 *мин* или 1 *час*. Следует обратить внимание на то, что распределение продолжительностей интервалов между событиями не подчиняется закону Пуассона. Их распределение будет обсуждаться в § 6.7, после того как мы познакомимся с распределением таких непрерывных переменных, как время.

5.7. Случайность

Примером распределения Пуассона в пространстве может служить распределение бомб по площади. Во время II мировой войны около 3000 самолетов-снарядов V-1 и ракет V-2 упало на площади около 300 *кв. миль* в Лондоне и его окрестностях. Хотя вероятность падения должна быть наибольшей вблизи точки прицеливания, которой, по-видимому, был центр этой обстреливаемой площади, достаточно большое количество снарядов и ракет, упавших вне этого района, позволяет в первом приближении предположить равномерное распределение вероятностей в пределах этого района. Другие допущения для распределения Пуассона выполнялись достаточно хорошо. Следовательно, мы можем предположить, что количество снарядов и ракет, падающих на площадь в 0,1 кв. мили, следовало распределению Пуассона с математическим ожиданием 1.

Если всю площадь в 300 кв. миль заблаговременно разделить на 3000 приблизительно равных участков, то в соответствии с табл. 5.2 следует ожидать, что примерно девять или десять из этих участков примут каждый по пяти снарядов и ракет и примерно два участка — по шести. Если разбивка площади на участки производилась после обстрела таким образом, что скопления упавших снарядов и ракет включались в один участок, и если, кроме того, была допущена некоторая подтасовка при попытках включить побольше упавших снарядов и ракет в один участок, то мы можем ожидать, что на некоторых участках площадью в 0,1 кв. милю будет наблюдаться падение значительно большего числа снарядов и ракет, чем это следует из наших расчетов.

Эти явления имели, конечно, место в действительности, и на них обращали внимание жители таких районов. Население говорило, что на некоторые определенные районы, охватывающие несколько кварталов, падало большое количество «фау» — значительно больше, чем этого следовало бы ожидать в среднем. Как мы показали, это именно то, чего и следует ожидать при случайном распределении; однако многие жители придерживались противоположного мнения: они утверждали, что такая группировка точек падения указывает на точное прицеливание по определенным пунктам.

Мы пытаемся продемонстрировать природу случайных явлений и показать, что одно из их свойств заключается в появлении событий, которые имеют такой вид, что неискушенный наблюдатель принимает их за неслучайные. Предположим, что имеется длинная последовательность цифр, якобы взятых из таблицы случайных чисел. Если мы выберем наугад последовательность в 10 цифр и окажется, что она содержит точно по одной каждой цифре от 0 до 9, то мы можем заподозрить, что таблица не была случайной, так как шансы против получения такой последовательности будут приблизительно 3000 к 1.

С другой стороны, если в 30 000 выборок отсутствует последовательность 10 цифр, содержащая точно по одной каждой цифре, мы опять могли бы заподозрить отсутствие случайности, так как здесь перед нами распределение Пуассона с математическим ожиданием 10 и шансы против получения 0 благоприятных исходов будут 22 000 к 1. Наконец, если последовательность десяти якобы случайных цифр содержит расположенные по порядку цифры от 0 до 9, мы должны с полным основанием заявить о неслучайности, так как шансы против этого будут 999 999 999 к 1. Заметим, однако, что вероятность *любой другой заранее выбранной* последовательности столь же мала.

ЛИТЕРАТУРА

См. гл. 8.

ЗАДАЧИ

См. гл. 8.

ГЛАВА 6

РАСПРЕДЕЛЕНИЯ НЕПРЕРЫВНЫХ ПЕРЕМЕННЫХ

До сих пор мы рассматривали только распределения переменных, принимающих дискретные значения. Однако теория вероятностей применима также и к непрерывным переменным. Рассмотрим хорошо сбалансированное колесо, обод которого при вращении проходит в непосредственной близости от стрелки-указателя пренебрежимо малой толщины.

Если окружность колеса разделена на n равных частей, то естественно предположить, что после остановки колеса стрелка имеет равную вероятность оказаться над каждой частью. Так как имеется n частей, а вероятность оказаться над какой-то частью равна 1, то вероятность для каждой части равна $1/n$.

Выберем сначала на окружности колеса точку, а затем разделим окружность на n частей таким образом, чтобы точка находилась внутри одного из этих интервалов; при необходимости условимся считать, что каждая разделяющая черта входит в состав интервала, находящегося от нее по ходу часовой стрелки. Затем будем все увеличивать n. Тогда вероятность остановки в интервале, содержащем заранее предписанную точку, будет становиться все меньше и меньше. В пределе вероятность остановки над точкой станет равна 0. Однако колесо будет останавливаться над какой-нибудь точкой. Вполне понятно, что такой точкой может оказаться и заранее нами выбранная.

Таким образом, рассматривая этот пример, мы столкнулись с новым явлением. До сих пор мы считали, что если вероятность наступления события равна 0, то это событие невозможно. Однако в случае непрерывной переменной нулевая вероятность не означает невозможности.

Это затруднение не заключено в самом понятии вероятности. Оно возникает всюду в математике в связи с понятием предела. Так, например, при операции интегрирования мы суммируем ординаты «нулевой» площади, чтобы получить конечную площадь. Подобно этому мы складываем бесконечно малые элементы вероятности, чтобы получить вероятность конечной величины.

6.1. Равномерное распределение

В примере с колесом видно без дальнейших выкладок, что вероятность остановки над интервалом длины x равна x/λ, если длина всей окружности колеса равна λ. Вероятность остановки над интервалом длины dx равна dx/λ. Таким образом, можно написать

$$P(x)\,dx=\frac{1}{\lambda}dx. \qquad (6.1a)$$

Функция без дифференциала (в нашем случае $1/\lambda$) называется *плотностью распределения вероятностей*, или *функцией плотности*, или *функцией распределения**. При умножении этой функции на dx мы получаем *элемент вероятности*. Так как для колеса эта вероятность одинакова для каждого интервала (от x до $x+dx$), то выражение (6.1а) называется *равномерным распределением*. Если окружность колеса развернуть в линию в масштабе 1 : 1 и на этой линии выбрать случайную точку, все сказанное выше останется в силе. На рис. 6.1 приведен график равномерного распределения.

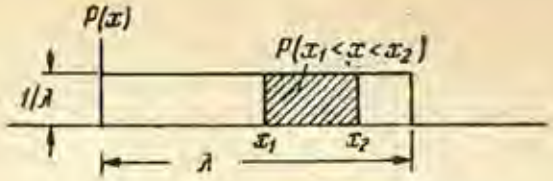


Рис. 6.1. Равномерное распределение.

По соображениям, вполне понятным из этого рисунка, такое распределение часто называют *прямоугольным*. В случае, когда оно нормируется так, что $\lambda=1$, его иногда называют *квадратным*. Для того чтобы найти вероятность случайно выбранной точке лежать в интервале $x_1\leqslant x\leqslant x_2$, мы должны сложить вероятности для каждого элемента этого интервала, т. е. проинтегрировать функцию плотности вероятностей. Таким образом,

$$P(x_1\leqslant x\leqslant x_2)=1/\lambda\int_{x_1}^{x_2}dx=\frac{x_2-x_1}{\lambda}.$$

Условия (5.17), ограничивающие функции дискретных переменных, которые могут быть распределениями вероятностей, применимы, с очевидными видоизменениями, к функциям непрерывных переменных, которые могут быть плотностями распределения вероятностей:

$$P(x)\geqslant 0 \quad \text{для всех } x \qquad (6.2a)$$

и

$$\int_{-\infty}^{\infty}P(x)\,dx=1. \qquad (6.2б)$$

Любая функция, удовлетворяющая этим условиям, может быть плотностью распределения вероятностей. Функция не обязательно должна быть непрерывной. Так, выражение (6.1) следовало бы, собственно говоря, записать в следующем виде:

$$\left.\begin{aligned} P(x)\,dx&=0 \quad \text{при } x<0\\ P(x)\,dx&=\frac{1}{\lambda}dx \quad \text{при } 0\leqslant x\leqslant\lambda\\ P(x)\,dx&=0 \quad \text{при } x>\lambda \end{aligned}\right\}. \qquad (6.1б)$$

* Под термином «функция распределения» иногда подразумевают *кумулятивную* функцию (рис. 6.7). Мы будем применять только вполне однозначный термин «*плотность распределения вероятностей*». — *Прим. авт.*

Но такая функция уже не является непрерывной.

В случае дискретной переменной условие $\Sigma P(k)=1$ влечет $P(k)\leqslant 1$ для всех k; выражение $P(k)>1$ не имело бы смысла. В случае непрерывной переменной таких ограничений для $P(x)$ не существует и $P(x)$ может принимать значения больше 1 *. Функция плотности вероятностей *не является* вероятностью; элемент же вероятности является вероятностью, а именно вероятностью события в интервале от x до dx.

6.2. Математическое ожидание и дисперсия

Мы определяем ожидаемое значение, математическое ожидание и дисперсию, как и раньше (§ 5.2), с очевидными видоизменениями:

$$E[f(x)]=\int_{-\infty}^{\infty} f(x)\,P(x)\,dx, \qquad (6.3)$$

$$\mu=E(x)=\int_{-\infty}^{\infty} xP(x)\,dx \qquad (6.4)$$

и

$$\sigma^2=E(x-\mu)^2=\int_{-\infty}^{\infty} x^2P(x)\,dx-\mu^2. \qquad (6.5)$$

Символ ожидаемого значения является, как и раньше, линейным оператором [формула (5.5)], так как таковым является символ интеграла. Для равномерного распределения

$$\mu=\int_{-\infty}^{\infty} xP(x)\,dx=\int_{-\infty}^{\infty} x\frac{dx}{\lambda}+\int_{0}^{\lambda}\frac{dx}{\lambda}=$$

$$=\frac{1}{\lambda}\left[\frac{x^2}{2}\right]_0^{\lambda}=\frac{\lambda^2}{2\lambda}=\frac{\lambda}{2},$$

как и следовало ожидать. Дисперсия равна

$$\sigma^2=\frac{1}{\lambda}\int_0^{\lambda} x^2dx-\mu^2=\frac{\lambda^2}{3\lambda}-\left(\frac{\lambda}{2}\right)^2=$$

$$=\frac{\lambda^2}{3}-\frac{\lambda^2}{4}=\frac{\lambda^2}{12}.$$

* Более того, $P(x)$ может равняться в практических случаях бесконечности. Так, например, кумулятивная вероятность обнаружения приближающегося самолета на дальностях вплоть до нулевой может достигать 99%, а оставшийся 1% падает на обнаружение на нулевой дальности. В этом случае кумулятивная функция имеет разрыв в точке $x=0$ (является ступенчатой функцией) и ее производная, функция плотности вероятности, в точке $x=0$ равна бесконечности. Случай такого рода изображен на рис. 23.2. — *Прим. авт.*

6.3. Моменты, медианы и моды

Если вычертить функцию плотности вероятности на картоне, имеющем всюду одинаковую толщину, и затем обрезать его по кривой $P(x)$ и оси x, вырезав из него некоторую фигуру, то можно убедиться, что эта фигура будет находиться в равновесии, если за точку опоры принять точку $x=\mu$, но не какую-либо другую точку на оси x. Это вытекает из определения математического ожидания, так как из механики мы знаем, что момент первого порядка относительно начала координат для тела, имеющего одно измерение и распределение массы вдоль оси x в соответствии с функцией $f(x)$, равен

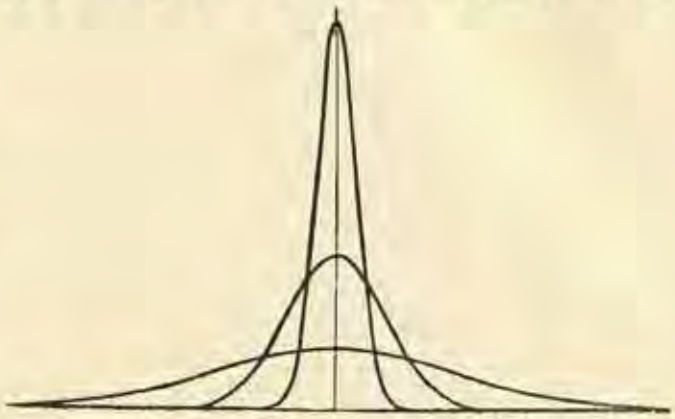


Рис. 6.2. Плотности распределения вероятностей с различными вторыми моментами.

$$M=\int_{-\infty}^{\infty} xf(x)\,dx.$$

Но если в этом выражении $f(x)$ заменить на $P(x)$, оно станет точным определением ожидаемого значения.

Моменты. Математическое ожидание распределения часто называют первым моментом относительно начала координат или, более кратко, первым моментом распределения. В общем случае величина $E(x^r)$ называется r-м моментом распределения относительно начала координат, а величина $E[(x-\mu)^2]$ называется r-м моментом относительно математического ожидания или, когда $r\geqslant 2$, просто r-м моментом **. Нулевой момент любого распределения вероятностей равен единице. Первый момент есть математическое ожидание и измеряет положение. Второй момент есть дисперсия и измеряет рассеяние. Высшие мо-

** Моменты относительно начала координат называют также *начальными моментами*, а моменты относительно математического ожидания — *центральными моментами*. — *Прим. ред.*

Рис. 6.3. Плотности распределения вероятностей с различными третьими моментами.

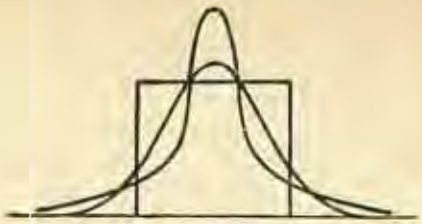


Рис. 6.4. Плотности распределения вероятностей с различными четвертыми моментами.

менты также измеряют характеристики распределения, и так как третий и четвертый моменты изредка встречаются в литературе, мы их кратко опишем.

На рис. 6.2—6.4 показаны различные распределения, которые имеют одинаковые математические ожидания, но отличаются одно от другого остальными, высшими моментами. Все три распределения, показанные на рис. 6.2, являются нормальными (см. § 6.4), с математическим ожиданием 0; отличаются они своей дисперсией. Так как нулевой момент в любом случае равен 1, кривая, которая около математического ожидания проходит выше других кривых, на краях должна спускаться ниже их. Эта разница в характере кривых выражается в самом определении дисперсии.

На рис. 6.3 приведены два распределения (одно из них нормальное), причем математическое ожидание и стандартное отклонение каждого равны единице. Однако эти кривые имеют различный вид: в то время как нормальное распределение является симметрическим относительно математического ожидания, второе распределение скошено в правую сторону.

С математической точки зрения это объясняется тем, что третий момент нормального распределения равен нулю, а третий момент другого распределения больше нуля. Для распределения Пуассона, например, все моменты численно равны математическому ожиданию. Мы говорим «*численно равны*» потому, что они имеют различную размерность. Так, в рассмотренном нами примере распределения Пуассона математическое ожидание было равно 60 автомашинам в 1 *сек*, дисперсия 60 *авто*^{2}/*сек*^{2}, а третий момент 60 *авто*^{3}/*сек*^{3}.

Чтобы исключить это неудобство, для измерения рассеяния обычно пользуются стандартным отклонением. В нашем примере оно имеет величину 7,75 автомашин в 1 *сек*. Подобно этому безразмерная мера, получаемая из третьего момента (путем деления третьего момента на дисперсию в степени $^3/_2$), используется для характеристики *скошенности*, или *асимметрии*, распределения.

На рис. 6.4 показаны три распределения, которые имеют одинаковые третьи моменты; они отличаются своим четвертым моментом, который при делении на квадрат дисперсии характеризует *островершинность, плосковершинность, крутость или куртозис* *. Равномерное распределение является менее островершинным, чем нормальное; третье распределение является еще более островершинным.

Медиана. Первый момент часто является основной мерой положения центра распределения, особенно в математической статистике; однако иногда, особенно в описательной статистике, применяются и другие меры, например медиана и мода. Медиана распределения вероятностей для непрерывной переменной определяется соотношением

$$\int_{-\infty}^{x_m} P(x)\,dx = \frac{1}{2} = \int_{x_m}^{\infty} P(x)\,dx.$$

Возможность распространения этого математического соотношения на распределения дискретной переменной достаточно очевидна без дополнительного пояснения, однако его вывод для этого случая достаточно кропотлив. При симметрическом распределении медиана и математическое ожидание одинаковы, но при асимметрическом характере распределения эти величины отличаются друг от друга.

Медиана является наиболее подходящим выражением среднего значения, когда мы не хотим, чтобы на результат слишком сильно влияли экстремальные отклонения. Так,

---

* Термин „*куртозис*" (kurtosis) образован от греч. κορτός „кривой, искривленный" и буквально значит: „кривизна"; по-русски его иногда передают словом „*крутость*". Наряду с куртозисом применяется также *коэффициент эксцесса* $\rho_4' = \rho_4 - 3$, где ρ_4 — куртозис. Для нормального распределения (см. ниже § 6.4) $\rho_4' = 0$; распределения с $\rho_4' > 0$ называются островершинными, распределения с $\rho_4' < 0$ — плосковершинными. Асимметрию можно обозначать через ρ_3. — *Прим. ред.*

в классической экономии, говоря о величине заработка, медианой определяют заработок «среднего человека», а математическое ожидание характеризует значительно более высокий заработок, так как на его величину влияют в этом случае немногие очень высокие заработки.

В некоторых случаях возникает потребность в применении своеобразного среднего значения, на величину которого экстремальные отклонения влияли бы даже больше, чем это имеет место в математическом ожидании. Таким свойством обладает среднеквадратическое значение *.

Мода. Модой распределения называется наиболее вероятное значение. При симметрических распределениях мода равна математическому ожиданию и медиане, однако при асимметрических распределениях она может значительно отличаться от них. Для того чтобы найти моду распределения непрерывной переменной, мы дифференцируем функцию плотности вероятности и приравниваем производную нулю. Для дискретных переменных необходимо исследовать отношение $P(k)$ к $P(k-1)$ для $k>0$ и определить, когда оно переходит из величины, большей единицы, в величину, меньшую единицы. Так, для биномиального распределения.

$$\frac{P(k)}{P(k-1)}=\frac{n!}{k!(n-k)!}p^k(1-p)^{n-k}:$$

$$:\frac{n!}{(k-1)!(n-k+1)!}p^{k-1}(1-p)^{n-k+1}=$$

$$=\frac{n-k+1}{k}\frac{p}{1-p}=\frac{(n+1)p-kp}{k-kp}. \qquad (6.6)$$

Это выражение принимает значение, превышающее единицу, равное единице или меньшее единицы, когда k принимает значение меньшее, равное или большее $(n+1)p$. Следовательно, мода биномиального распределения равна наибольшему целому числу, меньшему чем $(n+1)p$; если $(n+1)p$ — целое число, это значение также является модальным.

Для распределения Пуассона

$$\frac{P(k)}{P(k-1)}=\frac{e^{-\mu}\mu^k}{k!}:\frac{e^{-\mu}\mu^{k-1}}{(k-1)!}=\frac{\mu}{k}.$$

Таким образом, наибольшее целое число, меньшее чем математическое ожидание, является модой, и если математическое ожидание является целым числом, его также следует считать модальным. Сделанные нами выводы нельзя считать достаточно строгими, так как мы не доказали, что другие моды отсутствуют. Бимодальные распределения (рис. 6.5), встречающиеся иногда на практике, возникают обычно из-за воздействия на распределение двух независимых основных явлений.

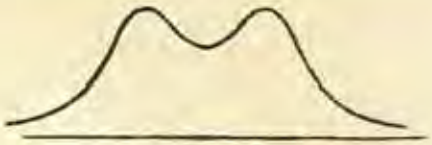


Рис. 6.5. Плотность бимодального распределения вероятностей.

6.4. Нормальное распределение

Наиболее важным распределением вероятностей в науке является нормальное распределение, или, как его часто называют, гауссово распределение (распределение Гаусса). Оно может быть получено несколькими различными способами. Рассмотрим сначала функцию вида

$$P(x)\,dx=C_1\exp[-C_2(x-C_3)^2]\,dx.$$

Попытаемся определить содержащиеся в ней постоянные, чтобы получить функцию плотности вероятности. Затем мы покажем, что полученная в результате вычислений функция, которую мы назовем *нормальной*, является предельным случаем многих других распределений (когда n достаточно велико) и обладает рядом других важных свойств.

Вывод. Для того чтобы определить постоянные C_1, C_2 и C_3, нам потребуется найти следующие три интеграла:

$$\int_{-\infty}^{\infty}e^{-y^2}dy,\quad \int_{-\infty}^{\infty}ye^{-y^2}dy \quad \text{и} \quad \int_{-\infty}^{\infty}y^2e^{-y^2}dy.$$

Совершенно очевидно, что первый и третий интегралы являются симметрическими относительно $y=0$, и в связи с этим мы можем заменить их соответственно на $2\int_0^{\infty}e^{-y^2}$ и $2\int_0^{\infty}y^2e^{-y^2}dy$.

Применив в обоих случаях подстановку

$$z=y^2 \text{ и } dy=\frac{1}{2}z^{1/2}dz,$$

получаем:

$$\int_{-\infty}^{\infty}e^{-y^2}dy=2\int_0^{\infty}1/2z^{-1/2}e^{-z}dz=\Gamma\left(\frac{1}{2}\right)=\sqrt{\pi} \qquad (6.7)$$

* Т. е. корень квадратный из математического ожидания квадрата рассматриваемой случайной переменной. — *Прим. ред.*

и

$$\int_{-\infty}^{\infty} y^2 e^{-y^2} dy = 2\int_0^{\infty} \frac{1}{2} z^{-\frac{1}{2}} z e^{-z} dz =$$

$$= \Gamma\left(\frac{3}{2}\right) = \frac{\sqrt{\pi}}{2}. \qquad (6.8)$$

Второй интеграл равен

$$\int_{-\infty}^{\infty} y e^{-y^2} dy = \frac{1}{2}\int_{y=-\infty}^{y=\infty} e^{-y^2} d(y^2) =$$

$$= -\frac{1}{2}\left[e^{-y^2}\right]_{y=-\infty}^{y=\infty} = 0 - 0 = 0. \qquad (6.9)$$

Для определения C_1 используем соотношение $\int_{-\infty}^{\infty} P(x)\,dx = 1$.

Тогда

$$C_1 = \frac{1}{\int_{-\infty}^{\infty} \exp[-C_2(x-C_3)^2]\,dx}.$$

Подставляя

$$y = \sqrt{C_2}\,(x - C_3) \quad \text{и} \quad dx = \frac{dy}{\sqrt{C_2}}, \qquad (6.10)$$

получаем

$$C_1 = \frac{1}{\left(\frac{1}{\sqrt{C_2}}\right)\int_{-\infty}^{\infty} e^{-y^2} dy} = \frac{\sqrt{C_2}}{\sqrt{\pi}}.$$

Следовательно,

$$P(x) = \frac{\sqrt{C_2}}{\sqrt{\pi}} \exp[-C_2(x-C_3)^2].$$

Совершенно очевидно, что эта функция симметрична относительно $x = C_3$; следовательно, C_3 является математическим ожиданием. Покажем это.

Согласно выражению (6.4)

$$E(x) = \frac{\sqrt{C_2}}{\sqrt{\pi}} \int_{-\infty}^{\infty} x \exp[-C_2(x-C_3)^2]\,dx.$$

Производя подстановку соотношений (6.10) и учитывая, что $x = \frac{y}{\sqrt{C_2}} + C_3$, получаем

$$E(x) = \frac{\sqrt{C_2}}{\sqrt{\pi}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{C_2}}\left(\frac{y}{\sqrt{C_2}} + C_3\right) e^{-y^2} dy =$$

$$= \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} \frac{y}{\sqrt{C_2}} e^{-y^2} dy + \frac{C_3}{\sqrt{\pi}} \int_{-\infty}^{\infty} e^{-y^2} dy.$$

Используя выражение (6.7) и (6.9), окончательно находим

$$E(x) = 0 + \frac{C_3}{\sqrt{\pi}}\sqrt{\pi} = C_3.$$

Следовательно,

$$C_3 = \mu \quad \text{и} \quad P(x) = \frac{\sqrt{C_2}}{\sqrt{\pi}} \exp[-C_2(x-\mu)^2].$$

Теперь, используя выражение (6.5), определяем дисперсию

$$\sigma^2 = \frac{\sqrt{C_2}}{\sqrt{\pi}} \int_{-\infty}^{\infty} x^2 \exp[-C_2(x-\mu)^2]\,dx - \mu^2.$$

Производя подстановку $y = \sqrt{C_2}\,(x-\mu)$, находим

$$\sigma^2 = \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} \left(\frac{y}{\sqrt{C_2}} + \mu\right)^2 e^{-y^2} dy - \mu^2 =$$

$$= \frac{1}{\sqrt{\pi}}\left(\int_{-\infty}^{\infty} \frac{y^2}{C_2} e^{-y^2} dy + 2\int_{-\infty}^{\infty} \frac{\mu}{\sqrt{C_2}} y e^{-y^2} dy + \right.$$

$$\left. + \int_{-\infty}^{\infty} \mu^2 e^{-y^2} dy\right) - \mu^2 = \frac{1}{\sqrt{\pi}}\left(\frac{1}{C_2}\frac{\sqrt{\pi}}{2} + \right.$$

$$\left. + 0 + \mu^2\sqrt{\pi}\right) - \mu^2 = \frac{1}{2C_2}.$$

Следовательно,

$$C_2 = \frac{1}{2\sigma^2}$$

и

$$P(x) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \equiv N(\mu, \sigma). \qquad (6.11)$$

Свойства нормального распределения. Формула (6.11) представляет собой обычную, знакомую форму для плотности нормального распределения вероятностей. Совершенно очевидно, что она симметрична относительно математического ожидания, так как для любого значения $x = \mu + w$ существует соответствующее значение $x = \mu - w$, которое дает то же значение $P(x)$. Следовательно, медиана должна равняться μ. Покажем, что это распределение является унимодальным (т. е. обладает одной модой) и что мода находится в точке $x = \mu$:

$$\frac{dP(x)}{dx} = \frac{2}{\sqrt{2\pi}\,\sigma}\left\{\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]\right\} \times$$

$$\times\left[-\frac{2(x-\mu)}{2\sigma^2}\right] = 0. \qquad (6.12)$$

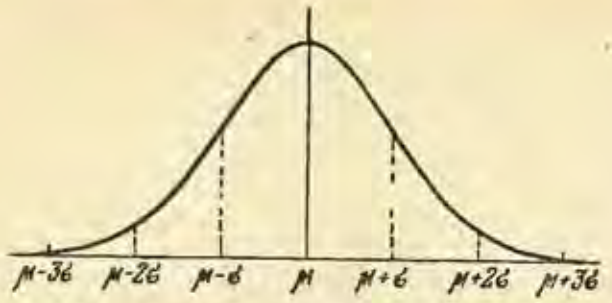


Рис. 6.6. Плотности нормального распределения вероятностей.

Уравнение (6.12) имеет решения при $x=-\infty$, $x=\infty$ и $x=\mu$. Первые два решения говорят о том, что функция у обоих краев асимптотически приближается к оси x. Последнее решение является модой.

Представляет также определенный интерес расположение точек перегиба кривой нормального распределения. Для их определения продифференцируем выражение (6.12) еще раз:

$$\frac{d^2P(x)}{dx^2}=\frac{1}{\sqrt{2\pi\,\sigma^3}}(x-\mu)\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]\times$$

$$\times\left(-\frac{x-\mu}{\sigma^2}\right)-\frac{1}{\sqrt{2\pi\sigma^3}}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]=0.$$

Разделив это выражение на

$$\left(\frac{1}{\sqrt{2\pi\sigma^3}}\right)\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

и устранив таким образом нежелательные решения для $x=\pm\infty$, получим

$$-\frac{(x-\mu)^2}{\sigma^2}+1=0,$$

или

$$x=\mu\pm\sigma.$$

Теперь мы можем вычертить кривую для нормального распределения (рис. 6.6).

Функция ошибок. Часто бывает целесообразным изменить координатные оси нормального распределения. Ось абсцисс, т. е. ось x, может быть смещена в сторону, или может быть изменен масштаб для отсчета по ней.

Обычно смещение производится таким образом, чтобы начало координат лежало в точке $x=\mu$, т. е. чтобы математическое ожидание было равно 0. Форма кривой, а следовательно, и дисперсия при таком смещении оси абсцисс остаются без изменения. Изменение масштаба заключается обычно в делении всех значений x на σ, в связи с чем стандартное отклонение становится равным единице. При таких преобразованиях стандартное отклонение делится на масштабный коэффициент, а дисперсия — на квадрат масштабного коэффициента. Получаемая при этом переменная, для которой математическое ожидание равно нулю, а дисперсия — единице, называется *нормированной переменной*. Формула (6.11) приобретает при этом простой вид

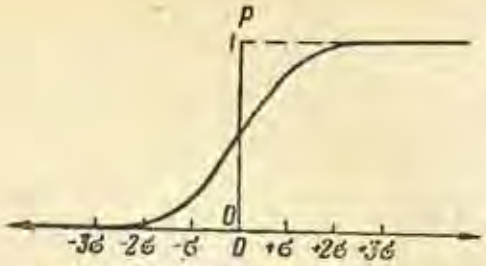


Рис. 6.7. Нормальная ожива

$$\int_{-\infty}^{x}\frac{1}{\sqrt{2\pi\sigma}}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]dx.$$

$$P(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}, \qquad (6.13)$$

и мы можем рассматривать рис. 6.6 как график этой функции.

Нормированный вариант нормального распределения называется иногда *функцией ошибок* (сокращенно erf); таблицы ее кумулятивных значений приводятся в книгах по статистике и в математических таблицах. Кумулятивная функция изображена на рис. 6.7. Кривые такой формы называются *оживами* (или *огивами*).

Любую данную функцию $N(\mu,\ \sigma)$ можно преобразовать в функцию ошибок с помощью преобразования $y=\frac{(x-\mu)}{\sigma}$. Этим способом мы можем найти любой нужный нам интеграл от нормальной функции:

$$\int_{x_1}^{x_2}N(\mu,\ \sigma)\,dx=\int_{y_1=\frac{x_1-\mu}{\sigma}}^{y_2=\frac{x_2-\mu}{\sigma}}N(0,\ 1)\,dy=$$

$$=\int_{-\infty}^{y_2}N(0,\ 1)\,dy-\int_{-\infty}^{y_1}N(0,\ 1)\,dy.$$

Таблицу функции ошибок можно составить различными способами; на рис. 6.8 показаны наиболее типичные из них. Применяя такие таблицы, следует ясно представлять себе, что именно в них содержится, и внимательно следить за знаком переменной. Обратим внимание на приведенные в таблицах сле-

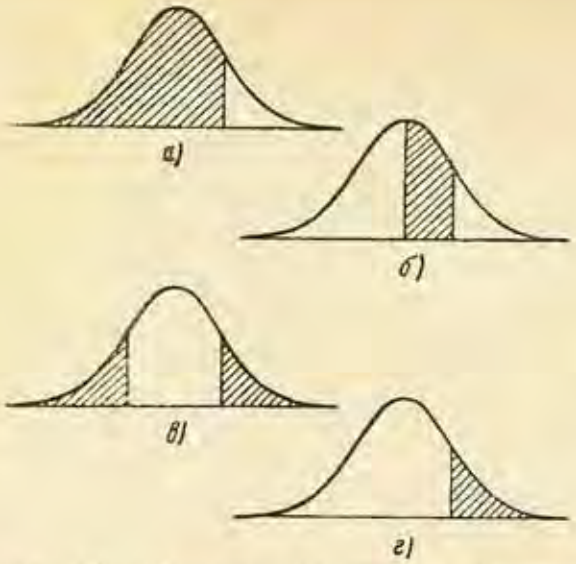


Рис. 6.8. Методы табулирования функции ошибок:

$$а)\ \int_{-\infty}^{x} N(0,\ 1)\,dx;\quad б)\ \int_{0}^{x} N(0,\ 1)\,dx;$$

$$в)\ 1-2\int_{0}^{x} N(0,\ 1)\,dx;\quad г)\ 0{,}5-\int_{0}^{x} N(0,\ 1)\,dx.$$

дующие особенно важные значения функции ошибок:

$$\int_{\mu-\sigma}^{\mu+\sigma} N(\mu,\ \sigma)\,dx=\int_{-1}^{+1} N(0,\ 1)\,dx=0{,}6827,$$

$$\int_{\mu-2\sigma}^{\mu+2\sigma} N(\mu,\ \sigma)\,dx=\int_{-2}^{+2} N(0,\ 1)\,dx=0{,}9545$$

и т. д.

Следующие приближенные числа целесообразно запомнить, так как они часто применяются при прикидочных расчетах:

1) отклонение от математического ожидания, превышающее $\pm\sigma$, имеет место приблизительно в одном случае из каждых трех;

2) отклонение, превышающее $\pm 2\sigma$, имеет место приблизительно в 5 случаях из 100;

3) отклонение, превышающее $\pm 3\sigma$, имеет место приблизительно в 3 случаях из 1000.

Заметим, что эти числа относятся к отклонениям *в обе стороны* от математического ожидания (рис. 6.8,*в*). Следовательно, вероятность превышения математического ожидания больше чем на стандартное отклонение (рис. 6.8,*г*) равна приблизительно 1/6.

Применения нормального распределения. Как будет показано в следующих главах, нормальное распределение находит очень широкое применение. Здесь мы покажем только, что оно является предельным случаем распределения Пуассона.

Как явствует из рис. 5.5, даже при таких небольших значениях математического ожидания, как 10, распределение Пуассона начинает становиться весьма похожим на нормальное распределение. Сейчас мы покажем аналитически, что по мере увеличения μ значение $P(k)$, предсказываемое распределением Пуассона, и значение $P(x)$, предсказываемое нормальным распределением, становятся одинаковыми.

Вероятность точного получения математического ожидания для распределения Пуассона (в предположении, что оно является целым числом) определяется выражением

$$P(\mu)=\frac{e^{-\mu}\mu^{\mu}}{\mu!}.$$

Полагая, что μ достаточно велико, мы можем применить приближенную формулу Стирлинга:

$$P(\mu)=\frac{e^{-\mu}\mu^{\mu}}{\sqrt{2\pi\mu}\,\mu^{\mu}e^{-\mu}}=\frac{1}{\sqrt{2\pi\mu}}.$$

Используя формулу (5.26), получаем

$$P(\mu)=\frac{1}{\sqrt{2\pi}\,\sigma}. \qquad (6.14)$$

Это выражение полностью совпадает с тем, которое мы получаем из формулы нормального распределения для случая $P(x=\mu)$. Аналогичным способом мы можем показать, что вероятность получения любого другого значения при распределении Пуассона полностью совпадает с вероятностью получения этого же значения при нормальном распределении, если только математическое ожидание достаточно велико и мы рассматриваем точку, лежащую не слишком далеко от математического ожидания. Определение $P(x)$ для точки, расположенной на расстоянии $\Delta\sigma$ от математического ожидания, где Δ обозначает небольшое число, легче произвести путем вычисления отношения значения $P(\mu)$, которое нам уже известно, к этой интересующей нас величине:

$$\frac{P(\mu)}{P(\mu+\Delta\sigma)}=\frac{e^{-\mu}}{e^{-\mu}}\,\frac{\mu^{\mu}}{\mu^{\mu+\Delta\sigma}}\,\frac{(\mu+\Delta\sigma)!}{\mu!}.$$

Применяя приближенную формулу Стирлинга, получаем

$$\frac{P(\mu)}{P(\mu+\Delta\sigma)}=\frac{1}{\mu^{\Delta\sigma}}\,\frac{(\mu+\Delta\sigma)^{\mu+\Delta\sigma+\frac{1}{2}}\,e^{-(\mu+\Delta\sigma)}\sqrt{2\pi}}{(\mu)^{\mu+\frac{1}{2}}\,e^{-\mu}\sqrt{2\pi}}=$$

$$=\frac{1}{e^{\Delta\sigma}}\left(\frac{\mu+\Delta\sigma}{\mu}\right)^{\mu+\Delta\sigma+\frac{1}{2}}.$$

Используя выражение (5.26), находим

$$\frac{P(\mu)}{P(\mu+\Delta\sigma)}=\left(1+\frac{\Delta}{\sigma}\right)^{1/2}\left(1+\frac{\Delta}{\sigma}\right)^{\Delta\sigma}\frac{\left(1+\frac{\Delta}{\sigma}\right)^{\sigma^2}}{e^{\Delta\sigma}}. \tag{6.15}$$

Рассмотрим теперь каждый из этих трех множителей в правой части равенства (6.15) в случае, когда σ становится значительно больше Δ. Первый множитель при этом становится равным 1. Второй множитель стремится к виду

$$\left[\left(1+\frac{\Delta}{\sigma}\right)^{\sigma}\right]^{\Delta}\to\left(e^{\Delta}\right)^{\Delta}=e^{\Delta^2}.$$

Для определения третьего члена, который мы обозначим через z, найдем его натуральный логарифм:

$$\ln z=\ln\frac{\left(1+\frac{\Delta}{\sigma}\right)^{\sigma^2}}{e^{\Delta\sigma}}=\sigma^2\ln\left(1+\frac{\Delta}{\sigma}\right)-\Delta\sigma.$$

Разложим этот логарифм в ряд, применив формулу $\ln(1+x)=x-\frac{x^2}{2}+\frac{x^3}{3}-\ldots$:

$$\left.\begin{aligned}\ln z&=\sigma^2\left(\frac{\Delta}{\sigma}-\frac{\Delta^2}{2\sigma^2}+\frac{\Delta^3}{2\sigma^3}-\ldots\right)-\\-\Delta\sigma&=-\frac{\Delta^2}{2}\left(1-\frac{2\Delta}{3\sigma}+\ldots\right).\end{aligned}\right\} \tag{6.16}$$

Следовательно, $z\to e^{-\frac{\Delta^2}{2}}$ и выражение (6.15) стремится к виду

$$\frac{P(\mu)}{P(\mu+\Delta\sigma)}\approx 1\times e^{\Delta^2}\times e^{-1/2\Delta^2}=e^{1/2\Delta^2},$$

откуда, используя выражение (6.14), получаем

$$P(\mu+\Delta\sigma)=\frac{P(\mu)}{e^{1/2\Delta^2}}=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2}\Delta^2}.$$

Это есть вероятность $k=\mu+\Delta\sigma$ удач при n испытаниях. Заменяя $\mu+\Delta\sigma$ на k и Δ на $(k-\mu)/\sigma$, получаем

$$P(k)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(k-\mu)^2}{2\sigma^2}\right]. \tag{6.17}$$

Функция (6.17) есть распределение вероятностей дискретной переменной. Однако математическое ограничение, которое не позволило нам рассматривать значения k, отличные от целых чисел, а именно наличие члена $k!$, снято с этого выражения применением приближенной формулы Стирлинга. Следовательно, вместо дискретной переменной k мы можем представить непрерывную переменную x, что эквивалентно проведению гладкой кривой через «полосатый» график на рис. 5.5,*в*. Эта подстановка дает выражение

$$P(x)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]=N(\mu,\ \sigma),$$

тождественное выражению (6.11). С помощью выражения (5.15) мы уже показали, по крайней мере частично, что то же самое мы можем получить непосредственно для биномиального распределения.

Нормальное распределение является почти самым хорошим приближением к распределению Пуассона и биномиальному распределению, какое можно получить с непрерывным распределением, симметрическим относительно математического ожидания. Асимметрия распределения Пуассона и биномиального распределения определяется соответственно выражениями $1/\sigma$ и $(1-2p)/\sigma$, в связи с чем нормальное распределение (с нулевой асимметрией) не может считаться хорошим приближением до тех пор, пока σ не примет достаточно больших значений, за исключением того частного случая, когда $1-2p$ очень близко к нулю.

В последнем случае (т. е. при биномиальном распределении с $p\approx 0{,}5$) нормальное распределение служит превосходным приближением, даже для столь малых n, как 25. В за-

Приближение	Ошибка приближения в % для:	
	$\sigma=5$, $\Delta\sigma=2$	$\sigma=100$, $\Delta\sigma=300$
1. Приближение Стирлинга эквивалентно $1+\frac{1}{12\sigma^2}+\ldots\approx 1$	0,3	0,001
2. $\left(1+\frac{\Delta}{\sigma}\right)^{\sigma}\approx e^{\Delta}$ эквивалентно $1+\Delta+\frac{\sigma-1}{\sigma}\frac{\Delta^2}{2!}+\ldots\approx$ $\approx\left(1+\Delta+\frac{\Delta^2}{2!}+\ldots\right)$	1	2
3. $\left(1+\frac{\Delta}{\sigma}\right)^{1/2}\approx 1$	4	1
4. $\left(1-\frac{3\Delta}{3\sigma}+\ldots\right)\approx 1$	6	1,4

ключение сведем в таблицу все приближенные выражения, которые были использованы при выводе нормального распределения из распределения Пуассона.

Действительная ошибка, получаемая при использовании выражения (6.17) в качестве приближения для (5.20), является сложной функцией величин, приведенных в этой таблице, но сравнительно невелика.

Пример. Согласно одному обзору рынка, посвященному сравнению двух видов продукции, покупатели закупили 4900 единиц продукции A и 5100 единиц продукции B. Какова вероятность такой ситуации, если выбор продукции производился покупателями случайно (т. е без предпочтения) и если каждая закупка была независимой от всех других?

Решение. Математически эту задачу можно решить, применив биномиальное распределение:

$$P(4900) = \frac{10\,000!}{(4900)!\,(5100)!}\left(\frac{1}{2}\right)^{4900}\left(\frac{1}{2}\right)^{5100}$$

Так как таблицы факториалов для таких больших чисел составляются очень редко, мы должны применить приближенную формулу Стирлинга, получив при этом $P=0{,}0010$. К несчастью, такой ответ не представляет интереса. В действительности нам нужна вероятность данного или *еще большего* отклонения от соотношения 5000 : 5000.

Для получения ответа на этот вопрос с помощью биномиального распределения потребовалось бы сложить 4901 член; в действительности только примерно сто членов будут влиять на точность и их можно найти с помощью рекурсивной формулы (6.6), но и эта работа будет весьма трудоемкой. Интересующую нас задачу можно решить значительно проще, применяя нормальное распределение. Стандартное отклонение, согласно формуле (5.12), равно $\sqrt{npq}=50$, а математическое ожидание равно $np=5000$. Таким образом, $k=5100$ (или 4900), $\mu=5000$, $\sigma=50$ и $(k-\mu)/\sigma=\pm 2$. Следовательно, мы хотим определить вероятность отклонения от математического ожидания нормального распределения на 2σ или более. Ответ, который мы можем получить из таблиц, равен 0,0227 или 0,0455 в зависимости от того, интересуемся ли мы $P(\leqslant 4900)$ или $P(\leqslant 4900)+P(\geqslant 5100)$. Таким образом, этот пример еще раз показал нам крайнюю важность правильной постановки вопроса в теории вероятностей.

6.5. Многомерные распределения

При рассмотрении в § 4.3 сложных и полных вероятностей мы говорили о двумерных дискретных вероятностях, т. е. о распределениях, содержащих две переменные. В других главах мы также в ряде случаев затрагивали вопросы, связанные с многомерными распределениями. Распространение этих понятий на непрерывные распределения в общем случае вполне очевидно.

Формула сложной вероятности (4.11) записывается следующим образом для непрерывных переменных:

$$P(x,\ y)\,dx\,dy = P(x)\,P_x(y)\,dx\,dy. \quad (6.18)$$

В случае независимых переменных она сводится к

$$P(x,\ y)\,dx\,dy = P(x)\,P(y)\,dx\,dy. \quad (6.19)$$

Знакомясь с одномерными распределениями, мы рассматривали событие как случайную точку на линии и вычерчивали график распределения вероятностей как плоскую кривую над этой линией. В случае двумерного распределения событие становится случайной точкой на плоскости, а распределение вероятностей изображается поверхностью, проходящей над плоскостью. Вероятность, что точка лежит в пределах площади A (раньше мы говорили об отрезке на линии), равна

$$P(x,\ y \text{ в } A) = \iint P(x,\ y)\,dx\,dy.$$

Любая функция двух переменных может быть плотностью распределения вероятностей, если она отвечает условиям, аналогичным формуле (6.2):

$$P(x,\ y) \geqslant 0 \text{ для всех значений } x,\ y, \quad (6.20а)$$

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x,\ y)\,dx\,dy = 1. \quad (6.20б)$$

Таким образом, элемент вероятности представляет собой прямоугольную призму длиной dx, шириной dy и высотой, определяемой функцией плотности.

При решении задач, связанных с многомерными распределениями, большое значение имеет уменье «взять» интеграл одной или нескольких переменных. Так, например, зная $P(x,\ y)$, можно найти $P(x)$ следующим образом:

$$P(x)\,dx = P(x \leqslant X \leqslant x+dx,\ -\infty < y < +\infty) =$$

$$= \int_{-\infty}^{\infty}\int_{x}^{x+dx} P(x,\ y)dx\,dy.$$

Мы знаем, что для любой функции $f(x)$

$$\int_{x}^{x+dx} f(x)\,dx = f(x)\,dx.$$

Следовательно,

$$P(x)\,dx = \left[\int_{-\infty}^{\infty} P(x,\ y)\,dy\right]dx$$

и

$$P(x) = \int_{-\infty}^{\infty} P(x,\ y)\,dy. \quad (6.21а)$$

Аналогично этому

$$P(y)=\int_{-\infty}^{\infty} P(x, y)\,dx, \qquad (6.21б)$$

причем функциональная форма P, вообще говоря, различна для $P(x)$ и $P(y)$. Распределения, описанные выражениями (6.21), называются *частными распределениями*. Но в силу (6.18) распределение условных вероятностей равно

$$P_{x\leqslant X\leqslant x+dx}(y_1\leqslant y\leqslant y_2)=$$

$$=\frac{P(y_1\leqslant y\leqslant y_2,\ x\leqslant X\leqslant x+dx)}{P(x\leqslant X\leqslant x+dx)}=$$

$$=\frac{\int\limits_{x}^{x+dx}\int\limits_{y_1}^{y_2}P(x, y)\,dy\,dx}{P(x)dx}=\frac{\left[\int\limits_{y_1}^{y_2}P(x,y)\,dy\right]dx}{P(x)\,dx}=$$

$$=\int_{y_1}^{y_2}\frac{P(x, y)}{P(x)}\,dy.$$

Таким образом, при $y_1=y$ и $y_2=y+dy$

$$P_x(y)\,dy=\frac{P(x,y)}{P(x)}\,dy. \qquad (6.22)$$

Если x и y — независимые переменные, то, применяя (6.19), можно преобразовать (6.22) к виду

$$P_x(y)\,dy=P(y)\,dy. \qquad (6.23)$$

Ожидаемое значение определяется как

$$E[f(x, y)]=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}f(x, y)\,P(x, y)\,dx\,dy, \qquad (6.24)$$

откуда следует, что и в этом случае E остается линейным оператором. Если $f(x, y)=xy$, то

$$E(xy)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}xyP(x, y)dxdy.$$

Если x и y независимые переменные, то $P(x, y)=P(x)\,P(y)$ и

$$E(xy)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}xy\,P(x)\,P(y)\,dxdy=$$

$$=\int_{-\infty}^{\infty}xP(x)\,dx\int_{-\infty}^{\infty}yP(y)dy=E(x)E(y). \qquad (6.25)$$

Если $f(x)$ есть функция только от x (т. е. $\partial f/\partial y=0$), то

$$E[f(x)]=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}f(x)\,P(x, y)dydx=$$

$$=\int_{-\infty}^{\infty}f(x)\left[\int_{-\infty}^{\infty}P(x, y)\,dy\right]dx=$$

$$=\int_{-\infty}^{\infty}f(x)\,P(x)\,dx. \qquad (6.26)$$

Следовательно, мы можем определять такие величины, как математическое ожидание и дисперсию частного распределения, таким же образом, как и раньше. Единственное различие состоит в том, что теперь для указания, к какой из переменных относятся эти величины, мы должны вводить соответствующий индекс. Так, выражение (6.5) принимает в этом случае вид

$$\sigma_x^2=E(x-\mu_x)^2=\int_{-\infty}^{\infty}x^2P(x)\,dx-\mu_x^2.$$

Математическим ожиданием двумерного распределения является точка $x=\mu_x$, $y=\mu_y$. Медиана не определена, за исключением частных и условных распределений. Модой или модами являются точки, в которых обе частные производные $\partial P(x, y)/\partial x$ и $\partial P(x, y)/\partial y$ равны нулю.

Кодисперсия и *коэффициент корреляции. Смешанный момент*, или *кодисперсия*, или *ковариация*, определяется выражением

$$\sigma_{xy}=E[(x-\mu_x)(y-\mu_y)].$$

Для непрерывных переменных это выражение получает вид

$$\sigma_{xy}=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x-\mu_x)(y-\mu_y)\,P(x, y)dxdy. \qquad (6.27)$$

Кодисперсия пары переменных характеризует степень связи между ними. Так, если x и y независимые переменные, интеграл в формуле (6.27) можно разбить на два интеграла, каждый из которых определяет первый момент одной переменной относительно соответствующего математического ожидания; следовательно, каждый из этих интегралов равен 0. Таким образом, кодисперсия независимых переменных равна нулю.

Обычно кодисперсия применяется в нормированном виде и называется в этом случае

коэффициентом корреляции, который определяется как

$$\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y}. \qquad (6.28)$$

Коэффициент корреляции не зависит от математического ожидания и дисперсии каждой из переменных, т. е. от выбора масштаба и начала координат. Когда переменные независимы, $\rho = 0$. Когда между переменными существует полная корреляция, $\rho = \pm 1$; в этом легко убедиться, положив $y = x$ в (6.27), так как совершенно очевидно, что x полностью коррелирован с x.

Абсолютная величина коэффициента корреляции не может превышать единицу. Если в (6.27) вместо y подставить $-x$, то мы найдем, что $\rho = -1$. Таким образом, отрицательное значение коэффициента корреляции указывает на тенденцию к взаимосвязи между большими значениями x и малыми значениями y. Значения коэффициента корреляции, лежащие между нулем и единицей, указывают на количественно изменяющиеся степени связи между переменными.

Двумерное нормальное распределение. Выражение для двумерного распределения можно вывести способом, который мы применяли при выводе аналогичного выражения для одномерного распределения. Однако в связи с большим количеством постоянных такой вывод оказывается очень громоздким и утомительным. Поэтому мы напишем без доказательства

$$P(x, y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{(1-\rho^2)}} \times$$

$$\times \exp\left\{\frac{-1}{2(1-\rho^2)}\left[\frac{(x-\mu_x)^2}{\sigma_x^2} - \right.\right.$$

$$\left.\left. - \frac{2\rho}{\sigma_x\sigma_y}(x-\mu_x)(y-\mu_y) + \frac{(y-\mu_y)^2}{\sigma y^2}\right]\right\}, \qquad (6.29)$$

где $\rho = \frac{\sigma_{xy}}{\sigma_x\sigma_y}$ — коэффициент корреляции. Если x и y независимы, то $\rho = 0$ и выражение (6.29) сводится к

$$P(x, y) = \frac{1}{2\pi\sigma_x\sigma_y} \times$$

$$\times \exp\left\{-\frac{1}{2}\left[\frac{(x-\mu_x)^2}{\sigma_x^2} + \frac{(y-\mu_y)^2}{\sigma_y^2}\right]\right\}. \qquad (6.30)$$

Это выражение можно достаточно просто вывести, воспользовавшись соотношением

$$P(x, y) = P(x)\,P(y).$$

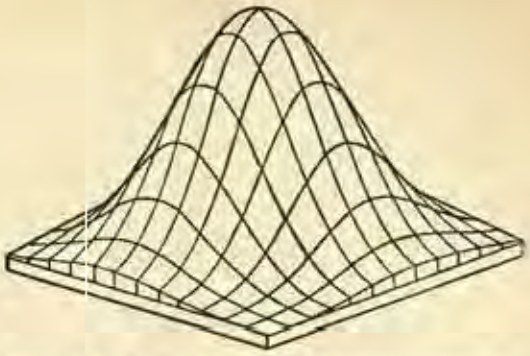


Рис. 6.9. Плотность двумерного нормального распределения вероятностей (по Юлу и Кендаллу [99]).

Двумерное нормальное распределение является поверхностью колоколообразной формы (рис. 6.9) с единственной модой в точке $x = \mu_x$, $y = \mu_y$. Любая вертикальная плоскость рассекает эту поверхность по кривой, дающей одномерное нормальное распределение. Кривые, показанные на рис. 6.9, изображают условные вероятности $P_y(x)$ и $P_x(y)$. Сечения рассматриваемой колоколообразной поверхности горизонтальными плоскостями суть эллипсы. Каждый такой эллипс имеет центр в точке $x = \mu_x$, $y = \mu_y$. Ориентации и эксцентриситеты всех этих эллипсов одинаковы и являются функциями коэффициента корреляции и относительных зна-

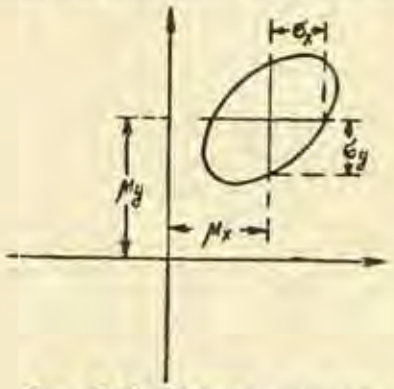


Рис. 6.10. Эллипс постоянной вероятности для двумерного нормального распределения с $\sigma_x = \sigma_y$, $\sigma_{xy} > 0$, $\mu_x > 0$, $\mu_y > 0$.

чений σ_x и σ_y. Эллипс превращается в круг тогда и только тогда, когда $\sigma_x = \sigma_y$ и $\rho = 0$.

На рис. 6.10 показан один такой эллипс для распределения, в котором $\sigma_x = \sigma_y$, но $\rho \neq 0$. Такие эллипсы очень интересны, так как представляют собой кривые постоянной вероятности для двумерного нормального распределения. Их можно получить из выражения (6.29), приравняв экспоненциальную функцию подходящей постоянной.

6.6. Преобразование переменных

В каждой задаче о вероятности задаются определенные вероятности и определенные условия, и требуется найти какие-то другие вероятности. В одном большом классе практических задач заданные вероятности представляют собой распределение некоторой переменной, условия задаются в виде функциональной зависимости между этой переменной и какой-либо другой, а искомые вероятности представляют собой распределение этой второй переменной. Мы уже познакомились с задачей такого рода при нормировании нормального распределения от вида (6.11) к виду (6.13). В следующем параграфе мы познакомимся с другим примером, в котором будет задано распределение переменной θ и потребуется найти распределение переменной $x=\operatorname{tg}\theta$.

В общем случае предположим, что нам задана плотность распределения вероятностей $P(\theta)$ и функциональная зависимость $x=f(\theta)$ и что мы хотим найти плотность распределения вероятностей $Q(x)$. Мы можем написать

$$P(\theta)\,d\theta = Q(x)\,dx.$$

Деля на dx, получаем

$$Q(x) = \frac{P(\theta)\,d\theta}{dx},$$

что чаще пишется в виде

$$Q(x) = \frac{P(\theta)}{dx/d\theta} = \frac{P(\theta)}{f'(\theta)}. \qquad (6.31)$$

Строгое дифференцирование приводит к такому же результату, который является весьма общим (при подходящих ограничениях на функцию f).

При желании распространить это выражение на распределения нескольких переменных, производная должна быть заменена якобианом. Так, например, если нам заданы $P(x, y)$ и функциональные зависимости $u=f(x,y)$ и $v=g(x,y)$, мы можем вычислить $Q(u,v)$, используя выражение

$$Q(u, v) = \frac{P(x, y)}{J}, \qquad (6.32а)$$

в котором якобиан задается определителем

$$J \equiv \frac{\partial(u, v)}{\partial(x, y)} \equiv \begin{vmatrix} \frac{\partial f}{\partial x} & \frac{\partial f}{\partial y} \\ \frac{\partial g}{\partial x} & \frac{\partial g}{\partial y} \end{vmatrix}. \qquad (6.33)$$

Якобиан обладает тем важным свойством, что

$$\frac{\partial(u, v)}{\partial(x, y)}\frac{\partial(x, y)}{\partial(u, v)} = 1.$$

Следовательно, можно применить обратный якобиан и записать

$$Q(u, v) = P(x, y)J^{-1} \qquad (6.32б)$$

6.7. Распределения некоторых других видов

Из всех функций, которые удовлетворяют условиям (5.17), (6.2) или (6.20) и, следовательно, могут быть функциями или плотностями распределения, в инженерной практике обычно встречаются только биномиальное, Пуассона, равномерное и нормальное распределения. Тем не менее в практической инженерной работе иногда приходится иметь дело и с другими функциями, удовлетворяющими указанным выше условиям. Две из этих функций мы сейчас рассмотрим.

Первая, экспоненциальная, рассматривается нами потому, что она имеет определенную взаимосвязь с распределением Пуассона и используется в теории массового обслуживания. Вторая, функция распределения Коши, рассматривается нами в связи с ее «патологическим» поведением. Она не обладает некоторыми определенными свойствами, о которых интуитивно можно думать, что ими должны обладать все «практические» функции распределения. В связи с этим становится ясной вся важность уточнения наших исходных предположений и (особенно в математической статистике) проверки или отрицания некоторых положений, которые на первый взгляд не вызывают сомнения и кажутся очевидными.

Экспоненциальное распределение. Рассмотрим следующую задачу. По дороге мимо определенной точки проезжают автомашины, распределение которых по времени подчиняется закону Пуассона с математическим ожиданием μ автомашин в единицу времени. Требуется найти распределение интервалов времени между автомашинами. Таким образом, мы хотим найти плотность распределения вероятностей $P(t)$, определяющую вероятность интервала времени длины t. Точнее, мы хотим найти элемент вероятности $P(t)dt$, определяющий вероятность того, что последующая автомашина пройдет более чем через t и менее чем через $t+dt$ единиц времени после проезда предыдущей автомашины.

Эта величина есть произведение двух вероятностей: вероятности, что за интервал между 0 и t не проедет ни одна автомашина, и вероятности, что за интервал между t и $t+dt$

проедет одна автомашина. Первый из этих членов можно найти произведя соответствующую подстановку в выражение (5.27); распределение имеет математическое ожидание μt, и нам нужно найти вероятность того, что за интервал длины t не произойдет ни одного благоприятного события:

$$P(k=0)\frac{e^{-\mu t}(\mu t)^0}{0!}=e^{-\mu t}.$$

По тем же соображениям вероятность того, что за интервал времени dt не произойдет ни одного благоприятного события, равна $e^{-\mu dt}$. Следовательно, вероятность, что за этот отрезок времени произойдет хотя бы одно благоприятное событие, равна $1-e^{-\mu dt}$. В пределе, когда dt стремится к нулю, это выражение определяет также вероятность наступления в точности одного благоприятного события. Из курса высшей математики мы помним, что для малых значений x

$$e^{-x}\approx 1-x,$$

откуда

$$1-e^{-\mu dt}\approx \mu dt.$$

Теперь мы можем написать выражение для элемента вероятности:

$$P(t)\,dt=\mu e^{-\mu t}\,dt, \qquad (6.34)$$

которое является формулой для экспоненциального (или *отрицательно экспоненциального*) распределения.

Эта функция является монотонной, ее мода лежит в точке $t=0$. Последнее утверждение на первый взгляд отнюдь не очевидно, так как может возникнуть предположение, что мода будет находиться близко к математическому ожиданию. Поэтому рассмотрим этот вывод подробнее.

Предположим, что ожидаемый темп движения автомашин равен одной машине в минуту. Считая за начало отсчета времени проезд предыдущей машины мимо контрольной точки, разделим последующий интервал времени на односекундные приращения. В течение какой секунды у нас больше всего оснований ожидать проезда следующей автомашины? Конечно, в течение первой секунды. Отсюда и следует, что мода соответствует $t=0$.

Такой путь решения задачи иллюстрирует также часто применяемый в математической статистике метод преобразования непрерывного распределения в дискретное распределение, когда последнее легче поддается расчетам. Этот метод обсуждается в § 12.7.

Для определения математического ожидания экспоненциального распределения используем выражение (6.4):

$$E(t)=\int_{-\infty}^{\infty}tP(t)\,dt=\int_{0}^{\infty}t\mu e^{-\mu t}dt=$$

$$=\frac{1}{\mu}\int_{0}^{\infty}\mu t e^{-\mu t}d(\mu t)=\frac{1}{\mu}\Gamma(2)=\frac{1}{\mu} \qquad (6.35)$$

Это ожидаемое значение находится в полном соответствии с нашей интуицией. Чтобы определить дисперсию, мы прибегнем к выражению (6.5) и снова используем гамма-функцию:

$$\sigma^2=\int_{-\infty}^{\infty}t^2P(t)\,dt-[E(t)]^2=$$

$$=\int_{0}^{\infty}t^2\mu e^{-\mu t}dt-\frac{1}{\mu^2}=\frac{1}{\mu^2}\int_{0}^{\infty}(\mu t)^2e^{-\mu t}d(\mu t)-$$

$$-\frac{1}{\mu^2}=\frac{1}{\mu^2}\Gamma(3)-\frac{1}{\mu^2}=\frac{1}{\mu^2}. \qquad (6.36)$$

Пример *. Рассмотрим то же распределение автомашин, что и раньше, но при условии, что мы начали наблюдать за их движением в случайный момент времени. Требуется определить ожидаемое значение времени ожидания проезда следующей автомашины.

Решение. Мы уже определили, что в случае, когда мы начинаем наблюдать за движением автомашин сразу же после проезда автомашины, ожидаемое значение времени ожидания равно $1/\mu$. В случае, когда мы начинаем следить в случайный момент времени, существует равная вероятность того, что наше наблюдение началось в конце интервала между машинами или в его начале. В среднем можно считать, что мы начинаем наблюдение в середине этого интервала времени. Таким образом, интуиция подсказывает нам, что ожидаемое значение времени ожидания должно составлять примерно $1/2\mu$. Однако *это неверно*. Ожидаемое значение и в этом случае равно $1/\mu$.

При выводе математических соотношений для распределения Пуассона мы условились считать, что наступление события не зависит от наступления предыдущего события, и, следовательно, оно не должно зависеть и от отсутствия предыдущего события, появления которого можно было ожидать. Поэтому ожидаемое значение остается одним и тем же вне зависимости от того, начали ли мы отсчет после отсутствия автомашин в течение периода времени, равного нескольким ожидаемым значениям, или непосредственно после проезда предыдущей автомашины.

Распределение Коши. На единичном расстоянии от берега в точке L находится маяк (рис. 6.11). Он посылает два луча света под углом в 180°. Один и только один из этих лучей встречает бесконечно длинную прямую берего-

* Заимствован из [33]. — *Прим. авт.*

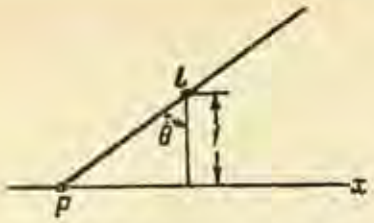


Рис. 6.11. Чертеж для вывода распределения Коши.

вую линию (которую мы назовем осью x) в точке P. Если световое устройство маяка вращается и останавливается в случайном положении таким образом, что угол θ распределен равномерно, то каким образом распределяется вдоль оси x точка P?

Центром распределения является, конечно, ближайшая к маяку точка берега. Если мы для упрощения поместим начало координат в эту точку, то наша задача сведется к определению распределения переменной $\operatorname{tg}\theta$ для случая равномерного распределения переменной θ. Применим формулу (6.31) с $f(\theta)=\operatorname{tg}\theta$ и $f'(\theta)=\sec^2\theta$.

Плотность вероятности $P(\theta)$ определяется соотношением

$$P(\theta)\,d\theta=\frac{1}{\pi}d\theta, \quad -\frac{\pi}{2}\leqslant\theta\leqslant\frac{\pi}{2}.$$

Следовательно,

$$P(x)=\frac{1}{\pi}\frac{1}{\sec^2\theta}.$$

Но для удобства это выражение необходимо преобразовать в явную функцию от x. С этой целью мы используем тригонометрическое тождество $\sec^2\theta=\operatorname{tg}^2\theta+1$, в результате чего получим

$$P(x)=\frac{1}{\pi}\frac{1}{\operatorname{tg}^2\theta+1}=\frac{1}{\pi(x^2+1)}$$

и окончательно

$$P(x)dx=\frac{dx}{\pi(x^2+1)}. \qquad (6.37)$$

Выражение (6.37) определяет распределение Коши. Форма этого распределения показана на рис. 6.12; можно подумать, что оно подобно нормальному распределению. Выражение (6.37) удовлетворяет условиям (6.2) для вероятностной функции, так как оно не может принимать отрицательных значений и его интеграл $\frac{\operatorname{arctg} x}{\pi}$, вычисленный в пределах от плюс до минус бесконечности, равен единице, что и требовалось. Определим математическое ожидание и дисперсию распределения Коши.

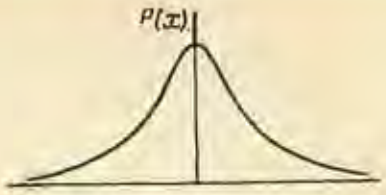


Рис. 6.12. Плотность распределения Коши.

Математическое ожидание дается формулой (6.4):

$$E(x)=\int_{-\infty}^{\infty}\frac{x\,dx}{\pi(1+x^2)}=\frac{1}{2\pi}\int_{x=-\infty}^{x=\infty}\frac{d(x^2)}{1+x^2}, \qquad (6.38)$$

$$E(x)=\frac{1}{2\pi}\left[\ln(1+x^2)\right]_{-\infty}^{\infty}. \qquad (6.39)$$

Величина в правой части равенства (6.39) не существует ни при одном из указанных пределов. Вследствие этого, а также ввиду других соображений [30], интеграл (6.38) взять нельзя. Следовательно, распределение Коши не имеет математического ожидания. Интеграл имеет так называемое *основное значение*, которое в данном случае равно нулю.

Это отвечает нашему интуитивному представлению о величине ожидаемого значения распределения. Распределение Коши имеет медиану и моду, причем как одна, так и другая равна нулю.

При определении дисперсии возникают более серьезные затруднения, если даже приписать конечное значение математическому ожиданию. Дисперсия определяется из выражения (6.5):

$$\sigma^2=\int_{-\infty}^{\infty}\frac{x^2dx}{\pi(1+x)^2}-[E(x)]^2=$$

$$=\frac{1}{\pi}[x-\operatorname{arctg} x]_{-\infty}^{+\infty}-[E(x)]^2.$$

Совершенно очевидно, что это выражение равно плюс бесконечности при одном пределе и минус бесконечности при другом. Следовательно, дисперсия равна бесконечности.

Пример. а) Предположим, что при проведении некоторого опыта с упомянутым выше маяком нам известно, что маяк находится на расстоянии 1 мили от бесконечно длинной береговой линии, но его положение по оси x нам неизвестно. Какова вероятность того, что при нашем первом наблюдении точка P (рис. 6.11) будет лежать в пределах 1 мили от истинного положения маяка на оси x?

б) Сколько наблюдений нужно сделать для того, чтобы получить среднее значение (математическое ожидание), имеющее вероятность 0,9 лежать в пределах 1 мили?

Решение

а) Вероятность, что при одном наблюдении точка P будет находиться в пределах 1 мили от центра распределения, равна

$$\int_{-1}^{+1} P(x)\,dx = 0{,}47.$$

б) Недостаточно никакого конечного числа. Методами, рассмотренными в следующей главе, может быть показано, что математические ожидания выборок из n наблюдений, каждое из которых взято из распределения Коши, имеют распределение

$$P(m)\,dm = \frac{dm}{\pi(1+m^2)}.$$

Это выражение тождественно выражению (6.37), независимо от значения n. Иными словами, математическое ожидание миллиона наблюдений не дает нам лучшего определения действительного центра распределения, чем одно наблюдение, — ввиду того, что одно или больше из миллиона наблюдений будет лежать настолько далеко от математического ожидания, что ответ потеряет всякий смысл.

Такие большие отклонения возможны и при других известных нам распределениях, таких, как нормальное или Пуассона, но в этих случаях они крайне маловероятны. При распределении Коши о них можно сказать, что они «умеренно» маловероятны. При нормальном распределении имеется меньше одного шанса из миллиона за то, что отклонение будет равно или расстоянию от центра до точки перегиба. Для распределения Коши это отношение справедливо лишь в том случае, если мы говорим об отклонении, равном примерно миллиону этих расстояний.

Это и является причиной того, что в распределении Коши дисперсия равна бесконечности.

ЛИТЕРАТУРА

См. гл. 8.

ЗАДАЧИ

См. гл. 8.

ГЛАВА 7

ХАРАКТЕРИСТИКИ И РАСПРЕДЕЛЕНИЯ СТАТИСТИК

Рассматривая биномиальное распределение, мы обратили внимание на тот факт, что при осуществлении n испытаний количество благоприятных исходов не всегда равно ожидаемому или наиболее вероятному значению. Даже если делалось N групп n испытаний, распределение результатов не было точно таким, как это следовало бы из биномиальной формулы. В самом деле, вероятность получения определенного результата сама имела биномиальное распределение. В этой главе мы займемся исследованием результатов, которые получаются, когда мы производим серию наблюдений над каким-либо исходным распределением. С такой постановкой задачи мы познакомились на примере в конце предыдущей главы, где нас интересовало распределение математических ожиданий выборок из распределения Коши.

Введем несколько новых терминов. Каждое значение переменной, которое мы можем или хотим наблюдать, мы называем *наблюдаемым значением* *. Исходное распределение, которое может содержать либо конечное, либо бесконечное число наблюдаемых значений, мы назовем *генеральной совокупностью* (некоторые авторы применяют для этого термин *«популяция»*). Наблюдаемые значения, выбираемые из генеральной совокупности при действительном наблюдении, будем обозначать буквами x_i и назовем их совокупность *выборкой;* количество наблюдаемых значений в выборке обозначим через n.

Числа, характеризующие генеральную совокупность, будут обозначаться греческими буквами, например μ и σ, и называться *параметрами*. Соответствующие числа, характеризующие выборку, будут обозначаться латинскими буквами **, например m и s, и называться *статистиками*. Таким образом, «статистикой» в этом смысле называется любая функция наблюдаемых значений, т. е. величина, которая может быть записана в виде

$$T = f(n, x_1, x_2, \ldots, x_n).$$

Так как единичные наблюдаемые значения имеют распределения, то статистика также будет иметь распределение, но не такое же, как наблюдаемые значения. Первое из этих распределений называется *исходным распределением*, а второе — *выборочным распределением*. Наша задача заключается в исследовании характеристик и распределений статистик.

* Термином «наблюдаемое значение» мы передаем английский термин «variate». Его передают также словами «количественный признак» и др. — *Прим. ред.*

** За исключением χ^2, где мы имеем дело с давно установившимся обозначением. — *Прим. авт.*

7.1. Математическое ожидание выборки

Математическое ожидание выборки (или выборочное среднее значение) определяется выражением

$$m \equiv \frac{\Sigma x_i}{n}. \qquad (7.1а)$$

Многие авторы обозначают математическое ожидание выборки символом $\bar{x}$. Мы можем узнать многое об этой важнейшей статистике, не зная вообще ничего об исходном распределении. В этом параграфе мы выведем формулы для математического ожидания и дисперсии математических ожиданий выборок из произвольных распределений. Для этой цели выражению (7.1а) удобно придать следующий вид:

$$m = \Sigma \frac{x_i}{n} = \frac{x_1}{n} + \frac{x_2}{n} + \ldots + \frac{x_n}{n}. \qquad (7.1б)$$

Здесь m представлено суммой n наблюдаемых значений, умноженных каждое на $\frac{1}{n}$.

Ожидаемое значение. Из выражения (5.5) следует, что ожидаемое значение любой линейной формы равно

$$E(\Sigma c_i x_i) = \Sigma c_i E(x_i).$$

Так как выражение (7.1б) является линейной формой, в которой каждое c_i равно $1/n$, то ожидаемое значение математического ожидания выборки равно

$$E(m) = E\left(\sum \frac{x_i}{n}\right) = \sum \frac{1}{n} E(x_i) = \sum \frac{1}{n} \mu_i.$$

Это выражение справедливо для математического ожидания выборки, каждое наблюдаемое значение которой взято из различного, и притом произвольного, распределения. Если каждое наблюдаемое значение берется из одного и того же распределения, то

$$E(m) = \mu. \qquad (7.2)$$

Таким образом, для произвольного распределения (предполагается только, что математическое ожидание генеральной совокупности существует) ожидаемое значение математического ожидания выборки равно математическому ожиданию генеральной совокупности.

Дисперсия. Если сделано много выборок объема n и если для каждой из них вычислено m, то m будут иметь распределение с математическим ожиданием μ и некоторой дисперсией σ_m^2. Для определения этой дисперсии мы снова прибегнем к выражению (7.1б) и попытаемся найти дисперсию линейной формы n переменных. Для упрощения задачи начнем с суммы двух переменных:

$$w = c_1 x_1 + c_2 x_2,$$

$$\sigma_w^2 = E[w - E(w)]^2 =$$

$$= E[c_1 x_1 + c_2 x_2 - c_1 E(x_1) - c_2 E(x_2)]^2 =$$

$$= E\{c_1[x_1 - E(x_1)] + c_2[x_2 - E(x_2)]^2\} =$$

$$= E(c_1 X_1 + c_2 X_2)^2,$$

где $X_1 = x_1 - E(x_1)$ и аналогично $X_2 = x_2 - E(x)$.

Раскрывая скобки, получаем

$$\sigma_w^2 = E(c_1^2 X_1^2 + c_2^2 X_2^2 + 2c_1 c_2 X_1 X_2) =$$

$$= c_1^2 E(X_1^2) + c_2^2 E(X_2^2) + 2c_1 c_2 E(X_1 X_2).$$

Но $E(X_1^2)$ по определению есть дисперсия для x_1, и аналогично $E(X_2^2)$ есть дисперсия x_2. Далее, $E(X_1 X_2)$ по определению есть кодисперсия для x_1 и x_2. Следовательно,

$$\sigma_w^2 = c_1^2 \sigma_{x_1}^2 + c_2^2 \sigma_{x_2}^2 + 2c_1 c_2 \sigma_{x_1 x_2}.$$

Для n переменных обобщение сразу же дает

$$\sigma_w^2 = \sum_i c_i^2 \sigma_{x_i}^2 + 2\sum_{ij} c_i c_j \sigma_{x_i x_j}. \qquad (7.3)$$

Это выражение является совершенно общим. Если применить его для определения дисперсии математического ожидания выборки и допустить, что наблюдаемые значения независимы, то член для кодисперсии будет равен нулю, каждый из коэффициентов c_i будет равен $1/n$, а каждая из дисперсий $\sigma_{x_i}^2$ будет равна σ^2:

$$\sigma_m^2 = \sum_{i=1}^{n} \frac{1}{n^2} \sigma^2 = \frac{\sigma^2}{n}. \qquad (7.4)$$

Таким образом, для любого распределения, если только существуют первый и второй моменты, дисперсия математического ожидания выборки независимых наблюдений равна дисперсии генеральной совокупности, деленной на число наблюдаемых значений в выборке. Этот вывод является чрезвычайно важным и служит основой для многих инженерных решений.

7.2. Дисперсия выборки

Дисперсия выборки определяется выражением

$$s^2=\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n}. \qquad (7.5)$$

Заметим, что в выражение входит параметр μ, в связи с чем его можно использовать только в том случае, когда значение этого параметра известно. Ожидаемое значение статистики s^2 равно

$$E(s^2)=\frac{E[\Sigma(x_i-\mu^2)]}{n}.$$

Используя (5.5) и (5.6), эту формулу можно привести к виду

$$E(s^2)=\frac{\Sigma[E(x_i^2)-2\mu E(x_i)+\mu^2]}{n}=\frac{\Sigma[E(x_i^2)-\mu^2]}{n}$$

и согласно (5.10) к виду

$$E(s^2)=\frac{\sum_{i=1}^{n}\sigma^2}{n}=\sigma^2. \qquad (7.6)$$

Так как, имея выборку, мы не всегда знаем математическое ожидание генеральной совокупности, встает вопрос об определении $E(s^2)$ через математическое ожидание выборки. Вполне естественно предположить, что для этого необходимо в (7.5) заменить μ на m. Временно обозначим получаемую при этом величину через s.

$$s=\frac{\Sigma(x_i-m)^2}{n}=\frac{\Sigma x_i^2}{n}-2m\frac{\Sigma x_i}{n}+$$

$$+\frac{\Sigma m^2}{n}=\frac{\Sigma x_i^2}{n}-m^2,$$

$$E(s^2)=E\left(\frac{\Sigma x_i^2}{n}-m^2\right)=E\left[\frac{\Sigma x_i^2}{n}-\left(\frac{\Sigma x_i}{n}\right)^2\right]=$$

$$=E\left[\frac{n\Sigma x_i^2-(\Sigma x_i)^2}{n^2}\right].$$

Представим теперь член в квадрате (второй член числителя) в виде произведения двух сумм, которые мы должны записать с самостоятельными индексами:

$$(\Sigma x_i)^2=\sum_i x_i\sum_j x_j=\sum_i x_i^2+\sum_{i\neq j}x_ix_j.$$

Тогда ожидаемое значение дисперсии принимает вид

$$E(s^2)=E\left(\frac{n\sum_i x_i^2-\sum_i x_i^2-\sum_{i\neq j}x_ix_j}{n^2}\right)=$$

$$=E\left[\frac{(n-1)\sum_i x_i^2-\sum_{i\neq j}x_ix_j}{n^2}\right]=$$

$$=\frac{n-1}{n^2}\sum_i E(x_i)^2-\frac{1}{n^2}\sum_{i\neq j}E(x_ix_j). \qquad (7.7)$$

Так как x_i и x_j независимы при $i\neq j$, то в соответствии с выражением (6.25) мы можем заменить $E(x_ix_j)$ на $E(x_i)\,E(x_j)$, причем каждый из двух членов этого произведения будет не что иное, как математическое ожидание генеральной совокупности. Кроме того, по формуле (5.10) мы можем заменить $E(x_i^2)$ на $\sigma^2+\mu^2$. При этом выражение (7.7) принимает вид

$$E(s^2)=\frac{n-1}{n^2}\sum_i(\sigma^2+\mu^2)-\frac{1}{n^2}\sum_{i\neq j}\mu^2. \qquad (7.8)$$

Теперь все члены под каждым знаком суммы являются постоянными величинами, и для суммирования нам необходимо только знать число этих членов. Первая сумма содержит n членов, а вторая — остальные из n^2 или $n(n-1)$ членов. Таким образом, выражение (7.8) принимает вид

$$E(s^2)=\frac{n-1}{n^2}n(\sigma^2+\mu^2)-\frac{1}{n^2}n(n-1)\mu^2=$$

$$=\frac{n-1}{n}\sigma^2. \qquad (7.9)$$

Полученный результат является неожиданным, так как при таком определении ожидаемое значение дисперсии выборки не равно дисперсии генеральной совокупности. Ясно, что это при отсутствии данных о математическом ожидании генеральной совокупности заставляет нас внести изменение в определение дисперсии выборки, для чего мы введем новый член

$$\hat{s}^2=\frac{\Sigma(x_i-m)^2}{n-1}, \qquad (7.10)$$

ожидаемым значением которого служит дисперсия генеральной совокупности. Знак $\hat{}$, поставленный над буквой s, часто используется в математической статистике для обозначения оценки; однако обычно мы будет писать просто s, подразумевая при этом величину $\hat{s}$, определяемую выражением (7.10).

Внимательное изучение причин применения $n-1$ вместо n показывает, что при вычислении дисперсии выборки мы использовали величину m, которая в свою очередь была вычислена на основании данных выборки. Таким образом, мы «потеряли степень свободы», или наложили ограничение на данные, так как только $n-1$ наблюдений являются независимыми. Следовательно, дисперсия становится больше и тем самым у нас уменьшается уверенность в нашей оценке (так как дисперсия является в известной степени мерой нашей неуверенности).

Ясно, что в больших выборках различием между n и $n-1$ можно пренебречь; однако в малых выборках это различие существенно. Такая ситуация — потеря степени свободы, эквивалентная потере одного независимого наблюдения, каждый раз, когда мы накладываем ограничение на получаемые данные, — возникает при работе со статистиками во многих случаях.

Найдя таким образом выражение для математического ожидания дисперсии выборки из произвольного распределения, мы могли бы приступить к определению дисперсии от дисперсии выборки и других характеристик математического ожидания, дисперсии или других статистик. Эти вопросы рассматриваются в учебной литературе [29], поэтому мы опустим их и сосредоточим свое внимание на распределениях статистик, получаемых из некоторых конкретных распределений. Однако перед этим мы должны познакомиться с более мощным методом исследования распределений, чем те, с которыми мы до сих пор имели дело.

7.3. Характеристическая функция

Так как статистика является функцией наблюдаемых значений, можно получить распределение любой статистики, рассматривая распределение наблюдаемого значения (т. е. исходное распределение генеральной совокупности) и осуществляя на нем подходящее преобразование переменной (§ 6.6). Для некоторых сложных статистик, нас интересующих, эта задача довольно трудна, и для ее решения был разработан ряд мощных математических методов. Некоторые из них лучше применять в одних случаях, другие — в других. Для нашей цели всего полезнее метод, известный под названием метода *характеристической функции*.

Характеристическая функция $\varphi(t; x)$ случайной переменной x определяется как

$$\varphi(t;\ x)=E(e^{ixt})=\int_{-\infty}^{\infty}e^{itx}P(x)dx \quad (7.11a)$$

или в случае дискретного распределения — как

$$\varphi(t;\ k)=E(e^{ikt})=\sum_{k=-\infty}^{\infty}e^{itk}P(k), \quad (7.11б)$$

где $i=\sqrt{-1}$. Переменная t всегда является непрерывной. При нахождении характеристической функции сама случайная переменная исключается благодаря интегрированию и остается лишь функция от t. Буква x в обозначении $\varphi(t; x)$ оставлена лишь для напоминания читателю о происхождении функции φ.

Познакомимся теперь с несколькими теоремами о характеристической функции.

Во-первых, характеристическая функция от функции случайной переменной есть

$$\varphi[t;\ f(x)]=E[e^{itf(x)}]=\int_{-\infty}^{\infty}e^{itf(x)}P(x)\,dx. \quad (7.12)$$

Характеристическая функция суммы случайных переменных определяется выражением

$$\varphi(t;\ x_1+x_2)]=E[e^{i(x_1+x_2)t}]=E(e^{ix_1t}e^{ix_2t}).$$

Если x_1 и x_2 независимы, это выражение преобразуется согласно (6.25) к виду

$$\varphi(t;\ x_1+x_2)=E(e^{ix_1t})E(e^{ix_2t})=$$
$$=\varphi(t;\ x_1)\varphi(t;\ x_2). \quad (7.13)$$

Если переменная x заменяется на постоянную величину a, характеристическая функция равна

$$\varphi(t;\ a)=E(e^{iat})=e^{iat}. \quad (7.14)$$

Используя соотношения (7.13) и (7.14), получаем

$$\varphi(t;\ x+a)=e^{iat}\varphi(t;\ x) \quad (7.15)$$

и, наконец,

$$\varphi\left(t;\ \frac{x}{a}\right)=E\left[e^{i\left(\frac{x}{a}\right)t}\right]=E\left[e^{ix\left(\frac{t}{a}\right)}\right]=\varphi\left(\frac{t}{a};\ x\right). \quad (7.16)$$

Выведем теперь характеристические функции для распределения Пуассона и нормального распределения.

Для распределения Пуассона

$$\varphi(t;\ k)=\sum_{k=0}^{\infty}e^{ikt}\frac{e^{-\mu}\mu^k}{k!}=$$
$$=e^{-\mu}\sum_{k=0}^{\infty}\frac{(e^{it})^k\mu^k}{k!}=$$
$$=e^{-\mu}\sum_{k=0}^{\infty}\frac{(\mu e^{it})^k}{k!}=e^{-\mu}e^{\mu e^{it}}=e^{\mu(e^{it}-1)}. \quad (7.17)$$

Для нормального распределения $N(0,1)$

$$\varphi(t;\ x)=\int_{-\infty}^{\infty} e^{itx}\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}dx=$$

$$=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\cos tx\, e^{-\frac{x^2}{2}}dx+\frac{i}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\sin tx\, e^{-\frac{x^2}{2}}dx.$$

Так как $e^{-\frac{x^2}{2}}$ симметрична относительно оси y и $\sin tx$ является нечетной функцией, второй интеграл уничтожается. Вычисляя первый интеграл, получаем

$$\varphi(t;\ x)=e^{-\frac{t^2}{2}}. \qquad (7.18)$$

Для $N(0, \sigma)$ с помощью (7.16) и (7.18) получаем

$$\varphi(t;\ \sigma x)=\varphi(\sigma t;\ x)=e^{-\frac{\sigma^2t^2}{2}}. \qquad (7.19)$$

Для $N(\mu, \sigma)$ с помощью (7.15) и (7.19) получаем

$$\varphi[t;\ \sigma(x+\mu)]=e^{it\mu}e^{-\frac{\sigma^2t^2}{2}}=e^{\left(it\mu-\frac{\sigma^2t^2}{2}\right)}. \qquad (7.20)$$

Большая ценность характеристической функции заключается в том, что с ее помощью обычно значительно проще производить математические преобразования, чем непосредственно с вероятностной функцией, и что от характеристической функции мы можем перейти к вероятностной функции. Можно показать, что как для дискретной, так и непрерывной переменных

$$P(x)=\frac{1}{2\pi}\int_{-\infty}^{\infty}e^{-ixt}\varphi(t;\ x)\,dt. \qquad (7.21)$$

Это значит, другими словами, что существует единственное преобразование между любой вероятностной функцией и ее характеристической функцией, хотя нахождение интеграла в выражении (7.21) не всегда легкая задача. Обратное преобразование для нормального распределения с математическим ожиданием 0 и стандартным отклонением 1 может быть найдено из выражений (7.18) и (7.21). Оно имеет вид

$$P(x)=\frac{1}{2\pi}\int_{-\infty}^{\infty}e^{-ixt}e^{-\frac{t^2}{2}}dt.$$

Применяя тот же процесс, как и при выводе формулы (7.18), получаем

$$P(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}=N(0,\ 1).$$

Мы будем использовать характеристическую функцию для вывода статистических распределений. Однако она используется и для других целей. Например, с помощью этого мощного орудия можно очень быстро получить нормальное распределение как предельный случай биномиального распределения и распределения Пуассона.

Из статистик мы рассмотрим здесь только математическое ожидание выборки, дисперсию выборки, отношение «хи-квадрат» и стьюдентово отношение t. В каждом из этих случаев будет предполагаться, что генеральная совокупность имеет нормальное распределение. Однако данный метод применим к распределениям других статистических величин, а также к другим распределениям.

7.4. Распределение математического ожидания нормальной выборки

Для того чтобы найти распределение математических ожиданий для выборок n наблюдаемых значений из нормальных генеральных совокупностей, мы сначала найдем распределение выборочных сумм. Математическое ожидание выборки равно, конечно, выборочной сумме, деленной на n. Примем, что x_i независимы друг от друга и что каждое x_i получено из различного нормального распределения со стандартным отклонением σ_i и математическим ожиданием 0.

Тогда, ввиду (7.13), для выборочной суммы

$$\varphi(t;\ x_1+\ldots x_n)=e^{-t^2(\sigma_1^2+\ldots+\sigma_n^2)/2}.$$

Следовательно, распределение выборочной суммы является нормальным с дисперсией $\Sigma\sigma_i^2$. Если x_i получены из нормальной генеральной совокупности с математическим ожиданием μ_i, то распределение выборочной суммы имеет математическое ожидание, определяемое формулой

$$E(x_1+\ldots+x_n)=E(x_1)+\ldots+E(x_n)=$$
$$=\mu_1+\ldots+\mu_n.$$

Таким образом, сумма наблюдаемых значений, полученных из нормальных распределений, имеет распределение

$$P(\Sigma x_i)=N(\Sigma\mu_i,\ \sqrt{\Sigma\sigma_i^2}) \qquad (7.22)$$

и математическое ожидание выборки имеет распределение

$$P(m)=N\left(\frac{\Sigma\mu_i}{n},\ \sqrt{\frac{\Sigma\sigma_i^2}{n^2}}\right), \qquad (7.23)$$

где μ_i и σ_i могут быть различными для каждого наблюдаемого значения. Это наглядно показывает большую общность нормального распределения — вопрос, который будет обсуждаться в следующей главе.

Если все наблюдаемые значения взяты из одной и той же нормальной генеральной совокупности, то

$$P(m)=N\left(\mu,\ \frac{\sigma}{\sqrt{n}}\right). \qquad (7.24)$$

Последнее соотношение есть частный случай уже нам известного важного принципа (7.4): дисперсия выборки уменьшается в той же мере, в какой возрастает объем выборки.

7.5. Распределение дисперсии нормальной выборки

Распределение дисперсии выборки сложнее тех распределений, с которыми мы встречались до сих пор, но оно также может быть найдено с помощью характеристической функции. Дисперсия выборки — определять ли ее выражением (7.5) или (7.10) — представляет собой сумму квадратов; из ее распределения можно вывести распределения других сумм квадратов. Применение двух распределений, выведенных из распределения дисперсии s^2, а именно χ^2 и F, обсуждается в гл. 12.

Рассмотрим выборку n независимых наблюдаемых значений, сделанную из нормальной генеральной совокупности с дисперсией σ^2; не теряя общности, мы можем принять, что математическое ожидание генеральной совокупности равно нулю, так как смещение оси координат не влияет на дисперсию генеральной совокупности или выборки. Характеристическая функция переменной x^2, взятой из распределения $N(0, \sigma)$, равна

$$\varphi(t;\ x^2)=\int_{-\infty}^{\infty} e^{itx^2}\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{x^2}{2\sigma^2}}dx.$$

Это выражение может быть проинтегрировано методами, аналогичными рассмотренным выше, в результате чего получается

$$\varphi(t;\ x^2)=(1-2\sigma^2 it)^{-1/2}.$$

Нас сейчас интересует характеристическая функция для x^2/n, а потому ввиду (7.16)

$$\varphi\left(t;\ \frac{x^2}{n}\right)=\varphi\left(\frac{t}{n};\ x^2\right)=\left(1-\frac{2it\sigma^2}{n}\right)^{-1/2}.$$

Наконец, чтобы определить характеристическую функцию для $s^2=\Sigma x_i^2/n$, найдем с помощью (7.13) сумму n независимых членов:

$$\varphi\left(t;\ \frac{\Sigma x^2}{n}\right)=\left(1-\frac{2it\sigma^2}{n}\right)^{-\frac{n}{2}}.$$

Из этого выражения с помощью (7.21) определяем искомую плотность распределения вероятностей

$$P(s^2)ds^2=P\left(\frac{\Sigma x^2}{n}\right)d\left(\frac{\Sigma x^2}{n}\right)=$$

$$=\frac{1}{2\pi}\int_{-\infty}^{\infty}e^{-ist^2}\left(1-\frac{2it\sigma^2}{n}\right)^{\frac{n}{2}}dt=$$

$$=\frac{\left(\frac{n}{2}\right)^{\frac{n}{2}}}{\sigma^n\Gamma\left(\frac{n}{2}\right)}e^{-\frac{ns^2}{2\sigma^2}}(s^2)^{\frac{n-2}{2}}ds^2. \qquad (7.25)$$

Здесь интегрирование производится в комплексной области; читатель отсылается за подробностями к любому углубленному учебнику по теории вероятностей [29 *b*]. Распределение дисперсии можно найти также и более элементарными методами, как это поясняется при решении задачи 8.7.

Если математическое ожидание генеральной совокупности нам неизвестно, то мы должны определять дисперсию выборки с помощью выражения (7.10); в этом случае распределение дисперсии s^2 является точно таким же, как и в выражении (7.25), с тем лишь исключением, что мы должны произвести замену $\nu=n-1$ (обозначает количество степеней свободы) всюду, где появляется n.

7.6. Распределение χ^2

Рассмотрим распределение величины «хи-квадрат», определяемой как отношение дисперсии выборки к дисперсии математического ожидания выборки из нормального распределения:

$$\chi^2=\frac{s^2}{\sigma_m^2}=\frac{ns^2}{\sigma^2}=\frac{\Sigma(x_i-\mu)^2/n}{\sigma^2/n}=\sum\left(\frac{x_i-\mu}{\sigma}\right)^2. \qquad (7.26)$$

Распределение этой статистики получается непосредственно из выражения (7.25) с помощью следующих подстановок:

$$\frac{ns^2}{2\sigma^2}=\frac{\chi^2}{2},\quad (s^2)^{\frac{n-2}{2}}=\left(\frac{\sigma^2}{n}\right)^{\frac{n-2}{2}}(\chi^2)^{\frac{n-2}{2}}$$

и

$$d(s^2)=\left(\frac{\sigma^2}{n}\right)d(\chi^2).$$

Итак,

$$P(\chi^2)\,d(\chi^2)=\frac{1}{2^{n/2}\Gamma\left(\frac{n}{2}\right)}e^{-\frac{\chi^2}{2}}(\chi^2)^{\frac{n-2}{2}}d(\chi^2). \quad (7.27а)$$

В математической статистике отношение «хи-квадрат» применяется обычно в тех случаях, когда математическое ожидание, или дисперсия генеральной совокупности, или обе эти величины неизвестны и приходится пользоваться соответствующими статистиками. В этом случае применяется та же самая формула, с той лишь разницей, что в ней число наблюдений n заменяется числом степеней свободы ν:

$$P(\chi^2)\,d(\chi^2)=\frac{e^{-\frac{\chi^2}{2}}(\chi^2)^{\frac{\nu-2}{2}}}{2^{\frac{\nu}{2}}\Gamma(\nu/2)}d(\chi^2). \quad (7.27б)$$

Таблицы для распределения χ^2 имеются в книгах по статистике и в математических таблицах; можно также найти его графики. На рис. 7.1 приведены графики формулы (7.27), вычерченные для различных значений ν. При ν, приблизительно большем 30, распределение χ^2 становится приблизительно нормальным. Другое интересное свойство распределения χ^2 заключается в том, что оно подобно нормальному распределению воспроизводит само себя: суммы независимых переменных, взятых из распределений χ^2, распределяются как распределение χ^2 с $\sum_{i=1}^{n}\nu_i$ степенями свободы.

Применение стастистики χ^2 рассматривается в § 12.7. Другой вывод [распределения χ^2 указан в задаче 8.7.

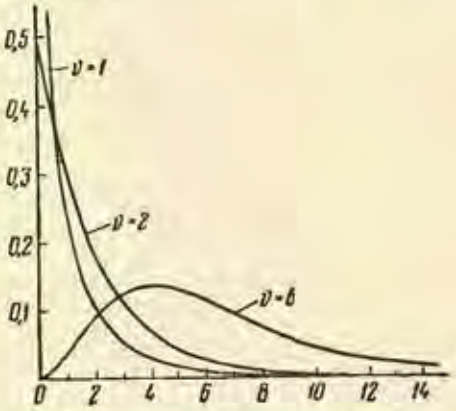


Рис. 7.1. Распределение χ^2 при различных степенях свободы (по Крамеру [43]).

7.7. Стьюдентово отношение t

Во многих случаях бывает желательно располагать единой статистикой, которая включала бы в себя как математическое ожидание, так и стандартное отклонение. Такой статистикой служит отношение $t=(x_i-m)/s$, где ве-

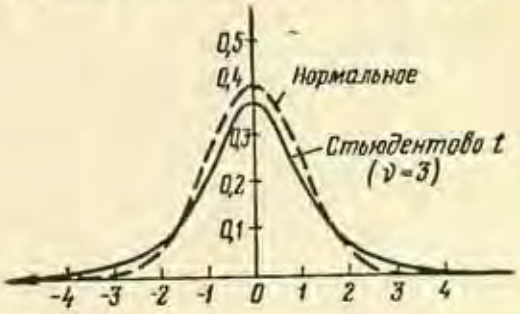


Рис. 7.2. Нормальное распределение и распределение стьюдентова отношения t (по Крамеру [43]).

личины x_i суть независимые наблюденные значения из нормального распределения. Эта статистика впервые была изучена в 1908 г. Госсеттом, который опубликовал свой результат под псевдонимом «Стьюдент». В связи с этим указанная статистика до сих пор носит название *стьюдентова отношения* t. Госсетт интуитивно предложил следующее распределение для t:

$$P(t)\,dt=\frac{\Gamma[(\nu+1)/2]}{\sqrt{\pi\nu}\,\Gamma(\nu/2)\left(1+\frac{t^2}{\nu}\right)^{\frac{\nu+1}{2}}}dt. \quad (7.28)$$

Строгое доказательство этого распределения было найдено только в 1926 г. Такое доказательство можно найти в большинстве современных учебников по статистике, и потому мы его здесь не приводим. Большая ценность стьюдентова отношения заключается в том, что и оно, и его распределение функционально независимы от параметров генеральной совокупности, а также в том, что оно учитывает воздействие как m, так и s.

Имеются таблицы распределения t для $\nu=1\div30$; при бо́льших значениях ν распределение Стьюдента почти не отличается от нормального распределения. На рис. 7.2 для сравнения приведены кривые для нормального распределения и распределения t при $\nu=3$. Одно практическое применение стьюдентова отношения t показано в § 12.5.

ЛИТЕРАТУРА

См. гл. 12.

ЗАДАЧИ

См. гл. 12.

ГЛАВА 8

УСТОЙЧИВОСТЬ И ЗАКОНЫ БОЛЬШИХ ЧИСЕЛ

Предположим, что незнакомец, играющий с вами в кости, бросил пару игральных костей и получил 7 очков, бросил их второй раз и снова получил 7 очков, бросил третий — снова 7 очков, четвертый — снова 7 очков, пятый — снова 7 очков, шестой раз — снова 7 очков, седьмой — снова 7 очков и, наконец, восьмой раз — еще раз 7 очков (мы повторяем эти слова несколько раз подряд с определенным умыслом, так как сказав просто «восемь раз подряд выбросил 7 очков», мы, по-видимому, не произведем такого сильного впечатления).

Читатель, вероятно, выразит сомнение в том, что такое явление математически возможно при использовании совершенно правильной пары игральных костей; тем не менее вероятность такого события составляет лишь несколько меньше 10^{-6}. В профессиональном игорном доме игроки могут бросать кости миллион раз в год; в этом случае мы можем ожидать приблизительно одного выпадения восьми последовательных семерок, несмотря на наше нерасположение к человеку, который окажется противником в этот момент.

Мы видим, что наша интуиция ненадежна. Она не делает различия между событиями довольно малой вероятности, скажем один на миллион или даже на сто, и событиями такой потрясающе ничтожной вероятности, при которой они были бы в сущности невозможны. Как пример последних можно взять упоминаемый ниже в задаче 8.1 случай разделения 10 000 почтовых мешков на две группы по 3000 и 7000 мешков; вероятность такого события при $p=0{,}5$ есть величина порядка 10^{-358}.

Наша интуиция не умеет также оценивать устойчивость больших чисел; выпадение восьми последовательных семерок при метании игральных костей, кажущееся в высшей степени невероятным, в действительности почти так же вероятно, как отклонение в 5σ при нормальном распределении (например, как выпадение 5250 или более гербов или решек при 10 000 последовательных бросаниях монеты), которое не слишком раздражает нашу интуицию.

На рис. 5.3 были приведены графики, иллюстрирующие для случая биномиального распределения тенденцию к сгущению частоты событий около математического ожидания при увеличении количества испытаний, хотя, как мы уже отмечали, вероятность получения в точности значения, равного математическому ожиданию, при этом уменьшается. Аналогичное явление, но уже применительно к распределению Пуассона иллюстрировалось примером в конце § 5.6.

Мы интуитивно верим в эту устойчивость, которую мы называем также *законом средних чисел*; однако, как мы уже имели возможность убедиться в связи с распределением Коши, наша интуиция при рассмотрении подобных вопросов может привести к ошибке. В настоящей главе мы проверим это положение количественно, для того чтобы осветить связь между теорией вероятностей и реальным миром.

8.1. Центральная предельная теорема

Термин *«центральная предельная теорема»* применяется к некоторым теоремам, которые показывают, что нормальное распределение является предельным распределением, получаемым несколькими различными путями при увеличении числа наблюдений n. Раньше мы уже показали, что нормальное распределение является предельным случаем для биномиального распределения и распределения Пуассона. Мы утверждали также, что распределения χ^2 и t в предельном случае являются нормальными распределениями (задача 8.8). В § 12.7 мы покажем, что многомерное нормальное распределение является предельным случаем полиномиального распределения.

Одна замечательная формулировка центральной предельной теоремы гласит, что при очень незначительных ограничениях (например, что первые и вторые моменты должны существовать) сумма (а следовательно, и математическое ожидание выборки) независимых наблюдаемых значений, взятых из одного и того же или из различных и притом совершенно произвольных распределений, является в пределе нормально распределенной. Однако еще более замечательной является скорость приближения к этому пределу.

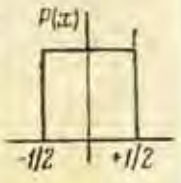


Рис. 8.1. Распределение наблюдений из равномерной генеральной совокупности.

Так, например, одно наблюдение, взятое из равномерного распределения, распределяется, конечно, в соответствии с равномерным распределением (рис. 8.1); сумма двух наблюдений из этого же распределения распределяется в соответствии с треугольным распределением (рис. 8.2);

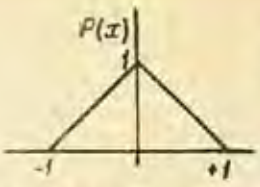


Рис. 8.2. Распределение суммы двух наблюдений из равномерной генеральной совокупности.

сумма трех наблюдений из равномерного распределения распределяется в соответствии с кривой на рис. 8.3, составленной из трех парабол. Кривая на рис. 8.3 уже очень похожа на кривую нормального распределения.

Природа ошибок такова, что при их анализе обычно применима центральная предельная теорема. Общая ошибка часто состоит из очень большого числа независимых составляющих, каждая из которых распределена случайно, причем типы распределений часто бывают неизвестны. Тогда инженер может предположить, что ошибки распределены в соответствии с нормальным законом, и это предположение чаще всего (но не всегда) оправдывается.

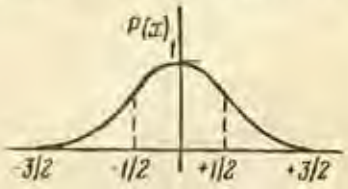


Рис. 8.3. Распределение сумм трех наблюдений из равномерной генеральной совокупности.

Таким образом, говоря о независимых ошибках, мы можем сделать два важных заключения. Первое из них состоит в том, что суммарная ошибка, как показано в § 7.4, равна квадратному корню из суммы квадратов составляющих ошибок. Второе заключение гласит, что согласно рассуждению из § 6.4 (которое относится только к распределению Пуассона, но является типичным для применений центральной предельной теоремы) мы можем достаточно уверенно судить о распределении ошибок в окрестности математического ожидания, но знаем меньше о «хвостах» распределения. К вопросу об ошибках мы вернемся еще раз в § 11.4.

Пример. Практическим примером непонимания этого принципа (нормального распределения ошибок) может служить случай, происшедший при проектировании одной системы большого масштаба, где к автопилоту предъявлялось требование удерживать заданную высоту полета с ошибкой, не превышающей 1 500 *м*. При расчете общей ошибки были изучены всевозможные причины ошибок и просуммированы все составляющие ошибок. Учитывались изменения рельефа местности, из-за которых показания барометрического высотомера не соответствовали истинной высоте полета; изменение температуры и барометрического давления; инструментальные ошибки высотомера; неотработка следящими системами данных с высотомера; неотработка самолетом данных со следящих систем. Многие из этих ошибок были, в свою очередь, комбинациями нескольких составляющих ошибок; так, инструментальные ошибки высотомера могли вызываться различными причинами.

В рассматриваемом нами случае для каждой перечисленной причины ошибок была определена максимальная величина ошибки и все эти составляющие ошибки были сложены, в результате чего была получена максимальная ошибка по высоте полета, превышающая 1 500 *м*. Отсюда было сделано заключение, что система управления высотой полета должна быть переспроектирована. Однако ясно, что такая оценка максимальной ошибки является неоправданной предосторожностью, так как вероятность одновременного возникновения всех перечисленных составляющих ошибок пренебрежимо мала.

Максимальную ошибку при инженерных расчетах обычно совершенно безопасно можно ограничить точками $\pm 3\sigma$ нормального распределения, хотя эти точки и могут быть истинными пределами диапазона равномерного распределения. Во всяком случае такие ошибки, если они независимы, не складываются линейно. Сумме даже двух наблюденных значений из равномерного распределения соответствует нулевая вероятность максимальной ошибки (рис. 8.2), а аналогичная вероятность для двух нормальных распределений составляет около 10^{-5}. Такими вероятностями можно наверняка пренебречь по сравнению, скажем, с вероятностью того, что вычисление произведено неправильно.

8.2. Теорема Чебышева

Рассмотрим произвольную плотность распределения вероятностей, например изображенную на рис. 8.4, для которой первые два момента существуют. Выберем произвольное конечное положительное число δ и будем считать его затем постоянным. Разделим площадь, ограниченную кривой, на четыре области [100], как показано на рисунке, с тем чтобы для всех x в областях, обозначенных буквой B, выполнялось условие

Рис. 8.4. Произвольное распределение вероятностей.

$$|x_B - \mu| > \delta\sigma, \tag{8.1}$$

а для всех x в областях, обозначенных буквой A, — условие

$$|x_A - \mu| \leqslant \delta\sigma. \tag{8.2}$$

Согласно определению [формула (6.5)]

$$\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 P(x)\,dx$$

и, следовательно,

$$\sigma^2 > \int_B (x-\mu)^2 P(x)\,dx, \tag{8.3}$$

где

$$\int_B = \int_{-\infty}^{\mu-\delta\sigma} + \int_{\mu+\delta\sigma}^{\infty}.$$

Подставляя (8.1) в (8.3), получаем

$$\sigma^2 > \int_B \delta^2\sigma^2 P(x)\,dx.$$

Деля обе части неравенства на положительную постоянную $\delta^2\sigma^2$, получаем

$$\frac{1}{\delta^2} > \int_B P(x)\,dx. \qquad (8.4)$$

Равенство (8.4) известно под названием *теоремы Чебышева*. Эта теорема говорит о том, что для любого распределения вероятностей, имеющего математическое ожидание и конечную дисперсию, вероятность отклонения от математического ожидания более чем на δ стандартных отклонений не превосходит единицы, деленной на δ в квадрате. Теорема справедлива для всех положительных значений δ, но для $\delta<1$ она, конечно, тривиальна. Можно показать, что более сильного утверждения такого рода для произвольных распределений сделать нельзя; мы можем высказать более сильные утверждения только для отдельных распределений. Так, в табл. 8.1 приведены вероятности отклонения от математического ожидания более чем на 1σ, 2σ и 3σ для нормального распределения, в сравнении с оценками по теореме Чебышева.

Таблица 8.1

Вероятности отклонений от математического ожидания, превышающих определенную величину $P(|x-\mu|>\delta\sigma)$

Распределение	δ		
	1	2	3
Нормальное	0,32	0,044	0,003
Произвольное	<1	$<0{,}25$	$<0{,}111$

8.3. Теорема Бернулли

Как показано раньше (§ 5.3), для биномиального распределения математическое ожидание относительной частоты k/n благоприятных исходов равно p, а дисперсия равна pq/n. Применяя теорему Чебышева, находим, что

$$P\left(\left|\frac{k}{n}-p\right|>\delta\sqrt{\frac{pq}{n}}\right)<\frac{1}{\delta^2}. \qquad (8.5)$$

Выберем теперь число ε, такое, что

$$\varepsilon=\delta\sqrt{\frac{pq}{n}}. \qquad (8.6)$$

Пусть это число остается постоянным. Тогда при изменении n число δ должно изменяться так, чтобы ε оставалось неизменным. Подставив (8.6) в (8.5) и заметив, что $1/\delta^2 = pq/n\varepsilon^2$, получим

$$P\left(\left|\frac{k}{n}-p\right|>\varepsilon\right)<\frac{pq}{n\varepsilon^2}. \qquad (8.7)$$

Неравенство (8.7) известно под названием *теоремы Бернулли*. Если предположить, что n беспредельно возрастает, то правая часть неравенства (8.7) стремится к нулю, в связи с чем левая часть этого неравенства также стремится к нулю независимо от того, насколько малым мы выберем ε.

Эта теорема, доказанная Яковом Бернулли в 1713 г., представляет собой одно из соотношений, называемых обычно *законами больших чисел*. Теорема говорит о том, что вероятность того, что частота событий с определенным исходом будет отличаться от своего ожидаемого значения больше чем на любое заданное число, как бы мало оно ни было, может быть сделана сколь угодно малой, если сделать достаточно много испытаний.

Теорема Бернулли подтверждает нашу интуицию. Однако всегда следует иметь в виду, что даже в том случае, когда вероятность отклонения, как бы мало оно ни было, приближается к нулю, нельзя считать, что некоторое отклонение, как бы оно ни было велико, является невозможным. Это утверждение имеет веские основания.

Рассмотрим ящик размером с Земной шар, наполненный черными шариками диаметром в 1 *см*. Положим один белый шарик в этот ящик и тщательно перемешаем его содержимое. Затем вынем из ящика наугад один шарик. Он наверное окажется черным, так как вероятность ему быть белым равна примерно 10^{-27}; однако вероятность, что мы вытащим именно тот шарик, который нам попал под руку, не больше этой величины.

В эксперименте с большим числом возможных исходов вероятность любого определенного исхода очень мала, и все же какой-то исход должен наступить: в достаточно длинной серии экспериментов можно ожидать появления даже самого невероятного исхода. По этой причине соотношения, подобные теореме Бернулли, не могут исключать возможности невероятного исхода. В самом деле, можно показать, что теорема Бернулли, подобно теореме

Чебышева, является самым сильным утверждением соответствующего характера, какое можно сделать. Ее можно распространить на другие распределения правильной формы (аналогичные нормальному). Кроме того, она достаточно сильна, чтобы служить основанием для математической статистики, т. е. для вывода заключений о реальном мире на основе теории вероятностей и достаточно большого числа наблюдений.

8.4. Введение в математическую статистику

Делом статистики является изучать серии наблюдений и выводить заключения. В некоторых случаях (относящихся к области исследований, называемой собственно *описательной статистикой*) заключения носят лишь характер простых сводок. Например, желая свести большие массивы данных к обозримым пропорциям, мы стараемся описать весь имеющийся материал несколькими подходящими числами, такими, как математическое ожидание, медиана, мода, дисперсия, скошенность, островершинность и целый ряд других полезных для специалистов-статистиков характеристик, включая такую тарабарщину, как *семиинтерквартильный диапазон*.

В других случаях (относящихся к области математической статистики, рассматриваемой нами в гл. 12) заключения носят характер выводов о природе механизма, лежащего в основе наблюдаемых нами явлений. При таком процессе индукции мы предполагаем справедливость законов теории вероятностей, и, в частности, законов больших чисел. Мы исследуем вероятность соответствия той или иной гипотезы опытным данным. Если при этом мы находим, что для одной гипотезы вероятность достаточно велика, а для всех других разумных гипотез она мала, то делаем заключение, что наша первая гипотеза, вероятно, справедлива. Мы также оцениваем степень нашего доверия к этому заключению. К несчастью, наука статистика в большинстве случаев не в состоянии сформулировать гипотезы, которые должны проверяться; это требует технических знаний, знакомства с исследуемой ситуацией и изобретательности.

ЛИТЕРАТУРА

Превосходной элементарной книгой по теории вероятностей является Фрей [33]. В ней рассмотрена значительная часть материала по части II нашей книги и некоторые дополнительные вопросы, и притом наиболее полезным для инженеров образом. В ней напечатаны также отличные таблицы биномиального распределения и распределения Пуассона; о других распределениях см. библиографию в гл. 12. Книга Успенского [44] является гораздо более изощренным изложением теории вероятностей. При желании ознакомиться с серьезным и строгим теоретическим изложением теории вероятностей рекомендуем воспользоваться Крамером [43].

ЗАДАЧИ

8.1. Почтовое отделение расположено между двумя главными почтовыми конторами и обслуживается этими конторами. Поступающая в отделение почта обрабатывается и укладывается в почтовые мешки. Как только накапливается 1 000 писем, почтовый мешок опечатывается и пересылается автобусом на одну из главных почтовых контор. Для ускорения отправки почтовый мешок отправляется с первым проходящим мимо почтового отделения автобусом, вне зависимости от того, в каком направлении он идет.

В конце года было установлено, что 7 000 почтовых мешков было направлено в почтовую контору, находящуюся севернее почтового отделения, и только 3 000 мешков в контору, расположенную к югу от отделения. Было также установлено, что каждый автобус, движущийся в северном направлении и останавливающийся у почтового отделения, возвращается затем на юг и снова останавливается у почтового отделения, и наоборот. Кроме того, как было установлено, письма поступают в почтовое отделение таким образом, что почтовые мешки становятся готовыми к отправке в совершенно случайные моменты времени.

Однако вероятность того, что разделение 7 000 : 3 000 возникло совершенно случайно, в действительности бесконечно мала. Как можно объяснить возникновение такого случая?

8.2. Определите вероятность $P(n)$ благоприятного исхода при n-м испытании в марковской цепи только с двумя возможными исходами. Условные вероятности для случая, когда предыдущий исход был благоприятным и неблагоприятным, равны соответственно α и β, $0<\alpha<1$, $0<\beta<1$. Определите безусловную вероятность $P(\infty)$.

8.3. а) При определенных условиях радиолампы выходят из строя с некоторой постоянной частотой m в единицу времени (например, 1 лампа из 50 000 ламп в час; $m=0{,}00002$). Какова вероятность, что данная лампа еще будет работать по истечении времени T?

б) Предположим, что радиоэлектронное устройство, например телевизионный приемник, содержит k таких ламп и что выход из строя одной лампы приводит к выходу из строя всего устройства в целом. Какова вероятность, что после T часов работы устройство будет исправно?

в) Изготовитель таких телевизионных приемников закупил долю d_1 этих ламп на заводе, для радиоламп которого частота отказов равна m_1; оставшуюся долю ламп $d_2=1-d_1$ закупил на другом заводе, радиолампы которого имеют частоту отказов m_2. Если d_1 из его телевизионных приемников было укомплектовано полностью радиолампами первого завода, а остальные d_2 из телевизоров — полностью радиолампами второго завода, то какова вероятность того, что случайно выбранный телевизор будет еще работать в момент времени T?

г) Какова вероятность того, что выбранный телевизионный приемник будет продолжать работать в мо-

мент времени T, если его изготовитель имеет такое же снабжение, но смешивает все купленные лампы и затем выбирает случайно для каждого приемника по k ламп?

Указание. Вычислите сначала условную вероятность того, что приемник, имеющий r ламп первого типа и $k-r$ ламп второго типа, будет работать в момент времени T; затем определите совместную вероятность того, что он будет иметь r ламп первого типа и что он окажется исправным в момент времени T. После этого вычислите полную вероятность суммированием по r.

д) Определите вероятность того же события, по следующим методом. Найдите вероятность того, что одна случайно выбранная лампа будет продолжать работать в момент времени T. Затем вычислите вероятность работы k таких ламп в этот момент времени.

е) Сравните результаты, полученные для случая смешения ламп (ответ *г* или *д*) с результатами, полученными для случая, когда смешения нет (ответ *в*). При каких условиях вероятность безотказной работы больше?

ж) Целесообразно ли на практике хранить порознь лампы разных заводов и комплектовать одни устройства лампами только первого типа, а другие устройства — лампами только второго типа?

8.4. Некоторое радиоэлектронное устройство содержит три радиолампы, каждая из которых имеет частоту отказов $m=0{,}0002$ в час. Если 1 000 таких устройств работает 1 000 *час*, то:

а) найдите ожидаемое количество работающих устройств;

б) определите стандартное отклонение для найденного количества работающих устройств;

в) вычислите вероятность работы 500 устройств;

г) вычислите вероятность работы 500 или более устройств.

8.5. Были выполнены наблюдения над N (10,1) с точностью до ближайшего целого числа. Определите $P(10)$, $P(9)$ $P(<9)$.

8.6. Над N (0,1) было сделано десять наблюдений с точностью до ближайшего целого числа. Одно из них меньше 9, два равны 9, три 10 и четыре 11. Какова вероятность получения такой выборки? (Нас также интересует вероятность получения этой выборки или любой другой менее вероятной выборки. Этот вопрос обсуждается в § 12.7).

8.7. Ошибки артиллерийского орудия по дальности имеют нормальное распределение со стандартным отклонением в 100 футов; ошибки этого орудия по направлению также имеют нормальное распределение со стандартным отклонением в 100 футов, не зависящее от ошибок по дальности. Каково распределение расстояний падений снаряда от цели, если считать, что систематическая ошибка отсутствует? Заметим, что это — задача на замену переменной. Даны переменные x с распределением N (0,100) и y с независимым распределением N (0,100) и требуется найти распределение переменной $r=\sqrt{x^2+y^2}$.

Указание. Сделайте преобразование $x=r\cos\theta$ и $y=r\sin\theta$, найдите совместное распределение для r и θ, затем исключите θ интегрированием.

8.8. Покажите, что при увеличении ν выражение (7.28) стремится к N (0,1).

8.9. Были выполнены наблюдения над N (17,32; 1,87). Насколько велика должна быть выборка для того, чтобы вероятность отклонения математического ожидания выборки от математического ожидания генеральной совокупности на 0,4 была меньше 1%?

8.10. Определите соотношение между распределением χ^2 при одной степени свободы и нормальным распределением.

Указание. Замените переменную χ^2 на χ.

8.11. Какова вероятность того, что два наблюдения, взятые из нормального распределения, отличаются больше чем на стандартное отклонение и больше чем на k стандартных отклонений?

8.12. Была подготовлена коммутационная система, предназначенная для случайного выбора единиц оборудования. При испытании, в котором одна за другой были выбраны все 100 единиц оборудования (коммутатор в каждом случае выбирал одну из оставшихся единиц), было замечено, что имело место в точности одно совпадение номера единицы оборудования с номером выбора, а именно шестьдесят третья единица оборудования была шестьдесят третьей выбранной. Указывает ли это на то, что работа механизма не носит случайного характера?

ЧАСТЬ 3

ВНЕШНЕЕ ПРОЕКТИРОВАНИЕ СИСТЕМ

ГЛАВА 9

НАЧАЛО РАБОТЫ И ПОСТАНОВКА ЗАДАЧИ

При проектировании системы, как и в других случаях, правильная постановка задачи в значительной мере определяет успех решения. Однако процессы, с помощью которых формулируется и решается системная задача, еще не настолько изучены, чтобы было возможно прямое перечисление необходимых и достаточных условий. Мы можем выделить элементы, необходимо входящие в правильную формулировку задачи, но их достаточность пока еще слишком зависит от специфики конкретной задачи. Мы различаем и последовательно рассматриваем четыре элемента, имеющиеся в каждой формулировке задачи, явно или неявно, а именно: 1) описание окружения задачи; 2) точка зрения, т. е. интересы, которым должно служить решение задачи; 3) общая область допустимых или желательных решений; 4) критерий (измеритель) эффективности, согласно которому должна производиться оценка предлагаемых решений.

9.1. Окружение

Описание окружения состоит в формулировке ряда определений терминов или ряда требований к желательным предметам рассмотрения. Такое описание создает одинаковую подготовку к пониманию задачи среди тех, кто связан с ее решением. Выработка такого описания в значительной мере является процессом самообучения проектировщика системы. Проектировщик подходит к задаче с определенными преимуществами: он обладает свежим взглядом на вопрос и знаком с орудиями системотехники, как они обрисованы в этой книге; однако он часто очень слабо знаком с той частной областью, в которой ему придется работать несколько месяцев или несколько лет. Поэтому ему приходится изучать основы ряда областей техники, например аэродинамики, телевидения или, скажем, работы фрезерных станков. Результаты такой самоподготовки не входят, конечно, в явном виде в описание окружения, но являются необходимой предпосылкой для его составления.

При изучении окружения проектировщик системы в большинстве случаев знакомится с уже существующей системой или со способом, посредством которого ее функция осуществляется в настоящее время. Эта работа подпадает под рубрику сбора данных и рассматривается в гл. 11. Проектировщик должен также отметить себе административные требования, «политические» соображения и другие аналогичные вопросы. Все это сделать легче, чем описать, и поэтому не будет здесь обсуждаться подробно

9.2. Точка зрения

В системотехнике, как и во всякой другой технике, решаемые задачи являются практическими задачами. Они решаются не из простого любопытства, а потому, что в этом возникает потребность. Такая потребность может быть связана с каким-либо индивидуумом или группой. Хорошим решением следует считать такое, которое удовлетворяет потребность этого индивидуума или группы. Если эта потребность не выявлена, забыта или оставлена без внимания, найденное решение может оказаться ошибочным. Следовательно, проектировщик системы в начале своей работы должен решить, какая потребность будет удовлетворяться.

Выдвигаемые требования обычно выражаются в форме вопросов. Рассмотрим следующий ряд вопросов. Как можно улучшить обороноспособность страны? Как можно улучшить противовоздушную оборону страны? Какой тип оружия является наилучшим для ПВО?

Какой тип реактивных снарядов является наилучшим для ПВО? Какая система наведения является наилучшей для реактивного снаряда А?

Каждый вопрос предполагает знакомство с общей областью рассмотрения, т. е. с характеристиками военных объектов, тактикой и т. д., но на разных уровнях детализации. Каждый вопрос означает определенную потребность, но для особой группы. Первый вопрос может ставиться, скажем, на уровне Совета безопасности *, а рассмотрение последнего вопроса наверняка могло бы ограничиться значительно более низким уровнем. Однако каждый вопрос представляет собой серьезную системную задачу. Как подчеркивается ниже, каждый вопрос влечет свой особый набор допустимых решений и свою особую меру эффективности выбранного решения.

Решение, принимаемое на высоком уровне, может снять задачу, стоявшую на низком уровне, а решение, найденное на низком уровне, может потребовать пересмотра постановки задачи на высоком уровне. Так, например, если Совет безопасности принимает решение, что единственным средством обороны страны является угроза ответного удара (т. е. наступательные действия), то целый ряд вопросов, следующих за этим первым, теряет всякий смысл. С другой стороны, разработка средства противовоздушной обороны, обладающего высокой эффективностью, может изменить точку зрения Совета безопасности.

Таким образом, каждый уровень должен непрерывно информироваться о позиции другого уровня. С другой стороны, каждый уровень имеет различные обязанности и различную ответственность. Как верхний уровень не должен пытаться рассматривать во всех деталях методы, которыми нижестоящая инстанция выполняет свою миссию, так и нижестоящая инстанция не должна считать, что только она одна несет основное бремя проектирования системы и что все то, что она считает безусловно правильным, должно обязательно изменять решения, ранее принятые на высшем уровне.

---

* Имеется в виду Совет национальной безопасности США (National Security Council, иногда переводится как «Национальный совет безопасности»), созданный в 1947 г. как совещательный орган при президенте США, где обсуждаются важнейшие вопросы внешней политики и военной стратегии США. В состав совета входят президент, вице-президент, государственный секретарь, министр обороны, директор управления по делам военной и гражданской мобилизации и несколько назначенных президентом лиц. В непосредственном подчинении совета находится Центральное разведывательное управление США. — *Прим. ред.*

Следует, однако, обратить внимание на большое число случаев, когда группа проектирования системы убеждается в том, что поставленная перед ней задача является в действительности неправильной задачей, так как рассматривается с ошибочной точки зрения.

Пример. В литературе [12] описывается подход исследовательской бригады Технологического института им. Кейса к проблеме запасов готовой продукции в одной химической компании Среднего Запада. Этой бригаде был поставлен вопрос: каким путем можно сократить запасы готовой продукции?

Казалось «совершенно очевидным», что задача состояла именно в сокращении запасов готовой продукции, так как эти запасы обходятся предприятию дорого по причине замораживания капитала, необходимости складского хранения и изменений в ценах и требованиях заказчиков. Однако бригада решила рассмотреть задачу с более широкой точки зрения. Она поставила перед собой вопрос: каковы оптимальные размеры запасов готовой продукции с точки зрения основного критерия эффективности — получения прибыли?

Рассматриваемое предприятие занималось производством продукции, к которому нельзя было применить поточные, конвейерные методы. Выпускаемое изделие при необходимости изготовлялось целой партией, и эта партия затем составляла запас готовой продукции, пока не расходилась по заказам, после чего изготовлялась новая партия.

Бригада исследовала количественно расходы при увеличении запасов готовой продукции в зависимости от перечисленных выше и некоторых других факторов. Она исследовала также экономию, которая могла бы быть получена при выпуске более крупных партий и менее частом изготовлении изделий. Оказалось, что экономия в этом случае больше затрат. Таким образом, решение проблемы запасов готовой продукции заключалось в действительности в их увеличении.

В этом случае первоначальная формулировка задачи не была неправильной, но просто задача ставилась с неправильной точки зрения. С точки зрения человека, несущего ответственность за запасы готовой продукции, задача заключалась в снижении себестоимости его собственной работы, что проще всего могло быть достигнуто путем сокращения этих запасов. С точки зрения совета директоров, следящего за получением максимальной прибыли с определенных капиталовложений, задача была совсем другой. С точки зрения промышленного диктатора (an industry czar), регулировавшего бизнес для правительства во время войны, задача могла бы быть опять совершенно другой.

Таким образом, проектировщик системы, получив задачу, должен быть уверен, что она с точки зрения лиц, давших эту задачу, поставлена правильно. Если у него возникает сомнение в правильности постановки задания, он должен обратить на это внимание заказчика. Это отнюдь не противоречит высказанному выше положению, что проектировщик не должен становиться на точку зрения, конфликтующую с нуждами заказчика.

Ошибочная постановка задачи может быть вызвана также выбором слишком узкой точки зрения. В таких случаях первоначальная,

неправильная постановка задачи может показаться совершенно нелепой с выгодной позиции предусмотрительности «задним числом», но факт остается фактом, что и весьма компетентные лица впадают в такие ошибки. Даже опытным наблюдателям иногда трудно бывает увидеть за деревьями лес.

Пример. В 1947 г. проектировалась одна система большого масштаба для обработки данных, получаемых при испытаниях систем управления реактивными снарядами. Первоначальное задание предусматривало обработку данных в реальном масштабе времени, т. е. данные должны были обрабатываться немедленно по мере их поступления. (В § 2.5 мы уже познакомились с аналогичными требованиями для системы обработки данных при испытаниях аппаратуры в аэродинамической трубе). Спустя год после начала работы, когда был завершен первоначальный проект, было установлено, что требуемое для этой цели оборудование будет чрезвычайно громоздким и дорогим. В связи с этим задание было пересмотрено и было принято решение, что в реальном масштабе времени должна осуществляться только запись поступающих данных, а обработка их может производиться уже позже. Это позволило в три раза уменьшить стоимость работ, оцениваемых во много миллионов долларов.

9.3. Допустимые решения

Постановка вопроса применительно к определенной точке зрения сразу же ограничивает класс допустимых решений. Так, например, вопрос: какой тип реактивного снаряда является наилучшим для ПВО?—исключает возможность рассмотрения истребителей как одного из ответов. Так как выбор решений не следует ограничивать без веских оснований, то первоначальную формулировку задачи следует продумать крайне внимательно. Однако класс допустимых решений ограничивается и другими условиями: состоянием техники, знанием исследуемой области, частными интересами, профессией проектировщика. Процесс решения, конечно, состоит в постепенном сужении того, что можно считать желательным решением. Однако указанные факторы оказывают свое воздействие главным образом уже после начала проектирования системы.

Как уже упоминалось в § 3.3 и будет еще раз подчеркнуто в § 27.7 и 31.3, проектирование систем должно основываться на существующей технике. Поэтому то, что уже существует, определяет и имеющиеся у нас решения. Если бы кто-нибудь приступил к проектированию системы обработки и индикации информации для фондовой биржи 50 лет назад, его мысли сосредоточились бы вокруг печатных бумажных лент, ручных телефонных коммутаторов, ручных вычислений и медленной междугородной связи. В наши дни при решении этой задачи проектировщик думал бы о магнитной ленте, электронных устройствах индикации, автоматических вычислениях и каналах связи с высокой скоростью передачи и большой пропускной способностью.

Последняя из этих двух систем не только была бы более эффективна — сама основа такой системы была бы совершенно иной. Возможность применения новых методов часто приводит к тому, что вещи, которые раньше нельзя было делать, оказываются сейчас вполне осуществимыми.

Перед тем как проектирование приведет к определенному выбору, желательно располагать как можно более широким классом допустимых решений. Следовательно, проектировщик системы должен стремиться к тому, чтобы его знания об имеющихся возможностях выбора как можно полнее совпадали с действительно имеющимися возможностями выбора. Незнание равносильно уменьшению количества научных разработок, на которые может опираться данный проект.

Ограничение класса решений из-за частных интересов может принимать многие формы. Такое положение дел, например, может иметь место в случае, когда проектировщик системы является сотрудником промышленной организации, изготовляющей компонент, который может быть применен в одном предлагаемом решении. В этом случае проектировщик будет стремиться использовать в разработке именно этот компонент, что исключит рассмотрение других решений.

Область допустимых решений может сужаться также слишком узкой профессиональной точкой зрения. Администратор может тяготеть к рассмотрению всех вопросов в первую очередь в свете организационных преобразований, технолог — в свете изменения методов и процедур, а инженер-системотехник — в свете автоматизации оборудования. Для инженера-системотехника тенденция ограничиться рассмотрением лишь вопросов автоматизации неоправданно сужает выбор решений, и нередко с большими издержками. «Чтобы делать это автоматически», — такой ответ еще не составляет основательной причины для проектирования системы большого масштаба. Этот довод в пользу проектирования новых систем обычно не излагается столь просто и может получить у проектировщика системы примерно следующую рационализированную форму: «То, что здесь делается человеком, можно представить функциями, содержащими частоты не выше нескольких герц. Однако мы, безусловно, можем создать штуковину (gadget), которая будет пропускать несколько сот герц. Совер-

шенно очевидно, что машина будет работать лучше человека».

Прежде чем принять этот аргумент, мы должны потребовать: 1) доказательства того, что человек действительно работает хуже машины при данном критерии эффективности; 2) если человек работает хуже, чем машина,— доказательства, что это связано с ограниченной полосой частот; 3) если это действительно так, — доказательства, что машинное решение не только имеет более широкие полосы частот, но и обеспечивает выполнение всех других функций, осуществляемых человеком попутно с его основной функцией (как, например, подача сигналов тревоги, гибкость).

Иногда автоматическая система внедряется потому, что кто-то другой, столкнувшись с аналогичной задачей, применил автоматическую систему и получил удовлетворительные результаты. Однако такая причина также может оказаться неосновательной. Положительные результаты, получаемые от автоматизации, которую мы копируем, могут оказаться лишь побочными продуктами производимой нами установки автоматического оборудования.

В ходе проектирования системы вся система полностью пересматривается. Это обычно приводит к существенным улучшениям работы системы и/или к снижению стоимости за счет изменения методов или процедур независимо от каких-либо выгод, приносимых самим автоматическим оборудованием. Выявление таких случаев и выдача соответствующих рекомендаций составляют часть обязанностей проектировщика; он без колебания должен заявить, что разработка новой системы не оправдана и что надлежащих результатов можно достичь путем изменения методов и процедур.

9.4. Критерий эффективности

Критерий (измеритель) эффективности является тем критерием, по которому будут оцениваться решения: предложенные решения, решения испытываемые и решения осуществленные. Этот критерий аналогичен показателю качества в технике компонентов. Чтобы объяснить это понятие, мы лучше всего приведем пример [11], когда критерий эффективности был выбран неправильно.

Пример. Во время II мировой войны английские торговые суда, плававшие в бассейне Средиземного моря, несли значительные потери от немецких пикирующих бомбардировщиков, базировавшихся на сухопутных аэродромах. Для защиты от них несколько сот таких судов было вооружено зенитными пушками. Это оснащение судов обходилось достаточно дорого, так как необходимо было не только установить орудия, но и иметь на каждом судне дополнительную команду для их боевого обслуживания. В связи с этим по истечении нескольких месяцев с момента установки орудий была проведена оценка эффективности мероприятия.

Оказалось, что вражеские самолеты сбивались только приблизительно в 4% всех атак. Такие результаты являются, конечно, невысокими, и было высказано предложение снять с торговых судов зенитные орудия и обслуживающие их расчеты. Однако лица, пришедшие к такому выводу, использовали неправильный критерий эффективности. Выбранный ими критерий (число сбитых самолетов) был связан, конечно, с вопросом: как можем мы наиболее эффективно уничтожать вражеские самолеты? Совершенно очевидно, что для этой цели существовало много других, более эффективных методов, чем установка зенитных орудий на торговых судах.

Основной задачей этих орудий было не уничтожение вражеских самолетов, а защита от них своих торговых судов. Если наличие на судах орудий заставляло пикирующие бомбардировщики держаться на большой высоте или применять противозенитные маневры и, как следствие этого, снижало результативность бомбометания, то зенитные орудия выполняли свою задачу. Для определения числа спасенных таким образом судов необходимо было сравнить число потопленных судов, имевших зенитные орудия, с соответствующими потерями судов, не имевших такой защиты. Это отношение есть критерий эффективности, относящейся к вопросу: как можно защитить торговые суда?

Подсчет показал, что из числа атакованных самолетами торговых судов с зенитными орудиями было потоплено только 10% судов, в то время как потери не защищенных орудиями судов составляли 25% от общего числа атакованных. Когда эти данные были собраны, стало совершенно ясным, что затраты на установку и эксплуатацию орудий окупаются с лихвой.

В рассматриваемом случае правильный выбор критерия эффективности по существу эквивалентен правильной формулировке задачи. Однако в большинстве случаев это дает больше, чем формулировку задачи, и в том числе выбор правильной точки зрения и правильной области допустимых решений. Эти понятия можно пояснить на примере проблемы уличного движения в городах. Отметим, однако, что эта проблема еще плохо изучена и, насколько нам известно, пока еще не было разработано достаточно общей ее формулировки; мы сомневаемся, существует ли хорошая постановка проблемы движения даже для какого-либо определенного города.

Формулировка проблемы уличного движения. В проблеме уличного движения имеется ряд соперничающих интересов, затрудняющих выбор точки зрения. Пешеход имеет одну точку зрения; лицо, которое работает в деловом центре города и должно ставить там свою автомашину на весь день, имеет другую точку зрения; лицо, которое едет на автомашине в деловой центр города для посещения магазинов или с деловым визитом и желает

поставить свою машину лишь на несколько минут, имеет третью точку зрения; лицо, которое должно проехать через город по сквозной магистрали, имеет четвертую точку зрения; водитель грузовой автомашины, которую нужно нагружать и разгружать, имеет пятую точку зрения; а владелец магазина, желающий обеспечить свободный подъезд автомашин покупателей к своему заведению, и фабрикант, желающий завозить и вывозить товары, имеют еще другие точки зрения. Над всем этим возвышаются разные группы должностных лиц и коммунальных предприятий, имеющие свои собственные точки зрения: полиция, которая должна регулировать уличное движение; пожарные, которые должны иметь быстрый доступ к любому пункту со своим громоздким снаряжением; администрация общественных работ, которая строит и ремонтирует дороги; предприятия общественного обслуживания, которые вынуждены периодически разрывать улицы для того, чтобы проложить под ними газовые и водопроводные трубы, электрические и телефонные кабели; транспортные компании, конкурирующие с автомобилями из-за пассажиров, а в случае наземного транспорта — также из-за пространства.

Правильная точка зрения будет некоторым взвешенным средним, дающим наибольший вес наибольшему числу, но пока совершенно неясно, как взять это среднее. Кто-нибудь может сказать: «Проблема уличного движения состоит в том, что транспорт движется слишком медленно. Если устранить левые повороты, скорость движения резко возрастет и проблема будет решена». Однако запрещение левых поворотов часто приводит к тому, что водитель должен проезжать лишний квартал после нужного ему перекрестка и затем выполнять три правых поворота. Это увеличивает общую длину пути при поездке и может в действительности не улучшить, а ухудшить положение дел. Кроме того, вполне возможно, что многие автомашины просто кружат в поисках стоянки и что наша проблема отнюдь не является проблемой скорости прежде всего.

Подобно этому, часто смешивают область допустимых решений с формулировкой задачи: «Проблема в том, что на перекрестках движение регулируется неправильно», или: «Недостаточно стоянок машин в переулках», или: «Чрезмерно узкие улицы», или: «Слишком большому количеству автомашин разрешается въезд в город» — и т. д. Некоторые группы могут быть убеждены, что решение одной из этих проблем решит проблему уличного движения; однако более подробное изучение показывает, что решение такой частной проблемы в действительности очень мало будет способствовать решению проблемы уличного движения или что можно ограничиться гораздо менее радикальными мерами.

Так, например, можно думать, что создание дополнительных автомобильных стоянок в переулках приведет к уменьшению стоянок в два ряда и к снижению количества автомашин, кружащих в поисках стоянок, и, таким образом, будет способствовать увеличению скорости движения транспорта; но с таким же основанием можно думать, что создание дополнительных стоянок в переулках привлечет больше автомашин в переполненную людьми и транспортом деловую часть города и ухудшит тем самым условия движения транспорта в этом районе.

Регулирование на перекрестках обсуждалось в § 2.1 и далее обсуждается в § 10.3. Два других предложенных выше решения (расширение улиц и ограничение свободного въезда автомашин в город) являются более радикальными. Ясно, что каждый из этих методов, если пойти достаточно далеко в его осуществлении, может разрешить непосредственно стоящую проблему уличного движения. Но отнюдь не ясно, насколько они допустимы: это опять-таки вопрос точки зрения. Мы хотим подчеркнуть, что такие решения нельзя ни отбрасывать без всякого рассмотрения, ни принимать без подробного исследования возможностей применения более простых методов.

Наконец, мы приходим к критериям эффективности: «Проблема в том, что слишком много заторов в уличном движении», или: «Слишком много аварий», или: «Слишком медленно движется транспорт» — и т. д. Некоторые из этих критериев мы исследуем, чтобы понять, какими необходимыми свойствами должен обладать критерий эффективности.

Характеристики критерия эффективности. Наиболее важная характеристика выбранного критерия эффективности состоит в том, что он должен измерять эффективность системы. Что этот кажущийся трюизм отнюдь не всегда можно считать само собой разумеющимся, мы уже убедились на примере с вооружением торговых судов зенитными орудиями.

Следующая по важности характеристика состоит в том, что критерий должен быть количественным — способным выражаться однозначно некоторым числом. Так, например, существование заторов транспорта (вызывающих полную остановку автомашин или резко снижающих их скорость движения) является хорошим критерием неблагополучия в улич-

ном движении, однако этот показатель является плохим критерием эффективности, так как его нельзя выразить количественно. Величина или продолжительность затора будет лишь немного лучшим критерием, так как их трудно определить. Число автомашин, проходящих через перекресток за час, среднее время задержки при переезде через перекресток или средняя скорость движения на определенном участке пути являются хорошими количественными критериями, и все они при необходимости находят то или иное применение.

Третья важная характеристика критерия эффективности заключается в том, что он должен быть эффективным в статистическом смысле (§ 12.4), т. е. должен обладать сравнительно небольшой дисперсией и, следовательно, определяться с достаточной точностью без больших затрат или потери времени. Так, одним из требований к транспортной системе является степень ее безопасности, и в качестве критерия такой безопасности можно было бы принять число долларов, теряемых ежегодно в результате несчастных случаев. Затем эту величину можно было бы сравнить непосредственно с дополнительными затратами на строительство более безопасной автомобильной магистрали.

К несчастью, природа автомобильных катастроф такова, что этот критерий имеет очень большую дисперсию. При одной катастрофе наемный рабочий может стать инвалидом на всю жизнь, и суд может присудить ему несколько сот тысяч долларов; при другой почти такой же аварии жертвы могут спастись с небольшими ранениями или, будучи ранены, могут иметь менее ограниченные средства и получат только одну десятую компенсации. Если рассматривать аварии, происшедшие в течение длительного периода времени, то закон больших чисел уравняет все эти факторы, однако другие критерии позволят получить ответ значительно быстрее.

Полнота является следующей желательной характеристикой нашего критерия. Так, простой подсчет числа аварий дает критерий безопасности, обладающий значительно меньшей дисперсией, чем рассмотренный выше критерий, и притом допускающий точное определение (при таких подсчетах исключаются аварии без ранений людей и аварии с материальным ущербом, не превышающим 50 долларов). Однако такой критерий не дает полной картины, так как одна серьезная авария может привести к значительно худшим последствиям, чем целая серия небольших аварий.

По этим же самым соображениям число аварий с человеческими жертвами также не дает полной картины, так как при этом не учитывается гораздо большее число аварий, в которых не было человеческих жертв, но последствия которых в своей совокупности могут оказаться более серьезными. Нам нужно какое-то взвешенное среднее всех этих аварий, которое давало бы нам более или менее полную картину того, насколько безопасна автомобильная магистраль (или перекресток, или город, или еще что-либо); но трудно найти критерий, который был бы и полным, и эффективным.

Желательно также, чтобы критерий эффективности обладал еще целым рядом других характеристик. Желательно, чтобы он был прост, когда это совместимо с требованием полноты, и имел физический смысл; в этом случае у нас меньше возможности впасть в ошибку при его применении. Кроме того, если критерий имеет физический смысл, часто удается довольно легко найти идеальную характеристику работы системы и сравнить ее с реальной характеристикой. Обычно бывает полезно знать, достигнут ли теоретический предел или еще существует значительный простор для улучшения. Так, при рассмотрении проблемы уличного движения критерий в виде средней скорости движения на определенном участке пути является достаточно простым, имеет реальный физический смысл и может сравниваться со свойствами идеальной транспортной системы.

При изучении проблемы междугородных автомобильных сообщений выбор такого критерия эффективности позволяет нам весьма успешно определять, какие улучшения может нам дать строительство первоклассной автострады, так как применительно к этому критерию большинство наших автострад обладает стопроцентной идеальной характеристикой. С другой стороны, в деловых районах некоторых наших городов реальная характеристика магистрали, с этой точки зрения, равна одной десятой идеальной характеристики или даже меньше нее, и достичь характеристики, близкой к идеальной, не удается, если только не прибегнуть к таким радикальным мерам, как строительство надземных магистралей над деловой частью города.

В тех случаях, когда реальную характеристику можно сравнивать с идеальной, иногда представляется целесообразным нормировать критерий эффективности, с тем чтобы он принимал значения от нуля (что соответствует самой плохой характеристике) до единицы (случай идеальной характеристики).

Распространенные критерии. Некоторые критерии эффективности применяются очень часто. Так, например, при исследовании коммерческих и индустриальных систем в качестве критерия эффективности почти всегда применяется стоимость; этот критерий довольно часто применяется и при оценке систем, не связанных с получением прибыли. При оценке различных военных систем почти во всех случаях в качестве критерия эффективности используется вероятность поражения цели и соотношение потерь. В невоенных системах, разрабатываемых или эксплуатируемых под контролем правительственных органов, на первом плане стоят такие факторы, как безопасность и качество обслуживания.

Там, где дело касается безопаснюсти или человеческой жизни, применяется один специальный коэффициент. Почти всегда можно обеспечить высокую безопасность системы за счет лишних долларов или худшей ее характеристики. Для эффективной оценки этих мер безопасности необходимо установить, какой суммы в долларах стоит человеческая жизнь или на сколько, например, миль в час мы согласны снизить скорость для того, чтобы спасти в среднем одну человеческую жизнь в год. Попытка спрятаться от прямой постановки этого вопроса не принесет никакой пользы и может расцениваться лишь как «политика страуса», как желание закрыть глаза на действительное положение вещей; отказ сделать такую оценку в явном виде означает только, что система будет проектироваться применительно к какому-то неявному эквиваленту, который, если сделать вещи явными, можно было бы улучшить.

Общие и местные критерии эффективности системы. Совершенно не обязательно, чтобы один и тот же критерий применялся ко всей системе в целом и к каждой ее части. Например, среднее число автомашин, проходящих через перекресток в течение одного часа, может служить отличным критерием эффективности для регулирования движения на данном перекрестке; однако из этого, конечно, не следует, что общее число автомашин, прошедших через все перекрестки, является критерием для оценки городской системы уличного движения в целом. С другой стороны, при проектировании частей системы нельзя не учитывать общего критерия эффективности, так как местный оптимум не всегда обеспечивает хорошую характеристику всей системы.

Всегда существует возможность улучшить условия движения на каком-либо определенном перекрестке посредством запрещения на нем поворотов, однако это, конечно, затрудняет решение проблемы движения на соседних перекрестках. Аналогичным образом, хотя на первый взгляд это менее очевидно, можно, улучшая условия движения на близлежащих перекрестках, сосредоточить столько автомашин на одном главном перекрестке (где сходятся не две, а три улицы), что там возникнет безнадежный затор; может оказаться невозможным устранить этот затор никакими другими методами, кроме как снижением скорости потока через главный перекресток.

Проектные критерии. Вообще говоря, о критерии эффективности думают как о чем-то, что надо увеличивать до максимума (или — в соответствующем случае — уменьшать до минимума) либо приближать насколько возможно к идеалу. Так, процент отказов оборудования на заводе-автомате должен быть как можно ниже; вероятность поражения цели для боевых систем должна быть как можно ближе к единице; дальность действия радиолокационной станции обычно должна иметь максимальное возможное значение. Однако существуют определенные случаи, когда критерий эффективности является скорее некоторой граничной характеристикой, которая должна быть получена, но превышение которой не приносит дополнительного выигрыша.

Так, радиолокационная станция сопровождения в системе наземного управления посадкой самолетов должна обеспечивать определение координат самолетов с вероятностью, практически равной 1,0, вплоть до максимальной дальности, с которой система управляет полетами самолетов. В то же время применение радиолокационной станции, способной определить координаты самолетов на расстоянии, в 10 раз превышающем это максимальное расстояние управления, не дает никаких особых преимуществ. Такие граничные характеристики мы называем *проектными критериями;* соответствие системы каждому такому проектному критерию является своего рода показателем эффективности системы.

Фактически эти проектные критерии представляют собой частный случай нелинейности определенных критериев эффективности. Например, если одна предложенная система осуществляет предупреждение гражданской ПВО через 1 час, а другая — через 2 часа, то ни одна из них не лучше другой, так как обе практически бесполезны; если одна система осуществляет предупреждение через 1 *сек*, а другая — через 2 *сек*, то ни одна из них не лучше другой, так как обе приемлемы при заданном проектном критерии; но если одна система осуществляет предупреждение через

1 *мин*, а другая через 2 *мин*, то первая явно лучше второй. Если критическая область измеряемой переменной (от 1 до 2 *мин*) соответствует резким изменениям требуемых усилий, то эта переменная должна быть введена в основной критерий эффективности.

Множественные критерии. В ряде случаев проектировщик системы может столкнуться с весьма большим количеством проектных критериев, причем на первый взгляд любой из них или любая их комбинация может быть выбрана как основа для критерия эффективности. Так, например, при разработке системы военного назначения, предназначенной для наблюдения обстановки за вражескими линиями (скажем, самолетной радиолокационной или инфракрасной системы), необходимо будет учесть по крайней мере следующие критерии: зону обзора (число квадратных миль, просматриваемых за один час), дальность действия (максимально достижимое расстояние за линией фронта), скорость (с которой данные разведки доставляются в соответствующие штабы), разрешающую способность (степень различимости отдельных деталей изображения), точность (ошибки в определении положения обнаруженных целей), пропускную способность (способность передавать большое количество данных), всепогодность (работа ночью, в туман, дождь, при дымке), уязвимость при воздействиях со стороны противника, уязвимость при контрмерах (применение помех, ложных целей, маскировки), скрытность (возможность перехвата противником данных разведки), тактическую осуществимость, техническую осуществимость, гибкость (применимость в различных климатических условиях, в различной местности, при различных видах боевых действий, возможность улучшения и т. п.), надежность, сложность, стоимость (затраты рабочей силы, материалов, денежных средств), мобильность (вес, обслуживание, требуемые аэродромы и т. п.), снабжение (дефицитные материалы).

Такой перечень, по словам одного крупного специалиста, «действует на разработку, как стрихнин на нервную систему». Этот перечень требований должен быть сокращен в такой мере, чтобы он стал практически выполнимым. Очень большую пользу приносит группировка критериев по рубрикам. Так, например, первые семь пунктов приведенного выше перечня мы могли бы сгруппировать под рубрикой «Характеристики работы». Все другие пункты можно было бы разбить на три группы под заголовками: «Осуществимость системы в действии», «Критерии разработки» и «Производственные критерии».

Ясно, что критерии работы следует рассмотреть в первую очередь. Если новая система не дает надежды на лучшую работу, чем существующая система, то она не оправдывает никаких дальнейших усилий. Остальные критерии на первом этапе должны подвергнуться краткому рассмотрению, чтобы среди них не оказалось такого, который делал бы создание всей системы невозможной. Однако многие из этих критериев не удастся надлежащим образом изучить до тех пор, пока не наступит основная фаза проектирования.

Когда наиболее важные критерии выбраны, проектировщик системы пытается определить взаимосвязь между ними. Так, например, рассматривая самолетную обзорную систему, можно принять, что для повышения разрешающей способности за счет уменьшения зоны обзора полеты должны осуществляться на небольшой высоте (или необходимо применить линзы с большим фокусным расстоянием в оптической системе, или же антенну бо́льших размеров, либо более короткую длину волн в радиолокационной системе).

Таким образом, размеры площади, просматриваемой за время одного обзора, и минимальные размеры объекта, характеризующие разрешающую способность системы, находятся в прямой линейной зависимости, и для определения этих отношений можно использовать один критерий — число строк, или «линий» (например, номинальный стандарт вещательных телевизионных станций* равен 525 строкам; однако следует помнить, что такие термины могут иметь различные определения и что такая же разрешающая способность считается в фотографии равной 262,5 строки).

Используя такие соотношения, иногда удается выявить немногие основные критерии, с помощью которых могут быть выражены все остальные. Конечная цель, правда не всегда достижимая, заключалась бы в получении одного критерия эффективности, включающего эффекты всех основных критериев. Однако для этого сначала требуется выявить все взаимосвязи, для чего необходимо построить математическую модель системы (гл. 10).

9.5. Исследование операций

Читатель, знакомый с методами исследования операций, заметит, что многое из сказанного выше является общим как для системотехники, так и для исследования операций.

---

* В США — *Прим. ред.*

Термин «исследование операций» имеет много определений. Морз [35] пишет: «Значительную часть времени ... конференции по исследованию операций затратили на определение исследования операций и на убеждение других также попытаться сделать это».

Морз и Кимбелл [11] называют исследованием операций «научный метод, дающий в распоряжение исполнительного органа количественные основания для принятия решения в управляемых ими операциях». Нас интересует не эта функция, а способ, с помощью которого таланты исследователя операций могут использоваться в бригаде проектирования системы; полученная этим человеком подготовка особенно полезна в вопросах внешнего проектирования системы. Имея это в виду, интересно посмотреть, какие характеристики ему даются в работах по исследованию операций.

Исследователь операций обладает академическим образованием в одной из научных дисциплин, обычно в одной из естественных наук, и почти всегда — серьезной математической подготовкой. Он имеет вкус к изучению различных вопросов и широкий кругозор. Он желает изучать операции и прекращает свою работу, как только приступают к непосредственной разработке аппаратуры (однако вновь включается в работу на последующих этапах для оценки созданной аппаратуры и определения наилучших методов ее использования).

В своей собственной сфере исследователь операций является орудием администратора. Область задач, над которыми он должен работать, может определяться администратором, однако довольно часто он сам открывает различные проблемы и просит разрешения работать над ними. После изучения и, возможно, решения проблемы он докладывает результаты своих исследований администратору, который затем принимает решения. Исследователь операций обычно не участвует в принятии этих решений, но следит за ними, оценивает их и в случае необходимости вновь докладывает свое мнение.

Этапы, выполняемые при решении той или иной задачи по исследованию операций, могут изменяться в зависимости от характера самой задачи. Часто порядок работы бывает следующим: подготовительное изучение общего характера задачи и тех людей и вещей, которые имеют к ней отношение; выбор одного или нескольких критериев эффективности; разработка математической модели, изображающей изучаемую систему; сбор данных; применение логических методов к работе с математической моделью; поиск решений задачи в свете выбранных критериев эффективности; наконец, в нужном случае оценка результирующих изменений. Орудиями исследователя операций служат теория вероятностей, теория игр, теория массового обслуживания, статистика, линейное программирование и много других орудий проектировщика систем.

Совершенно понятно, конечно, что квалифицированный исследователь операций может составить ценное добавление к бригаде проектирования системы. Когда он работает над задачами системотехники, все сказанное выше остается в силе, с тем лишь исключением, что он работает как член бригады проектирования системы, приобретает интерес к аппаратуре и оказывает помощь в принятии решений по проекту. Каждый из этапов, нормально используемых им при исследовании операций, также необходим и при проектировании систем и обсуждается в соответствующих разделах настоящей книги.

Как легко заметить из предыдущего, исследование операций является очень широкой областью. Оно не похоже на другие орудия проектирования систем, так как, во-первых, оно не вполне определено и, во-вторых, включает в себя многие из этих орудий. В самом деле, в последнее время обозначилась тенденция настолько расширять определение исследования операций, что оно практически становится синонимом системотехники. Однако при этом существует коренное отличие в подходе: исследователь операций стремится в первую очередь изменить процедуру, методы работы, в то время как проектировщик системы хочет в первую очередь изменить аппаратуру. По всем этим причинам нам нет необходимости останавливаться более детально на исследовании операции как таковом.

ЛИТЕРАТУРА

Первая книга об исследовании операций, написанная Морзом и Кимбеллом [11], еще остается превосходным трудом по этому вопросу. Обширную библиографию по исследованию операций и его применениям дают Мак-Клоски и Трефетен [36].

ГЛАВА 10

МАТЕМАТИЧЕСКИЕ МОДЕЛИ

В последние годы XIX в. лорд Кельвин заявил, что он не может понять ни одного явления до тех пор, пока не представит себе механическую модель, обладающую подобными свойствами. Это заявление неоднократно цитировалось и большей частью порицалось, как пример ненаучного консерватизма. Тем не менее в нем содержится значительная доля истины. Сегодня мы не настаиваем, чтобы модель была механической, поскольку это слово предположительно указывает на ньютоновскую механику, которая не является универсально применимой. Однако остается верным, что понимание некоторой модели необходимо для полного понимания любого явления — и в особенности таких сложных явлений, какие происходят в системах большого масштаба.

Математическая модель позволяет нам также изучать за один прием по одной части системы; без нее нам пришлось бы решать задачи, обладающие слишком большой сложностью и характеризуемые чрезвычайным разнообразием. Предположим, что в нашем распоряжении (как результат боевых донесений военных лет) имеется 100 сообщений о дальностях обнаружения вражеских подводных лодок самолетной радиолокационной станцией некоторого определенного типа и нам необходимо охарактеризовать параметры этой радиолокационной станции и предсказать ее дальность действия по тем или иным целям, чтобы установить возможность ее использования в качестве компонента системы большого масштаба, имеющей военное назначение. Подобные задачи возникают, конечно, для многих систем в отношении диапазона входов.

Допуская, что в нашем распоряжении имеются все возможные данные (конечно, на практике этого никогда не бывает), исследуем зависимость дальности обнаружения от таких переменных, как состояние (волнение) моря, размеры подводной лодки, ее ракурсы и направление ветра по отношению к линии наблюдения лодки с самолета, качество технического ухода за радиолокационной станцией, бдительность и подготовка оператора радиолокационной станции. Однако даже и в этом случае существуют некоторые данные, которые мы не можем надеяться получить, например такие, как количество вражеских подводных лодок, вошедших в зону обнаружения, но еще не обнаруженных радиолокационной станцией.

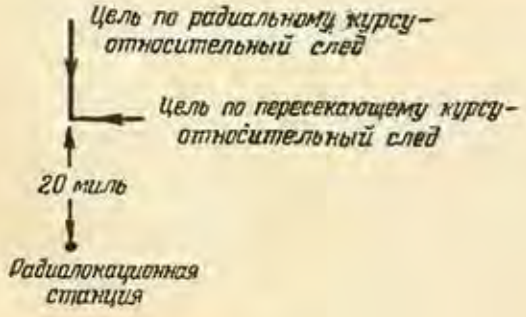


Рис. 10.1. Цели, движущиеся радиальным и пересекающим курсом относительно линии наблюдения.

И если мы сможем определить нужные функциональные зависимости (что совершенно невозможно при небольшом объеме наличных данных), определение дальности действия станции при различных условиях ее работы все же остается весьма сложной задачей. Например, как можем мы сравнить кумулятивные вероятности обнаружения двух изображенных на рис. 10.1 целей в момент, когда они приблизятся к радиолокационной станции на расстояние в 20 миль? Ясно, что цель, идущая пересекающим курсом, имеет бóльшую вероятность обнаружения в этой точке, но насколько больше эта вероятность? И если мы сумеем ответить на эти вопросы, как можно будет распространить полученные данные на подводные лодки других типов?

Один из методов подхода заключался бы в получении дополнительных данных путем проведения экспериментов, предусматривающих подходы подводной лодки в различных условиях и определение дальностей обнаружения. Однако такие опыты обходятся крайне дорого. Кроме того, при таких опытах бдительность оператора радиолокационной станции имеет большое значение, и потому потребовались бы большие выдержки времени перед приближением подводной лодки на дальность обнаружения, в связи с чем в течение одного дня удалось бы получить лишь немногие данные. А для того чтобы наблюденные значения дальности обнаружения можно было сгруппировать в классы для статистического анализа, требуется очень и очень много таких данных. Представляется очевидным, что нам нужна математическая модель, чтобы сделать правильные выводы из имеющихся в нашем распоряжении данных.

Модели радиолокационного обнаружения. Во II мировой войне для оценки отдельных радиолокационных станций часто использовалась рабочая характеристика, называемая *максимальной дальностью.* Для ее определения использовалась в качестве математической модели *формула дальности радиолокации.* Эта формула устанавливает, что мощность принимаемого сигнала равна произведению следующих величин:

— мощности передаваемого сигнала;

— усиления передающей антенны;

— затухания, связанного с квадратическим ослаблением сигнала при его распространении от станции к цели;

— радиолокационного сечения (эффективной площади рассеяния) цели;

— затухания сигнала на обратном пути от цели к станции;

— площади приемной антенны.

Если нам известна мощность передаваемого сигнала, минимальная мощность принимаемого сигнала, при которой он может быть еще обнаружен, и параметры антенны, то мы можем найти по этой формуле *максимальную дальность обнаружения* радиолокационной станцией цели с определенным радиолокационным сечением.

Однако в действительности эта характеристика не имеет существенного практического значения. Не говоря уже об атмосферных явлениях, которые в какой-то период времени могут сделать видимыми цели на чрезвычайно больших дальностях, заметим, что одинаковые по своему типу цели отнюдь не обнаруживаются на одном и том же расстоянии одной и той же радиолокационной станцией при одинаковых условиях погоды. Дальности обнаружения колеблются в соответствии с некоторым распределением вероятностей. Поэтому необходимо применить другую математическую модель, которая учитывала бы эти факторы.

Один очевидный и непосредственный метод подхода мог бы состоять в нахождении распределения каждой из переменных, входящих в формулу дальности радиолокации. Так, радиолокационное сечение цели изменяется во времени (явление, часто называемое *мерцанием* цели), и это изменение можно определить теоретически или экспериментально; путь распространения радиоволн изменяется таким способом, который можно описать через рассмотрение детальных метеорологических условий; шумы, среди которых должен быть выделен отраженный сигнал, можно описать (гл. 28) с помощью нормального распределения, параметры которого распределяются в соответствии с другими величинами, включая количество пятен на Солнце и высоту ионосферы. К несчастью, этот метод оказывается в применении к радиолокационному обнаружению слишком сложным и громоздким, чтобы приносить пользу в большинстве практических случаев.

Другой метод, используемый часто в ВВС США и особенно пригодный для устройств автоматического обнаружения, основан на предположении, что обнаружение происходит всякий раз, когда сигнал, сложенный с шумом, превысит некоторое определенное значение.

Амплитуды сигнала и шума могут тогда определяться независимо одна от другой; приближенно сигнал можно принимать даже за некоторую постоянную величину, определяемую формулой дальности. Решение при этом представляется в виде двух вероятностей: вероятности того, что некоторая определенная цель будет обнаружена на некоторой определенной дальности (т. е. что сигнал, сложенный с шумом, будет превышать пороговый сигнал), и вероятности того, что в течение определенного интервала времени только шум, при отсутствии реального сигнала, создает сигнал ложной тревоги.

Третий метод, используемый в ВМФ США, основан на предположении о существовании определенной вероятности того, что сигнал появится в какой-либо определенный период обзора; во-вторых, определенной вероятности того, что оператор заметит сигнал, если он появится; и, в-третьих, определенной вероятности того, что оператор сообщит об обнаружении на основании одиночного наблюдения сигнала (во многих случаях оператор ожидает появления в следующий период обзора подтверждающего сигнала и сообщает об обнаружении цели только в том случае, если сигнал от нее появляется при двух последовательных обзорах).

Если теперь подводную лодку направить по маршруту, известному оператору, то вторую и третью вероятности можно сделать заранее известными (а именно, равными единице в каждом случае). При этом оператор точно знает, в какой точке экрана должен появиться сигнал, и при каждом обзоре пространства может сообщать, наблюдает ли он сигнал или нет. (Так как в этом случае оператор хорошо знает положение цели, то при опыте должны быть приняты специальные меры, исключающие возможность фиксации воображаемого сигнала; см. § 11.5.)

На практике подводная лодка направляется от радиолокационной станции ра-

дальным курсом и каждые несколько секунд фиксируется одно наблюдение (наличие или отсутствие обнаружимого сигнала при данном обзоре на известной дальности). После того как цель вышла из области обзора и удалилась от нее на значительное расстояние (т. е. после того, как сигнал не наблюдался в течение ста или более последовательных обзоров), она разворачивается и входит в область обзора под неизвестным азимутом, а мы производим одну серию наблюдений по обнаружению. После обнаружения цели регистрация наблюдаемых результатов осуществляется до тех пор, пока сигналы не будут появляться при каждом обзоре, после чего приступают к следующему циклу испытаний.

Эти опыты по обнаружению цели дают и вероятность фиксации сигнала оператором, причем достаточно сравнительно небольшого числа опытов, так как вероятность появления обнаружимого сигнала при этих опытах уже известна (кстати сказать, эти опыты используются также для подтверждения теоретических положений, на которых основана математическая модель). Вероятность того, что потребуется только один сигнал, может быть найдена на основании субъективной оценки оператором своих действий, а подтверждением может служить форма кривой обнаружения.

Типы моделей. Математическая модель системы будет полезна на всем протяжении проектирования системы. Мы говорим о ней сейчас потому, что она очень важна при внешнем проектировании систем и служит соединительным звеном между внешним и внутренним проектированием. Модель каждой части системы часто может создаваться более или менее независимо от моделей других частей системы, и улучшение различных моделей часто осуществляется почти непрерывно на всем протяжении проектирования системы.

Будем различать четыре типа моделей: аналитическая жесткая, численная жесткая, аналитическая вероятностная и численная вероятностная, называемая также моделью «Монте Карло». Примером аналитической жесткой модели является формула дальности радиолокации.

Другие рассмотренные модели радиолокационного обнаружения являются аналитическими вероятностными моделями. Численное интегрирование дифференциального уравнения является численной жесткой моделью; этот тип редко встречается в проектировании систем, и на нем мы в дальнейшем останавливаться не будем.

Модели «Монте Карло» также использовались при разработке радиолокационной аппаратуры; такими моделями являются опыты по обнаружению, в которых видеосигналы имитируются устройствами, генерирующими соответствующие случайные сигналы.

В большинстве случаев оказывается возможным создавать модели более чем одного из перечисленных типов; выбор типа модели определяется при этом простотой решения. Жесткая модель (под этим в дальнейшем будет подразумеваться аналитическая жесткая модель) применяется в случаях, когда различные действия можно описать не прибегая к распределениям. Из этого не обязательно следует, что лежащие в основе явления не суть явления статистической природы; это говорит только о том, что мы хотим иметь дело со средними значениями, а не с целыми распределениями. Так, ньютоновская механика и термодинамика суть жесткие модели, верные, как мы знаем, только для средних значений, но тем не менее очень полезные.

Существуют, однако, некоторые задачи, для которых более приемлемыми оказываются методы статистической механики; в этом случае перед нами просто аналитическая вероятностная модель тех же самых явлений. В сложных ситуациях вероятностные модели часто с трудом поддаются аналитическому рассмотрению, и тогда можно воспользоваться более утомительными, но зато более простыми методами «Монте Карло».

10.1. Жесткие модели

Механика Ньютона, законы Кирхгофа, законы сохранения массы и энергии и все другие знакомые аналитические орудия инженера суть жесткие модели, и нам не нужно много говорить о методике, формулировке и решении их. Однако следует особо отметить, что жесткие модели часто могут применяться с большой точностью к явлениям, которые на первый взгляд кажутся чисто случайными. Несколько интересных примеров этого описано в литературе [13] и [14]. Так, например, следующая формула описывает большое число социальных явлений:

$$R^n S = M. \qquad (10.1)$$

Здесь все объекты некоторого класса располагаются по рангу в соответствии с их размерами; R — ранг определенного объекта, S — его размер, а n и M — постоянные.

При $n=1$ эта формула описывает распределение населения 3464 городов Соединенных

Штатов, имевших (по переписи 1940 г.) более 2 500 жителей. Если данные определенным образом изменить, ориентируясь на зоны метрополий*, а не на политическое (применяемое при переписи) деление, то соответствие будет еще более поразительным. График на рис. 10.2 иллюстрирует это соответствие.

Уравнение (10.1) оказывается применимым и к результатам всех других переписей населения городов США, проводившихся в период с 1790 г. до наших дней. По-видимому, эта формула изображает результат какого-то определенного социального давления; например, она не выполняется для городов Великобритании, но выполняется для всех городов континентальной Европы. Та же формула, но при $n=0{,}5$ характеризует доходы населения в Соединенных Штатах; при $n=1$ она описывает частоту употребления различных слов в английском языке (точнее, сколько раз определенное слово будет встречаться, скажем, на один миллион слов).

Другое интересное соотношение, подробно описанное в указанных двух статьях и в другой литературе, приводимой в них, касается так называемого потенциала населения: действие различных центров населения в некоторой точке пропорционально отношению числа жителей этих центров к расстоянию последних от рассматриваемой точки. Например, число студентов, посещающих Гарвардский университет, но проживающих в каком-либо другом штате, чем Массачусетс, равно произведению некоторой постоянной на число жителей штата, деленному на его расстояние от Кембриджа**. То же правило, но с другими постоянными применимо к Принстонскому университету и Массачусетскому технологическому институту. Количество экземпляров газеты «Сент-Луис Стар-Таймс», покупаемых в любом пригороде или другом городе, находящемся на удалении примерно до 150 миль от Сент-Луиса, пропорционально количеству жителей соответствующего пункта, деленному на его удаление от Сент-Луиса.

То же правило применимо к количеству заказов, сделанных фирме Джордан Марш и К°, Бостон, в различных городах и поселках Новой Англии.

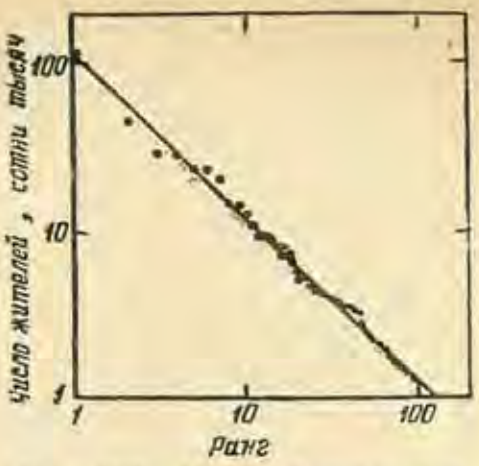


Рис. 10.2. 100 наибольших зон метрополий США в 1940 г. (по Ципфу [13]).

В случаях, когда между двумя городами происходит взаимообмен того или другого вида, объем взаимообмена пропорционален произведению количеств жителей в этих городах, деленному на расстояние между ними; это относится к телефонным переговорам, перевозкам багажа железнодорожной компанией «Рейлвей Экспресс», движению грузовых автомашин и автобусов и ко многим другим вещам.

Одна хорошо известная жесткая модель, применяемая к случайным процессам, получила название *уравнений Ланчестера*. Наиболее известное из них указывает, что эффективность боевой группы пропорциональна квадрату ее величины. Например, 100 солдат уничтожат 50 солдат противника, потеряв только 25 своих. Квадратическое уравнение Ланчестера оказалось удивительно хорошо применимым ко многим условиям боевых действий в период II мировой войны [15].

Ясно, что такие формулы могут оказать значительную помощь при организации большого массива данных и составлении предсказаний. Конечно, такие правила имеют много исключений, часть из которых заранее можно предвидеть. Например, прием студентов из разных городов в «национальные университеты» будет описываться приведенной выше формулой более точно, если мы будем рассматривать не все население штата, а только белое население; формула совсем неприменима к Мексике и Канаде.

10.2. Аналитические вероятностные модели

Взаимосвязь между аналитическими вероятностными моделями и жесткими моделями можно показать на простом примере из статистической механики. Мы можем по-

---

* В США зоной метрополии города (metropolitan district, metropolitan area) называется территория города вместе с его пригородной зоной, хотя бы эти пригороды и были административно самостоятельны. — *Прим. ред.*

** Имеется в виду американский город Кембридж, штат Массачусетс, в котором расположен Гарвардский университет. — *Прим. ред.*

строить жесткую модель на наивном предположении, что все молекулы движутся с одинаковой скоростью, и, используя простейшие алгебраические соотношения, вывести закон Бойля в виде

$$PV=\frac{2E}{3}, \qquad (10.2)$$

где E — полная кинетическая энергия молекул в объеме V.

Более строгий подход к этому вопросу, предполагающий, что скорости молекул распределяются в соответствии с законом Максвелла — Больцмана, требует гораздо более сложной математики; такова аналитическая вероятностная модель, приводящая, как оказывается, к тому же самому результату — формуле (10.2). Конечно, последняя модель позволяет также предсказывать с высокой точностью другие свойства, которые не могут быть выведены из жесткой модели.

Мы уже говорили в § 6.6, что в любой задаче по теории вероятностей нам даются некоторые вероятности и некоторые правила и затем требуется вычислить другие вероятности. Аналитическая вероятностная модель устанавливает те правила, по которым надо делать эти расчеты; другими словами, она описывает зависимости между двумя или более распределениями вероятностей. Два простых примера этого приводились в § 6.7.

В одном из них нам было дано, что появление некоторых событий происходит в соответствии с распределением Пуассона, и требовалось определить характер распределения интервалов времени между появлением этих событий. Ответом служило экспоненциальное распределение. Во втором примере нам требовалось найти распределение $\operatorname{tg}\theta$ при равномерном распределении угла θ. Ответом служило распределение Коши. Большое число задач такого типа было рассмотрено в гл. 7. Аналогичный пример был рассмотрен также в начале этой главы, когда рассматривалось соотношение между кумулятивной вероятностью обнаружения и составляющими вероятностями. Мы будем находить много еще более сложных примеров этого в системотехнике; в частности, мы посвятим целую главу теории массового обслуживания, которая является простым развитием аналитической вероятностной модели, основанным на предполагаемом распределении входов при различных перестановках каналов, различных буферных накопителях и различных распределениях времен занятия.

Сила аналитической вероятностной модели демонстрируется следующим примером. Предположим, что в некоторой системе военного назначения боевые средства атакуют цель. Известно, что вероятность поражения при единоборстве между одной боевой единицей и одиночной целью равна p. Ожидается, что множественные цели будут встречаться сразу с группой боевых единиц, в связи с чем возникает задача распределения боевых средств. Если принять, что боевые единицы должны выбирать цели для атаки самостоятельно, т. е. децентрализованно, то мы должны иметь в виду возможность «переуничтожения» целей, так как значительное число боевых единиц будет атаковать одну и ту же цель, в то время как другие цели окажутся неатакованными.

Применение аппаратуры централизованного управления боевыми единицами повысит эффективность их распределения по целям, что приведет к уничтожению большего числа целей. С другой стороны, долларовые средства, которые придется при этом затратить на разработку и производство аппаратуры управления, можно было бы обратить на закупку дополнительного количества боевых единиц применяемого типа или на разработку и приобретение нового боевого оружия, обладающего большей вероятностью p. Какая из этих трех альтернатив желательна?

Пусть m — количество целей и n — количество боевых единиц в одной отдельной задаче распределения боевых средств (т. е. при одном вражеском рейде). Для простоты предположим, что отношение n/m является целым числом; результирующая ошибка при этом будет небольшая. Рассмотрим две модели распределения боевых единиц:

Случай 1. Полное управление, или распределение на каждую цель n/m боевых единиц.

Случай 2. Случайное распределение, когда каждая боевая единица имеет вероятность $1/m$ атаки любой данной цели.

Случай 1 служит примером хорошего управления действиями своих боевых средств, а случай 2 — примером отсутствия всякого управления.

Конечно такой подход — слишком большое упрощение реальной задачи. Фактически вероятность p не является постоянной величиной и может быть некоторой функцией конкретной комбинации боевых единиц и целей. Если вероятность p не является постоянной величиной, то полное управление в том смысле, как оно здесь определено, нельзя, конечно, считать наилучшим методом управления. Подобно этому, случай 2 не является наихудшим из всех возможных методов управления.

Возможен случай, когда все боевые единицы атакуют одну из нескольких целей, а все другие цели не подвергаются нападению. При реальном проектировании систем все эти соображения должны, конечно, приниматься во внимание, но сейчас мы их опустим.

Для решения задачи мы оценим два выражения — ожидаемое количество уничтоженных целей в первом и втором случае. В случае 1 каждая цель атакуется n/m боевыми единицами, в связи с чем мы, используя формулу (5.14), можем написать, что вероятность уничтожения i-й цели равна

$$P_i = 1 - (1-p)^{n/m},$$

где $i = 1, 2, \ldots, m$. Так как ожидаемое количество уничтоженных целей для каждой цели равно $P_i \times 1$, то общее ожидаемое значение в случае 1 составляет

$$E_1 = m[1 - (1-p)^{n/m}]. \qquad (10.3)$$

Для случая 2 вероятность направления на i-ю цель первой боевой единицы равна $1/m$, а вероятность ее уничтожения этой боевой единицей составляет p/m. Вероятность, что она не будет уничтожена этой боевой единицей, равна $1-p/m$, а вероятность, что она не будет уничтожена всеми боевыми единицами, равна $(1-p/m)^n$. Следовательно, вероятность того, что i-я цель будет уничтожена, равна

$$P_i = 1 - \left(1 - \frac{p}{m}\right)^n,$$

откуда следует, что

$$E_2 = m\left[1 - \left(1 - \frac{p}{m}\right)^n\right]. \qquad (10.4)$$

Это выражение можно вывести для более общих условий, умножив вероятность уничтожения при данном способе распределения боевых единиц (k_i боевых единиц нацеливаются на i-ю цель) на вероятность этого распределения (здесь имеет место полиномиальное распределение вероятностей) и просуммировав эти произведения по всем возможным способам распределения, для которых $\Sigma k_i = n$. В итоге мы получаем выражение

$$E_2 = m^{-n} \sum_{\Sigma k_i = n} \frac{n!}{\Pi k_i!} \sum_{i=1}^{m} [1 - (1-p)^{k_i}],$$

которое преобразуется в выражение (10.4). Посредством этого выражения при необходимости можно также определить дисперсию числа уничтоженных целей.

Формулы (10.3) и (10.4) можно нормализовать делением каждого выражения на m, чтобы получить числа, лежащие в пределах от 0 до 1. При этом наша математическая модель принимает вид:

Случай 1.

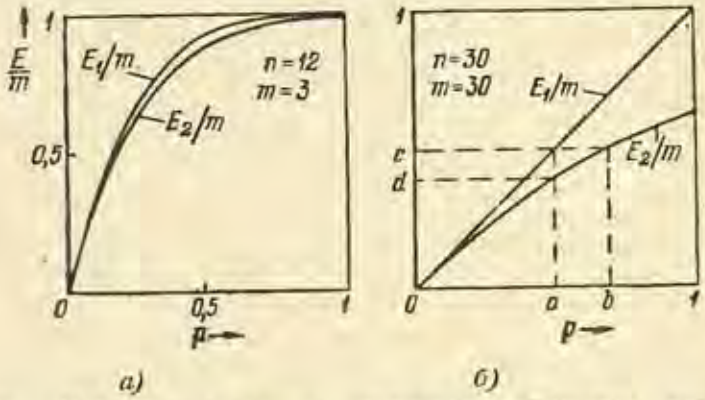
$$\frac{E_1}{m} = 1 - (1-p)^{n/m}. \qquad (10.5)$$

Случай 2.

$$\frac{E_2}{m} = 1 - \left(1 - \frac{p}{m}\right)^n. \qquad (10.6)$$

Если начертить графики выражений (10.5) и (10.6) как функций от p, получатся кривые такого вида, как на рис. 10.3. Можно показать, что преимущества, получаемые при централизованном управлении боевыми средствами, удивительно невелики, за исключением того частного случая, когда m приблизительно равно n, а p велико (случай, когда $n < m$, не рассматривается, однако интуиция подсказывает, что случайное и регулируемое распределения боевых средств приведут при этом, по существу, к одному и тому же результату).

В этом частном случае мы можем получить также ответы на вопросы, с которых начинали. Что лучше: применить аппаратуру управления, или затратить те же деньги на закупку большего количества существующего оружия, или затратить те же деньги на разработку лучшего оружия? Предположим, что мы имеем 30 боевых единиц, обладающих некоторым постоянным значением вероятности уничтожения цели (a на рис. 10.3,$б$). При



а) *б)*

Рис. 10.3. Аналитическая вероятностная модель для системы управления. Верхние кривые относятся к регулируемому нацеливанию: $\frac{E_1}{m} = 1 - (1-p)^{n/m}$.

Нижние кривые относятся к случайному нацеливанию: $\frac{E_2}{m} = 1 - (1 - p/m)^n$.

отражении рейда 30 целей мы можем повысить эффективность с d до c, применив аппаратуру управления. Если эти же деньги затратить не на аппаратуру управления, а на создание лучшего оружия, то мы смогли бы действовать с таким же успехом в том и только в том случае, если сумели бы увеличить вероятность уничтожения цели с a до b. Наконец, если бы мы решили израсходовать деньги на закупку Δn дополнительных боевых единиц существующего типа, то мы были бы должны, используя формулу (10.6), определить ожидаемое количество уничтоженных целей при применении $n+\Delta n$ боевых единиц и сравнить эту величину с c.

Теперь можно сделать количественную оценку различных возможных решений. Конечно, в реальном случае нам пришлось бы вычислять эти сравниваемые показатели для многих значений m и производить оценку ожидаемого количества уничтоженных целей применительно к распределению вероятностей для количества целей, участвующих в рейде. Однако даже в сложных случаях такие упрощенные модели оказываются весьма могущественными вне видимых сфер их применимости и оказывают неоценимую помощь при системном мышлении.

10.3. Модели «Монте Карло»

Во многих случаях вероятностная модель получается слишком сложной для аналитического решения. Если, например, имеется несколько параллельных очередей, завершаемых несколькими последовательными операциями, и притом каждая очередь со своим распределением времени занятия, то модель может оказаться безнадежно сложной. Одна из возможных альтернатив в таких случаях состоит во введении приближений, каков, например, отказ от учета существующих в действительности распределений времени занятия или исследование распределения только в той одной точке цепочки, которая представляется узким местом системы. Другая альтернатива заключается в применении метода «Монте Карло» (метода статистических испытаний), особенно полезного в тех дискретных случаях, которые более всего не поддаются аналитической трактовке. Метод «Монте Карло» мы лучше всего поясним с помощью двух примеров.

Модель уличного движения, разработанная в Мичиганском университете [84]. Модель имитирует контролируемый светофором перекресток двух улиц с двумя полосами движения в каждой улице. Полный цикл работы светофора занимает 60 *сек*, причем в направлении одной улицы зеленый свет горит 37 *сек*, желтый 3 *сек* и красный 20 *сек*; в направлении перпендикулярно расположенной улицы продолжительность горения огней светофора составляет соответственно 20, 3 и 37 *сек*. Все эти параметры можно при необходимости изменять.

Автомашины подходят к перекрестку с каждого из четырех направлений с интенсивностью 360 (или 720, или 1 080) автомашин в час. Моменты их подхода к перекрестку распределяются во времени в соответствии с законом Пуассона. Процент поворачивающих автомашин может быть различным (например, в одном случае 10% машин поворачивает налево, 10% поворачивает направо, а 80% движется через перекресток в прямом направлении). Скорость движения машин принимается постоянной и равной 11 футам в $^1/_4$ *сек* (30 миль в час), если не возникает каких-либо препятствий их движению. Исследованию подвергается влияние продолжительности рабочего цикла светофора, интенсивности движения в каждой полосе и процента поворачивающих машин на задержку автомашин при переезде через перекресток.

При имитации положение автомашин проверяется один раз за каждую $^1/_4$ *сек* реального времени и затем производится перемещение всех машин, которые по условиям движения могут двигаться. Каждая из четырех подходящих к перекрестку полос движения разделена на несколько «точек», отстоящих одна от другой на 11 футов; для поворачивающих машин введены дополнительные точки, расположенные с меньшими интервалами между собой. Автомашина, находящаяся на перекрестке, должна быть в одной из этих точек. Возможность движения каждой из автомашин, находящихся в той или иной полосе, оценивается в течение $^1/_4$ *сек*, и каждая из них перемещается на одну точку вперед, если две точки впереди нее оказываются свободными и еще две точки впереди не занимаются движущимися автомашинами.

Машины могут «генерироваться» следующим образом. Предположим, что принятая интенсивность движения в определенной полосе составляет 1 080 автомашин в час, или 0,075 автомашины в $^1/_4$ *сек*. Тогда количество автомашин, включающихся в эту полосу в течение каждой $^1/_4$ *сек*, определяется распределением Пуассона с математическим ожиданием, равным 0,075. Вероятности появления нуля, одной и двух машин составляют соответственно 0,9278, 0,0696 и 0,0026.

Теперь возьмем из таблицы случайных

чисел четыре цифры и образуем четырехзначное число. Если это число лежит в пределах от 0000 до 9277, то мы говорим, что ни одна машина не выехала на перекресток; если оно лежит в пределах от 9278 до 9973, мы говорим, что на перекресток выехала одна машина; и, наконец, если это число лежит в пределах от 9974 до 9999, мы говорим, что на перекресток выехали две автомашины. На практике, если требуется меньшая точность, можно использовать только двузначные случайные числа и исключить из рассмотрения вероятность выезда в течение четвертьсекундного интервала времени двух автомашин. Заметим попутно, что в реальной модели случайные числа выдаются цифровой вычислительной машиной и никогда не преобразуются из двоичной формы в десятичную.

Все созданные таким методом «автомашины» добавляются к *общему итогу*. При записи общего итога автомашины не фиксируются по их положению, а просто подсчитываются. Затем общий итог считывается, и если он будет содержать одну или более автомашину, причем две последние точки ряда свободны, а две следующие за ними точки не заняты движущимися автомашинами, то общий итог уменьшится на единицу и одна автомашина переместится в следующую точку.

Когда автомашина достигает первой точки (ближайшей к перекрестку), ее последующие действия определяются другой группой случайных чисел. Например, можно выбрать наугад одну десятичную цифру и исследовать ее: если эта цифра равна 0, автомашина поворачивает налево; если она лежит в пределах от 1 до 8, автомашина движется прямо; и если она равна 9, автомашина поворачивает направо.

В каждом из трех случаев исследуются точки, через которые автомашина должна пройти, а также другие соседние точки, чтобы убедиться в возможности соответствующего движения машины. Последующее движение автомашины, находящейся в первой точке, определяется также сигналом светофора. Если сигнал зеленый, автомашина движется (если ее путь свободен); если сигнал красный, автомашина останавливается; если сигнал желтый, то автомашина либо движется, либо останавливается в зависимости от числа четвертей секунд, в течение которых этот свет уже горит. Для случая поворота автомашины налево предусмотрена специальная серия правил, касающихся желтого света.

Имеется ряд дополнительных усовершенствований этой модели, которые мы здесь не рассматриваем, а также большое число усовершенствований, которые можно было внести в модель. Моделирование осуществлялось на вычислительной машине с масштабом времени приблизительно в одну треть реального.

Модель уличного движения, разработанная в Калифорнийском университете. Калифорнийским университетом, расположенным в г. Лос-Анжелосе, была разработана модель «Монте Карло» для уличного движения, в основном аналогичная описанной выше модели, но совершенно отличная от нее по исполнению [16]. Имитация движения, осуществляемая по методу Калифорнийского университета, требует применения большой специализированной цифровой вычислительной машины, но при этом можно имитировать поток движущихся автомашин через большую сеть перекрестков в масштабе времени, в несколько тысяч раз более быстром, чем реальный. Эта модель обладает также большой гибкостью, так как позволяет изменять пути между перекрестками любым желательным способом.

Предложенная специализированная вычислительная машина состоит из большого числа простых цифровых вычислительных элементов, соединенных в значительной части так же, как обычно соединяются аналоговые вычислительные элементы (гл. 18), так что они работают параллельно. Некоторые из этих элементов изображают перекрестки, другие — улицы между перекрестками, третьи — оборудование управления (светофоры, решающие элементы, управляющие поворотами машин, и т. д.), четвертые — появление новых автомашин на улицах, и пятые — измерительные устройства для оценки результатов моделирования.

Элементы, имитирующие перекрестки, представляют собой наиболее сложную часть модели. При имитации перекрестка двух улиц с двумя полосами движение в каждой (т. е. когда по каждому из направлений: южному, северному, западному и восточному — движение может осуществляться только в одну полосу) каждое направление изображается счетчиком и соответствующей электрической схемой. Автомашины имитируются импульсами. Автомашина, выезжающая на перекресток, увеличивает число в счетчике на единицу, а машина, покидающая перекресток, уменьшает это число на единицу. Выехать с перекрестка машина может только в том случае, если сигнал светофора зеленый (желтый свет не был включен в рассмотрение, но при необходимости его можно было бы учесть очевидными методами).

Выехав с перекрестка, автомашина попадает на элемент, который решает (в соответствии с заранее заданной вероятностью), поедет ли машина прямо или повернет. Если

окажется, что она должна двигаться прямо, то ее пропускают на следующую улицу; если же окажется, что она должна повернуть, то она подходит к следующему решающему элементу, который определяет, повернет ли машина направо или налево. Если машина поворачивает направо, то она подходит к счетчику, который может считать только до единицы. Пока в счетчике стоит эта единица от машины, все движение позади машины останавливается; машина едет дальше в том случае, если движению не препятствует поток пешеходов.

Аналогично осуществляется и левый поворот, с тем лишь исключением, что автомашина, ожидающая поворота налево, может быть задержана либо пешеходами, либо автомашинами, движущимися в противоположном направлении. Это означает, что от счетчика, изображающего автомашины, движущиеся, например, в северном направлении, должен поступать сигнал на вход счетчика левого поворота автомашин, движущихся в южном направлении.

С помощью дополнительных схем в модель можно ввести ряд усовершенствований. Например, можно разрешить правый поворот на красный свет. Это потребует дополнительного счетчика и «запрещающего» сигнала с выхода счетчика автомашин, движущихся в южном направлении, на счетчик правого поворота автомашин, движущихся в восточном направлении.

Имитацию улиц с четырехполосным движением можно осуществить введением отдельного счетчика для каждой полосы движения; тогда левый поворот возможен только из внутренней полосы, а правый — только из внешней. При этом необходимо одно дополнительное усложнение; так как движение некоторых автомашин во внутренней полосе будет тормозиться машинами, находящимися впереди них и ожидающими левого поворота, то такие автомашины должны иметь право перейти из внутренней полосы во внешнюю, при условии достаточных интервалов между машинами во внешней полосе.

Каждая улица в этой модели изображается задержкой требуемой длительности, отвечающей длине улицы. Должна существовать также определенная связь между улицей и ее выходным перекрестком, чтобы можно было предпринимать надлежащие действия в случаях, когда на улице сосредоточивается такое число автомашин, что вся улица оказывается занятой автомашинами вплоть до следующего перекрестка. При упрощенной имитации эти «надлежащие действия» могли бы заключать-

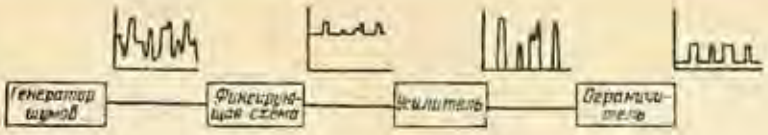


Рис. 10.4. Генератор случайных импульсов и выходные сигналы.

ся во включении светового сигнала и констатации, что скопление автомашин стало нетерпимым.

Роль светофоров выполняют просто датчики времени, которые приостанавливают на определенные периоды времени потоки автомашин, движущихся в направлениях север — юг или восток — запад. Соответствующим соединением этих реле времени можно имитировать любую последовательность действия светофоров. Решающие устройства, определяющие повороты машины, помехи со стороны пешеходов и т. п., показаны на рис. 10.4. Выходной сигнал такого устройства представляет собой цуг прямоугольных импульсов с фиксированной амплитудой, случайной длительностью и случайными паузами, но с известной вероятностью наличия импульса в любой данный момент времени, определяемой уровнем фиксации сигнала в фиксирующей схеме.

Теоретически для каждого решения по каждой полосе на каждом перекрестке требуется свое независимое решающее устройство. На практике можно думать, что на несколько перекрестков можно применять одно решающее устройство при условии, что эти перекрестки расположены не слишком близко друг к другу. Входом в систему служат надлежащие генераторы импульсов на каждой улице, ведущей в уличную сеть. В качестве таких генераторов можно использовать генераторы случайных импульсов, использующие генераторы шумов, или таблицы случайных чисел, или генераторы импульсов с постоянной частотой повторения, или даже ленты, на которых записаны действительные данные об уличном движении. Выходами системы служат просто провода, по которым «автомашины» (импульсы) выводятся из системы.

Характер измерительной аппаратуры определяется в значительной мере выбранными для системы критериями эффективности. Можно, например, постепенно увеличивать частоту повторения импульсов, до тех пор пока не создастся затор (все изменения уличного движения на протяжении многих часов можно проимитировать всего за несколько секунд, так как каждый период повторения импульсов, занимающий лишь несколько микросекунд, может изображать около секунды

реального времени; поэтому можно очень быстро достичь стационарного, установившегося режима).

Или же можно измерять скорость потока движения через всю уличную сеть или через какой-либо перекресток, или же можно измерять общее время проезда через уличную сеть. Любое из этих измерений требует знания числа автомашин в счетчиках, или частоты повторения выходных импульсов, или и того и другого вместе. Следует, однако, заметить, что движение одиночной автомашины через уличную сеть проследить нельзя. Это не только не предусмотрено, но такая имитация и не имела бы особого смысла. Одиночный импульс, если бы за ним и можно было следить, стал бы, весьма возможно, делать бесконечные бесцельные повороты и вышел бы в конце концов из сети где-нибудь вблизи точки своего запуска. Однако пока модель «Монте Карло» дает правильный процент машин, поворачивающих в каждой точке, нам безразлично, какие конкретно автомашины поворачивают; мы все равно получим правильные результаты.

Необходимость эксперимента. Все эти модели имеют небольшую ценность без численных значений, а численные значения обычно приходится определять из опыта. Например, процент автомашин, делающих повороты, процент времени, в течение которого пешеходы блокируют правый поворот, и фактически самое наличие некоторых явлений, таких, как смена полосы едущей машиной, должны вводиться в модель на основе наблюдения реального мира. Аналогичные замечания можно сделать и относительно каждой из рассмотренных нами моделей. Поэтому в следующей главе мы обратим наше внимание на вопросы сбора информации.

ГЛАВА 11

ПЛАНИРОВАНИЕ ЭКСПЕРИМЕНТОВ. СБОР ДАННЫХ

При проектировании системы данные собираются в двух целях: во-первых, для подстановки в общую схему математической модели, т. е. для конкретизации математического описания системы, и, во-вторых, для оценки работы системы в соответствии с выбранным критерием эффективности, чтобы определить увеличение эффективности в результате предлагаемых или уже осуществленных изменений.

В обоих случаях нужно до начала измерения хорошо изучить и понять систему, чтобы осуществить сбор именно тех данных, которые необходимы. Но отсюда следует, что еще до сбора данных должна быть создана подробная математическая модель системы, а это нельзя осуществить до тех пор, пока мы не будем иметь в своем распоряжении данных о системе. Круг такого рода обычен при проектировании систем, но он не является столь порочным, как это может показаться с первого взгляда. Все этапы проектирования системы представляют собой непрерывные действия, требующие непрерывного взаимодействия между собой и постоянных переходов от общего к деталям и от деталей к общему; все это приводит к непрерывному улучшению понимания и описания системы.

11.1. Другие источники, помимо измерения

Сбор данных обычно понимается как сбор численных данных. Действительно, сбор численных данных—наиболее важная часть работы по сбору данных вообще. Однако проектировщик системы должен будет собрать также некоторые сведения, которые нелегко выразить в числовой форме, например данные о природе входов и выходов, о различных функциях и внутренних связях системы. Наиболее очевидным и наиболее важным источником данных служит измерение; однако существуют еще три других источника, о которых также не следует забывать: документы, беседы и личное участие.

Под документами мы понимаем руководства, наставления, производственные бюллетени, корреспонденцию и т. п. Для большинства систем большого масштаба ощущается острый недостаток нужной документации, хотя может иметься масса побочного материала для просмотра. Этот недостаток может вызываться тем, что информация является чьей-то собственностью, или же тем, что никто еще не пытался сделать необходимые описания на бумаге.

Беседы полезны по многим соображениям, особенно если нужных документов мало. Даже если собрана обильная документация, есть вещи, которые можно узнать только устно и которые никогда не фиксировались на бумаге. Однако при этом всегда следует иметь в виду следующее предупреждение Морза и Кимбелла [11]:

«Всегда следует иметь в виду необходимость непредвзятого и неограниченного знания *фактов*, а не мнений или суждений... Часто задают вопрос: «За-

чем непременно нужно быть свидетелем операции (или получить подробное сообщение об операции), если такой-то может вам все рассказать о ней?». Весь опыт научных исследований за последние три столетия говорит о том, что подобной точки зрения следует избегать, если нужно получить результаты, имеющие научную ценность».

Данные, полученные из документов и бесед, должны считаться лишь предварительными и подлежат дальнейшей проверке через опыт и наблюдение: они могут оказаться весьма устаревшими, особенно данные в документах, и, кроме того, оба источника могут отражать теоретические взгляды, существенно отличающиеся от практики. Однако, если они указывают на такие противоречия, их следует считать достойными внимания, и их изучение может выявить много скрытых фактов или интересных точек зрения.

Личное участие является наилучшим способом «прочувствовать» систему. Морз и Кимбелл говорят, что исследователь операций «должен летать на бомбардировщике, путешествовать в автобусе, работать на радиолокационной станции дальнего обнаружения воздушных целей или делать покупки в магазине, смотря по обстоятельствам». Мы подчеркнули бы в связи с этим, что инженер-системотехник должен как можно больше лично знакомиться с разными сторонами разрабатываемой системы и что эта работа не всегда будет приятна. Наш инженер, конечно, захочет совершить полет в кабине пилота или побыть в пункте управления, но ему следует также помнить о том, что он может получить очень много полезных для него сведений, работая в составе наземной ремонтной бригады.

Сведения (и контакты), полученные в процессе такого личного участия, будут полезны не только тем, что вооружат проектировщика данными; они во многом помогут ему также при организации рабочих испытаний как для дальнейшего сбора данных, так и для оценки нового оборудования и методов его применения. Во время таких личных ознакомлений проектировщик системы должен сохранять широкий подход к задаче. Оператор, за которым он пристально наблюдает каждое мгновение, имеет свою узкую, пристрастную точку зрения и может считать, что его работа является основной, критической для всей системы. Проектировщик системы может и должен помнить, в какой мере работа оператора действительно влияет на работу всей системы в целом.

11.2. Измерение и эксперимент

Основным источником данных служит измерение, так как получаемые при этом данные являются количественными и — при должном внимании — объективными. Измерение предполагает эксперимент, т. е. опыт, причем мы различаем два рода экспериментов, важных для проектировщика систем. Наиболее типичной формой экспериментов первого рода являются *лабораторные эксперименты* ученого-физика; назначением этих опытов при проектировании системы является получение чисел для подстановки в математическую модель. Наиболее типичной формой экспериментов второго рода являются *рабочие испытания*, назначение которых при проектировании системы заключается в получении чисел, характеризующих качество работы системы. Слова «наиболее типичная форма» здесь выражают то обстоятельство, что между этими двумя родами экспериментов существует переходная область, в которой эти типы уже нельзя четко отличить друг от друга.

При лабораторных экспериментах — в идеальном случае — нами контролируются все независимые переменные. На практике это значит, что делается попытка перечислить все существенные (действительно влияющие на дело) переменные и осуществить контроль над ними. При рабочих испытаниях контролировать все существенные независимые переменные (если даже они действительно известны) не представляется возможным и, что еще важнее, такой контроль над ними вообще нежелателен.

Чтобы сделать более ясным различие между экспериментами этих двух родов, вернемся еще раз к системе ближней телевизионно-радиолокационной навигации (система «Телеран», § 2.1). На самолете установлен телевизионный приемник, и нам необходимо убедиться, что он получает сигнал надлежащего уровня. Например, для снижения аэродинамической нагрузки на самолет мы могли бы установить антенну таким образом, чтобы она не выступала за внешние контуры корпуса самолета; если, однако, интенсивность сигнала будет недостаточна при определенных удалениях и ориентациях самолета, то потребуется установка внешней антенны или увеличение габаритов внутренней антенны.

Для определения необходимых параметров мы проводим лабораторные эксперименты, исследуя эффективный коэффициент усиления антенны при всех возможных ориентировках последней и при контролируемой переменной интенсивности сигнала (под «контролируемой» мы подразумеваем переменную, известную нам и фиксированную на определенном значении по крайней мере в течение одной серии экспериментов и затем изменяющуюся, если

только она вообще изменяется, к новому, но заранее выбранному значению, которое сохраняется неизменным в течение следующей серии экспериментов). Все другие переменные, которые предположительно являются существенными, также должны контролироваться. Аналогичные контролируемые эксперименты необходимы для многих других частей системы.

Однако в конце концов должен наступить момент, когда мы спросим: работает «Телеран» или нет? Тогда мы берем самолет или группу самолетов, оборудованных аппаратурой этой системы, и даем им указание осуществить посадку, используя эту аппаратуру. Так как теперь нам приходится иметь дело со многими неизвестными, то возникает большая дисперсия и мы вынуждены повторять опыт много раз. Проведя эти испытания многократно, мы надеемся получить случайную выборку из действительных условий, но мы не пытаемся контролировать эти условия. Наша задача заключается в определении набора чисел, характеризующих работу системы; эти числа будут подвержены статистическому изменению, и нас в первую очередь интересуют их средние значения, хотя нам будет нужно знать и их распределения.

К параметрам этого рода могут относиться ожидаемые отклонения самолета от выбранного маршрута полета, частота грубых ошибок и скорость, с которой самолет может заходить на посадку. Если бы даже и можно было фиксировать температуру, освещение, ветер и т. п., мы не захотели бы сделать это (хотя мы захотели бы провести по крайней мере некоторые из наших испытаний при предельных условиях). Для получения чисел, которыми можно было бы пользоваться на практике, необходимо позволить событиям идти неконтролируемым путем в надежде получить тем самым представительную (репрезентативную) выборку из условий работы.

Итак, мы проводим различие между экспериментами для подстановки чисел в модель, когда независимые переменные контролируются, и экспериментами для оценки работы системы, когда независимые переменные не контролируются. Мы уже отмечали, что эксперименты первого рода обычно являются лабораторными и малого масштаба, а эксперименты второго рода — рабочими и большого масштаба, однако это справедливо не всегда. Иногда эксперименты первого рода приобретают характер рабочих испытаний.

Например, после того как рассмотренная выше антенна будет спроектирована, нам может потребоваться испытать ее на самолете в реальных условиях полета. Эксперимент еще остается контролируемым. Мы будем делать измерения при различных ориентациях антенны (относительно линии визирования передатчика) и при фиксированной дальности и неизменных условиях погоды, затем мы захотим сделать измерения при других заданных дальностях и других условиях погоды, и т. д.

Случаи, когда группа проектирования системы приходит к выводу, что существующие научные теории не дают ответов на интересующие ее вопросы и что необходимо проведение опытов лабораторного типа, которые на первый взгляд могут показаться чисто академическими, возникают довольно часто. Однако на вопросах планирования таких лабораторных испытаний мы останавливаться не будем, так как они достаточно хорошо рассмотрены во многих других трудах [37, 46]. Мы обратим внимание читателя лишь на несколько положений, специально касающихся рабочих испытаний (имея, конечно, в виду, что некоторые из этих положений применимы к экспериментам обоих типов).

11.3. Замечания о планировании рабочих испытаний

Почти всегда рабочие эксперименты и испытания над системами большого масштаба обходятся дороже и занимают больше времени, чем предполагается первоначально; почти всегда по окончании работ приходят к выводу, что до начала этих испытаний можно было бы затратить больше усилий на их подготовку и лучше спланировать их проведение; почти всегда недооценивается проблема обработки полученных данных. Если дюжина кинокамер фиксирует состояние стольких же предметов (циферблатов, шкал и т. д.) с обычной скоростью 24 кадра в секунду, то за один час накапливается миллион кадров.

Если обработка данных требует просмотра киноленты вручную кадр за кадром, как это часто бывает, то большая часть заснятой киноленты будет, по-видимому, обречена на мертвое лежание на полке, без какой-либо возможности использования.

Введение контролируемых переменных в рабочие испытания было рассмотрено нами в предыдущем параграфе. Когда размеры вводимого контроля могут выбираться по желанию, следует иметь в виду, что контроль над большим числом переменных делает каждое наблюдение более точным в том смысле, что за одно наблюдение можно достичь большей точности в определении искомой функциональной зависимости (см. гл. 12). Но точность

можно увеличить также увеличением числа наблюдений, и потому выбор наилучшего способа планирования эксперимента может потребовать обширного анализа.

Прежде чем начинать какие-либо широкие рабочие испытания, проектировщик системы стремится выявить как можно больше существенных независимых переменных и создать себе хотя бы общее представление о порядке величин, которые он собирается измерять.

Привести простой, но убедительный пример неожиданно выявляющейся существенной переменной довольно затруднительно, так как в большинстве случаев на первый взгляд кажется, что все переменные можно заранее предвидеть. Такие неожиданные существенные переменные лучше всего выявляются в процессе рабочих испытаний, если брать возможно более разнообразные условия, какие только есть. Путем тщательного наблюдения рабочих испытаний и соответствующего документирования их в ряде случаев можно выявить некоторые из таких неожиданных существенных переменных. Например, может оказаться, что какое-либо из других радиоэлектронных устройств самолета создает помехи испытываемому телевизионному приемнику системы «Телеран» в то время, когда самолет находится в полете, хотя при проведении наземных испытаний не было никакого указания на это. Рабочие испытания обычно обнаруживают такие «дефекты» системы, и это является одной из причин, почему рабочие испытания должны быть максимально реалистическими, максимально приближены к реальным условиям.

Когда существенные переменные не контролируются, их значения по возможности будут измеряться; однако это не обеспечивает наиболее благоприятного выбора точек на исследуемой кривой, и потому результаты будут менее эффективны, чем в случае оптимального множества точек. Более того, даже тогда, когда переменные контролируются, мы, вообще говоря, не сможем держать все, кроме одной, независимые переменные фиксированными все то время, пока производим измерения по всему диапазону данной одной переменной.

Ввиду этого анализ полученных результатов окажется менее простым (§ 12.8) и, если зависимости сложны, менее точным. Наконец, будут еще некоторые существенные переменные, о которых мы не будем знать или которые мы не сможем измерить; их влияние должны учитываться в наблюдаемой дисперсии суммарно (в предположении, что они не являются систематическими; см. ниже).

Необходимость иметь хотя бы грубое представление о порядке измеряемой величины совершенно очевидна. Если экспериментатор отправляется с секундомером на измерение продолжительности какого-нибудь события, а это событие в действительности займет 30 *мсек*, то он не получит никаких полезных результатов. Морз и Кимбелл [11] подчеркивают важность того, что они называют «мышлением в гемибелах» *: «При предварительном анализе операций обычно бывает достаточным определять значение характеристики с точностью до множителя 3». Обычно это можно сделать при беглом наблюдении. Однако на пути к основному эксперименту можно сделать промежуточный шаг и провести эксперимент с точностью до «одной значащей цифры». Такой эксперимент осуществим сравнительно быстро и дешево и оказывает большую помощь при планировании основного эксперимента не только тем, что намечает диапазоны изменения переменных, но и тем, что освещает некоторые (хотя отнюдь не все) проблемы, с которыми придется встретиться.

Однако следует помнить, что это лишь предварительный эксперимент, который дол-

---

* Как известно, во многих случаях оказывается весьма удобным измерять или считать величины не в обычной, линейной шкале, а в логарифмической; в частности, логарифмическая шкала нередко значительно облегчает вычерчивание графиков. Для счета по логарифмической шкале применяются особые единицы, и в частности бел и его подразделения.

Бел — единица логарифмической шкалы, соответствующая множителю 10. Гемибел — единица логарифмической шкалы, равная половине бела и потому соответствующая множителю $\sqrt{10}$, т. е. приближенно 3. Таким образом, множитель 1 — это логарифмически 0 гемибел; множитель $1\sqrt{10}\approx 3$ — это логарифмически 1 гемибел; множитель 10 — это логарифмически 2 гемибела; множитель $\sqrt{10^3}\approx 30$ — это логарифмически 3 гемибела; множитель 100 — это логарифмически 4 гемибела; и т. д. Морз и Кимбелл [11] указывают, что использование гемибела в качестве единицы счета чрезвычайно удобно в ряде областей науки и что многие американские инженеры и ученые очень широко ее применяют.

Приставка «геми» (ἡμι) по-гречески означает «полу..», «наполовину». В русском переводе книги Морза и Кимбелла применяется неудачное написание «хемибел», применительно к английскому произношению этого слова (hemibel). Мы полагаем, что греческие терминологические приставки лучше транскрибировать согласно давно сложившимся в русской литературе традициям передачи греческих терминов, и потому приняли в настоящей книге написание «гемибел».

Другое широко употребляемое подразделение бела — это децибел (*дб*), одна десятая бела, соответствующая множителю $\sqrt[10]{10}$. Гемибел равен 5 *дб*. — *Прим. ред.*

жен быть быстро закончен. К несчастью, существует тенденция расширять объем и цель таких экспериментов, вводя импровизации и утончения в надежде получить точные данные без больших затрат. В результате таких попыток часто получают неточные данные и притом дорогой ценой.

Факториальное планирование. Когда приходится иметь дело с большим количеством независимых переменных, как это бывает при рабочих испытаниях, большую помощь может оказать методика планирования, разработанная первоначально для сельского хозяйства и получившая название *факториального* (или *многофакторного*) *планирования.*

Предположим, что мы проводим испытания системы посадки самолетов по приборам и сравниваем между собой различные способы, с помощью которых летчик может определять положение самолета относительно заданной траектории полета (например, визуальный, звуковой и комбинированный визуально-звуковой). Измеряемыми (зависимыми) переменными будут горизонтальное и вертикальное отклонение самолета от заданной траектории полета. Существенными могут быть очень много независимых переменных (в дополнение к способам индикации), например: тип самолета, направление и сила ветра, наличие или отсутствие предметов, затрудняющих заход самолета на посадку и вынуждающих летчика выводить самолет по сложной кривой. Было бы желательно провести по одной серии экспериментов для каждой возможной комбинации условий; соответствующее планирование таких повторяющихся экспериментов называется *факториальным планированием,* а математические методы их анализа получили наименование *дисперсионного анализа* (§ 12.8).

Такие серии экспериментов имеют то дополнительное преимущество, что позволяют нам исследовать эффекты взаимодействий различных порядков. Как пример взаимодействия второго порядка укажем, что звуковые сигналы могут оказаться более эффективными в тяжелых самолетах при сильном ветре, а в легких самолетах — в безветренную погоду, хотя в среднем сколько-нибудь заметная разница отсутствует.

В рамках приемлемых расходов обычно нельзя бывает провести полные серии таких экспериментов во всех возможных комбинациях. Однако наиболее важные цели факториального планирования еще могут достигаться методами, известными под названием *частичного воспроизведения.* При этих методах определенные комбинации экспериментов пропускаются, исключаются из испытаний с таким расчетом, что мы теряем лишь немного информации о непосредственных функциональных зависимостях, но совсем не получаем информации о взаимодействиях очень высоких порядков. Это особенно полезно в тех случаях, когда предполагается, что некоторые переменные несущественны.

Например, в связи с тем, что в рассмотренном нами примере посадка самолетов осуществляется по приборам, мы можем предположить, что точность приземления не зависит от времени суток; однако нам потребуется проверить это предположение экспериментально, включив время суток (день, ночь) в число переменных при факториальном планировании. Так как эта переменная, по-видимому, не является существенной, можно вполне здраво предположить, что ее взаимодействиями высшего порядка с другими переменными можно пренебречь. По факториальному планированию и частичному воспроизведению имеется большая литература [47].

Классификация ошибок. Главные цели планирования эксперимента заключаются в достижении максимальной точности и минимальной стоимости эксперимента. Так как увеличение числа наблюдений приводит, с одной стороны, к уменьшению дисперсии (т. е. $\sigma^2_m = \sigma^2/n$), а с другой стороны, увеличивает стоимость испытаний, то стоимость и точность являются до некоторой степени взаимозаменимыми величинами. В предыдущем разделе мы говорили о снижении стоимости путем тщательного выбора экспериментов. Если в наших рассуждениях, мы будем считать стоимость постоянной величиной, то основной задачей станет уменьшение ошибок в наших результатах.

Ошибки обычно делятся на два класса: случайные и систематические. Любое наблюдение можно рассматривать как состоящее из двух частей: ожидаемого значения наблюдения и отклонения от этого значения, причем это отклонение обладает распределением в смысле теории вероятности. Если ожидаемое значение наблюдения отличается от значения оцениваемой нами переменной (например, если для взвешивания применяются весы, показывающие всегда на 2 фунта больше), то присутствует систематическая ошибка, или *смещение.* Независимо от наличия или отсутствия систематической ошибки, отклонения от ожидаемого значения наблюдений образуют случайную ошибку.

11.4. Случайные ошибки

Вопросам определения случайных ошибок посвящена почти вся гл. 12. В настоящем же

параграфе отметим некоторые характеристики случайных ошибок, которые могут быть приняты во внимание уже в самом начале планирования экспериментов. Случайные ошибки удобны для обработки: мы имеем представление об их величине и тем самым можем установить для них доверительные пределы (§ 12.5) или уменьшить эти ошибки путем увеличения размеров выборки.

Рассматривая относительные влияния независимых случайных ошибок, мы можем определить относительные усилия и стоимости, обусловливаемые различными частями эксперимента. Из равенства (7.3) мы знаем, что

$$\sigma_0^2 = \sum c_i^2 \sigma_i^2, \qquad (11.1)$$

где σ^2_0 — дисперсия суммарных ошибок, σ_i^2 относится к составляющим ошибкам. Например, мы можем пытаться измерить направление полета самолета с помощью наземной радиолокационной станции. Тогда σ_1 может относиться к ошибкам измерения положения антенны, $\sigma_1{}^2$ — к ошибкам вследствие рефракции радиолокационного луча в атмосфере, σ_3 — к ошибкам вследствие мерцания (флюктуации) отраженного сигнала и т. д.

Формула (11.1) в пределах своей применимости приводит к важному выводу. Предположим, что ошибка Δx_0 есть сумма двух ошибок Δx_1 и Δx_2, что все другие ошибки пренебрежимо малы и что σ_1 в четыре раза больше, чем σ_2. Тогда

$$\sigma_0^2 = \sigma_1^2 + (0{,}25\sigma_1)^2 = 1{,}06\sigma_1^2,$$

$$\sigma_0 = 1{,}03\sigma_1. \qquad (11.2)$$

Другими словами, ошибкой, равной одной четверти наибольшей ошибки, можно пренебречь — при условии, что таких ошибок не слишком много; ясно, что мы должны посвятить основные усилия тем источникам ошибок, которые значительно влияют на σ_0. Исключения могут составить лишь случаи, когда небольшие ошибки можно легко уменьшать.

В качестве примера такой ситуации рассмотрим ошибки округления, возникающие при вычислениях. Проводя вычисление с одним добавочным десятичным знаком, мы уменьшаем стандартное отклонение ошибки округления в 10 раз и согласно формуле, аналогичной (11.2), уменьшаем влияние этой ошибки на σ_0 в 200 раз. Это стоит сделать, но добавление еще одного дальнейшего знака было бы нецелесообразно, если даже дополнительный труд был бы весьма невелик. Отсюда вытекает хорошо известное эмпирическое правило, что вычисление должно проводиться обычно с добавочной значащей цифрой по сравнению с наблюденной величиной. Количество значащих цифр должно, конечно, определяться в конце вычисления; если вычисление достаточно долгое и содержит кумулятивные (накапливающиеся) ошибки округления, распространяющиеся на несколько разрядов сверх наименьшей значащей цифры, то с начала вычисления необходимо использовать добавочные разряды.

Чтобы формула (11.1) имела силу, ошибки должны быть случайными и независимыми. Например, ошибки калибровки являются систематическими, и формула (11.1) к ним неприменима. Кроме того, можно предполагать, что при очень низких углах места антенна будет «клевать», и с полной уверенностью можно утверждать, что при низких углах места рефракционные ошибки будут возрастать. В этом случае σ_1 и σ_2 не будут независимыми и формулу (11.1) опять нельзя применять.

В рассмотренном выше примере все ошибки были аддитивными. Однако может оказаться, что мы измеряем несколько параметров в сложной функции и хотим знать, в какой мере ошибка каждого измерения будет влиять на конечный результат. Точный анализ в таких случаях может оказаться невероятно затруднительным, однако существует одно полезное приближение. Для любой функции двух или более переменных, скажем $f(x, y)$, полный дифференциал дается формулой

$$df = \frac{\partial f}{\partial x}\,dx + \frac{\partial f}{\partial y}\,dy. \qquad (11.3)$$

Подставляя вместо бесконечно малых приращений их конечные эквиваленты, получаем

$$\Delta f \approx \frac{\partial f}{\partial x}\,\Delta x + \frac{\partial f}{\partial y}\,\Delta y, \qquad (11.4)$$

что является хорошим приближением, если величины приращений малы и/или функция является приближенно линейной в точке, для которой определяются производные. Обычно эти предположения выполняются достаточно хорошо, когда этими Δ служат ошибки, и формула (11.4) тогда справедлива безотносительно к независимости ошибок.

Если, кроме того, мы можем предположить, что ошибки независимы, то формулы (11.1) и (11.4) можно объединить, приняв частные производные в рассматриваемой точке за коэффициенты c_i, в результате чего получим выражение

$$\sigma_f^2 = \left(\frac{\partial f}{\partial x}\right)^2 \sigma_x^2 + \left(\frac{\partial f}{\partial y}\right)^2 \sigma_y^2. \qquad (11.5)$$

Так как составляющие ошибки — величины Δx и Δy из (11.4) — обычно имеют нор-

мальное распределение с математическим ожиданием, равным нулю, приближенно можно считать, что Δf также имеет нормальное распределение, причем параметры этого распределения определяются выражением (7.22), т. е. математическое ожидание распределения равно нулю, а дисперсия определяется выражением (11.5).

11.5. Систематические ошибки

Систематические ошибки трудно выявлять. Большая трудность при нахождении этих ошибок состоит, конечно, в том, что мы не знаем об их наличии. Здесь-то и становятся особенно важными изобретательность, опыт, терпеливость и другие качества, отличающие талантливого экспериментатора.

Полезно поразмыслить об одном случае, который произошел со студентами, проходившими курс электрических измерений. Преподаватель имел несколько «метровых стержней» с нанесенными на них обычным способом «сантиметровыми» и «миллиметровыми» делениями; однако фактически длина каждого стержня составляла только 97 *см*. Студенты очень тщательно провели серию экспериментов, включавшую ровно одно линейное измерение. Сделав затем статистический анализ полученных результатов, они определили, что вероятная ошибка* составляет около 0,1%. Более того, «независимые» проверки различных студентов дали тот же порядок величины.

Конечно, результаты были отягощены совпадающими 3-процентными систематическими ошибками, обусловленными дефектом измерительного устройства. Сходство с реальными экспериментами здесь беспокоящее близкое. Единственная мораль отсюда — относиться ко всем калибровкам с подозрением.

Существует несколько методов исключения или уменьшения систематических ошибок. Один из них заключается в проверке репрезентативности полученной выборки; другой — в устранении наблюдательского смещения; третий — в замене, где только можно, абсолютных измерений относительными; четвертый — в проведении эксперимента в условиях, изменяющихся максимально широко, т. е. при максимально возможном изменении тех факторов, которые считаются не влияющими на ход эксперимента; пятый — в исключении влияния неизвестных переменных с помощью «контрольных» экспериментов.

Репрезентативная выборка. Важность получения репрезентативной (представительной) выборки совершенно очевидна; возникающие трудности делаются ясными из трех примеров, в которых были взяты нерепрезентативные выборки.

Наиболее известный из них касается пробного подсчета голосов на президентских выборах в 1936 г. журналом «Литерари Дайджест» (Literary Digest). Этот подсчет предсказывал с солидным запасом победу Лендона, в то время как в действительности победу с подавляющим числом голосов одержал Рузвельт. Причиной такой ошибки была нерепрезентативная выборка: фамилии избирателей были взяты из телефонных книг, списка подписчиков «Дайджеста» и аналогичных источников. Выбранные таким образом люди были по своему благосостоянию значительно выше среднего уровня, а в 1936 г. имелась значительная корреляция между благосостоянием и стремлением голосовать за республиканцев.

Из этого, однако, не следует, что все выборки должны браться обязательно из всего населения целиком; это также могло бы привести к ошибочным выводам, так как не все население голосует и те, кто не участвуют в голосовании, могли бы сделать выборку смещенной. Если бы кто-нибудь захотел произвести выборочный опрос населения в связи с системой «Резервизор», то он должен был бы сделать акцент на более состоятельных слоях населения (в правильной пропорции), так как более состоятельные слои населения пользуются воздушным транспортом чаще, чем бедные слои. Важно обеспечить репрезентативность выборки именно для той группы, которая является собственно предметом изучения.

Второй пример касается одного агрономического опыта, цель которого заключалась в выборочном определении урожая зерновых с некоторого поля. Чтобы быть уверенными в случайности выборки, экспериментатор забрасывал в поле кольцо и собирал стебли, которые оказывались в этом кольце. Конечно, результат выборки был смещен в сторону высоких стеблей.

Третий пример касается переписи населения. Было признано желательным обрабатывать некоторые данные переписи только для некоторой выборки из населения, и наиболее удобная из предложенных выборок состояла из первых фамилий на каждой странице отчетов, представленных теми, кто проводил пе-

* Вероятная ошибка равна 0,674 стандартного отклонения серии результатов, т. е. дает 50-процентный доверительный интервал для регистрируемого математического ожидания, если предполагается, что случайные ошибки подчиняются нормальному закону и систематическая ошибка равна нулю (см. гл. 12). — *Прим. авт.*

репись. К счастью, служащие статистики Бюро переписей обнаружили, что переписчики начинали обычно страницу записью новой семьи, причем глава семьи при этом всегда записывался первым.

Наблюдательское смещение. Влияние наблюдателя при проведении опыта является тонким фактором, обесценившим многие эксперименты. Его нельзя исключить усилием воли. Вильсон [46] говорит: «Никакой человек не свободен даже приближенно от таких субъективных влияний; честный и осведомленный исследователь организует эксперимент так, чтобы его собственные предубеждения не могли повлиять на результат. Только наивный или нечестный человек утверждает, что его объективность является достаточной гарантией».

Один пример уже упоминался в гл. 10: радиолокационная станция сопровождает цель, координаты которой известны, мы должны принять меры, гарантирующие, что оператор не примет мнимого сигнала за истинный. Единственным вполне удовлетворительным решением явилась бы замена оператора автоматическим устройством, но и при отсутствии последнего мы будем в состоянии уменьшить ошибку, если отдаем себе отчет в возникающих трудностях. Одно из преимуществ более сложного эксперимента (при котором позиция цели не известна оператору) заключается в том, что такой эксперимент уже не допускает этого возможного источника ошибок.

В общем случае ошибки такого рода можно исключить только при условии, что человек не знает совершенно ничего о всех относящихся к эксперименту факторах в то время, когда он принимает решение, требующее суждения. Считывание показаний измерительного прибора не требует суждения, но решения о том, что нечто «удовлетворительно», «видно», «значимо» и т. п., подвержены сильному наблюдательскому смещению. Это означает, например, что для определения значимости наблюдаемого события мы до начала сбора данных должны выбрать определенный уровень значимости (α и β, § 12.2).

Один из методов исключения наблюдательского смещения в рассматриваемом опыте радиолокационного наблюдения цели мог бы состоять в замене индикатора кругового обзора измерительным прибором со стробированием сигнала по дальности. Тогда оператор просто бы считывал интенсивность сигнала от известного положения цели. Если при этом ввести соответствующую поправку на шумы, то такой метод давал бы количественную оценку наличия или отсутствия обнаружимого сигнала. Указанный метод обладал бы и другим преимуществом; можно было бы заменить шкалу с двумя градациями: «наличием» и «отсутствием» обнаружимого сигнала (где слову «обнаружимый» трудно дать определение) — шкалой с несколькими градациями. Если даже получаемая при этом точность окажется очень низкой, так что мы сможем эффективно использовать только три или четыре градации, дисперсия получаемых данных будет значительно меньше, чем при применении шкалы с двумя градациями. В общем, при записи наблюдений всегда хорошо определить количественно, если можно, туманную область между «да» и «нет».

Относительные измерения. Проведение относительных измерений вместо абсолютных иногда означает немногим больше, чем аккуратная калибровка. В этих случаях калибровка должна производиться по известному стандарту, как можно более подобному измеряемому объекту.

Рассмотрим экспериментальное определение эффективной отражающей поверхности цели. При проведении относительных измерений нам требуется только измерить амплитуду принимаемого сигнала, отраженного от интересующей нас цели, и амплитуду сигнала, отраженного от какой-либо цели с известной отражающей поверхностью (например, от сферы), когда обе цели имеют те же самые координаты (и приблизительно одинаковую эффективную отражающую поверхность). При измерении абсолютных значений мы должны знать мощность передатчика, усиление антенны, ослабление сигнала в фидерной системе и множество других факторов. Ошибки определения этих последних факторов, по-видимому, будут значительно больше, чем ошибка, которая могла бы быть получена при относительных (калиброванных) измерениях.

В данном частном случае абсолютное значение эффективной отражающей поверхности сферы известно из теории электромагнитного поля; если бы это было не так, все же нам было бы полезно сделать относительные измерения эффективной отражающей поверхности целей с различной интересующей нас конфигурацией, так как эти относительные (сравнительные) измерения могут быть очень точными. В дополнение к этому мы могли бы пожелать измерить и записать абсолютные значения, понимая при этом, что они имели бы, по-видимому, значительно бо́льшую ошибку.

Изменение условий. Призыв изменять условия проведения эксперимента как будто находится в противоречии с основными положениями «научного метода», поскольку считается, что этот метод допускает изменение только

двух величин, а именно тех зависимой и независимой переменных, функциональная зависимость которых исследуется. К несчастью, однако, никогда не удается держать все другие переменные постоянными.

Чтобы взять простейший возможный пример, предположим, что мы измеряем два значения зависимой переменной при двух различных значениях независимой переменной. Это требует двух опытов, и они не могут быть совершены точно в одном и том же месте и в одно и то же время — мы можем провести их последовательно в одном и том же месте или одновременно в различных местах. В обоих случаях мы прилагаем все усилия, чтобы эти два эксперимента были тождественными, но не всегда наши действия будут удачны.

Так как мы всегда должны повторять наши эксперименты (с целью снижения случайной ошибки и устранения возможности грубой ошибки), то нам надлежит половину раз менять эти предположительно несущественные переменные. Если предположительно несущественная переменная (например, порядок проведения экспериментов) в действительности является существенной, то это обнаружится в наших результатах. Если же она окажется действительно несущественной, то мы можем ее игнорировать и усреднить результаты всех экспериментов, не ухудшая их дисперсии.

Контрольные эксперименты. Контрольным называется эксперимент, который экспериментатор на основе своих знаний и опыта считает совершенно тождественным интересующему нас эксперименту, за исключением отличий в отношении одной существенной переменной. Такие контрольные опыты стали стандартным методом в биологической и медицинской работе, ввиду сложного характера неконтролируемых переменных. По тем же самым причинам контрольные эксперименты следует применять и при рабочих испытаниях. Вильсон говорит: «Если кто-либо сомневается в необходимости контрольных испытаний, пусть подумает над следующим утверждением: „Сотнями экспериментов бесспорно доказано, что удары в гонг возвращают солнце на небо после затмения"».

Применительно к испытанию системы посадки самолетов по приборам, о чем мы говорили в § 11.3, контрольный эксперимент заключался бы в установке трех различных испытываемых систем и одной существующей системы на совершенно одинаковых самолетах, пилотируемых одними и теми же летчиками. Если это невозможно (а часто так и бывает), наилучший метод заключается в том, чтобы разделить пилотов на пары по какому-либо подходящему признаку (например, четыре опытных и четыре неопытных летчика) и затем выбрать случайным образом, кто из четырех летчиков в каждой группе будет пилотировать самолет с визуальной аппаратурой, кто — со звуковой, кто — с комбинированной визуально-звуковой и кто — с существующей системой. При этом «случайный выбор» отнюдь не есть выбор мгновенный, без размышления, который может быть подвержен неосознанному смещению. «Случайность» выбора означает, что выбором управляет бросание монеты, таблица случайных чисел или какой-либо другой истинно случайный механизм. В противном случае предположения, на основании которых делаются выводы, окажутся необоснованными и экспериментатор пойдет по ложному пути. Теперь мы переходим к изучению таких статистических выводов.

ЛИТЕРАТУРА

Среди лучших книг по этому вопросу можно назвать книгу Фишера [47], которая обладает универсальной применимостью, хотя и написана применительно к планированию сельскохозяйственных экспериментов, и книгу Вильсона [46], которая является исключительно смелой и удивительно удачной попыткой рассмотреть вопросы научного исследования в одном томе.

ГЛАВА 12

АНАЛИЗ ЭКСПЕРИМЕНТОВ. МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

Математическая модель описывает зависимости между интересующими нас переменными. Если модель полная, то она показывает, от каких других переменных зависит каждая существенная переменная, каковы функциональные соотношения при каждой такой зависимости, какие постоянные входят в каждое функциональное соотношение и как результаты распределяются около ожидаемых значений. Догадки о некоторых или обо всех этих вещах могут более или менее успешно делаться уже на ранних этапах, однако при окон-

чательном анализе их надо получить для математической модели из физической модели, т. е. из эксперимента.

Чтобы спланировать действенный эксперимент, необходимо располагать математической моделью системы, которая указывала бы, какие переменные следует измерять и какие значения необходимо выбрать для контролируемых переменных. Затем результаты эксперимента должны быть использованы для совершенствования математической модели, что, в свою очередь, позволит спланировать лучшие эксперименты. Однако в этой петле обратной связи имеется один пробел. Мы знаем из гл. 7, что исходы экспериментов распределены в смысле теории вероятностей. Когда мы проводим эксперимент и получаем в результате определенные числа (статистики), эти числа никогда не бывают точно равны тем числам, которые нам нужны (параметрам). Математическая статистика и была разработана для заполнения этого пробела.

Теория вероятностей является дедуктивной наукой о случае; она предсказывает результаты, вытекающие из некоторого набора предположений. Математическая статистика является индуктивной наукой о случае; она позволяет делать выводы о характере исходного распределения и оценивать его параметры на основе нашего знания исходов эксперимента. Как и во всякой индуктивной науке, здесь никогда нельзя быть абсолютно уверенным, что в основе наблюдаемого эффекта лежит та или иная определенная причина. Однако математическая статистика допускает численные выводы и оценки и указывает при этом каждый раз число, характеризующее степень неопределенности полученного результата (или степень нашего доверия к своему заключению).

Две основные задачи математической статистики состоят в получении индуктивных выводов и оценок; мы начнем с задачи о выводах. Мы уже знаем, что в выборках из «хороших» распределений некоторые функции наблюдаемых значений имеют общее свойство: дисперсия значений этих функций при увеличении n, т. е. числа наблюдаемых значений в выборке, уменьшается. Следовательно, увеличивая n, мы всегда можем повысить точность нашей индукции. К несчастью, увеличение n обходится обычно весьма дорого, а иногда оказывается невозможным, в связи с чем мы будем заниматься вопросами эффективности статистических процедур, в частности вопросом о том, сколько испытаний необходимо провести и, обратно, какая точность или уровень значимости могут быть получены при фиксированном числе испытаний.

12.1. Теорема Бейеса

Проблемы, возникающие в процессе получения вывода, носят как математический, так и философский характер, и потому мы начнем с рассмотрения исторического развития некоторых теорий вывода. Следует отметить, что многое из того, что сейчас широко используется в математической статистике, было разработано только в последние годы, хотя теория вероятностей была вполне развита уже 150 лет назад.

Пример. Имеется три урны, и в одной из них содержится два серебряных шара, в другой — один серебряный и один золотой шар, а в третьей — два золотых шара. Наудачу выбирается одна из этих урн; затем из нее наудачу выбирается шар, и этот шар оказывается серебряным. Какова вероятность того, что другой шар в этой урне тоже серебряный?

Решение. Шар не мог быть вынут из третьей урны. Многие интуитивно считают, что обе остальные урны здесь одинаково вероятны, и соответственно делают вывод, что ответом будет $^1/_2$; однако это заключение неправильно (хотя к нему и пришел математик Декарт). Раз мы знаем, что вынут серебряный шар, это влияет на вероятность того, что он взят из урны № 1 или № 2, точно так же, как это влияет на вероятность того, что он взят из урны № 3. Действительно, это делает урну № 1 в два раза более вероятной, чем вторую. Следовательно, правильным ответом будет $^2/_3$.

Рассмотрим теперь общую задачу. Вероятность, что шар был взят из урны № 1 (урны с двумя серебряными шарами), равна вероятности выбора урны № 1, умноженной на условную вероятность того, что (коль скоро выбрана урна № 1) вынут серебряный шар. Однако она также равна вероятности, что вынут серебряный шар, умноженной на условную вероятность того, что (коль скоро вынут серебряный шар) он взят из урны № 1. Символически

$$P(U_1, S) = P(U_1)\, P_{U_1}(S) = P(S)\, P_S(U_1),$$

или

$$P_S(U_1) = \frac{P(U_1)\, P_{U_1}(S)}{P(S)}. \qquad (12.1)$$

Левая часть формулы (12.1) является искомой величиной; величины в числителе правой части формулы нам известны. Величина, стоящая в знаменателе, может быть найдена из формулы (4.17):

$$P(S) = P(U_1)\, P_{U_1}(S) + P(U_2)\, P_{U_2}(S) + \\ + P(U_3)\, P_{U_3}(S).$$

Последний член здесь равен нулю. Поэтому, исключая его и подставляя оставшиеся члены в выражение (12.1), получаем

$$P_S(U_1)=\frac{P(U_1)P_{U_1}(S)}{P(U_1)P_{U_1}(S)+P(U_2)P_{U_2}(S)}. \quad (12.2)$$

В рассмотренном частном примере численное решение имеет вид

$$P_S(U_1)=\frac{1/3\times 1}{1/3\times 1+1/3\times 1/2}=\frac{2}{3}.$$

Формулу (12.2) мы можем обобщить на случай двух событий, каждое из которых имеет несколько возможных исходов. Если исходы первого события суть $i=A, B, \ldots$, а исходы второго суть $j=\alpha, \beta, \ldots$, то формула (12.1) принимает вид

$$P_j(i)=\frac{P(i)P_i(j)}{\sum_i P(i)P_i(j)}. \quad (12.3)$$

Формула (12.3), предложенная впервые Бейесом в XVIII в., известна как *теорема Бейеса*. Она получила столь большую известность, что разным ее компонентам были присвоены даже особые названия. Величина, стоящая в левой части формулы, называется *апостериорной вероятностью*; величины, подобные первому множителю числителя, т. е. $P(i)$, получили название *априорных вероятностей*; а величины, подобные второму множителю числителя, т. е. $P_i(j)$, называются *производящими вероятностями*. Хотя сам Бейес и имел некоторые опасения относительно этой теоремы, многие другие считали ее решением проблемы вывода.

Теорема, конечно, строго верна. Если только дано распределение вероятностей исходного параметра (т. е. априорные и производящие вероятности), существует способ найти вероятность того, что выборка была взята из определенной генеральной совокупности. Вся трудность в том, что мы почти никогда не знаем необходимого распределения. В большинстве случаев исходный параметр даже не подчиняется законам теории вероятностей.

Предположим, например, что мы взяли выборку из генеральной совокупности и нашли, что ее математическое ожидание равно 10. Если выборка была достаточна велика, мы можем быть уверены, что математическое ожидание генеральной совокупности не отличается значительно от 10; предположим, однако, что выборка невелика и мы хотим сделать некоторые выводы касательно математического ожидания генеральной совокупности, например определить апостериорную вероятность того, что математическое ожидание генеральной совокупности не превышает 7. На вопросы такого рода мы и пытаемся ответить в математической статистике, однако теорему Бейеса в этом случае применить нельзя, так как мы не можем ввести в формулу численное значение априорной вероятности того, что математическое ожидание генеральной совокупности равно 7 (или меньше 7).

Обычно мы не знаем этих априорных вероятностей, если не считать искусственных случаев, вроде классических задач с шарами и урнами. Единственное исключение составляет область теории информации, где нам могут быть известны необходимые априорные и производящие вероятности (задача 12.1).

12.2. Критерий значимости

Если мы не можем установить вероятность, относящуюся к значению интересующего нас параметра, то как вообще можем мы делать заключение об этом параметре? На этот вопрос ученые-статистики дают несколько ответов, которые обсуждаются в этой главе. Однако читатель должен быть заранее предупрежден, что ни один из них не является безупречным в философском отношении.

Нулевая гипотеза. Предположим, что имеется биномиальная генеральная совокупность с неизвестным параметром p. Мы сделали из нее выборку в n наблюдений и зарегистрировали k удач. При отсутствии другой информации значение параметра p, приводящее с максимальной вероятностью к появлению наблюденной частоты, есть $p=k/n$. Предположим теперь, однако, что нас интересует некоторое определенное значение $\mathfrak{p}$ параметра p (например, пусть мы проверяем игральную кость, чтобы убедиться, что она падает цифрой 5 вверх с правильной частотой; в этом случае $\mathfrak{p}=1/6$).

Тогда с помощью формулы (5.1) мы можем вычислить вероятность получения нашей выборки в случае, если бы гипотеза $p=\mathfrak{p}$, так называемая *нулевая* или *основная гипотеза*, была действительно справедлива:

$$P(k)=C_k^n\,\mathfrak{p}^k(1-\mathfrak{p})^{n-k}.$$

Если n велико, эта вероятность будет очень мала (она была бы очень мала и в том случае, если бы $\mathfrak{p}$ было равно k/n). Однако в действительности мы хотим знать вероятность того, что наблюденная выборка отклонится от

ожидаемого значения на наблюденную величину или более. Если $\mathfrak{p} > k/n$, эта последняя вероятность равна

$$P(\leqslant k) = \sum_{i=0}^{k} C_n^i \mathfrak{p}^i (1-\mathfrak{p})^{n-i}. \qquad (12.4)$$

Предположим теперь, что мы согласны ошибаться только в одном случае из 20. Тогда мы подставляем численные значения в (12.4) и сравниваем полученный результат с 0,025 (половина от одной двадцатой, так как k может быть больше $\mathfrak{p}n$). Если полученное значение меньше 0,025, мы отвергаем нулевую гипотезу.

Эта процедура называется *критерием значимости*, а дробь (которую мы взяли равной одной двадцатой) — *уровнем значимости* * и обозначается символом α. Критерий такого рода был впервые предложен Р. А. Фишером [101] в начале XX в. Если мы регулярно принимаем решения на этой основе, то критерий гарантирует, что в среднем мы будем отвергать нулевую гипотезу только приблизительно в 5% всех тех случаев, когда она оказывается правильной.

Ошибки I и II родов. На практике, когда мы отвергаем одну гипотезу, мы принимаем другую, альтернативную гипотезу. Если мы отвергаем гипотезу, что игральная кость сделана правильно, симметрично, мы соглашаемся с гипотезой, что игральная кость сделана неправильно, несимметрично, но это утверждение не имеет смысла до тех пор, пока мы не сможем сказать, насколько она неправильна и несимметрична.

Нам хотелось бы принять нулевую гипотезу [что $P(5) = 1/6$], когда она справедлива, и альтернативную гипотезу [скажем, что $P(5) = \theta_1$], когда она справедлива. Однако при выборе гипотезы мы можем допустить ошибки двух родов. Мы можем отвергнуть нулевую гипотезу (и, следовательно, принять альтернативную гипотезу), когда нулевая гипотеза является в действительности правильной; эта ошибка получила название *ошибки I рода*. Мы можем также принять нулевую гипотезу, когда она не является параметром генеральной совокупности, и совершить, таким образом, *ошибку II рода*. Эти термины были введены Нейманом и Пирсоном в 30-х годах нашего века и всегда применяются в соответствии

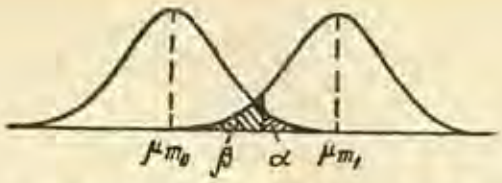


Рис. 12.1. Вероятности ошибок I и II рода ($n = \text{const}$)

с приведенными выше определениями (хотя наименования одной гипотезы «нулевой», а второй «альтернативной» являются произвольными).

Предположим теперь, что мы располагаем фиксированным числом n наблюдений из генеральной совокупности с известной дисперсией; это задает $\sigma_m = \sigma/\sqrt{n}$. Мы устанавливаем уровень значимости, который определяет вероятность совершения ошибки I рода; это автоматически определяет вероятность совершения ошибки II рода (соответствующий уровень значимости для этой последней обозначается символом β). Вместо этого мы могли бы установить значения как α, так и β и определить, насколько большая при этом требуется выборка.

Эти соотношения графически изображены на рис. 12.1. Нулевая гипотеза есть распределение с математическим ожиданием μ_{m_0}; а альтернативная гипотеза — распределение с математическим ожиданием μ_{m_1}. Кривые представляют распределения выборочных математических ожиданий m_0 и m_1 для соответствующих гипотез. Сплошная вертикальная линия указывает критическую точку выбора: если наше математическое ожидание выборки лежит справа от этой линии, мы принимаем альтернативную гипотезу и принимаем вероятность, что мы сделаем ошибку I рода в α случаях (так как вся площадь, ограничиваемая кривой, равна единице, а заштрихованный участок равен α). Если математическое ожидание выборки лежит слева от этой точки, мы принимаем нулевую гипотезу и принимаем вероятность, что мы сделаем ошибку II рода в β случаях. Сдвигая положение линии влево, мы можем уменьшить β за счет увеличения α. Мы можем также увеличить размер выборки, что уменьшит дисперсии двух распределений и тем самым уменьшит как α, так и β.

Значения α и β обычно выбираются с таким расчетом, чтобы ожидаемые значения стоимости были равны. Ожидаемое значение стоимости равно (в простом случае) произведению вероятности совершить ошибку (α или β) на стоимость ошибки в случае, когда она совершена. Это определяет относительные ве-

* В русской литературе по статистике вместо слова «значимость» в этом смысле употребляется иногда слово «существенность». При этом вместо критериев значимости и уровней значимости говорят соответственно о критериях существенности и уровнях существенности. — *Прим. ред.*

личины α и β; абсолютные значения должны, вообще говоря, определяться стоимостью их уменьшения, что может быть достигнуто увеличением n. Таким образом, увеличение размера выборки будет увеличивать стоимость эксперимента, но в то же время уменьшать ожидаемое значение стоимости ошибки. Если уменьшение стоимости ошибки превосходит увеличение стоимости эксперимента, то размеры выборки следует увеличить. Как и при других вопросах проектирования систем, оптимальным решением будет то, при котором ожидаемое значение становится максимальным. В нашем случае «максимизируется» минимальный выигрыш (§ 24.2 и 24.6).

С таким положением вещей мы сталкиваемся на практике почти каждый раз, когда принимаем решение. Если мы решаем, перейти ли улицу сейчас или подождать, пока не появятся более продолжительные интервалы между автомашинами, то ошибкой I рода будет перейти улицу слишком рано, а ошибкой II рода — перейти улицу позже, чем необходимо; стоимостью ошибки I рода будет смерть под колесами автомобиля, стоимостью ошибки II рода — потеря времени. Ясно, что в этом случае мы будем регулировать уровень значимости до тех пор, пока α не станет весьма малой, но мы никогда не сможем сделать ее равной нулю — или мы никогда не перейдем улицу. Конечно, такие случаи не поддаются математическому анализу, так как α и β и их стоимости нельзя определить численно.

Мощность критерия. Предположим, у нас возникли подозрения, что при бросании пары игральных костей слишком часто выпадает десятка. Мы хотим испытать одну игральную кость и посмотреть, не выпадает ли пятерка чаще, чем это должно быть, т. е. посмотреть, будет ли $P(5)>{}^1/_6$. В этом случае нулевой гипотезой является $P(5)={}^1/_6$, и мы должны выбрать какую-нибудь конкретную альтернативную гипотезу. Ситуация графически показана на рис. 12.2. По оси абсцисс откладывается величина $P(5)$, которая может изменяться от 0 до 1; по оси ординат — величина $1-\beta$, т. е. вероятность принятия альтернативной гипотезы (т. е. вероятность, что математическое ожидание выборки будет лежать справа от вертикальной линии на рис. 12.1), когда эта альтернативная гипотеза правильна.

Предположим теперь, что мы бросили игральную кость 100 раз, фиксировав тем самым кривые на рис. 12.1, за исключением правой кривой, которая может смещаться вправо или влево. Затем установим значение α (например, пусть $\alpha=0{,}2$), фиксировав тем самым положение вертикальной линии на рис. 12.1. Если теперь в качестве альтернативной гипотезы мы примем $P(5)=1$, то ясно, что β будет по существу равна нулю, а $1-\beta$ будет по существу равна единице.

Рис. 12.2. Мощность критерия.

Если, последовательно уменьшая значения $P(5)$, принимать эти значения за альтернативную гипотезу и при этом сохранять значения n и α неизменными, то разность $1-\beta$ должна постепенно уменьшаться (сплошная кривая на рис. 12.2). Для определения общего характера кривой посмотрим, что получается, когда, выбирая альтернативную гипотезу, мы достигаем значения $P(5)={}^1/_6$.

В вырожденном случае, когда обе гипотезы одинаковы, мы не можем сделать между ними различия, кроме чисто случайного предпочтения. Следовательно, мы будем отвергать нулевую гипотезу α раз и альтернативную гипотезу в остальных случаях, т. е. $1-\alpha$ раз. Таким образом, $1-\beta=\alpha$, и в этой точке наш график проходит через α. Наконец, так как мы испытывали игральную кость, желая определить, не будет ли $P(5)$ больше ${}^1/_6$, то кривая должна идти еще дальше вниз, хотя эта часть кривой имеет небольшое практическое значение.

Пунктирная кривая на рис. 12.2 изображает ту же ситуацию с возросшим n. Критерий в этом случае во всех точках интересующей нас области стал равномерно более мощным, а во всех точках области, где $P(5)<{}^1/_6$, стал равномерно менее мощным.

Обе эти кривые изображают односторонние критерии. Штриховая кривая представляет собой соответствующую кривую для двустороннего критерия. В этом случае мы подозреваем, что вероятность выпадения пятерок отличается от ${}^1/_6$, но мы не знаем, больше она или меньше. Наша нулевая гипотеза тогда заключалась бы в том, что $P(5)={}^1/_6$, а альтернативная гипотеза могла бы состоять в том, что $P(5)$ лежит вне интервала $0{,}16\leqslant P(5)\leqslant 0{,}175$.

Взамен увеличения мощности в области $P(5)<{}^1/_6$ критерий стал менее мощным в области $P(5)>{}^1/_6$. В случае этого двустороннего критерия увеличение n сделало бы критерий равномерно более мощным по всей области от

0 до 1. Наконец, при одностороннем критерии мы могли бы взять другое значение α. Если α будет меньше, то критерий будет менее мощным и справа от $P(5)=1/6$ он будет выглядеть подобно штриховой кривой на рис. 12.2. Штриховая кривая в действительности и была получена таким способом, а именно α была разбита на две части, по одной на каждом краю распределения.

В более сложных ситуациях можно вывести критерии, более мощные в одной области и менее мощные в другой. Одна из задач математической статистики заключается в нахождении критериев, которые являются наиболее мощными в интересующей нас области.

12.3. Последовательный анализ

Рассмотрим снова классический метод проверки статистических гипотез, описанный выше. Мы имеем некоторую нулевую гипотезу H_0, которую хотим проверить; например, H_0 может быть гипотезой, что некоторая система работает удовлетворительно; гипотезой, что операторы не влияют на качество работы системы; гипотезой, что эффективная отражающая поверхность цели составляет точно 100 кв. футов, или гипотезой, что игральная кость падает цифрой 5 вверх не чаще чем одну шестую времени. Если мы отвергнем нулевую гипотезу, то мы должны принять некоторую определенную альтернативную гипотезу. Если игральную кость нельзя считать правильной, симметричной, значит она неправильна, несимметрична, но насколько? Будет ли частота 0,16668 неправильностью? На практике это бы не сочли неправильностью. Будет ли частота 0,1668 неправильностью? 0,168? 0,18? 0,2? Чтобы эксперимент имел смысл, мы должны указать, какова же альтернативная гипотеза H_1.

Так как мы никогда не можем быть абсолютно уверены в результатах какой бы то ни было статистической проверки, мы должны принять некоторую отличную от нуля вероятность ошибки. Это заставляет ввести еще два числа: α — вероятность совершить ошибку I рода (отвергнуть H_0, когда H_0 — правильная гипотеза) и β — вероятность совершить ошибку II рода (принять H_0, когда H_1 — правильная гипотеза).

Определив эти числа, мы готовы начать планирование нашей проверки. При классическом методе проверки статистических гипотез мы условливаемся сделать выборку заранее определенного размера n и измерить надлежащую статистику этой выборки, после чего эта статистика сравнивается с H_0 и H_1; кроме того, мы условливаемся принять результат этой проверки. Значение n можно вычислить по четырем другим числам (H_0, H_1, α и β). В других случаях мы можем выбрать другой набор четырех чисел и вычислить пятое; так, например, мы часто выбираем H_0, H_1, α и n и вычисляем результирующее значение β.

В случае испытания подозрительной игральной кости, которая будто бы падает цифрой 5 вверх слишком часто, предположим, что мы согласны, что любое значение $P(5)$, лежащее ниже 0,175, указывает на действительную симметрию кости; это задает H_1 [а именно, что $P(5)>0{,}175$]. Выбрав надлежащие значения α и β (они не обязательно должны быть одинаковыми), мы имеем возможность определить n — число наблюдений, которые мы должны сделать. Предположим, что мы определили $n=1\,000$.

Имеется, однако, ряд признаков того, что этот метод проверки не является таким действенным, каким он мог бы быть. Мы не обращаем внимания на порядок, в котором появляются интересующие нас данные, хотя этот порядок также может служить источником ценной для нас информации. Кроме того, мы заранее определяем время, когда мы будем изучать собранные в процессе испытаний данные, хотя, может быть, будет целесообразнее пересматривать это принятое нами решение в ходе самого эксперимента.

Предположим, что во всех первых 10 бросаниях кость упала цифрой 5 вверх; в этом случае вряд ли будет иметь смысл продолжать эксперимент. Или еще предположим, что совершено 1 000 бросаний и оказалось, что цифра 5 выпала 176 раз; это соответствует частоте 0,176. Мы условились браковать игральную кость при таких обстоятельствах (и мы должны быть готовы выполнить это обещание, чтобы не вкралось наблюдательское смещение, как говорилось в гл. 11). Однако это решение было произвольным: мы выбрали 0,175 в значительной мере потому, что это число круглое. Другими словами, в первом из рассмотренных нами двух случаев мы могли бы прекратить наш эксперимент при меньшем числе наблюдений, а во втором случае было бы целесообразно увеличить число наблюдений.

Другой метод проверки мог бы состоять в том, что мы условливаемся сделать 1000 бросаний игральной кости. Если при этом 5 выпадет 170 или меньше раз, признаем кость правильной; если 5 выпадет 180 или больше раз, забракуем кость; если цифра 5 выпадет от 171 до 179 раз, повторим эксперимент. Подобный подход к вещам часто встречается в реальной жизни; мы говорим: «Мы не имеем достаточных данных, чтобы отвергнуть или

принять гипотезу с надлежащей уверенностью, и нам необходимо получить дальнейшие данные». Однако как формальный метод статистической проверки такой подход не предлагался до 1929 г. [25] и не был принят до 1940 г.

Но, зайдя столь далеко, разумно, конечно, пойти еще дальше. Мы не будем принимать никаких предварительных решений о числе испытаний n, но будем считать это число случайной переменной. Условимся закончить эксперимент и принять один из двух возможных образов действия в любой момент времени, когда полученные результаты позволяют нам принять одну из двух гипотез в соответствии с установленным доверительным уровнем; условимся также продолжать опыт до тех пор, пока не окажется возможным такое окончание его. Эта процедура называется *последовательным анализом*.

При последовательном анализе по окончании каждого наблюдения мы применяем какое-нибудь правило, говорящее нам, что́ из трех мы должны сделать: прекратить эксперимент и принять нулевую гипотезу (действие 1), прекратить эксперимент и принять альтернативную гипотезу (действие 2) или сделать еще одно наблюдение и затем применить то же самое правило. Правило, которым мы при этом руководствуемся, называется *планом выборки*.

Последовательный анализ отнюдь не отменяет требования выбора двух определенных (и, быть может, произвольных) гипотез и двух уровней значимости. Однако он позволяет повысить результативность эксперимента в том смысле, что ожидаемое число испытаний при хорошем плане выборки оказывается меньше, чем при классическом методе выбора числа n заранее. С другой стороны, при заданном ожидаемом значении n мы можем получить более высокие уровни значимости при тех же самых гипотезах или проверить более близкие гипотезы при тех же самых уровнях значимости.

Метод последовательных отношений правдоподобия. Термин *«последовательный анализ»* часто применяется только для обозначения частного метода, разработанного Вальдом [24] и называемого собственно *планом выборки с последовательными отношениями вероятностей* или *методом последовательных отношений правдоподобия*. Этот метод во многих случаях позволяет уменьшить приблизительно на 50% (по сравнению с классическим методом предварительного выбора n) ожидаемое число наблюдений, которые должны быть выполнены для достижения определенных уровней значимости при принятии или отклонении определенных гипотез.

Метод Вальда, описываемый ниже, применим только в определенной ситуации (которая является, однако, весьма общей). Видоизменения этого метода для других ситуаций описаны в книге [24]. Мы намереваемся принять решение о двух возможных образах действия, один из которых мы должны выбрать в результате испытаний. Например, нам может потребоваться решить, оставить ли систему работающей (или остановить ее на ремонт), ввести ли определенную программу обучения операторов (или не вводить), построить ли более мощную радиолокационную станцию (или нет), забраковать ли подозрительную игральную кость (или оставить).

Мы будем принимать наше решение в зависимости от значения параметра θ распределения некоторой наблюдаемой величины x. Такой наблюдаемой величиной могла бы быть эффективность системы, измеряемая каким-нибудь подходящим образом, например радиолокационное сечение цели или частота выпадения пятерок; параметром же обычно (но не всегда) бывает математическое ожидание исходного распределения. Функциональная форма распределения известна, как и все параметры, кроме θ. Например, мы можем предположить, что критерий эффективности нашей системы имеет нормальное распределение с известной дисперсией или что вероятность выпадения пятерок подчиняется биномиальному распределению.

Желательность выбора действия 1 (в отличие от действия 2) является непрерывной и монотонной функцией от θ (рис. 12.4). Итак, существует некоторое значение параметра θ, скажем θ', при котором нам безразлично, какой из двух образов действия принять. При $\theta < \theta'$ мы предпочитаем действие 1, и чем меньше θ, тем важнее для нас выбрать действие 1. Наоборот, при $\theta > \theta'$ мы предпочитаем действие 2, и чем больше θ, тем важнее для нас выбрать действие 2. Тогда будет существовать некоторая область $\theta_0 < \theta < \theta_1$, в которой ошибка в принятии правильного решения будет приводить к незначительной разнице. Однако при $\theta < \theta_0$ мы хотели бы выбрать действие 1, и мы желаем принять план выборки, при котором мы не будем делать более чем α раз ошибку выбора при этих условиях действия 2; аналогично мы хотим, чтобы вероятность выбора действия 1 была меньше β, если θ будет больше θ_1. Эти четыре числа (θ_0, θ_1, α и β) должны быть заданы еще до установления плана выборки (как в методе Неймана—Пирсона мы должны были задать четыре числа: H_0, H_1, α и β).

Короче говоря, имеется наблюдаемое значение x, для которого мы знаем функциональную

форму распределения и все параметры, кроме одного, θ. Чем меньше значение θ, тем больше мы будем стремиться к тому, чтобы выбрать действие 1, и обратно; в частности, мы хотим выбирать действие 1 с вероятностью по меньшей мере $1-\alpha$ при $\theta<\theta_0$ и действие 2 с вероятностью по меньшей мере $1-\beta$ при $\theta>\theta_1$. Из всех возможных планов выборки, которые удовлетворяют этим условиям, мы хотим использовать тот, который дает наименьшее ожидаемое значение числа требуемых наблюдений.

План выборки. В случае плана выборки с последовательными отношениями вероятностей мы должны после каждого наблюдения вычислить вероятности появления при каждой из двух гипотез (θ_0 и θ_1) полного множества наблюденных до сих пор значений и взять отношение этих вероятностей. Если это отношение меньше числа B или больше числа A, то эксперимент прекращается и предпринимаются соответственно действия 1 и 2; если же это отношение лежит между B и A, то делается еще одно наблюдение. Числа A и B определяются выражениями

$$A=\frac{1-\beta}{\alpha} \text{ и } B=\frac{\beta}{1-\alpha} \qquad (12.5)$$

и всегда удовлетворяют условию $0<B<1<A$.

Вычисление осуществляется следующим образом. Все параметры распределения величины x нам известны, кроме θ. Если мы теперь выберем некоторое значение θ, а именно θ_0, то функция распределения величины x, т. е. $P_0(x)$, будет полностью определена. Осуществив m наблюдений, мы получим множество наблюденных значений x, которые мы обозначим x_1, $x_2, \ldots, x_m$. Вероятность иметь в точности это множество наблюдений при гипотезе $\theta=\theta_0$ равна

$$p_0(m)=P_0(x_1)\,P_0(x_2)\ldots P_0(x_m). \qquad (12.6)$$

Аналогично этому мы можем вычислить вероятность $p_1(m)$ того, что мы получим в точности это множество m наблюдений при гипотезе $\theta=\theta_1$. Отношение вероятностей, используемое Вальдом, есть $p_1(m)/p_0(m)$.

Это отношение вероятностей вычисляется последовательно (т. е. после каждого наблюдения) до тех пор, пока оно не будет равно или больше постоянной A, либо равно или меньше постоянной B. На практике удобнее вычислять логарифмы постоянных A и B и отношения вероятностей, так как отношение вероятностей для m получается из отношения вероятностей для $m-1$ умножением на коэффициент $P_1(x_m)/P_0(x_m)$. Если x — непрерывная переменная, производим вычисления на основании соответствующей функции плотности вероятностей.

Перед тем как приступить к выводу формулы (12.5), необходимо ввести понятие *рабочей кривой*. Это просто график функции $L(\theta)$ от θ для любого данного плана выборки, где $L(\theta)$ — вероятность того, что в результате испытаний будет предпринято действие 1. Идеальная рабочая кривая изображена на рис. 12.3, но она, конечно, не может быть получена при конечном ожидаемом значении числа наблюдений. Нашему выбору чисел θ_0, θ_1, α и β отвечает более реалистическая рабочая кривая такого вида, как на рис. 12.4. Так как мы согласились допускать неправильное решение (а именно, предпринимать действие 1) в β из тех случаев, когда θ в действительности равно θ_1, то

$$L(\theta_1)=\beta \qquad (12.7a)$$

и аналогично

$$L(\theta_0)=1-\alpha. \qquad (12.7б)$$

Для вывода формул (12.5) заметим, что, по определению,

$$\frac{p_1(m)}{p_0(m)}\leqslant B \qquad (12.8)$$

всякий раз, когда предпринимается действие 1. Это значит: вероятность того, что данная выборка в m наблюдений могла быть взята из генеральной совокупности с параметром θ_1, меньше постоянной B, умноженной на вероятность того, что она могла быть взята из генеральной совокупности с параметром θ_0. Следовательно, вероятность предпринять действие 1 при $\theta=\theta_1$ меньше, чем постоянная B, умноженная на вероятность предпринять действие 1 при $\theta=\theta_0$. Однако эти последние две вероятности ввиду (12.7) суть не что иное, как β и $1-\alpha$. Следовательно,

$$\frac{\beta}{1-\alpha}\leqslant B, \qquad (12.9)$$

что является нижней границей для B.

Таким же путем может быть выведена аналогичная верхняя граница для A. Эти неравенства являются очень хорошими приближениями к формулам (12.5) и были бы даже тождественны с ними, если бы m было непре-

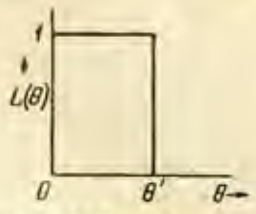


Рис. 12.3. Идеальная рабочая кривая.

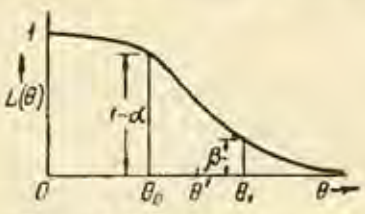


Рис. 12.4. Типичная рабочая кривая.

рывно, а не ограничивалось целыми значениями. Таким образом, степень ошибки при использовании формул (12.5) вместо (12.9) равна величине, на которую отношение вероятностей превышает A (или на которую B превышает это отношение вероятностей) при том наблюдении, на котором испытание прекращается.

Кривая среднего числа наблюдений. Ожидаемое значение числа испытаний является, конечно, важной характеристикой любого плана выборки. Очевидно, что это ожидаемое значение будет зависеть от неизвестного значения θ. Если $\theta \ll \theta_0$ или $\theta \gg \theta_1$, то план выборки с последовательными отношениями вероятностей укажет на это очень быстро, что и составляет одно из преимуществ этого плана. Однако в более вероятном случае, когда θ лежит в области $\theta_0 < \theta < \theta_1$ или вблизи от неё, потребуется большое число наблюдений. Типичная кривая среднего числа наблюдений показана на рис. 12.6.

Для любого данного плана выборки и любого данного критерия можно построить кривую такого типа [24], но при этом могут потребоваться длительные вычисления. Вальд вывел формулу нижней границы числа $E(n)$ для любого плана выборки при $\theta = \theta_0$ и формулу этой границы при $\theta = \theta_1$. Он показал, что план выборки с последовательными отношениями дает по существу эти значения $E(n)$ при θ_0 и θ_1. Никакой один план выборки не может быть оптимальным [т. е. давать минимальное $E(n)$] при всех значениях θ; указанный же план никогда не отходит далеко от оптимального и является оптимальным в этих двух критических точках.

Таблица 12.1

Средний процент экономии при плане выборки с последовательными отношениями вероятностей, по сравнению с классическим планом (по Вальду [27])

Неизвестный параметр есть математическое ожидание нормального распределения; $\theta = \theta_1$ (для $\theta = \theta_0$ необходимо поменять местами α и β)

β \ α	0,01	0,02	0,03	0,04	0,05
0,01	58	60	61	62	63
0,02	54	56	57	58	59
0,03	51	53	54	55	55
0,04	49	50	51	52	53
0,05	47	49	50	50	51

Таблица 12.1 показывает экономию по отношению к классическому методу в типичном случае. В таблице даются ожидаемые значения, и при той или иной конкретной проверке нельзя быть уверенным в том, что n в этом случае не будет больше, чем при классическом методе. Однако можно показать, что при $n \gg E(n)$ вероятность $p(n)$ пренебрежимо мала и эксперимент будет всегда оканчиваться, если $P(x)$ — «разумное» распределение и не выбраны «неразумные» значения (например, $\alpha = 0$).

Непрерывное испытание. Последовательный анализ особенно удобен при автоматическом контроле работы системы. Предположим, например, что речь идет о заводе-автомате, выпускающем конденсаторы с гарантированным номинальным напряжением 100 *в*. Мы выбираем каждый сотый конденсатор и испытываем его на пробой, т. е. увеличиваем напряжение на его зажимах до тех пор, пока не пробьется диэлектрик. Из предыдущего эксперимента мы знаем, что напряжения пробоя будут иметь нормальное распределение со стандартным отклонением 20 *в*. Мы считаем, что конвейер должен быть остановлен всякий раз, когда математическое ожидание напряжения пробоя выходит из интервала $140\ в = \theta_0 \leqslant \theta \leqslant \theta_1 = 150\ в$.

Полагаем $\beta = 0{,}1$ (мы хотим сделать эту вероятность высокой, так как не возражаем против остановки производства в случае, когда мы не находимся действительно в опасной зоне, но очень близко подошли к ней) и $\alpha = 0{,}02$ (мы хотим быть твердо уверены в том, что не выпускаем негодных конденсаторов). Затем вычисляем $A = 9{,}8$ и $B = 0{,}022$. Остальное может сделать автоматически машина *. Такая вычислительная машина, используя введенные в ее запоминающее устройство значения нормального распределения вероятностей, могла бы, например, вычислять для каждого испытанного конденсатора логарифм отношения вероятностей и сравнивать его с $\log B$ и $\log A$. Если указанный логарифм переходит через одно из этих значений, то в соответствии с этим немедленно начинается выпуск новой серии конденсаторов или же останавливается конвейер.

Применение биномиального распределения. В рассмотренном примере мы могли бы несколько изменить ситуацию и испытывать каждый конденсатор при 100 *в* (или, чтобы иметь определенный запас прочности, при 120 *в*), бракуя те, которые откажут, и принимая остальные. В этом случае x равен либо 0, либо 1,

* Две машины для биномиального распределения были построены под руководством Гуда и Гилмена в 1945 г.: одна — в Гарвардском университете, другая — в Массачусетском технологическом институте. — *Прим. авт.*

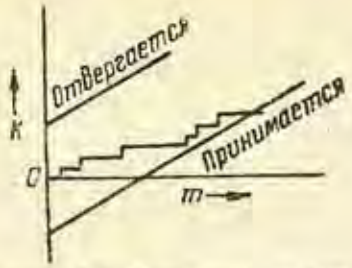


Рис. 12.5. Типичный результат последовательного анализа. Количество бракованных изделий приемлемо.

а θ обозначает вероятность отбраковки. Получаем биномиальное распределение, в котором число дефектных изделий k при n испытаниях задается формулой (5.2), где $p=\theta$ и $q=1-\theta$. Установив значения $p_0=\theta_0$ и $p_1=\theta_1$, получим

$$p_0(m)=p_0^k q_0^{m-k} \text{ и } p_1(m)=p_1^k q_1^{m-k}. \quad (12.10)$$

План выборки с последовательными отношениями вероятностей задается выражением

$$A \leqslant \frac{p_1(m)}{p_0(m)} \leqslant B.$$

Подставляя (12.5) и (12.10) в это выражение и решая его относительно k, получаем

$$r \log \frac{\beta}{1-\alpha} + rsm < k < r \log \frac{1-\beta}{\alpha} + rsm, \quad (12.11)$$

где

$$\frac{1}{r}=\log \frac{p_1}{p_0} - \log \frac{q_1}{q_0} \text{ и } s=\log \frac{q_0}{q_1}.$$

На рис. 12.5 изображен график для выражения (12.11), причем область неравенства

Рис. 12.6. Кривая среднего числа наблюдений для биномиального распределения.

ограничена двумя прямыми линиями равного наклона. Кривая среднего числа наблюдений для этого случая приведена на рис. 12.6.

12.4. Оценка параметра

Мы хотим оценить по наблюденной выборке математическое ожидание рассматриваемой генеральной совокупности и, возможно, также дисперсию, а если мы подозреваем, что распределение не является нормальным, то и некоторые другие параметры. Эти оценки будут использоваться для ответа на такие, например, вопросы: на какой дальности будет происходить радиолокационное обнаружение цели? как часто пересекают автомобили перекресток? как часто будет проходить сигнал в интервал времени, когда реле находится в процессе переключения?

Какой статистикой нам воспользоваться для оценки математического ожидания генеральной совокупности в случае выборки из распределения, про которое мы знаем, что оно нормальное? Мы уже показали, что математическое ожидание выборки равно математическому ожиданию генеральной совокупности, но мы могли бы также показать, что медиана выборки равна медиане генеральной совокупности и что для нормального распределения математическое ожидание и медиана равны. Мы привыкли использовать математическое ожидание выборки (и, как покажем в дальнейшем, для нормального распределения это действительно наилучшая оценивающая функция), однако интуитивно совсем не очевидно, что это лучше, чем применять медиану (или моду, или какую-либо другую статистику). Действительно, для распределения Коши математическое ожидание большой выборки является, конечно, очень плохой оценивающей функцией, в то время как медиана — вполне хорошей.

Желательные характеристики оценивающей функции. Желательно, чтобы оценивающая функция обладала четырьмя характеристиками: несмещенностью, состоятельностью, эффективностью и достаточностью.

Статистика является несмещенной оценивающей функцией данного параметра, если ее ожидаемое значение равно значению параметра. Мы, например, показали, что математическое ожидание выборки из любого распределения является несмещенным; с другой стороны, как мы уже убедились, дисперсия выборки является смещенной, если только вместо n мы не будем употреблять $n-1$. Подобно этому медиана асимметрического распределения не была бы несмещенной оценкой математического ожидания.

Независимо от своей несмещенности или смещенности статистика является состоятельной, если при $n \to \infty$ она имеет предел (в смы-

сле теории вероятностей), равный тому параметру, для оценки которого она применяется. Очевидно, что все несмещенные статистики состоятельны. Однако дисперсия выборки с n в знаменателе состоятельна, хотя и смещена.

Статистика является эффективной оценивающей функцией, если для оценки с заданным уровнем значимости она требует сравнительно малой выборки; иными словами, если она имеет малую дисперсию. Например, мы могли бы брать выборку из нормального распределения, группировать наблюденные значения по классам (рис. 12.8) и затем для оценки центральной точки (т. е. математического ожидания) распределения использовать моду. Совершенно очевидно, что это менее эффективный метод, чем применение математического ожидания выборки. Однако интуитивно совсем не очевидно, что является более эффективной оценивающей функцией: математическое ожидание или медиана.

Если распределение данной статистики может быть записано как произведение двух сомножителей, из которых один содержит оцениваемый параметр и эту статистику, но никакой другой функции наблюдаемых значений, а другой содержит наблюдаемые значения, но не параметр, то говорят, что эта статистика достаточна. В этом случае «оценка содержит всю информацию (в некотором определенном смысле) о значении параметра, имеющуюся в выборке» [37]. Достаточные статистики существуют не всегда.

Оценка наибольшего правдоподобия. Метод наибольшего правдоподобия представляет собой метод вывода оценивающей функции для любого данного параметра известного распределения. Этот метод в тех случаях, когда он применим, позволяет найти такое значение интересующего нас параметра, которое, если бы было правильно, давало бы наблюденную статистику с максимальной вероятностью. Можно показать, что в тех случаях, когда этот метод применим, он дает оценивающую функцию, которая является состоятельной, эффективной (в том смысле, что она имеет наименьшую дисперсию), достаточной, если вообще существует хотя бы одна достаточная статистика, и стремится к нормальности при больших n. Это — замечательное собрание достоинств для одного метода.

Рассмотрим пример, касающийся нормального распределения $N(\mu, \sigma)$. В этом применении метода мы сначала находим совместную вероятность появления определенной группы n наблюдаемых значений, взятых из известного распределения, и затем производим частное дифференцирование относительно интересующего нас параметра. Приравнивая эту производную нулю, определяем экстремальное значение, которое, как можно показать, дает оценивающую функцию по методу наибольшего правдоподобия.

Вероятность совместного появления нескольких независимых наблюдаемых значений при нормальном распределении равна произведению их отдельных вероятностей. Следовательно,

$$P(x_1, x_2, \ldots, x_n)\,dx_1 dx_2 \ldots dx_n = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left[-\frac{\Sigma(x_i-\mu)^2}{2\sigma^2}\right] dx_1 \ldots dx_n. \tag{12.12}$$

Мы можем собрать все постоянные (которые затем будут исключены) в одну постоянную k (заметим, что σ есть постоянная для этого дифференцирования и что множители dx не изменяются). Частное дифференцирование дает

$$\frac{\partial P}{\partial \mu} = k\sum (x_i-\mu)\exp\left[-\frac{\Sigma(x_i-\mu)^2}{2\sigma^2}\right].$$

Чтобы найти $\hat{\mu}$, т. е. оценивающую функцию для μ, приравняем это выражение нулю:

$$k\sum (x_i-\hat{\mu})\exp\left[-\frac{\Sigma(x_i-\hat{\mu})^2}{2\sigma^2}\right] = 0.$$

Так как экспоненциальный член всегда положителен, то другой член выражения должен быть равен нулю. Следовательно,

$$\sum (x_i-\hat{\mu}) = 0, \quad \sum x_i - n\hat{\mu} = 0$$

$$\hat{\mu} = \frac{\Sigma x_i}{n} = m. \tag{12.13}$$

Итак, мы показали, что математическое ожидание выборки является «наилучшей» оценивающей функцией центра нормального распределения в том смысле, в каком метод максимального правдоподобия дает наилучшую оценивающую функцию. Можно показать, что дисперсия медианы выборки приблизительно на 25% больше дисперсии математического ожидания выборки.

12.5. Доверительные интервалы

В § 12.1 указывалось, что теорема Бейеса строго верна, но применима отнюдь не всегда, так как мы не знаем, какими числами выражаются вероятности отдельных значений параметра. Идея доверительных интервалов есть

Рис. 12.7. Доверительные интервалы

попытка обойти это логическое затруднение; удачна ли она, об этом можно спорить, но проектировщику систем часто придется слышать об оценках с доверительными интервалами и следовало бы знать, что означает этот термин.

Предположим, что мы берем выборки объемом $n=9$ из нормальной генеральной совокупности, стандартное отклонение которой нам известно и равно 3; математическое ожидание генеральной совокупности неизвестно и должно быть оценено. Стандартное отклонение математического ожидания выборки равно

$$\sigma_m=\sqrt{\frac{\sigma^2}{n}}=\sqrt{\frac{9}{9}}=1.$$

Далее, пусть в качестве уровня значимости мы взяли $\alpha=0{,}05$, что соответствует точке двух сигм. Мы можем начертить (рис. 12.7) две линии:

$$m_1=\mu-2\sigma_m=\mu-2$$

и

$$m_2=\mu+2\sigma_m=\mu+2. \qquad (12.14)$$

Для *любого* значения μ вероятность того, что наблюденное значение m будет лежать между этими линиями, равна 0,95. Теперь возьмем некоторую определенную выборку и вычислим математическое ожидание m_s. Ордината, проведенная из точки m_s, пересекает наши две линии в двух точках, и расстояние между этими двумя точками есть 95%-ный доверительный интервал для μ. Как и в случае индуктивного вывода, когда для нулевой гипотезы допускалась возможность неправильного суждения в долю α всего времени, так и сейчас мы можем утверждать, что, если мы делаем такие оценки достаточно часто, истинное значение математического ожидания будет лежать в этом интервале в $1-\alpha=0{,}95$ всего времени (95% всех случаев).

В общем случае возьмем некоторый параметр θ и соответствующую статистику T. Выбрав значение α так, чтобы разность $1-\alpha$ являлась требуемым уровнем значимости, можно найти функции от θ

$$\left.\begin{aligned} T_1&=f_1(\theta)\\ T_2&=f_2(\theta)\end{aligned}\right\}, \qquad (12.15)$$

которые соответствуют уравнениям (12.14) и определяются соотношением

$$\int_{-\infty}^{f_1(\theta)} P(T)\,dT=\int_{f_2(\theta)}^{\infty} P(T)\,dT=\frac{\alpha}{2} \qquad (12.16а)$$

или в более общем виде соотношением

$$\int_{-\infty}^{f_1(\theta)} P(T)\,dT+\int_{f_2(\theta)}^{\infty} P(T)\,dT=\alpha. \qquad (12.16б)$$

Графики для уравнений (12.15), вообще говоря, уже не будут прямыми линиями, как это имело место на рис. 12.7. Площадь между этими линиями называется *доверительной зоной*. Теперь мы возьмем некоторую определенную выборку, характеризуемую значением T_s, и решим уравнения (12.15) аналитически или графически, чтобы определить соответствующие значения θ; последние обозначаются символами $\bar{\theta}$ и $\underline{\theta}$ для верхней и нижней границ соответственно. Тогда

$$P[\underline{\theta}\leqslant\theta\leqslant\bar{\theta}]=1-\alpha, \qquad (12.17)$$

$\bar{\theta}$ и $\underline{\theta}$ определяют доверительный интервал с уровнем $1-\alpha$ для любого значения T. Если функция не является симметрической, то не всегда очевидно, как будет распределена α между этими предельными значениями. Тогда можно использовать формулу (12.16а); иногда также используется дру ая функция, минимизирующая ширину доверительного интервала.

Доверительные интервалы обычно сопровождают всякую оценку параметра, например оценку математического ожидания генеральной совокупности на рис. 12.7; они могут быть помещены вокруг любой оценки или предсказания. Например, пусть мы располагаем выборкой из нормальной генеральной совокупности с математическим ожиданием m и стандартным отклонением s и хотим выразить нашу уверенность в том, что следующее наблюдение из этой же генеральной совокупности будет находиться в определенных пределах. В соответствии с (12.17) мы хотим найти статистику T, для которой

$$P(m-Ts<x<m+Ts)=1-\alpha$$

и которая определяла бы, что с вероятностью $1-\alpha$

$$T \geqslant \left| \frac{x-m}{s} \right|. \qquad (12.18)$$

Однако величина в правой части этого неравенства является в точности стьюдентовым отношением t, и потому для любого данного значения α мы можем посмотреть искомую величину в таблицах.

Пример. При проектировании одной самолетной системы была выведена формула для предсказания веса аппаратуры. Эта формула была проверена на 15 уже построенных системах. При этом в указанных 15 случаях были обнаружены следующие ошибки (в процентах): $-1{,}4$; $+7{,}4$; $-1{,}4$; $+18{,}8$; $+4{,}6$; $-7{,}8$; $-8{,}4$; $-5{,}1$; $+2{,}0$; $-13{,}6$; $-0{,}5$; $-18{,}1$; $+7{,}7$; $-5{,}9$ и $+4{,}6$ (данные взяты из литературы [42]). Желательно абстрагировать из этих данных числа, выражающие вероятную точность, с которой формула будет предсказывать вес следующей оцениваемой системы. Найти:

а) 95%-ные доверительные границы, т. е. границы ошибки, внутри которых с вероятностью 0,95 будет находиться наше следующее предсказание;

б) доверительные границы для 10%-ной ошибки, т. е. вероятность того, что наше следующее предсказание будет ошибочным менее чем на $\pm 10\%$.

Решение. Эта задача допускает не один ответ; как нередко бывает в математической статистике, точный ответ зависит от дополнительных предположений, которые мы выбираем, а эти предположения в свою очередь зависят от других факторов, часть из которых может оказаться неизвестной. В рассматриваемом случае размер выборки достаточно велик, так что некоторые из наших предположений будут лишь незначительно влиять на результаты; если бы выборка была меньше, могло бы оказаться даже более важным получить дополнительную информацию.

Итак, наблюдаемыми значениями x_i являются относительные ошибки, измеряемые в процентах, а $n=15$. Мы должны сначала предположить, что все эти значения относятся к одной и той же генеральной совокупности. Допустимость этого отнюдь не очевидна, так как, вообще говоря, относительные ошибки не остаются постоянными в большой области изменения переменной, т. е. нет никаких априорных оснований считать, что та же самая относительная ошибка будет сделана при оценке веса малого и большого самолета. В рассматриваемом случае мы предположим, что мы исследовали формулу и нашли, что можно считать дисперсию относительной ошибки постоянной.

Естественно предположить, что ошибки имеют нормальное распределение. Быструю проверку этого предположения можно сделать, построив гистограмму числа ошибок, попадающих в области шириной в 4% (рис. 12.8), а затем проанализировав эту гистограмму. Гистограмма не показывает, что гипотезу о нормальном распределении ошибок следует отвергнуть. Существует и более сложная аналитическая методика для проверки таких гипотез, однако размеры нашей выборки недостаточно велики, чтобы оправдать ее применение.

Мы можем определить m и s, но нам неизвестны μ или σ, и так как в нашем распоряжении имеется менее 30 наблюдений, нам следует применить стьюдентово отношение t [(выражение 12.18)] с числом степеней свободы, равным 14. Тогда

$$m = \Sigma x_i/n = -1{,}14$$

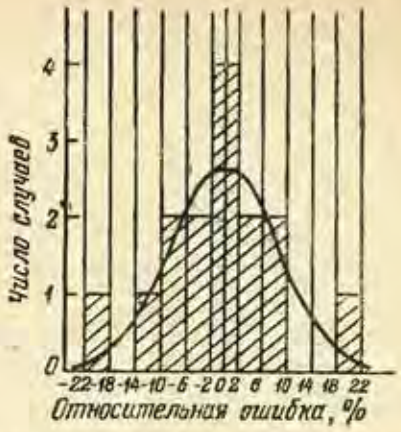


Рис. 12.8. Нормальная кривая, выравнивающая гистограмму ошибок веса.

и

$$s = \Sigma (x_i - m)^2/(n-1) = 9{,}28.$$

а) Мы ищем в таблицах t для $\nu = 14$ и $\alpha = 0{,}05$ (отклонение, превышающее 0,025) и находим $t = 2{,}145$. Тогда

$$-1{,}14 - (2{,}145 \times 9{,}28) \leqslant x_i \leqslant -1{,}14 + (2{,}145 \times 9{,}28),$$

$$-21{,}04 \leqslant x_i \leqslant +18{,}76.$$

б)

$$t = \frac{-10 - (-1{,}14)}{9{,}28} = 0{,}955,$$

вероятность (отклонения $\geqslant t$) равна 0,18;

$$t = \frac{+10 - (-1{,}14)}{9{,}28} = 1{,}200,$$

вероятность (отклонения $\geqslant t$) равна 0,13;

$$\alpha = 0{,}18 + 0{,}13 = 0{,}31, \quad 1 - \alpha = 0{,}69.$$

Следовательно, если применять рассматриваемую формулу к оценке веса 100 других систем, взятых из той же генеральной совокупности, что и первые 15, то можно ожидать появления ошибки от $+10$ до -10% приблизительно в 69 случаях и ошибки от $+19$ до -21% приблизительно в 95 случаях (если мы только не используем вновь получаемые данные для пересмотра наших оценок величин μ и σ).

Можно было бы предположить, что математическое ожидание генеральной совокупности равно нулю (применительно к ошибкам это предположение часто оказывается справедливым). В нашем случае мы можем получить лучшую оценку стандартного отклонения выборки, использовав формулу $s = \Sigma (x_i - \mu)^2/n = 9{,}145$. Затем мы ищем в таблицах t для $\nu = 15$ и находим:

а) $t = 2{,}131$, откуда

$$0 - (2{,}131 \times 9{,}145) \leqslant x_i \leqslant 0 + (2{,}131 \times 9{,}145),$$

$$-19{,}5 \leqslant x_i \leqslant +19{,}5;$$

б) $t = \dfrac{10{,}0}{9{,}145} = 1{,}092,$

вероятность (отклонения $\geqslant t$) равна 0,147,

$$\alpha = 0{,}294 \text{ и } 1 - \alpha = 0{,}706.$$

Обратите внимание на следующее обстоятельство. Доверительные интервалы стали несколько уже (а это говорит о некотором улучшении формулы) благодаря дополнительной информации о том, что математическое ожидание равно нулю.

12.6. Наименьшие квадраты

Когда собраны данные о функциональной зависимости между переменными, обычно бывает желательно выразить ее в виде равенства, определяющего значение зависимой переменной через значения независимых переменных. Во многих случаях функциональная форма зависимости будет сразу же ясна из собранных данных (например, она будет линейной) и наша задача будет сводиться к определению постоянных в уравнении. В этом случае можно с некоторыми ограничениями применять метод наименьших квадратов. Иногда функциональная форма не очевидна, и может потребоваться значительное воображение. Вклад в науку таких людей, как Кеплер и Планк, и состоял в нахождении таких функциональных форм, и для повторений их достижений не существует простой формулы.

Часто оказывается полезным представить данные различными графиками (x как функция от y, $\log x$ как функция от $\log y$ и т. п.). Понимание действующих физических сил помогает определить функциональную зависимость, и, конечно, справедливо также обратное. Существует несколько правил, позволяющих принимать такие решения, как выбор между полиномами 3-го или 4-го порядка. Они кратко упомянуты ниже.

Линейная функция, одна переменная. Пусть мы имеем дело с переменной y, зависящей от одной независимой переменной x, и пусть нам известно, что эта функциональная зависимость имеет вид

$$y = ax + b. \tag{12.19}$$

Мы располагаем серией наблюдений x_i, Y_i, относительно которых мы считаем, что значения x_i являются точными, а значения Y_i содержат некоторую ошибку наблюдения ε_i. Тогда

$$Y_i = y_i + \varepsilon_i,$$

или

$$\varepsilon_i = Y_i - y_i = Y_i - ax_i - b. \tag{12.20}$$

Постоянные a и b определяют соответственно наклон и начало прямой (12.19).

Метод наименьших квадратов состоит в нахождении таких значений этих постоянных, которые приводили бы к минимуму функцию

$$G = \sum_i \varepsilon_i^2 = \sum_i (Y_i - ax_i - b)^2.$$

Для этой цели мы берем частные производные от G и приравниваем их нулю:

$$\left.\begin{aligned} \frac{\partial G}{\partial a} &= \sum_i 2(Y_i - ax_i - b)(-x_i) = \\ &= -2\sum_i x_i Y_i + 2a\sum_i x_i^2 + 2b\sum_i x_i = 0 \\ \frac{\partial G}{\partial b} &\sum_i 2(Y_i - ax_i - b)(-1) = \\ &= -2\sum_i Y_i + 2a\sum_i x_i + 2b\sum_i 1 = 0 \end{aligned}\right\}. \tag{12.21}$$

Перенося члены и подставляя вместо последней суммы в формуле (12.21) (суммирование единиц) ее эквивалент n, т. е. число наблюдений, получаем:

$$a\sum x_i^2 + b\sum x_i = \sum x_i Y_i, \tag{12.22а}$$

$$a\sum x_i + bn = \sum Y_i. \tag{12.22б}$$

Уравнения (12.22) называются *нормальными уравнениями*; решение их позволяет определить требуемые значения a и b.

Теперь предположим, как и раньше, что наблюденные значения x_i являются точными и что, кроме того, ε_i (т. е. ошибки в y_i) имеют нормальное распределение с математическим ожиданием, равным нулю, и стандартным отклонением σ. Тогда элемент вероятности получить определенное наблюденное множество ошибок ε равен

$$P(\varepsilon_1, \ldots, \varepsilon_n)\, d\varepsilon_1 \ldots d\varepsilon_n =$$

$$= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left(-\frac{\Sigma\varepsilon_i^2}{2\sigma^2}\right) d\varepsilon_1 \ldots d\varepsilon_n.$$

Дифференцируя функцию плотности вероятностей и приравнивая производную нулю, мы можем, согласно методу наибольшего правдоподобия (§ 12.4), найти значения параметра, которые давали бы с наибольшей вероятностью наблюденные статистики:

$$\left[\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left(-\frac{\Sigma\varepsilon_i^2}{2\sigma^2}\right)\left(-\frac{1}{2\sigma^2}\right)\right]\left(\frac{\partial\Sigma\varepsilon_i^2}{\partial a}\right) = 0;$$

и аналогично для b. Выражение в квадратных скобках не может равняться нулю, и другой сомножитель приводит к тому же резуль-

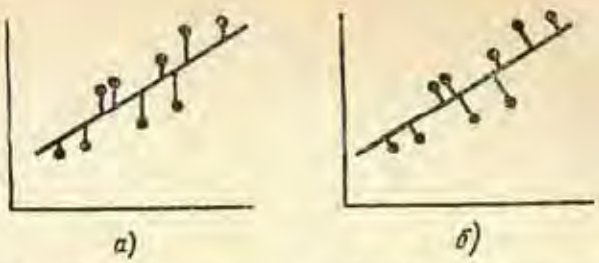
тату, что и выше, — к нормальным уравнениям (12.22).

Это и служит оправданием широко используемого метода наименьших квадратов. Если выбрана правильная функциональная форма; если измерения независимой переменной не содержат ошибки; если ошибки измерения зависимой переменной независимы, нормально распределены и относятся все к одному распределению (т. е. имеют одну дисперсию); и если математическое ожидание распределения равно нулю — то метод наименьших квадратов дает наилучшую возможную оценку согласно критерию наибольшего правдоподобия.

В этой формулировке имеется много «если», но метод наименьших квадратов часто используется даже в тех случаях, когда не все эти условия выполняются. Три из последних четырех (что ошибки независимы и нормально распределены с математическим ожиданием, равным нулю) обычно выполняются. Требование относительно дисперсии является более серьезным затруднением, особенно если измерения охватывают большой диапазон. Если переменная изменяется от 100 до 1 000, а стандартное отклонение ошибок у нижнего предела составляет 10, то стандартное отклонение у верхнего предела может иметь значение 10 (постоянная абсолютная ошибка), или 100 (постоянная относительная ошибка, измеренная в процентах), или какое-либо другое значение.

О выборе правильной функциональной формы мы уже говорили. Иногда можно считать, что независимая переменная является точной, в особенности если она является датой, результатом счета или какой-либо другой переменной, не содержащей в себе ошибки. Тогда минимизация сумм квадратов ошибок для зависимой переменной равносильна минимизации сумм квадратов линий, изображенных на рис. 12.9,*а*. Если независимая переменная подвержена ошибкам, которые также имеют распределение $N(0, \sigma)$, то мы могли бы вместо этого минимизировать суммы квадратов линий, изображенных на рис. 12.9,*б*, а это значительно более сложная процедура.

Линейная функция, многие переменные. Если мы имеем дело с m независимыми переменными, то i-е наблюдение содержит $m+1$ порций данных, а именно Y_i, x_{i1}, x_{i2}, . . . , x_{ij}, . . . x_{im}. Предположим, как и раньше, что функциональная зависимость имеет вид $y = \Sigma a_j x_j$ (чтобы ввести в эту формулу постоянный член, будем считать, что одно значение x тождественно равно 1).



а) б)

Рис. 12.9. Выравнивание наблюденных данных линейной функцией.

Как и раньше, предположим, что значения x_{ij} являются точными и что y_i имеет ошибку ε_i. Тогда, рассуждая как и раньше, получаем

$$\sum_i \varepsilon_i^2 = G = \sum_i \left(Y_i - \sum_j a_j x_{ij}\right)^2.$$

Следующий шаг состоит в дифференцировании этой функции по одному из a; чтобы отличить выбранное нами a от других a_j, введем новый индекс k:

$$\frac{\partial G}{\partial a_k} = 2\sum_i \left(Y_i - \sum_j a_j x_{ij}\right)(-x_{ik}) = 0,$$

$$\sum_i x_{ik} \sum_j a_j x_{ij} = \sum_i Y_i x_{ik}, \qquad (12.23)$$

где i изменяется от 1 до n, j — от 1 до m, а k — от 1 до m. Если мы суммируем по j, то x_{ik} в левой части равенства (12.23) есть постоянная величина и, следовательно, все коэффициенты могут быть перенесены под второй знак суммы; кроме того, если мы суммируем по i, то a_j есть постоянная величина. Итак, мы можем переписать (12.23) следующим образом:

$$\sum_j a_j \sum_i x_{ik} x_{ij} = \sum_i Y_i x_{ik}. \qquad (12.24)$$

Существует m уравнений вида (12.24), и из них мы записали k-ое. Эти m уравнений суть нормальные уравнения, определяющие коэффициенты a_j; число коэффициентов a_j также равно m.

*Полиномиальная функция**. Уравнение (12.24) непосредственно переходит в уравнение (12.22), если принять, что $m=2$ и что j пробегает значения 0 и 1, причем $a_0 = b$, $x_0 = 1$, $a_1 = a$ и $x_1 = x$. Для полиномов более высоких степеней мы можем воспользоваться таким же

* Полиномиальной функцией авторы называют полином (многочлен) как функцию его переменной. В математической литературе эту функцию обычно называют *«целой рациональной функцией»*. — *Прим. ред.*

Рис. 12.10. Данные, которые следует выравнивать полиномиальной функцией.

приемом, положив, что x_j в уравнении (12.24) равны x^j.

Собрание данных в графическом изображении может, например, выглядеть подобно рис. 12.10. Сразу видно, что эти данные не подчиняются линейному закону; однако нельзя сразу сказать, могут ли они быть «выравнены» * многочленом 2-й степени или многочленом высшей степени. Метод наименьших квадратов естественным образом дает критерий для такого выбора. Сумму квадратов приводят к минимуму, и эта минимальная сумма квадратов, деленная на число степеней свободы, дает оценку дисперсии наблюденных данных около выбранной функции. Иными словами,

$$s^2 = \frac{\min(\Sigma \varepsilon_i^2)}{n-m} . \qquad (12.25)$$

Это число можно использовать как меру качества выравнивания. Кроме того, эту оценку дисперсии для многочлена 2-й степени можно сравнить с такой же оценкой дисперсии для многочлена 3-й степени, и если последняя не оказывается заметно меньшей, то можно сделать вывод о нецелесообразности дополнительного усложнения выражения. Многочлен можно написать в таком виде, что влияние члена 3-й степени можно будет проверять независимо [39], выявляя тем самым его значимость. При тщательных исследованиях поэтому, прежде чем оборвать ряд, убеждаются, что члены двух степеней не имеют существенного значения.

12.7. Качество выравнивания

В предыдущих параграфах мы познакомились с методами оценки параметров, например математического ожидания и стандартного отклонения, и методами принятия или отрицания частных гипотез относительно таких параметров. Во многих из этих случаев мы предполагали, что нам известна функциональная форма исходного распределения, и, действительно, она нам часто известна. Однако часто может возникать необходимость в проверке гипотезы, что распределение имеет ту или иную определенную форму: что, например, ошибки имеют нормальное распределение, уличное движение происходит в соответствии с распределением Пуассона или что входы некоторого рода имеют равномерное распределение. В этих случаях мы выставляем нулевую гипотезу, что действует некоторое определенное распределение, предсказываем результаты и затем сравниваем это предсказание с наблюденной выборкой.

Предположим, например, что мы подсчитали число вызовов в секунду, поступающих на телефонную станцию, для каждой секунды в течение получасового периода. Предположим, что общее число вызовов составляет 7 200 (математическое ожидание равно 4 вызовам в секунду) и что мы составили таблицу числа секунд, в течение которых не было ни одного вызова (k_0), числа секунд, в течение которых было по одному вызову (k_1), и т. д. вплоть до k_{13}, причем все более высокие k равны нулю.

Итак, мы располагаем собранием 14 чисел k, где 14-е k есть число секунд с 13 или более вызовами. Мы хотим проверить гипотезу, что эти вызовы распределяются по закону Пуассона с математическим ожиданием 4 вызова в секунду. Из формулы (5.20) мы можем вычислить p_i для каждого i, причем $\Sigma p_i = 1$; если каждое p_i умножить на n (в нашем случае $n = 7\,200$), то получится число, равное ожидаемому значению k_i. Конечно, мы не ожидаем, что наблюденное k_i будет точно равно теоретическому np_i; наша задача — определить по данному множеству m наблюдений, настолько ли это множество невероятно, чтобы мы с полным правом могли отвергнуть нулевую гипотезу.

Если распределение непрерывно, то распределение произвольно делится на небольшое число m интервалов (как на рис. 12.8), чтобы получилось множество m наблюденных частот k_i, которое можно было бы сравнивать с множеством m теоретических частот np_i.

Используя формулу (5.6) для полиномиального распределения, можно вычислять вероятность $P(k_1, k_2, \ldots, k_m)$, что мы в точности получим наблюденное множество чисел k_i. Однако, как мы уже отмечали раньше, такой ответ не представляет интереса; мы хотим знать вероятность получения наблюденного множества или любого другого, которое менее вероятно. С аналогичной задачей мы уже встречались, когда знакомились с биномиальным распределением и рассматривали задачу, в которой требовалось определить,

* Под *выравниванием* (или *согласованием*) данных понимается подбор для них наиболее вероятной (в некотором особом смысле, см. ниже § 12.7) формы функциональной зависимости, или, говоря геометрически, подбор наиболее вероятной (в том же смысле) кривой. — *Прим. ред.*

насколько маловероятно деление на две части по 4 900 и 5 100 элементов при предположении, что $p=0{,}5$ (§ 6.4); в этом случае мы использовали нормальное распределение как приближение к биномиальному распределению. Мы смогли тогда определить вероятность, что некоторая определенная статистика этого распределения (а именно, математическое ожидание выборки) будет отклоняться от своего ожидаемого значения на наблюденную величину. В самом деле, мы смогли найти эту вероятность в таблице кумулятивных значений нормального распределения.

В рассматриваемом сейчас случае мы можем вывести аналогичное приближение полиномиального распределения; это приближение оказалось бы многомерным нормальным распределением. Однако это еще не дало бы полного решения стоящей перед нами задачи. Нам нужна такая статистика выборки, которая была бы монотонным распределением единственной переменной. Эта единственная переменная будет функцией отклонений $k_i - np_i$ реальных частот от их ожидаемых значений.

Как отмечалось в § 5.4, каждое из k_i имеет биномиальное распределение с математическим ожиданием np_i и дисперсией $np_i(1-p_i)$. Подобное распределение имеют и отклонения $k_i - np_i$, с тем отличием, что их математическое ожидание равно нулю. Кроме того, эти наблюденные значения не являются независимыми; так как мы имеем множество $\Sigma k_i = n$, то

$$\Sigma (k_i - np_i) = 0. \qquad (12.26)$$

Ввиду такого отсутствия независимости многомерное нормальное распределение, являющееся предельной формой полиномиального распределения, когда n становится большим, будет в этом случае содержать ковариантные члены и окажется довольно сложным.

Представим себе на момент, что все эти отклонения независимы, нормальны и имеют нулевое математическое ожидание и единичную дисперсию (это даже не является хорошим приближением, но позволит нам лучше понять сущность задачи). В этом случае сумма квадратов отклонений определялась бы ввиду (7.26) как статистика χ^2, а распределение χ^2, задаваемое формулой (7.27), удовлетворяет нашим требованиям. Таким образом, мы могли бы вычислить сумму квадратов отклонений, сравнить полученное значение с таблицей кумулятивной вероятности распределения χ^2 (для m степеней свободы) и тем самым определить точно искомую вероятность, т. е. вероятность того, что в случае справедливости нулевой гипотезы получилась выборка, отличающаяся от предсказанной выборки настолько же, как и наша наблюденная выборка, или больше.

Оказывается, что из полиномиального распределения вероятностей m биномиальных наблюдаемых значений можно получить многомерное распределение m нормальных наблюдаемых значений, из которых $m-1$ являются независимыми. Из этого последнего распределения затем можно получить еще одно многомерное нормальное распределение $m-1$ независимых наблюдаемых значений с нулевым математическим ожиданием и единичной дисперсией. Затем мы можем вычислить χ^2 при $m-1$ степенях свободы для нашего сравнения.

Мы начнем с полиномиального распределения, определяемого выражением (5.16):

$$P(k_1, \ldots, k_m) = \frac{n!}{k_1! \ldots k_m!} p_1^{k_1} \ldots p_m^{k_m},$$

и сделаем преобразование

$$x_i = \frac{k_i - np_i}{\sqrt{np_i}}, \qquad (12.27)$$

где новые наблюдаемые значения x имеют нулевое математическое ожидание и дисперсию $1-p_i$. После ряда длинных выкладок, которые приводятся в стандартных книгах по статистике и в которых используется формула (12.26), мы приходим к формуле

$$P(x_1, \ldots, x_m) = Ce^{-\Sigma x_i^2}, \qquad (12.28)$$

где C — постоянная, которую совсем не обязательно оценивать. Это и есть формула распределения m нормальных наблюденных значений с нулевым математическим ожиданием и единичной дисперсией.

Она справедлива не для всякого произвольного множества наблюденных значений x_i, а только для таких множеств, которые удовлетворяют ограничению, вытекающему из формулы (12.26), т. е. что только $m-1$ из x_i являются независимыми. Однако Карл Пирсон показал, что можно найти следующее замечательное преобразование величин x_i в некоторое множество величин y_i: одна из величин y_i тождественно равна нулю; остальные $m-1$ величин y_i независимы, нормально распределены и имеют нулевое математическое ожидание и единичную дисперсию; якобиан преобразования равен единице. От-

сюда следует, что суммы квадратов величин y_i распределяются как χ^2 при $m-1$ степенях свободы и, так как якобиан равен единице, суммы квадратов величин x_i имеют такое же распределение.

Следовательно, используя выражение (12.27), мы можем по нашей выборке вычислять суммы квадратов величин x_i и сравнивать результаты с таблицей распределения χ^2 для $m-1$ степеней свободы. Эта таблица даст нам искомую вероятность получения нашей выборки при нулевой гипотезе. Сумма Σx_i^2 обычно тоже называется χ^2.

Этот критерий, называемый *критерием качества выравнивания по* χ^2,* вообще говоря, оставляет желать многого. При выводе формулы (12.28) необходимо сделать несколько приближений, наиболее важные из которых состоят в том, что k_i являются большими числами, все x_i — небольшими и что числами порядка $1/(np_i)^{1/2}$ можно пренебречь. В предельном случае, когда n стремится к бесконечности, эти предположения, конечно, справедливы; в практических же случаях они не вносят заметной ошибки (после суммирования), если все k_i умеренно велики (скажем, равны 20 или более) и если все теоретические частоты классов (np_i) приблизительно равны наблюденным частотам классов (k_i).

Последнее предположение не будет нарушаться, если согласие действительно хорошее; первое же предположение часто будет нарушаться на практике в большей или меньшей степени. Следовательно, мы не должны применять этот критерий к небольшим выборкам (таким, как данные на рис. 12.8). Если же выборка довольно велика, то данный критерий оказывается достаточно хорошим, в случае, когда он дает высокую вероятность; если, однако, применять этот критерий для опровержения какой-нибудь гипотезы при доверительном уровне в 5 или 10%, как это часто делается, то мы, вероятно, получим оправдание для опровержения гипотезы, хотя мы и не знаем в действительности, на каком уровне отвергаем ее. Успенский [44] по этому поводу высказал даже следующее: «Недостаток информации по поводу ошибки, вносимой использованием приближенного выражения для Q_n, делает применение этого изобретенного Пирсоном „критерия χ^2“ в некоторой степени сомнительным». Однако этот критерий очень легко применять, и он почти повсеместно используется специалистами-статистиками.

При применении критерия каждый из m классов данных должен содержать довольно большое число наблюдений (Кендалл [29] в консервативном духе рекомендует минимум 20 наблюдений); если какой-либо класс содержит меньше наблюдений, его следует объединить с одним из соседних (причем величины k_i и np_i складываются перед вычислением x_i). Для большинства задач число классов должно быть в пределах $5 \leqslant m \leqslant 20$. Число степеней свободы равно числу классов m минус число наложенных связей (ограничений). Число классов должно определяться после проведения всех объединений, необходимых для того, чтобы размеры каждого класса были бы достаточно большими.

Минимальное число наложенных связей равно единице, как в случае (12.26). Этот минимум, например, имеет место, когда мы проверяем выборку, желая убедиться в том, что она взята из нормальной генеральной совокупности с определенным математическим ожиданием и дисперсией. С другой стороны, если мы по выборке вычисляем как математическое ожидание, так и дисперсию и затем проверяем согласие этой выборки с генеральной совокупностью, имеющей эти параметры, то число степеней свободы должно быть $m-3$.

Как грубый эмпирический прием можно указать, что если $\chi^2 \approx \nu - 1$ (где ν — число степеней свободы), то вероятность того, что выборка, взятая из некоторого предполагаемого распределения, будет давать лучшее согласие, чем наблюденная выборка, равна приблизительно $1/2$. Следовательно, если $\chi^2 \leqslant \nu - 1$, то предполагаемое распределение можно оценивать как правдоподобное, в то время как при $\chi^2 \gg \nu - 1$ гипотезу, что мы имеем дело с предполагаемым распределением, обычно отвергают.

Пример. Данные в табл. 12.2 иллюстрируют типичный случай вычисления χ^2. Величины k_i суть 7200 телефонных вызовов, упоминавшиеся выше; вероятности p_i можно вычислить с помощью приведенной в таблице формулы, однако проще поискать их в таблице распределения Пуассона. Последние четыре класса объединены между собой, чтобы в каждом классе было хотя бы по 20 элементов; таким образом, число классов равно $m=12$. Так как мы вычислили математическое ожидание по выборке, то мы должны наложить еще одну дополнительную связь сверх той, которая предусматривается выражением (12.26); число степеней свободы поэтому равно 10.

Разыскивая 9,19 в таблице распределения χ^2 для 10 степеней свободы, находим, что $P(>\chi^2)=0{,}52$. Это значит, что в 52% случаев, когда данные действитель-

* В нашей литературе по статистике этот критерий часто называется также «критерием согласия» χ^2.— *Прим. ред.*

Таблица 12.2

Вычисление χ^2

i	k_i	$p_i=\frac{e^{-4}4^k}{k!}$	np_i	k_i-np_i	$(k_i-np_i)^2$	$x_i^2=(k_i-np_i)^2/np_i$
0	119	0,0183	132	—13	169	1,28
1	511	0,0733	529	—18	324	0,64
2	1 075	0,1465	1 055	20	400	0,38
3	1 465	0,1954	1 406	59	3 481	2,48
4	1 351	0,1954	1 406	—55	3 025	2,15
5	1 115	0,1563	1 125	—10	100	0,09
6	782	0,1042	750	32	1 024	1,36
7	429	0,0595	429	0	0	0,00
8	206	0,0298	214	—8	64	0,30
9	92	0,0132	95	—3	9	0,09
10	34	0,0053	38	—4	16	0,42
11	13	0,0019	21	0	0	0,00
12	6	0,0006				
13	2	0,0002				
>13	0	0,0001				
Σ	7 200	1,0000	7 200	0	...	$9,19=\chi^2$

$m=12$ $\nu=10$

Таблица 12.3

Влияние мастерства операторов на работу полуавтоматической системы

Оператор	1	2	3
	11 7 8 4 5	6 1 2	8 7 9 4
n_i	5	3	4
m_i	7	3	7
s_i	2,7	2,6	2,2

но относятся к распределению Пуассона с математическим ожиданием 4, мы имели бы отклонения, более значительные, чем это. Таким образом, у нас нет оснований отклонять нулевую гипотезу. Заметим, что основная часть суммы χ^2 приходится только на два из 12 классов; это типично.

12.8. Дисперсионный анализ

Дисперсионный анализ есть методика рассортировки влияний нескольких переменных, действующих полностью или в любой комбинации на результат единичного измерения. Некоторые из этих переменных могут контролироваться и измеряться, тогда как другие переменные остаются неконтролируемыми. Дисперсионный анализ позволяет нам также оценить численно вероятность того, что измеренные переменные действительно оказали влияние на результаты.

Возьмем крайне простой случай, когда мы эксплуатируем полуавтоматическую систему и хотим исследовать, влияет ли или нет разница в мастерстве операторов на эффективность работы этой системы. С этой целью мы составляем таблицу полученных результатов для нескольких разных операторов, используя подходящий критерий эффективности (табл. 12.3).

Мы видим, что при работе первого и третьего операторов мы получаем в среднем цифру 7, а при работе второго средняя цифра составляет только 3; с другой стороны, выборка в нашем случае невелика, и при работе каждого оператора получается значительная дисперсия, в связи с чем такое множество данных могло быть получено случайно. Грубое представление об этом можно получить, обратив внимание на тот факт, что разница между значениями математических ожиданий при работе различных операторов имеет тот же порядок величины, как стандартное отклонение каждой выборки; однако мы еще не знаем, являются ли эти различия значимыми или нет.

Поэтому мы хотим проверить две статистические гипотезы: 1) что математические ожидания генеральных совокупностей одинаковы; 2) что они не одинаковы. Первую гипотезу мы примем за нулевую гипотезу и не будем указывать никакой альтернативной гипотезы; таким образом, мы будем иметь дело только с ошибками I рода. Если нулевая гипотеза будет отвергнута (из-за любого переменного фактора), то мы сможем только сказать, что имеет место влияние этого фактора, но не сможем сказать, как оно велико.

Если три наши выборки взяты из тождественных генеральных совокупностей, предположительно нормальных, с равными математическими ожиданиями и равными дисперсиями, то различия между этими тремя выборками вызваны случайными флюктуациями, которые всегда возможны, когда дисперсия больше нуля. Мы можем составить себе представление о размерах этих случайных флюктуаций, исследуя коллективную дисперсию всех выборок; затем мы можем сравнить найденную коллективную дисперсию с подходящими функциями дисперсии, наблюдаемой в пределах каждой выборки или между одной выборкой и другой.

Таким образом, мы осуществляем сравнение дисперсий, а не математических ожиданий, которые лежат в основе нашей нулевой гипотезы. С точки зрения математики это удобнее, чем непосредственная проверка математических ожиданий. Однако такая методика требует новой статистики F, которая представляет собой отношение двух оценок одной и той же дисперсии.

Таким образом, дисперсионный анализ заключает в себе два этапа. Во-первых, мы должны отделить дисперсию, вызываемую контролируемыми переменными (или — в общем случае — дисперсии, вызываемые каждой из контролируемых переменных), от остальной дисперсии, вызываемой случайными флюктуациями (называемой *взаимодействием, расхождением* или просто *ошибкой*). Во-вторых, мы должны сравнить отношение этих двух дисперсий с табличным значением F, подобно тому, как при оценке согласия мы сравнивали вычисленное значение хи-квадрата с табличными значениями.

Статистика F определяется отношением

$$F = \frac{s_1^2}{s_2^2}, \qquad (12.29)$$

где каждое из s^2 есть несмещенная оценивающая функция дисперсии генеральной совокупности, определяемая выражением (7.10). Так как обе выборки взяты из одной и той же нормальной генеральной совокупности, эта статистика обладает сравнительно регулярными свойствами. Она имеет распределение, определяемое отношением двух распределений χ^2, а именно

$$P(F)\,dF = \frac{\Gamma[(\nu_1+\nu_2)/2]\,\nu_1^{\frac{\nu_1}{2}}\,\nu_2^{\frac{\nu_2}{2}}\,F^{\frac{\nu_1-2}{2}}}{\Gamma(\nu_1/2)\,\Gamma(\nu_2/2)\,(\nu_2+\nu_1 F)^{\frac{\nu_1+\nu_2}{2}}}\,dF, \qquad (12.30)$$

где $\nu_1 = n_1 - 1$ есть число степеней свободы в первом хи-квадрате и аналогично ν_2.

В нашем примере, когда имеется только одна контролируемая переменная с $k=3$ значениями и n_i наблюдениями для i-го значения, полная дисперсия внутри групп дается выражением

$$(n-k)\,s_1^2 = \sum_{i=1}^{k}\sum_{j=1}^{n}(x_{ij} - m_{i\cdot})^2, \qquad (12.31)$$

где точка указывает на то, что производилось суммирование по индексу, замененному точкой. Здесь n — полное число наблюдений, $(n-k)$ — число степеней свободы и $m_{i\cdot}$ — математическое ожидание переменных в i-й группе:

$$m_{i\cdot} = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij}.$$

Дисперсия между группами дается выражением

$$(k-1)\,s_2^2 = \sum_{i=1}^{k} n_i (m_i - m)^2,$$

где m — математическое ожидание всех n наблюдений. Полная дисперсия определяется, конечно, соотношением

$$(n-1)\,s_3^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - m)^2.$$

Дисперсия между группами часто вычисляется по формуле

$$s_1^2(n-k) + s_2^2(k-1) = s_3^2(n-1).$$

Вывод этих формул и доказательство независимости s^2_1 и s^2_2 приводятся в стандартных книгах по статистике.

В сложных случаях, когда приходится иметь дело с несколькими контролируемыми переменными и большим количеством наблюдений, вычисления весьма длительны и трудоемки. Вычисления для простого случая табл. 12.3 показаны в табл. 12.4.

Таблица 12.4

Вычисления при дисперсионном анализе

i	j	x_{ij}	m_i	$x_{ij}-m_i$	$(x_{ij}-m_i)^2$	n_i	m_i-m	$(m_i-m)^2$	$n_i(m_i-m)^2$
1	1	11	7	4	16				
1	2	7	7	0	0				
1	3	8	7	1	1	5	1	1	5
1	4	4	7	−3	9				
1	5	5	7	−2	4				
2	1	6	3	3	9				
2	2	1	3	−2	4	3	−3	9	27
2	3	2	3	−1	1				
3	1	8	7	1	1				
3	2	7	7	0	0				
3	3	9	7	2	4	4	1	1	4
3	4	4	7	−3	9				
		72			58				36

Здесь $n=12$; $k=3$; $k-1=2$ равно числу степеней свободы для s_2^2; $n-k=9$ равно числу степеней свободы для s_1^2.

$$F = \frac{36}{2} \div \frac{58}{9} = 2{,}79.$$

Обратившись к таблице значений F для двух и девяти степеней свободы, находим, что $F_{2,9} = 4{,}26$ при 5%-ном уровне, т. е. $P(F_{2,9} \geqslant 4{,}26) = 0{,}05$. Следовательно, нулевую гипотезу нельзя отвергнуть при 5%-ном уровне, и мы приходим к выводу, что наличие значимого влияния операторов на работу системы не было доказано.

Другая статистика z, определяемая соотношением

$$z = \frac{1}{2} \ln F,$$

анализировалась Фишером еще до того, как Снедекор выдвинул статистику F. Ее распределение представляет определенный интерес при статистических исследованиях и тоже табулировалось. Хотя фишерово z также можно использовать в дисперсионном анализе, снедекорово F вычисляется несколько легче и стало статистикой, повсеместно применяемой для этой цели.

12.9. Корреляция

В § 6.5 мы рассмотрели параметр ρ — коэффициент корреляции между парой переменных. Соответствующая статистика, измеряющая корреляцию между двумя множествами наблюденных значений, называется *выборочным коэффициентом корреляции* и обозначается через r. Методы вычисления этого коэффициента по выборке и оценки его значимости изложены в стандартных книгах по статистике. Достаточно хорошее общее представление о порядке величины коэффициента корреляции проще всего получить из *диаграммы разброса* *. Если нам требуется определить, коррелированны ли или нет две переменные, мы можем изобразить данные графиком, как на рис. 12.10. Если они скучиваются вокруг какой-то прямой или вокруг какой-то простой кривой, как это наблюдается на диаграмме рис. 12.10, то имеет место сильная корреляция; если же они кажутся разбросанными более или менее случайно, то коэффициент корреляции, вероятно, незначительно отличается от нуля.

При переходе к многомерному случаю коэффициенты корреляции можно определять для любой пары переменных. Определяется также *множественный коэффициент корреляции* R между одной и двумя или несколькими другими переменными. Этот коэффициент R никогда не принимает отрицательных значений и даже в случаях небольших связей принимает значения, близкие к единице [29a].

Проектировщик систем должен быть очень осторожен при выводе заключений даже из явно значащих, отличных от нуля коэффициентов корреляции. Следующий пример, хотя и анекдотический, сделает вопрос более ясным. Аист — весьма почитаемая птица в Стокгольме, и каждый год там производится точная перепись аистов. Для 73-летнего периода, относящегося к XIX в., количество аистов, живших в Стокгольме в каждом году, было сравнено с рождаемостью в Стокгольме в том же году и было установлено, что коэффициент корреляции между этими двумя переменными составляет более 0,9. Такой высокий коэффициент корреляции при такой большой выборке является подавляющим свидетельством функциональной зависимости.

Мораль сей истории в том, что функциональная зависимость не обязательно является прямой (т. е. причинной); если прямой связи нет, то обе переменные, вероятно, имеют прямую связь с одной или несколькими другими переменными (в рассматриваемом случае — с общим ростом и экономическим благополучием города Стокгольма), и без уяснения природы этих других связей нельзя делать веских предсказаний. Перед тем как делать выводы из высокого коэффициента корреляции, следует поискать случаев, в которых предполагаемая контролирующая переменная имеет предельные, крайние значения; эти случаи могут быть вырожденными (в нашем примере это был бы город, где нет аистов; мы заметили бы, что люди там еще имеют детей) **.

ЛИТЕРАТУРА

Двумя лучшими книгами следует считать Крамера [43] и Кендалла [29]. Вторая из этих книг чрезвычайно велика и подробна, и обе они трудны для лиц, не имеющих серьезной математической подготовки. По специальным вопросам статистики имеется несколько хороших книг, например книга Вальда [24] о последовательном анализе. Из более элементарных книг по статистике мы рекомендуем Мида [48] и Дикстона и Мэсси [49]; последняя из этих двух элементарнее. Обе эти книги содержат полезные таблицы статистик, рассмотренных в этой главе.

ЗАДАЧИ

12.1. В книге Шеннона [40] по теории информации постулируется трехбуквенный алфавит, в котором вероятность появления буквы полностью определяется

* Такие диаграммы разброса называются также корреляционными полями. — *Прим. ред.*

** Следует вообще иметь в виду, что полноценные, соответствующие действительности статистические выводы могут быть получены только при правильном сочетании количественного и качественного методов, исходя из понимания вероятностных связей как разновидности объективных закономерных связей между явлениями. Многих представителей зарубежной математической статистики критиковали в советской литературе именно за пренебрежение или недооценку качественного подхода. — *Прим. ред.*

предыдущей буквой (это не является слишком искусственным: если взять английский язык, то мы можем дать частичное описание его, определив вероятность появления каждой буквы вслед за каждой данной буквой).

Шеннон приводит следующие численные значения для $P(i, j)$, т. е. для вероятности того, что в паре букв первой будет некоторая определенная буква i, а второй — некоторая определенная буква j. (Например, вероятность, что пара букв будет представлять собой *AC в этом порядке*, равна 1/15.)

Вероятности $P(i, j)$ упорядоченных пар букв

i	j		
	A	*B*	*C*
A	0	4/15	1/15
B	8/27	8/27	0
C	1/27	4/135	1/135

Найти:

а) априорные вероятности $P(i)$;

б) априорные вероятности $P(j)$;

в) производящие вероятности $P_i(j)$ (Шеннон *называет их переходными вероятностями*);

г) апостериорные вероятности $P_j(i)$.

12.2. Используя методы из § 12.4, покажите, что s^2 есть оценивающая функция по методу наибольшего правдоподобия для статистики σ^2 нормального распределения.

12.3. Для одной программы моделирования желательно определить вероятность некоторого маловероятного события. При 10 000 испытаниях это событие произошло 10 раз.

а) Сколько требуется провести испытаний, чтобы иметь 99,7% доверия, что вероятность события определена в пределах ±10%?

б) Если проведено столько испытаний, то каковы были бы 95%-ные доверительные границы для ошибки?

12.6**. С помощью метода наименьших квадратов выровнять следующие данные параболой вида $Y=a+bx+cx^2$:

x	0	5	10	15	20	25	30
y	81	84	88	104	134	148	170

** Из книги [143].

12.7. Получен заказ на 1 000 конденсаторов по цене 60 центов за штуку. Требуется принять или отклонить эту партию на основе простых статистических испытаний, в процессе которых n конденсаторов проверяется на пробой. Предположим, что пробивное напряжение имеет нормальное распределение со стандартным отклонением 100 *в*. Партию конденсаторов следует забраковать, если истинное математическое ожидание пробивного напряжения составляет менее 1 000 *в*, и принять, если оно превышает 1 020 *в*. Если оно лежит в пределах от 1 000 до 1 020 *в*, мы по своему усмотрению можем принять или отклонить партию.

*а) Рассмотрите стоимость ошибок I и II рода, а также увеличения n.

б) Напишите инструкцию по проведению испытаний (т. е. предположите, что значения α, β и n выбраны, и изложите признаки, по которым следует принимать решение о забраковке или приемке партии после испытания n конденсаторов).

в) Выберите разумное и совместимое множество значений для проверочных параметров α, β и n.

12.8. В одном промышленном производстве получается около 15% брака. Контроль при приемке осуществляется на больших партиях, каждая из которых оказывается или хорошей, или плохой. Когда брак составляет 10%, желательно отвергать в среднем партию только в одном случае из 100 ($\alpha=0{,}01$), а когда брак повышается до 20%, желательно принимать партию только в 5 случаях из 100 ($\beta=0{,}05$). Вычислите линии последовательной проверки и решите, стоило ли принимать или браковать продукцию и когда, если наблюдения привели к следующим результатам (g=«хорошая продукция», d=«брак»):

gggdgdggdgggddgdgddgdgddgdgd.

12.4* Было сделано 100 наблюдений при определенном распределении и получены следующие результаты:

Область изменения переменной x	$x<90$	$90\leqslant x<100$	$100\leqslant x<110$	$110\leqslant x<120$
Число наблюдений	8	15	21	23

Область изменения переменной x	$120\leqslant x<130$	$130\leqslant x<140$	$140\leqslant x$
Число наблюдений	16	9	8

а) Найти математическое ожидание выборки.

б) Найти стандартное отклонение выборки.

в) Нормальна ли генеральная совокупность?

12.5*. При типовых испытаниях поточной линии для производства точных сопротивлений было найдено, что для 15 сопротивлений, изготовленных одной машиной, дисперсия выборки составляет 17 *ом*^{2}, а для 12 сопротивлений, изготовленных второй машиной, дисперсия выборки составляет 29 *ом*^{2}. Можно ли отклонить с 10%-ным уровнем значимости гипотезу, что машины работают одинаково?

* Из книги [49].

ГЛАВА 13

ПРИМЕР ВНЕШНЕГО ПРОЕКТИРОВАНИЯ СИСТЕМЫ

До сих пор нам приходилось рассматривать проектирование систем применительно к небольшим частям больших задач; действительно, рассмотрение всей задачи проектирования системы требует обсуждения большого материала. К счастью, Белловская телефонная система довольно подробно документировала свою работу в печати, и мы смогли найти статью, дающую примеры сразу для многих рассмотренных выше методов. Эта статья [102], воспроизводимая здесь почти дословно (с небольшими сокращениями), посвящена вопросам внешнего проектирования при переходе к координатной системе № 5 (см. § 2.2). Однако в ней нельзя было избежать некоторого вхождения в вопросы внутреннего проектирования системы при окончательном выборе численных значений для математической модели — взаимосвязь, о которой говорилось в гл. 3.

Эта статья дает превосходный пример для иллюстрации не только потому, что представляет собой необычно полную и компетентную работу, но и потому, что рассматриваемая в ней система действительно имеет все характеристики систем большого масштаба. Это система оборудования, назначение которой — соединять телефоны абонентов. Она велика по масштабу; добавление одного лишнего маркера на каждой автоматической телефонной станции страны стоило бы многих миллионов долларов. Она сложна и имеет много петель обратной связи. Она работает автоматически и использует вычислительные устройства. Ее входы множественны и изменяются случайно во времени. Что касается состязательных сторон, то, хотя никто и не пытается сознательно разрушить систему, стремление некоторых абонентов набирать номер, не дожидаясь сигнала готовности станции, приводит к аналогичному эффекту. Как можно будет убедиться из статьи, преждевременный набор номера представляет одну из серьезных проблем.

При чтении статьи следует использовать материал первых двенадцати глав. Краткий анализ статьи дается в конце настоящей главы.

ОЦЕНКА ТРЕБОВАНИЙ К АВТОМАТИЧЕСКИМ ТЕЛЕФОННЫМ СТАНЦИЯМ

Уоррен О. Тэрнер

После того как 30 лет назад автоматические телефоны заменили ручные телефоны, в Белловской системе значительная часть ответственности за хорошую работу телефона перешла от телефонисток к инженерам. Излагаемый материал касается одного этапа исследований, проводившихся в Белловских телефонных лабораториях с целью помочь инженерам-телефонистам справиться с этими растущими обязанностями. Рассмотрение будет ограничено задачей проектирования лишь той части оборудования автоматических телефонных станций, которая посылает абоненту сигнал готовности станции к набору номера и соединяет его линию с другими коммутационными устройствами на телефонной станции.

В первых автоматических телефонных станциях телефонные линии от абонентов подключались группами к «линейным стативам», обычно по 200 линий в каждой (рис. 13.1). Каждая группа линий имела определенное число устройств, называемых «линейными искателями». Функция этих устройств заключалась в соединении вызывающих линий с другими коммутационными устройствами телефонной станции, которые в свою очередь продолжали бы это соединение вплоть до вызываемого телефона. Любой линейный искатель в группе мог обслуживать любую линию этой группы. По окончании разговора линейный искатель освобождался и становился доступным для нового вызова.

Количество линейных искателей, требующихся на каждую группу в 200 линий, зависело от количества поступивших вызовов на одну линию (частота вызовов) и длительности использования каждой линии при каждом разговоре (время занятия). Чтобы установить на станциях надлежащее количество оборудования, нужно было знать указанные величины для каждой станции и разработать метод для расчета искомого количества линейных искателей по этим данным.

Если инженеры назначат слишком мало линейных искателей на группу линий, то в более загруженные периоды дня какие-то абоненты, сняв трубки, не услышат знакомого «ровного гудка». Если они начнут набирать номер, не дожидаясь этого сигнала, их вызовы все равно не будут осуществлены, так как все линейные искатели в это время заняты обслуживанием других вызовов. С другой стороны, если на станциях установить слишком много линейных искателей, то какие-то деньги телефонной компании будут потрачены зря.

Таким образом, возникла задача найти разумный критерий обслуживания абонентов и разработать на основе этого критерия метод расчета, который позволял бы определять надлежащее количество линейных искателей на каждой телефонной станции, независимо от изменения характеристик использования аппаратуры от станции к станции.

При введении в действие первых автоматических телефонных станций не имелось никаких опытных данных, которые позволяли бы выбрать критерий хорошего обслуживания; однако ввиду того что абоненты

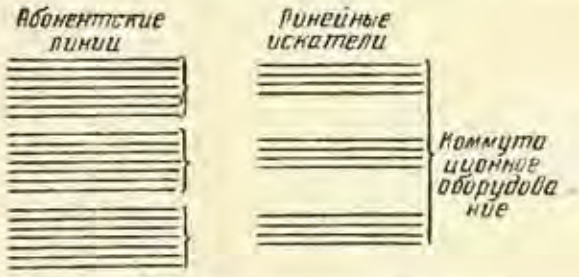


Рис. 13.1. Подключение абонентов в АТС первоначального типа.

впервые должны были сами набирать номера для своих вызовов, не обращаясь более за обслуживанием вызова к телефонистке, было решено, что этот новый способ работы следует сделать как можно более легким. Казалось правильным поставить такую задачу: даже в наиболее загруженные часы дня абонент, желающий сделать вызов, не должен ждать сигнала готовности перед тем, как начать набор номера.

В течение нескольких лет инженеры Белловских лабораторий изучали проблемы телефонной коммутации в Белловской системе и установили, что для определения количества телефонных устройств, необходимых для обслуживания известного числа телефонных вызовов при любой желательной вероятности задержки, можно пользоваться математической формулой, называемой формулой Пуассона. Эта формула и была принята для расчета количества линейных искателей, причем в качестве критерия обслуживания была выбрана задержка одного из каждой тысячи вызовов в наиболее загруженный час дня.

Рабочие таблицы для инженеров были составлены таким образом, что, зная среднее количество вызовов, поступивших от группы абонентов в загруженный час, и среднюю продолжительность разговора в секундах, можно легко определить нужное количество линейных искателей. После того как было введено в действие несколько телефонных станций, рассчитанных по этому методу, специалистам по телефонии стало ясно, что некоторое сокращение оборудования не приведет к заметному ухудшению обслуживания абонентов. В связи с этим были составлены новые расчетные таблицы, основанные на задержке одного вызова из ста в наиболее загруженный час дня (табл. 13.1).

Таблица 13.1

Загрузка линейных искателей на основе формулы Пуассона

Вероятность задержки = 0,01

Количество линейных искателей	Количество абонентских линий	Занятость, %
2	5	6,9
3	16	14,8
4	30	20,8
5	46	25,5
10	149	41,4
15	269	49,9
20	399	55,5
25	535	59,5
30	675	62,5
35	818	65,0
40	964	66,9
45	1 112	68,7
50	1 261	70,1
60	1 565	72,5
70	1 872	74,3
80	2 184	76,0
90	2 499	77,1
100	2 816	78,2

Этот критерий обслуживания оказался удовлетворительным и использовался в течение примерно двух десятилетий, вплоть до II мировой войны. Во время войны Управление по военному производству призвало телефонную промышленность содействовать экономии материалов путем сокращения расходов дефицитного сырья на телефонных заводах. В результате переговоров с Управлением была определена возможность снижения потребления материалов, используемых при изготовлении оборудования телефонных станций, примерно на 5%. Применительно к оснащению станций линейными искателями это новое правило означало возрастание вероятности задержки до двух вызовов на каждые сто в наиболее загруженный час дня.

Вследствие очень большой потребности в телефонной связи во время и после войны многие телефонные станции оказались перегруженными. В ряде городов качество обслуживания вызовов упало ниже нормы, согласованной с Управлением по военному производству в качестве приемлемой при данных обстоятельствах. Чтобы держать положение под контролем, были разработаны программы периодической проверки качества обслуживания. Такая проверка выполнялась с помощью контрольных вызовов, осуществляемых в наиболее загруженный час дня, и хронометрирования задержки в получении сигнала готовности станции. Инструкции предусматривали запись общего числа сделанных контрольных вызовов и числа случаев, когда сигнал готовности станции не был слышен в течение 3 *сек*.

Этот метод оказался весьма успешным и позволил вскрывать слабые звенья, требующие усиления. В результате вместо старого критерия, основанного на проценте вызовов, встретивших задержки какой бы то ни было величины, был предложен новый критерий, основанный на проценте вызовов, встретивших задержки в 3 *сек* или более. Исходя из существовавших условий, руководство Белловской системы решило, что 1,5% вызовов с задержками в 3 или более секунд в наиболее загруженный час дня следует считать приемлемой нормой. На этой основе были составлены новые расчетные таблицы, которые оказались по счету четвертыми в серии расчетных таблиц, применявшихся в Белловской системе с 1920 г. Сравнение этих четырех отличных друг от друга таблиц может представить интерес (табл. 13.2).

Таблица 13.2

Количество линейных искателей в группе

Количество линий в группе	На основе пропорции задержек			На основе 1,5% задержек свыше 3 *сек*
	0,001	0,01	0,02	
160	13	11	10	10
200	15	12	12	12
240	17	14	13	13
280	18	16	15	15
320	20	18	16	16
360	22	19	18	18
400	23	20	19	19
440	25	22	21	21
480	27	23	22	22
520	28	25	24	23
560	30	26	25	25
600	31	28	26	26

Сразу же по окончании II мировой войны Белловские телефонные лаборатории приступили к разработке новой телефонной коммутационной системы, воплощающей многие радикально новые черты. (Эта система известна в телефонной промышленности под названием координатной системы № 5.)

В этой системе «линейные искатели» не применяются. Когда абонент снимает трубку своего телефонного аппарата, одно из группы устройств, называемых «маркерами» и обслуживающих все линии на станции,

выбирает один незанятый «регистр» и подключает вызывающую линию к этому регистру. Затем маркер освобождается приблизительно через 0,5 *сек* и готов к обслуживанию других вызовов. Регистр посылает сигнал готовности к вызывающему абоненту и остается подключенным к абонентской линии лишь такое время (приблизительно 12 *сек*), чтобы абонент окончил набор вызываемого номера и чтобы информация об этом вызываемом номере прошла к другому оборудованию, которое осуществляет соединение.

Было признано, что разработка расчетных методов для определения надлежащего числа маркеров и регистров в новой коммутационной системе потребует значительных исследований и анализа, так как совершенно новая конструкция системы приводит к новым проблемам, не встречавшимся ранее, при применении старых систем. В связи с этим в Беллловских телефонных лабораториях была организована небольшая группа, на которую была возложена задача разработки расчетных методов для этой новой системы.

Было бы естественно и логично стремиться к тому, чтобы разрабатываемая новая система обеспечивала такое же качество обслуживания, какое оказалось в общем приемлемым для миллионов абонентов, обслуживаемых другими системами. Однако было ясно, что от новой системы нельзя получить во всех отношениях то же качество обслуживания, как и в старых системах. Для объяснения этого потребуется краткая экскурсия в область теории группообразования.

Теория группообразования соединительного оборудования

В телефонии широко применяются математические формулы, позволяющие предсказать вероятность задержки любой длительности при получении какого-либо незанятого устройства, если известна нагрузка, приходящаяся на группу устройств, количество устройств в группе и средняя продолжительность каждого занятия устройства. Кривые, построенные по таким формулам, показаны на рис. 13.2.

Заметим, что при увеличении «занятости» (т. е. относительного объема сообщения, приходящегося на устройство, или процента времени, в течение которого устройство занято) вероятность задержки также увеличивается. Это, конечно, вполне естественно. Заметим также, что продолжительность задержки показана на чертеже в виде кратных от времени занятия, т. е. при

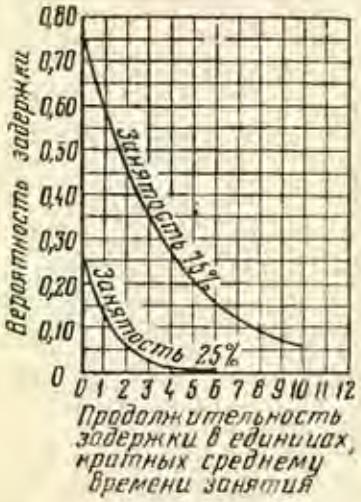


Рис. 13.2. Продолжительность задержки в единицах, кратных среднему времени занятия.

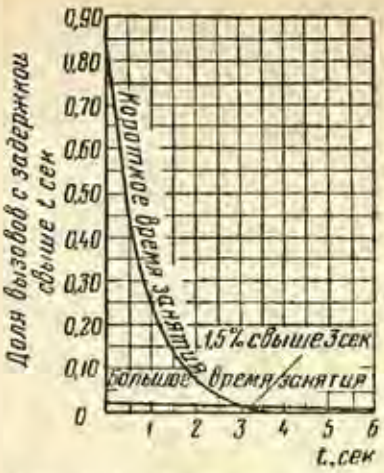


Рис. 13.3. Типичное распределение задержек для устройств с коротким и долгим временем занятия.

прочих равных условиях, задержка тем больше, чем больше среднее время занятия. Это тоже понятно...

Если интересующая нас норма задержки есть лишь небольшая доля от среднего времени занятия, то почти все задержанные вызовы будут задерживаться по крайней мере до достижения нашей нормы. Так и было в старых телефонных системах, где трехсекундная задержка, являвшаяся критерием качества работы системы, составляла только около 2% среднего времени занятия применяемых устройств. Это время занятия устройства определялось средней продолжительностью разговора и могло лежать в пределах от 120 до 240 *сек*. При таких условиях продолжительность большинства задержек составляла не менее 3 *сек* и потому не представляло особой важности, выбирать ли в качестве нормы задержки 3 *сек* или более короткий интервал.

Это легко можно заметить по кривой рис. 13.2. При режиме работы, изображаемом верхней кривой, 75 вызовов из ста встретят какую-то задержку, и практически все эти задержанные вызовы будут задержаны не менее чем на 2% от среднего времени занятия.

В новой системе устройства, участвующие в посылке сигнала готовности станции, отличаются очень коротким временем занятия. При таких условиях число вызовов, задержанных на 2 *сек* или менее, может во много раз превышать число вызовов, задержанных на 3 *сек* или более, как видно из рис. 13.3. Эти кривые построены по тем же самым формулам, как и кривые на предыдущем рисунке, но при других предположениях относительно занятости и числа применяемых устройств. Выбран также и другой масштаб для оси абсцисс: если раньше на этой оси откладывались кратные от времени занятия, то сейчас на ней отложены секунды.

В случае устройства с продолжительным занятием дело сводится к тому, чтобы взять очень небольшой отрезок на левом конце обычной кривой задержки и сильно его увеличить, так что наклон этого отрезка кривой станет едва заметным. Кривая для устройств с небольшим временем занятия сохраняет привычный крутой наклон. Как легко заметить, пересечение кривых происходит в традиционной точке выбора нормы обслуживания, при 1,5% всех вызовов свыше 3 *сек*. Из этого чертежа хорошо видно, что новую систему, использующую аппаратуру с небольшим временем занятия, нельзя сделать так, чтобы она обладала во всех отношениях таким же качеством обслуживания, как и старая система; действительно, хотя эти системы можно сделать сравнимыми по задержкам какой-то одной

выбранной величины, однако распределение задержек около этой выбранной нормы будет крайне различным.

Пробное сравнение ожидаемого качества работы старой и новой систем приведено в табл. 13.3. Эта таблица составлена в предположении, что в обеих системах будет одинаковое количество задержек в 3 и более секунды и что в новой системе половина таких задержек будет вызвана маркерами и половина — регистрами. В то время как вероятность задержек в 3 *сек* и более для обеих систем одинакова, количество задержек в

Таблица 13.3

Задержки в старой и новой системах

Пробное сравнение № 1

	Занятость, %	Вероятность задержки, превышающей			
		0 *сек*	1,5 *сек*	2 *сек*	3 *сек*
Старая система	58	0,02	0,018	0,017	0,015
Новая система, маркеры	92	0,80	0,080	0,035	0,0075
Новая система, регистры	85	0,11	0,028	0,018	0,0075
Новая система, полностью	—	0,82	0,106	0,042	0,015

2 *сек* и более в новой системе будет более чем в два раза превышать число таких задержек в старой системе, а количество задержек, превышающих 1,5 *сек*, будет в новой системе в 6 раз больше, чем в старой.

Встал вопрос о пересмотре пригодности старого эталона для измерения качества обслуживания. Старый эталон — процент вызовов, задержанных более чем на 3 *сек*,— был выбран потому, что такой период времени удобно измерять секундомером. Как уже показано, норма задержки в 2 *сек* или менее дала бы почти такие же результаты в старых системах, но в новой системе результаты изменялись бы значительно в зависимости от выбранного интервала.

Рис. 13.4. Поведение абонентов при наборе номера. Время с момента снятия трубки до посылки первого импульса в случае хорошего обслуживания вызовов сигналами готовности станции:

4 городские АТС, 8446 вызовов, 1945—1946 гг.; — — — подстанция 3 (Бруклин), 2535 вызовов, 1949 г.; ——— Омаха, 1923 г.; ——— Нью-Хейвен, 1266 вызовов, 1942 г.; - - - - Мидия, 3217 вызовов, 1948—1949 гг.; ▬ результирующая кривая.

Имелись данные, собранные в нескольких различных городах за многолетний период, показывающие нормальное распределение интервалов времени от снятия телефонной трубки до начала набора номера абонентами, привыкшими к хорошему обслуживанию сигналами готовности (рис. 13.4). Эти данные ясно говорили о том, что в качестве критерия качества обслуживания было бы всего логичнее выбрать задержки в получении сигналов готовности, равные 2 *сек*, так как для большинства людей момент начала набора номера лежит обычно в полосе, ширина которой равна одной секунде, а середина совпадает с ординатой 2 *сек*. Чтобы быть уверенным в том, что наличные данные не устарели вследствие каких-либо недавних существенных изменений в привычках абонентов, были проведены новые измерения на нескольких телефонных станциях. Как видно по рис. 13.4, новые данные поразительно хорошо совпадали с распределением старых.

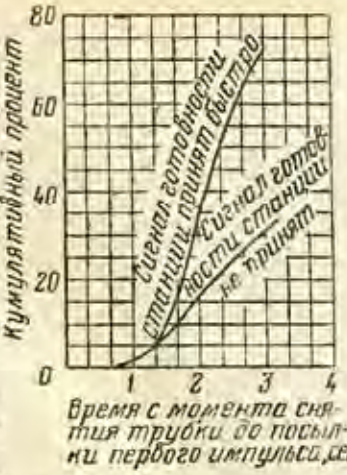


Рис. 13.5. Поведение абонента при наборе номера. Время с момента снятия трубки до посылки первого импульса (г. Мидия, штат Пенсильвания, 1948—1949 гг).

В свете этих исследований казалось правдоподобным, что задачу создания новой системы с качеством обслуживания, признанным приемлемым для старой системы, было бы легче всего решить, если бы новая и старая системы были сделаны сравнимыми в отношении задержек в 2 *сек* и более. Однако с практической точки зрения необходимость изменения установленного критерия на такую небольшую и, по-видимому, малозначащую величину, как одна секунда, вызывала большие сомнения. Если абоненты действительно ждут сигнала готовности станции, как это им рекомендуют телефонные компании, то, конечно, незначительная задержка с началом набора номера не имеет никакого значения и вызовы не будут существенно задержаны.

Возможность подвергнуть этот вывод проверке представилась во время исследования, которое проводилось при первом введении в действие новой системы в г. Мидия, штат Пенсильвания*. В ходе этого исследования сигналы готовности для определенной части вызовов задерживались на различные промежутки времени и затем регистрировалось поведение абонентов при наборе номера.

Результаты исследования приводятся на рис. 13.5. Верхняя кривая представляет собой кумулятивный график поведения тех жителей Мидии, которые обслуживались сигналами готовности хорошо. На предыдущем рисунке было показано некумулятивное распределение этих интервалов. Для сравнения на рис. 13.5 приведена также вторая кривая, показывающая процент наблюденных случаев, когда абоненты при задержках сигнала готовности начинали набирать номер, не дожидаясь появления сигнала. Несмотря на большой процент абонентов, умышленно увеличивавших интервал времени между снятием трубки и началом набора номера, все же количество лиц, начинавших набирать номер еще до по-

* Мидия — пригород Филадельфии, небольшой город в 20 *км* к западу от нее, около 6 тыс. жителей, является административным центром графства Делавер штата Пенсильвания. — *Прим. ред.*

Таблица 13.4

Задержка в старой и новой системах

Пробное сравнение № 2

	Занятость, %	Вероятность задержки, превышающей		
		0 *сек*	2 *сек*	3 *сек*
Старая система	58	0,02	0,017	0,015
Новая система, маркеры	89	0,72	0,0085	0,0009
Новая система, регистры	82	0,07	0,0085	0,0031
Новая система, полностью	—	0,74	0,017	0,0040

явления сигнала готовности, ввиду чего их вызовы оставались без последствий, оказалось весьма значительным. Очевидно, что для таких абонентов задержка сигнала готовности свыше 2 *сек* означала плохую работу телефонной станции.

Если бы новая система должна была сравняться со старой по качеству обслуживания при задержке в 2 *сек* и более, то табл. 13.3 необходимо было бы переделать, как это показано в табл. 13.4. Заметим, что в отношении трехсекундной задержки новая система давала бы почти в четыре раза лучшее обслуживание, чем старая: только 0,4% задержек такой продолжительности против 1,5% в старой системе.

Рассмотрим теперь более подробно цифры, приведенные под заголовком «занятость» в табл. 13.4. Из них следует, что если проектировать систему согласно этой таблице, то маркеры и регистры в наиболее загруженные часы дня будут работать с эффективностью соответственно в 89% и 82%. Другими словами, каждый маркер был бы загружен 80% часа, а каждый регистр — 82% часа.

Достаточно даже небольшой практики в телефонном деле, чтобы убедиться, что было бы ненадежно проектировать систему с таким небольшим запасом, как 12%. Число телефонных переговоров изменяется изо дня в день, из сезона в сезон и из года в год, в зависимости от экономического цикла, общественных привычек абонентов, погоды и многих других факторов. Инженерные оценки традиционно основывались на типично загруженных днях в загруженном сезоне года, однако в наиболее загруженные часы «пиковых» дней число переговоров может превышать и действительно значительно превышает среднее расчетное число переговоров.

Требовались статистические данные, характеризующие изменения числа переговоров изо дня в день. Соответственно на ряде телефонных станций были приняты меры для записи числа телефонных переговоров в наиболее загруженный час всех рабочих дней года. Из рис. 13.6 видно, что эти изменения в действительности намного превосходят указанный небольшой запас в 12% и в крайних случаях даже превышают 50%.

Рис. 13.6. Ожидаемые отклонения вызовов в загруженные часы сверх нормальных уровней загруженного сезона.

Ясно, что использовать устройства станции с полной эффективностью в средние дни нельзя, если мы хотим обеспечить приемлемое обслуживание абонентов в пиковые дни. Здесь было бы необходимо сократить занятость в средние дни, обеспечивая тем самым лучшее качество обслуживания в эти дни по сравнению со старыми системами. Однако критерий обслуживания для нормальных дней, применявшийся при проектировании старых систем, отнюдь не обязательно применять также и к «пиковым» дням, которые могут возникать только несколько раз в году. В эти редкие периоды работы можно допускать другое, несколько худшее качество обслуживания.

Решение вопросов технической политики, мы полагаем, не входит в обязанности исследователя операций, а этот вопрос — сколько истратить денег и какое при этом обеспечить качество обслуживания наших абонентов — является, несомненно, вопросом технической политики. Тем не менее руководство всегда предпочитает рассматривать конкретное предложение. Поэтому, исходя из данных, представленных на рис. 13.6, было предложено спроектировать новую систему на основе занятости 71,3% в часы с нормальной загрузкой, что обеспечивало бы 40%-ный резерв, способный поглотить все пики нагрузки, кроме, разве, самых высоких, в какие-нибудь два дня в году. Сравнение этой предложенной и старой системы по качеству обслуживания в течение года при 2-*сек* задержках дано в табл. 13.5.

Новая система, спроектированная предложенным образом, обеспечивала бы лучшее обслуживание в течение всего периода работы, за исключением семи пиковых дней в году.

Указанное сравнение двух систем было сделано с точки зрения относительной возможности для абонентов встретить затруднения в случае набора номера до появления сигнала готовности станции. Мы видели

Таблица 13.5

Задержки в 2 *сек* в старой и новой системах

	Число дней	Вероятность задержки в 2 *сек* или более	
		Старая система	Новая система
Нормальные дни	—	0,017	Пренебрежимо малая
Тяжелые дни:			
25% или менее сверх нормального . . .	33	0,016—0,12	0,08 или менее
25—35% сверх нормального	4	0,12—0,22	0,08—0,55
35% или более сверх нормального . . .	3	0,22 или более	0,55 или более

Таблица 13.6

Задержки в 10 *сек* в старой и новой системах

	Число дней	Вероятность задержки в 10 *сек* и более	
		Старая система	Новая система
Нормальные дни	—	0,01	Пренебрежимо малая
Тяжелые дни:			
25% или менее сверх нормального . . .	33	0,01—0,09	Пренебрежимо малая
25—35% сверх нормального	4	0,09—0,18	0,01—0,08
35% или более сверх нормального . . .	3	0,18 или более	0,08 или более

(рис. 13.5), что многие абоненты будут ждать сигнала готовности достаточно долго, прежде чем приступят к набору номера, вследствие чего они смогут эффективно использовать телефон даже в тех случаях, когда телефонная станция сильно перегружена. Ввиду этого руководство было бы заинтересовано в сравнении старой и новой систем с точки зрения тех абонентов, которые в редких случаях ненормально большой загрузки станции согласны ждать сигнала готовности некоторое умеренное время.

Для оценки быстроты ответа от телефонисток, работающих за ручными телефонными коммутаторами, в Белловской системе в течение многих лет используется нормативный интервал времени 10 *сек*. Если 10 *сек* дают приемлемый критерий для оценки потери времени на ожидание ответа телефонистки, то это значение можно считать также подходящим критерием потери времени на ожидание сигнала готовности автоматической телефонной станции. Из табл. 13.6 следует, что абоненты, которые согласны ждать сигнала готовности 10 *сек*, будут значительно лучше обслуживаться новой системой в течение всего времени, за исключением двух или трех дней в году.

На этом этапе работы мы решили, что имеем верное, логически корректное предложение для руководства, основанное на статистически надежных данных о поведении абонентов. Однако мы должны были доказать правильность теории, использованной при предсказании качества обслуживания абонентов новой телефонной системой. С этой целью были приняты меры к наблюдению и измерению задержек, действительно происходящих на новой АТС в г. Мидия. Рисунок 13.7 позволяет сравнить наблюденные и теоретически предсказанные задержки в условиях загрузки, имевших место в один определенный день. Были собраны данные за многие дни при различных условиях загрузки и разных количествах действовавшего оборудования, и всюду было обнаружено одинаково хорошее соответствие между наблюденными данными и теорией ...

Возникновение и формулировка задачи. Интересно проследить изменение постановки задачи с ходом исследования. Данная конкретная задача возникла в связи с появлением нового изобретения — координатной системы № 5. Однако до этого существовала старая система, и естественно было провести аналогию между этими двумя системами и начать исследование с предположения, что задачи этих систем одинаковы. Таким образом, первоначально поставленная задача заключалась в устранении задержек сигнала готовности в течение наиболее загруженного часа среднего дня. Однако в окончательной постановке задача уже заключалась в устранении задержек сигнала готовности в течение наиболее загруженного часа наиболее загруженного дня в году. Применительно к старой системе эти две задачи очень схожи между собой, но для новой системы вследствие кратких времен занятия они существенно различны.

Как указывалось раньше, критерий эффективности очень тесно связан с постановкой задачи, но не тождествен с ней. В табл. 13.2 мы находим различные критерии эффективности, примененные к старой системе, с несколько различным количеством потребного оборудования в зависимости от выбранного критерия и его минимального допустимого значения. В табл. 13.5 используется третий критерий (основанный на окончательной постановке задачи), для которого также можно принять различные минимальные значения. Следует отметить, что при сравнении двух систем вывод о том, какая из них «лучше», зависит от используемого критерия.

Заметим еще, как точка зрения проектировщика системы влияет на формулировку задачи: с началом войны критерий эффективности изменился. К тому же автор интересуется только центральной станцией автоматической телефонной системы, и даже только той ее частью, о которой он говорит в первом абзаце статьи.

Интерес представляет также степень влияния другого оборудования телефонной станции. Из табл. 13.3 следует, что короткие задержки вызываются в основном маркерами (среднее время занятия 0,5 *сек*), в то время как длительные задержки (свыше 3 *сек* в нашем случае) вызываются в основном регистрами (среднее время занятия 12 *сек*). Другие устройства (например, промежуточные и соединительные линии, § 2.2), время занятия которых составляет 2 или 3 *мин*, будут определять вероятности очень больших задержек при необычно резком увеличении количества переговоров. Наконец, в статье не исследовалась гибкость системы, т. е. легкость или трудность изменения системы (например, уста-

новки дополнительного маркера) в случае, если в будущем нагрузка станции увеличится. Все эти вопросы, являющиеся частями всей реальной задачи, были, несомненно, рассмотрены, но за недостатком места не вошли в приведенную нами статью.

Первоначальный критерий эффективности (никогда никакой задержки) не имеет смысла. Последующие критерии (вероятность любой задержки в наиболее загруженный час; вероятность задержки свыше заданной длительности в наиболее загруженный час; вероятность задержки свыше заданной длительности в наиболее загруженный час наиболее загруженного дня) постепенно усложняются, но в то же время становятся и более полными. Окончательно выбранный критерий представляет собой в итоге некоторый комплекс вероятностей задержек различных длительностей (2 *сек* и 10 *сек*) в два или три наиболее загруженных дня в году. Этому критерию не достает только простоты. Он является количественным (в такой степени, что даже можно определить понятие «наиболее загруженный день»), имеет небольшую дисперсию (рис. 13.7), является полным, обладает физическим смыслом, характеризует идеальную работу системы (которая соответствует нулевой вероятности, по крайней мере для продолжительных задержек), а также непосредственно связан с измеряемыми переменными.

Математическая модель. Рассмотрение математической модели занимает бо́льшую часть статьи. Эта модель является аналитической вероятностной моделью, основанной на законе Пуассона (гл. 5) и теории массового обслуживания (гл. 23). (Хотя автор и не пишет об этом, Белловские телефонные лаборатории использовали также модель «Монте Карло» системы № 5 перед ее установкой в эксплуатацию. Так как в то время универсальные цифровые электронные вычислительные машины не могли быть использованы, то для имитации появления вызовов и различных коммутационных операций была сконструирована электромеханическая аналоговая машина. Управление этой машиной осуществлялось четырьмя операторами, которые помогали ей непрерывно изображать состояния занятия и освобождения в последовательные моменты времени для 10 тыс. абонентских линий и нескольких тысяч несущих информацию каналов и имитировать соответствующую работу аппаратуры управления и сигнализации.)

Сила аналитической вероятностной модели иллюстрируется графиками на рис. 13.2, 13.3 и 13.4. Картина на рис. 13.2 носит настолько общий характер, что ее можно применить

Рис. 13.7. Характеристики задержек сигнала готовности для АТС координатной системы № 5.

Уровень нагрузки — 2570 вызовов в загруженный час, занятие маркеров 65%, занятие регистров 70%. Кривые: *M* — задержка из-за маркеров, *R* — задержка из-за регистров, *C* — результирующие задержки. Данные: 150 контрольных вызовов, 10.00—11.00 до полудня, 31/I 1949 г., Мидия, Пенсильвания.

к другим задачам по массовому обслуживанию (т. е. она не ограничена только телефонной «теорией группообразования соединительного оборудования»), при условии, что имеется только одна очередь (линия ожидания).

При наличии нескольких параллельных очередей кривые сдвинутся вниз. Так, по графику на рис. 13.2 можно предсказать, что при 75%-ной занятости маркер (время занятия 0,5 *сек*) даст 28%-ную вероятность двухсекундной задержки; однако из табл. 13.5 видно, что при занятости приблизительно в 75% и пяти параллельных очередях (т. е. пяти маркерах) вероятность такой задержки пренебрежимо мала. Рисунок 13.3 дает представление о различных влияниях продолжительных и коротких времен занятия. Рисунок 13.7 показывает замечательную способность предсказания экспериментальных результатов по теоретической модели.

Сбор данных. Была проделана большая экспериментальная работа. Некоторые из полученных данных использовались для подстановки в математическую модель и касались в первую очередь работы системы, другие опыты подпадали под рубрику технической психологии (гл. 30) и измеряли привычки абонентов.

Некоторые более ранние опыты подразумевались, но не были описаны, а именно опыты, доказывающие, что вызовы возникают в соответствии с распределением Пуассона. Необходимо было также провести экс-

периментальные работы для определения математического ожидания этого распределения применительно к конкретной телефонной станции и для выявления изменений этого математического ожидания от часа к часу и изо дня в день. Наконец, было необходимо определить поведение абонентов при наборе номера. Чтобы получить данные, приведенные на рис. 13.5, потребовалось нарушить в некоторой степени работу системы, однако при сборе остальных данных необходимости в этом не было.

Следует обратить внимание на сбор данных методом «инспектирования». Чтобы оценить пригодность оборудования во время войны, действовавшие компании получили точные данные в соответствии с принятым критерием эффективности путем подачи контрольных вызовов в наиболее загруженные часы, не полагаясь на чье-либо мнение, что обслуживание было хорошим или плохим. Такой тип контроля весьма желателен при эксплуатации системы, чтобы определять, какие улучшения действительно необходимы.

Анализ данных. Так как автор не мог предполагать, что его читатели знакомы с математической статистикой, он избегал специальных терминов. Однако следует отметить, что уже в четвертом абзаце он определил рассматриваемые им ошибки I и II рода: с одной стороны, плохое обслуживание, с другой, потеря денег на ненужное оборудование. Точной стоимости ошибки каждого рода в статье не приводится, по-видимому, в связи с тем, что автор считает, что он не может правильно оценить стоимость ошибки I рода, так как при этом пришлось бы обсуждать такие отдаленные вопросы, как возможность воздействия со стороны правительства при ухудшении качества обслуживания.

Однако автор представляет руководству конкретное предложение (что всегда следует делать); мы полагаем, что это предложение должно было бы включать объяснение того, каким образом оно сводит к минимуму ожидаемые потери из-за ошибок обоих родов, т. е. оно должно было бы указывать стоимость и вероятность ошибки каждого рода, с тем чтобы руководство могло учесть различные причины, явившиеся основанием для выбора предлагаемого варианта проектируемой системы, и оценить их в свете своего собственного опыта.

Обратим внимание на рис. 13.4: поражает замечательная стабильность данных о человеческом поведении, разбросанных по периоду в 25 лет и расстоянию в 1000 миль. В интерпретации этих данных допущена интересная логическая ошибка. Задержка в 2 *сек* выбрана потому, что большинство абонентов «начинает набирать номер в моменты времени, лежащие в пределах односекундной полосы времени, серединой которой служит двухсекундная ордината». Однако не это является предметом рассмотрения. Вопрос заключается в том, сколько абонентов начинает набирать номер до окончания двухсекундного периода. (График на рис. 13.5 показывает, что число таких абонентов составляет одну треть, т. е. достаточно большой процент.) Это еще не значит, что 2-*сек* критерий был выбран неправильно; ведь выбор, например, односекундного критерия мог оказаться неоправданно дорогим, а с другой стороны, этот критерий в любом случае применяется только к нескольким нечетко определенным «наиболее загруженным часам». Однако это лишь объяснение, а не настоящее оправдание.

ЧАСТЬ 4

ТЕОРИЯ ВЫЧИСЛИТЕЛЬНЫХ МАШИН — ОСНОВНОЕ ОРУДИЕ ВНУТРЕННЕГО ПРОЕКТИРОВАНИЯ СИСТЕМ

ГЛАВА 14

ВВЕДЕНИЕ В ТЕОРИЮ ВЫЧИСЛИТЕЛЬНЫХ МАШИН

Мы определяем вычислительную машину как машину, выполняющую физические операции, которые могут быть описаны посредством математических операций, и используемую для осуществления этих математических операций*. Вычислительная машина является не только одним из главных элементов в любой системе большого масштаба — она является также одним из главных орудий при проектировании систем. Связь между системами большого масштаба и вычислительными машинами является настолько тесной, что в прошлом иногда наблюдалась тенденция пренебрегать целостным подходом к системам ради гипнотизирующего сосредоточения на вычислительных машинах, составляющих центральный управляющий элемент системы.

14.1. Применение вычислительных машин

В настоящее время вычислительные машины выполняют многие другие важные функции в системной технике; мы различаем четыре главных применения вычислительных машин: вычисление, управление, моделирование и контроль.

Вычисление. Это наиболее очевидное применение вычислительных машин имеет две формы: решение формальных математических задач и обработка информации. Примерами первой формы являются: вычисление таблиц сферических тригонометрических функций для навигационных задач и таблиц сфероидальных волновых функций для определения радиолокационных сечений, решение систем неравенств в линейном программировании, обращение матриц и определение полюсов рациональных функций при проектировании следящих систем.

Обработка информации имеет особо важное значение при эксплуатационных испытаниях систем. Даже сравнительно простые испытания могут привести к сбору и записи миллионов элементов данных; при помощи вычислительных машин эта масса данных быстро сводится к небольшому набору чисел или функций. Даже такая простая задача, как вычисление дисперсии результатов большой выборки, может занять один час ручной работы; более сложные функции, например связанные с дисперсионным анализом, вообще были бы неприменимы к большим массивам данных, если бы не было быстродействующих вычислительных машин.

Управление. В § 3.3 мы провели различие между элементами логического управления и элементами рефлексивного управления. Этими элементами в обоих случаях служат вычислительные машины. В первом случае вычислительная машина является «мозгом» системы. В центре каждой системы большого масштаба находится некоторый решающий (распорядительный) механизм. Если функция логического управления системы не автоматизирована, то этим механизмом является человеческий мозг (например, диспетчер в транспортной системе). Если система автоматизирована, то обычно большинство решений принимает вычислительная машина, хотя в некоторых системах (например, в телефонной системе) вычислительная машина может быть совершенно непохожа на обычные универсальные вычислительные машины, которые вошли в обиход лабораторий.

В случае рефлексивного управления также существует аналогия с человеком. Только в этом случае вычислительная машина не при-

* В русской технической литературе вычислительные машины называются также «счетными машинами», «математическими машинами» и «счетнорешающими устройствами». — *Прим. ред.*

нимает логических решений, а выполняет рефлексное действие. Вычислительные машины в этом случае сравнительно просты и в большинстве случаев являются аналоговыми устройствами. Примером такой вычислительной машины может служить автопилот. В автопилоте замеренное отклонение самолета от заданного курса преобразуется в напряжение, которое, действуя на сервомотор, вращающий руль поворота, возвращает самолет на заданный курс.

Моделирование. Эта функция вычислительных машин является менее очевидной и принимает несколько форм. После того как сформулирована математическая модель, любую часть системы можно представить на вычислительной машине. Выходы (выходные величины) из одной части системы используются как входы (входные величины) в одну или несколько других частей с надлежащим учетом петель обратной связи. Таким путем целая сложная система может быть представлена на вычислительной машине сравнительно просто.

В этой книге описано много видов моделирования, особенно в гл. 10, 18 и 23. При моделировании системы ее параметры могут варьироваться почти произвольно. Шум (помехи) также может довольно легко моделироваться, и его амплитуда также может варьироваться. Или же может быть промоделирована серия опытов без шума и затем сравнена с серией опытов с шумом. Существует почти бесконечное разнообразие методов, которые могут быть использованы при разных видах моделирования; некоторые из этих методов рассмотрены ниже. Существует важное различие между моделированием, осуществляемым в «реальном» времени, и моделированием, осуществляемым в «замедленном» или «ускоренном» времени; в первом случае модель системы выполняет все операции с той же скоростью, с какой они выполняются в действительной системе.

По мере того как с ходом проектирования изготавливаются отдельные компоненты системы, мы можем найти желательным подвергнуть испытанию один или несколько из них, не дожидаясь окончания изготовления всей системы. Это обычно оказывается возможным, если только система может быть промоделирована в реальном времени. А именно, мы удаляем ту часть моделирующего оборудования, которая соответствует испытуемому компоненту, и включаем реальный компонент в цепь моделирования, используя на его входе и выходе соответствующие преобразователи.

Такие испытания часто оказываются полезными даже при наличии полностью изготовленной системы: во-первых, потому, что действительные испытания в реальных эксплуатационных условиях обходятся очень дорого, а во-вторых, благодаря большому удобству получения при моделировании широкого диапазона изменения параметров. Пусть, например, мы желаем испытать сервомотор, управляющий движением элерона самолета. Можно установить этот сервомотор на реальном самолете и провести испытания в воздухе, но это дорого и ненадежно. Кроме того, мы должны будем испытать сервомотор в особо трудных условиях, например в условиях, возникающих при аварии или в воздушном бою, но мы не хотим подвергать самолет и летчика таким опасностям. И наконец, возможно, что мы пожелаем провести испытания в широком диапазоне условий окружения (например, при тропических температурах), которые нелегко встретить в месте испытаний.

Важным компонентом, который часто испытывается таким образом, является сам человек. Очень трудно предсказать заранее количество ошибок и задержек, которые будут совершены такими операторами, как диспетчер в транспортной системе, или оператор радиолокатора в системе наземного управления посадкой самолетов, или кассир в системе резервирования мест.

И особенно трудно предсказать тенденции появления этих ошибок и задержек в тех случаях, когда работу оператора затрудняет густое движение, шум в окружающей среде или другие факторы. В этих случаях реального оператора можно поместить перед реалистическими органами индикации и управления, моделируя всю остальную систему на вычислительной машине.

Системы большого масштаба, согласно нашему определению этого термина, имеют множественные входы; возьмем, например, вызовы, поступающие на телефонную станцию от абонентов, или цели (истинные и ложные), появляющиеся в военной системе. После завершения постройки такие системы всегда требуют испытания перед пуском в настоящую работу, так как в них всегда имеются «недоделки», которые должны быть устранены. Такие испытания часто производятся путем моделирования входов системы, и вследствие множественности и разнообразия входов для моделирования их требуется большая вычислительная машина. Хотя некоторые погрешности системы выявляются немедленно при таком испытании, однако

прежде чем достигается уверенность, что система работает правильно, часто требуется полностью обработать результаты испытаний, а это может занять месяцы.

Кроме испытания системы, должна быть произведена подготовка (тренировка) операторов для нового оборудования, и в случае систем большого масштаба такая подготовка требует больших усилий. Подготовка операторов может быть совмещена с испытаниями системы; операторы, которые управляют системой во время ее испытаний, могут получить при этом подготовку, необходимую для действительной работы на системе после ее пуска в действие.

В военных оборонительных системах при отсутствии вражеских нападений такие испытания должны производиться периодически как для подготовки новых пополнений, так и для поддержания у остального персонала способности работать при пиковых нагрузках. Некоторые невоенные системы можно считать в этом смысле «оборонительными» против природы; так, например, система наземного управления посадкой самолетов будет сильно загружена только при особо плохой погоде, но ее операторы должны быть натренированы для таких пиковых нагрузок.

В меньших системах знаменитый аэротренажер «Линк» и менее известные, но весьма совершенные современные тренажеры экипажей реактивных самолетов являются примерами тренировочных устройств, в которых применяются вычислительные машины для моделирования входов вместе с надлежащей обратной связью от действий оператора. Операторы радиолокационных станций также нуждаются в частой тренировке с помощью моделирующих устройств, так как только таким путем они могут сохранить умение распознавать сигналы от разнообразнейших целей, которые могут им встретиться на практике.

Контроль. Одним из недостатков автоматики является ее негибкость перед лицом частичной аварии или непредусмотренных условий или входов. Для борьбы с этим недостатком некоторые системы большого масштаба снабжаются контрольной аппаратурой, которая и сама по необходимости также является автоматической. Для такого контроля может быть использован механизм логического управления в центре системы, однако обычно оказывается более целесообразным использовать для этого отдельную вычислительную машину.

Контрольная аппаратура составляет также часть рабочего испытательного оборудования, где она используется для записи результатов; записанные результаты могут использоваться как входы для будущих испытаний, а также как материал для последующего изучения возможностей дальнейшего улучшения системы. Примером контрольной аппаратуры является упомянутая ранее система автоматического ремонта в телефонной системе.

14.2. Определение аналогового и цифрового устройств

Если говорить возможно проще, то цифровое устройство считает, а аналоговое устройство измеряет. Это различие действительно носит принципиальный характер, вытекая из математического различия между дискретными (цифровыми) и непрерывными (аналоговыми) величинами. Это математическое различие приводит к такой разнице в методах, как между суммированием и интегрированием или между уравнениями в конечных разностях и дифференциальными уравнениями. Столь же большую разницу мы найдем и между цифровыми и аналоговыми вычислительными машинами в их возможностях и в их операциях.

Примерами цифровых вычислительных машин (в порядке возрастающей сложности) могут служить обычные счеты, настольная счетная машина, счетно-перфорационные машины и современная электронная цифровая вычислительная машина. Последняя из них представляет для нас главный интерес. Хотя она отличается от более простых цифровых устройств только количественно, тем не менее это различие столь велико, что приводит на практике к важным качественным различиям, подобно тому как количественное различие в длине волны между радиоволнами и светом приводит к очевидным качественным различиям в их свойствах. В цифровых вычислительных машинах обычно основными операциями являются сложение и вычитание, все остальные математические операции получаются из надлежащих повторений этих основных операций.

Примерами аналоговых вычислительных машин могут служить счетная линейка, механический дифференциальный анализатор, электромеханическая аналоговая вычислительная машина и электронная аналоговая вычислительная машина. Для нас главный интерес представляет электромеханическая вычислительная машина, хотя мы отметим также особенности работы полномеханических и полноэлектронных устройств.

Каждое аналоговое вычислительное устройство основано на каком-нибудь физическом принципе, математическое описание которого нам известно; физическая реализация используется тогда для выполнения математических операций. В счетной линейке физический принцип состоит в сложении длин: два числа, которые мы хотим перемножить, преобразуются в длины, пропорциональные их логарифмам, и сумма этих логарифмов есть логарифм произведения. Таким путем мы осуществляем физическую аналогию сложения логарифмов или умножения антилогарифмов.

В рассматриваемых ниже аналоговых устройствах числа представляются или в виде электрических напряжений, или в виде угловых поворотов валов. Однако в качестве аналога числа может быть использовано любое физически измеримое явление. Были построены практические аналоговые вычислительные машины, основанные на законах химии, гидравлики, магнетизма, электричества, электроники, оптики, механики и акустики.

Многие важные математические операции допускают непосредственную реализацию на аналоговых машинах, в зависимости от используемого физического закона. В электромеханических аналоговых машинах основными операциями являются сложение, интегрирование (ограниченное интегрированием по времени), умножение и образование синуса, косинуса, арксинуса и арккосинуса; некоторые другие функции могут быть образованы несколько сложнее. Возможность производить такие операции непосредственно, а не в виде длинных последовательностей сложений и вычитаний, как в цифровых вычислительных машинах, имеет большое значение. Особенно большое значение имеет простота выполнения операции интегрирования на таких машинах *.

14.3. Системы счисления

В некоторых областях системотехники — главным образом в технике цифровых вычислительных машин, но также в теории информации и в технике связи — числа представляются не в обычной десятичной, а в двоичной системе счисления. Некоторые другие системы счисления, хотя более редки, но также достаточно важны для того, чтобы познакомить с ними проектировщика систем.

Вообще, можно использовать любую систему записи чисел, лишь бы она была однозначна. Вместо обычной арабской десятичной системы можно было бы пользоваться даже римскими цифрами, однако римские цифры крайне неудобны при чтении и выполнении арифметических операций. Мы рассмотрим некоторые системы, в которых, как и в обычной десятичной системе, значение каждой цифры определяется ее положением, но в которых используется другое основание, чем 10; мы рассмотрим также одну систему, в которой, как и в римской системе, положение играет другую роль.

Основание системы счисления (например, число 10 в десятичной системе) обозначается символом r. В системе должно быть r различных символов (*цифр*), каждому из которых придается одно из r значений от 0 до $r-1$ (например, от 0 до 9 в десятичной системе). Число записывается в виде последовательности k разрядов («знаков»), каждый из которых есть один из этих r символов. Эти разряды располагаются справа налево от самого младшего разряда к самому старшему; самый младший разряд является нулевым, самый старший является $(k-1)$-м. Такая последовательность цифр обозначает собой число

$$x=\sum_{i=0}^{k-1} a_i r^i, \qquad (14.1а)$$

где каждое a_i должно иметь значение одного из наших r символов.

Неправильные дроби можно записать по формуле

$$x=\sum_{i=-j}^{k-1} a_i r^i. \qquad (14.1б)$$

При этом положение разряда a_0 в последовательности цифр отмечается запятой, которая ставится справа от такого разряда. В десятичной системе эта запятая называется *десятичной запятой*, а в двоичной системе — *двоичной запятой*. Так, например, в десятичной системе символ 201,1 обозначает число

$$x=2\cdot 10^2+0\cdot 10^1+1\cdot 10^0+1\cdot 10^{-1}=$$
$$=2\cdot 100+0\cdot 10+1\cdot 1+1\cdot \frac{1}{10},$$

но в троичной системе (основание 3) этот символ имеет следующее десятичное значение:

$$x=2\cdot 3^2+0\cdot 3^1+1\cdot 3^0+1\cdot 3^{-1}=$$
$$=2\cdot 9+0\cdot 3+1\cdot 1+1\cdot {}^1/_3=19{,}333\ldots, \qquad (14.2)$$

* Цифровые машины называются также «дискретными машинами» или «машинами дискретного счета», а аналоговые машины — «моделирующими устройствами» или «машинами непрерывного действия». — *Прим. ред.*

а в двоичной системе (основание 2) символ 201,1 вообще не имеет смысла.

Таблица 14.1

Первые двадцать одно целое число в шести различных системах счисления

Десятичная	Восьмеричная	Троичная	Двоичная	Рефлексный код	Двоично-десятичная
00	00	000	00000	00000	000000
01	01	001	00001	00001	000001
02	02	002	00010	00011	000010
03	03	010	00011	00010	000011
04	04	011	00100	00110	000100
05	05	012	00101	00111	000101
06	06	020	00110	00101	000110
07	07	021	00111	00100	000111
08	10	022	01000	01100	001000
09	11	100	01001	01101	001001
10	12	101	01010	01111	010000
11	13	102	01011	01110	010001
12	14	110	01100	01010	010010
13	15	111	01101	01011	010011
14	16	112	01110	01001	010100
15	17	120	01111	01000	010101
16	20	121	10000	11000	010110
17	21	122	10001	11001	010111
18	22	200	10010	11011	011000
19	23	201	10011	11010	011001
20	24	202	10100	11110	100000

В табл. 14.1 показаны первые 21 целое число в системах с основаниями 10, 8, 3 и 2. Последние две колонки таблицы объясняются ниже. К любому из чисел можно приписать слева дальнейшие разряды (все нули), не изменяя значения числа.

Равенства (14.1) могут послужить основой для вывода алгорифма перевода чисел из одной системы счисления в другую. Так, перевод из троичной системы в десятичную иллюстрируется равенством (14.2). Тот же метод можно использовать и для перевода чисел из десятичной системы в двоичную, однако он оказывается неудобным. Например, чтобы выразить десятичное число 139 в двоичной системе, мы пишем $139=100+30+9$. Затем производим сложение в двоичной системе по тем же правилам, что и в десятичной системе:

десятичное	100 = двоичному	1 100 100
десятичное	30 = двоичному	11 110
десятичное	9 = двоичному	1 001
		10 001 011

Проверка: двоичное число 10 001 011 = десятичному числу $128+8+2+1=139$.

Обычный алгорифм для перевода десятичных чисел в другие системы счисления состоит в последовательном делении десятичного числа на основание новой системы счисления и выписывании получающихся остатков деления в виде последовательности цифр справа налево; эта последовательность и есть искомое число. Например, десятичное число 139 переводится в двоичную и троичную системы следующим образом:

В двоичную систему

$$\begin{array}{r|l} 139 & 2 \\ \hline 1 & 69 \end{array} \qquad \begin{array}{r|l} 69 & 2 \\ \hline 1 & 34 \end{array} \qquad \begin{array}{r|l} 34 & 2 \\ \hline 0 & 17 \end{array} \qquad \begin{array}{r|l} 17 & 2 \\ \hline 1 & 8 \end{array} \qquad \begin{array}{r|l} 8 & 2 \\ \hline 0 & 4 \end{array} \qquad \begin{array}{r|l} 4 & 2 \\ \hline 0 & 2 \end{array}$$

В троичную систему

$$\begin{array}{r|l} 139 & 3 \\ \hline 1 & 46 \end{array} \qquad \begin{array}{r|l} 46 & 3 \\ \hline 1 & 15 \end{array} \qquad \begin{array}{r|l} 15 & 3 \\ \hline 0 & 5 \end{array} \qquad \begin{array}{r|l} 5 & 3 \\ \hline 2 & 1 \end{array}$$

Проверка: троичное число 12 011 = десятичному числу $81+2\cdot 27+3+1=139$.

Для перевода десятичных дробей используется несколько иной алгорифм. Часть числа, стоящая справа от запятой, последовательно умножается на основание новой системы. При этом в каждом таком произведении часть, стоящая слева от запятой, отбрасывается перед следующим умножением и записывается в виде последовательности слева направо.

Например, дробь $^5/_8$ переводится из десятичной формы в двоичную и троичную следующим образом:

В двоичную форму

$$\begin{array}{r} 0{,}625 \\ \times\quad 2 \\ \hline 1{,}25 \\ \times\quad 2 \\ \hline 0{,}5 \\ \times\quad 2 \\ \hline 1{,}0 \end{array}$$

Результат: двоичное число

$$0{,}101 = \frac{1}{2}+\frac{1}{8}.$$

В троичную форму

$$\begin{array}{r} 0{,}625 \\ \times\quad 3 \\ \hline 1{,}875 \\ \times\quad 3 \\ \hline 2{,}625 \\ \times\quad 3 \\ \hline 1{,}875 \end{array}$$

Результат: троичное число

$$0{,}121212\ldots = \frac{1}{3}+\frac{2}{9}+\frac{1}{27}+\frac{2}{81}+\ldots$$

Вместо этого мы могли бы разделить 5 на 8 в соответствующей арифметике по обычным правилам деления. Например, в троичной системе:

```
12,0  | 22
22    |-----
------| 0,121
 210
 121
-----
 120
  22
 ===
```

Основные правила арифметики в любой системе счисления могут быть записаны в виде таблицы сложения и таблицы умножения. Эти таблицы оказываются особенно простыми для двоичной системы, как это видно из табл. 14.2.

Таблица 14.2

Таблицы арифметических действий в двоичной системе счисления

Таблица сложения				Таблица умножения			
0	0	1	1	0	0	1	1
+0	+1	+0	+1	×0	×1	×0	×1
0	1	1	10	0	0	0	1

Следует отметить, что счет есть абсолютный процесс, не зависящий от какой бы то ни было системы символов, использованной для представления его результатов. Число атомов в молекуле воды равно трем независимо от того, представляем ли мы это число символом 3 в десятичной системе, символом 10 в троичной системе или символом 11 в двоичной системе; значение числа не может измениться от выбора символов. Когда мы применяем алгебру, то буквы обозначают именно эти неизменные значения интересующих нас чисел; для вычисления этих чисел мы должны применить какую-нибудь арифметическую символику; любая из них годится, если только она однозначна, т. е. непротиворечива.

Основание 10, выбранное для обычной практической деятельности благодаря анатомическим особенностям человека, достаточно удобно, хотя некоторые другие основания, как 8 или 12, подошли бы без сомнения еще лучше. Восьмеричная система счисления иногда используется в выходных устройствах вычислительных машин благодаря тому, что перевод из двоичной системы в восьмеричную очень прост, а восьмеричные индикаторы легко читаются операторами, привыкшими к десятичной системе. В десятичной системе не слишком трудно запомнить такое, например, число, как 1950. В двоичной системе это число получает вид 11 110 011 110 и содержит 11 двоичных разрядов, или «битов» *.

При основании 60 (вавилонская система счисления, остатки которой еще видны в подразделении суток на часы, минуты и секунды) число 1950 потребовало бы всего двух разрядов, однако нам пришлось бы выучить 60 различных цифр, в результате чего таблицу умножения большинство людей не смогло бы запомнить и арифметика была бы доступна только для высокообразованных людей.

С другой стороны, для применения в цифровых вычислительных машинах двоичная система оказывается более удобной, чем десятичная, по трем причинам. Во-первых, благодаря простоте арифметики (табл. 14.2). Во-вторых, существует много электронных устройств с двумя устойчивыми состояниями, способными изображать цифры 1 и 0 соответственно. В-третьих, известно, что для представления одних и тех же чисел двоичная система может потребовать меньше оборудования, чем десятичная. Это обстоятельство демонстрируется в табл. 14.3, изображающей световое табло, на котором в каждой колонке зажигается по лампочке (изображаемой крестиком) для представления соответствующего разряда.

Легко видеть, что десятичное табло потребует 60 лампочек для представления 1 миллиона чисел, тогда как двоичное табло потребует только 40 лампочек для представления более чем 1 миллиона чисел. Таким образом, двоичная система требует примерно на одну треть меньше оборудования, чем десятичная, для представления столь же больших чисел. Кроме того, в каждом табло из табл. 14.3 можно обойтись одной строкой меньше, изображая 0 гашением всех лампочек. В этом случае мы получим 50 лампочек для представления 1 миллиона чисел в десятичной системе и 20 лампочек для представления более чем 1 миллиона чисел в двоичной системе — экономия на 3/5.

Таким образом, мы имеем некоторое основание утверждать, что вычислительная машина с внутренней логикой, основанной на двоичной системе счисления, требует меньше оборудования, чем вычислительная машина

---

* Бит — распространенное название двоичного разряда от английского сокращения bit (*bi*nary digit), предложенного К. Э. Шенноном в его работах по теории информации (см. ниже, гл. 28). В русской литературе бит часто называют также «двоичной единицей информации». — *Прим. ред.*

Таблица 14.3

Десятичные и двоичные матрицы для 1 миллиона чисел

Десятичная

			Тысячи	Сотни	Десятки	Единицы
0	x	x				x
1			x			
2						
3						
4						
5					x	
6						
7						
8						
9				x		

Двоичная

																	Восьмерки	Четверки	Двойки	Единицы
0	x	x	x	x	x	x	x	x	x	x					x	x					x
1											x	x	x	x			x	x	x	x	

Примечание. Десятичная матрица показывает числа от 0 до 999 999, двоичная — от 0 до 1 048 575. В обеих стоит число 1950.

с внутренней логикой, основанной на десятичной системе счисления.

Рефлексный код. Все системы счисления, рассмотренные до сих пор, удовлетворяли равенству (14.1). Однако, как уже указывалось выше в связи с римской нумерацией, это ограничение отнюдь не обязательно при построении однозначной системы счисления, и по крайней мере одна такая система находит применение в технике систем большого масштаба. Это так называемый *рефлексный двоичный код*, или *циклически переставленный код*, или *код Грея*. Для рефлексного кода равенство (14.1) заменяется равенством

$$x=\sum_{i=0}^{k-1} \pm a_i(r^{i+1}-1), \qquad (14.3)$$

где основание r всегда равно 2 и перед самым старшим разрядом стоит плюс, перед следующим справа разрядом — минус и т. д., так что каждое дальнейшее a_i, отличное от нуля, имеет новый знак. Первые 21 число в рефлексном коде приведены в табл. 14.1. Например, десятичное число 13 записывается в рефлексном коде как 1011, что интерпретируется как $15-3+1=13$.

Равенство (14.3) может быть использовано для перевода чисел из рефлексного кода в двоичную или десятичную форму. Приведем алгорифм перевода из двоичной формы в рефлексный код: j-й разряд в рефлексном коде равен нулю, если j-й и $(j+1)$-й разряды двоичного числа одинаковы; если же они различны, то j-й разряд в рефлексном коде равен единице. Перевод чисел из двоичной формы в рефлексный код и обратно легко может быть автоматизирован (§ 15.1, рис. 15.11 и 15.12).

Рефлексный код имеет следующее замечательное свойство: при переходе от любого числа к следующему за ним большему числу изменяется одна и только одна цифра кода Грея — либо 0 на 1, либо 1 на 0. Это свойство рефлексного кода определяет одно его применение, о котором говорится в § 19.1.

Восьмеричная и шестнадцатеричная системы. Подобно восьмеричной системе шестнадцатеричная система (основание 16) обладает легкой взаимопереводимостью с двоичной системой и образует весьма удобный язык для связи между цифровыми вычислительными машинами и внешним миром. В качестве шести новых цифр (10, 11, 12, 13, 14, 15) часто используются буквы a, b, c, d, e, f, соответственно.

Двоично-десятичная система. Часто бывает желательным хранить десятичные числа, по крайней мере короткое время, в цифровом устройстве, которое по существу двоично. Это делается с помощью так называемой *двоично-десятичной системы счисления*. В этой системе каждому десятичному разряду придается четыре двоичных разряда и 10 десятичных

цифр условно кодируются десятью из 16 возможных 4-разрядных двоичных чисел. Эта система не экономична, но удобна. Так, например, для того чтобы хранить в регистре число от нуля до 999, потребуется 10 битов в случае двоичной системы и 12 битов в случае двоично-десятичной; десятичное число 384, записываемое в двоичной системе как 0 110 000 000, записывается в двоично-десятичной системе как 0011 1000 0100.

Плавающая запятая. В цифровых вычислительных машинах часто применяется так называемая *плавающая запятая.* В десятичной системе любое число может быть записано в виде произведения числа, лежащего между 0,1 и 1, и некоторой целой степени числа 10. Например: $384=0{,}384\times10^{3}$ и $0{,}0071=0{,}71\times10^{-2}$. Подобно этому в двоичной системе с плавающей запятой любое двоичное число записывается без запятой и понимается как число, лежащее между 1/2 и 1 (в двоичном обозначении — между 0,1 и 1) и умноженное на целую степень от 2, показатель которой указывается несколькими последними разрядами числа. Например, число 384, равное $0{,}75\cdot2^{9}$, записывается как 11 (для 0,75), за которым следует столько нулей, сколько это диктуется точностью («разрядностью») устройства; за ними, в свою очередь, следуют цифры 1001 (для показателя 9), предшествуемые серией нулей, количество которых зависит от максимальной величины показателя степени, который должен запоминаться.

Кроме того, в представлении числа имеются два знаковых разряда: один — для знака числа, другой — для знака показателя. В нашем примере оба знаковых разряда были бы равны 1, указывая плюс.

ЛИТЕРАТУРА

Широкий и многосторонний обзор состояния вычислительной техники на 1953 г. можно найти в специальном выпуске журнала «Proceedings of the Institute of Radio Engeneers» [31], посвященном вычислительной технике.

ЗАДАЧА

14.1. Возьмите два трехзначных десятичных числа, переведите их в двоичную систему и сложите, переведите сумму обратно в десятичную систему и проверьте результат. Переведите эти числа в троичную систему и сложите, переведите сумму в десятичную систему. Поделите большее из этих чисел на меньшее по обычному способу, но используя двоичную систему, и переведите частное в десятичную систему для проверки; затем проделайте то же самое в троичной системе.

ГЛАВА 15

КОМПОНЕНТЫ ЭЛЕКТРОННЫХ ЦИФРОВЫХ ВЫЧИСЛИТЕЛЬНЫХ МАШИН

Предположим, что мы хотим вычислить и протабулировать функцию $x^2+2x+10/x$ для некоторой последовательности значений x и что мы используем обычную настольную счетную машину. Промежуточные и конечные результаты вычислений мы, очевидно, стали бы записывать с помощью карандаша и бумаги, и запись в рабочей таблице выглядела бы примерно как на рис. 15.1. Мы можем начертить блок-схему, изображающую наши действия (рис. 15.2). На этой блок-схеме стрелки, идущие к человеческому мозгу (человеческие руки, выполняющие действия, опущены), изображают считывание (исходного числа, результата, который надо списать со счетной машины, и т. д.), стрелки, выходящие из мозга, изображают команды и действия, а остальные стрелки изображают очевидные шаги при решении задачи. На рис. 15.3 изображена по существу та же самая блок-схема, но названия в ней изменены в соответствии с принятой терминологией для цифровых вычислительных машин.

Вход x	Промежуточные значения			Выход $x^2+2x+10/x$
	x^2	$2x$	$10/x$	
1	1	2	10	13,00
2	4	4	5	13,00
3	9	6	3,33	18,33
4	16	8	2,5	26,50
5	25	10	2	37,00
6	36	12	1,66	49,66
7	49	14	1,42	64,42
8	64	16	1,25	81,25

Рис. 15.1. Рабочая таблица простого вычисления.

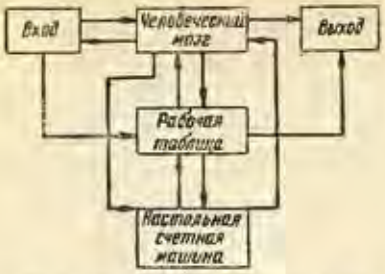


Рис. 15.2. Блок-схема простого вычисления.

Цифровая вычислительная машина решает эту задачу и вообще большинство задач по существу тем же самым путем, как решает их человек, вооруженный настольной счетной машиной.

Чтобы создать цифровую вычислительную машину, мы должны подобрать компоненты для выполнения всех необходимых функций каждой из этих пяти основных частей машины. Ниже перечисляются наиболее очевидные из этих функций:

1. Входное устройство. Оно должно быть способно по порядку:

а) считать с некоторого носителя набор чисел и

б) передать эти числа в надлежащую часть машины (обычно в запоминающее устройство).

В зависимости от конструкции машины входные данные могут поступать в десятичной форме с пишущей машинки, или даже — в аналоговой форме, когда машина способна сама выполнить перевод этих входов на свой внутренний язык (например, двоичный), или же может оказаться необходимым заранее закодировать входные данные на внутреннем языке машины и подготовить их специальным образом, как это имеет место в случае перфокарт.

2. Выходное устройство. Оно должно быть способно по порядку:

а) принять числа;

б) отпечатать или каким-либо другим способом наглядно отобразить эти числа.

3. Запоминающее устройство

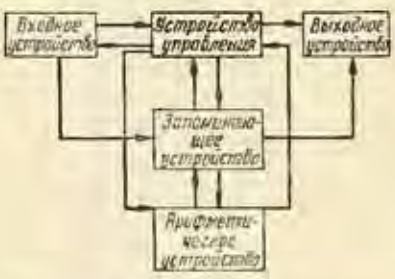


Рис. 15.3. Блок-схема цифровой вычислительной машины.

(«память» машины). Оно должно быть способно по порядку:

а) принять и удержать числа, различая их друг от друга;

б) передать определенные из этих чисел в надлежащие части машины (арифметическое устройство, устройство управления, выходное устройство) *.

4. Арифметическое устройство. Оно должно быть способно по порядку:

а) принять два числа, различая их друг от друга;

б) выполнить простые арифметические или логические операции и

в) передать результат в надлежащую часть машины (обычно в запоминающее устройство).

Набор простых арифметических операций, упоминаемых в п. 4б, включает как практический минимум сложение и вычитание или его эквиваленты. Почти всегда включается умножение и деление, а часто и другие операции. Далее, для работы автоматической цифровой вычислительной машины весьма существенна способность выполнять логическую операцию *сравнения* (определить, какое из двух чисел больше); сравнение обычно рассматривается как функция арифметического устройства, однако оно требует более сложного управления, чем простое вычитание.

5. Устройство управления. Функции, выполняемые этим устройством, зависят в определенной степени от конструкции машины. Как минимум оно должно быть способно:

а) вырабатывать основные синхронизирующие сигналы;

б) управлять коммутацией между основными устройствами машины;

в) начинать каждый процесс и обнаруживать его окончание;

г) размещать результаты в запоминающем устройстве и уметь находить их в нем снова;

д) принимать решения относительно следующего шага вычислений на основании результата последней операции и любой храни-

---

* Для обозначения функции этого устройства в русской технической литературе применяется несколько конкурирующих друг с другом синонимов: «запоминание», «хранение», «запасание», «накопление». Соответственно изменяется и название самого устройства: «запасающее устройство», «накопитель» и т. д. Термин «запоминание» вызывал ряд возражений ввиду как бы свойственного ему некоторого антропоморфизма, но, в конце концов, остался самым популярным. С учетом этого мы и остановились при переводе в основном на нем, используя также в подходящих случаях термин «хранение». — *Прим. ред.*

мой в запоминающем устройстве информации, имеющей отношение к этому решению;

е) получать и расшифровывать хранимые в запоминающем устройстве инструкции о том, как выполнять все эти функции.

15.1. Логика машины

Все перечисленные функции могут быть синтезированы логически из четырех основных операций: из логических операций И, ИЛИ, НЕ и единичной задержки *. В свою очередь, эти основные операции могут быть осуществлены с помощью электронных схем.

Представим себе устройство с двумя входными проводами A и B и одним выходным проводом C. Это устройство «выполняет» логическую операцию И, если импульс на его выходе появляется тогда и только тогда, когда импульсы поступают одновременно на вход A и на вход B. В дальнейшем будем предполагать, что выходной импульс появляется одновременно с входными импульсами, хотя на практике (см. § 15.4) наблюдается некоторая задержка. В электронных двоичных цифровых вычислительных машинах состояние 1 часто изображается наличием импульса в определенный момент времени, а состояние 0 — отсутствием импульса в этот момент. Соответственно в табл. 15.1 утверждения о присутствии и отсутствии импульсов изображаются цифрами 1 и 0.

Аналогично, наше устройство выполняет операцию ИЛИ, если импульс на выходе появляется в том случае, когда импульс поступает на вход A или на вход B, но не в каком-либо другом случае. В логике слово ИЛИ означает «включающее или» (A, или B, или оба), если только специально не оговаривается другое толкование. Ниже будет показано, что «исключающее или» (A или B, но не оба), как и все другие логические операции, может быть получено из трех перечисленных выше основных логических операций.

Операция НЕ выполняется нашим устройством, если импульс на входе появляется в том случае, когда импульс поступает на вход A и не поступает на вход B, но не в каком-либо другом случае **. Эта операция также показана в табл. 15.1. Операция «B и не A», которая также называется операцией НЕ, в таблице не показана.

Часто бывает желательно задержать

Таблица 15.1

Логические таблицы истинности***

Вход A	Вход B	Выход для: И	ИЛИ	НЕ (A и не B)	Исключающее или
0	0	0	0	0	0
0	1	0	1	0	1
1	0	0	1	1	1
1	1	1	1	0	0

импульс на некоторое время. Задержку удобно измерять в единицах времени, равных длительности одного импульса. Устройство единичной задержки имеет один входной провод и один выходной провод. Импульс на входе вызывает импульс на выходе не немедленно, а лишь спустя одну единицу времени.

Условные обозначения для этих четырех основных функций приведены на рис. 15.4 ****.

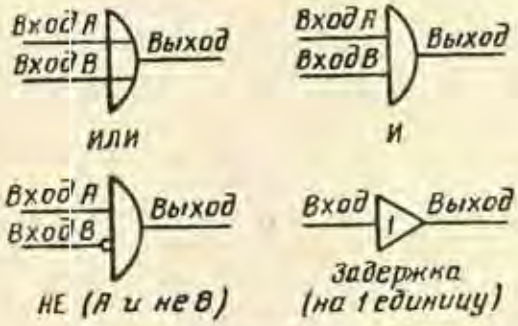


Рис. 15.4. Логические символы.

Электронные схемы, используемые для реализации функций И, ИЛИ и НЕ, называются *клапанами* (клапан «открывается» и позволяет импульсу пройти через него, если выполняются соответствующие условия таблицы 15.1). Вход B в клапане НЕ называется

* Возможны и другие, в том числе более краткие, наборы основных операций. — *Прим. авт.*

** В формальной логике под операцией НЕ, или операцией *отрицания*, понимается только часть «не B» нашей операции «A и не B». — *Прим. авт.*

*** Таблицей истинности сложного высказывания (в нашем случае высказываний «A и B», «A или B» и т. д.) называется таблица «значений истинности» этого высказывания для всех комбинаций значений истинности входящих в него простых высказываний (в нашем случае высказываний A и B). «Значение истинности» высказывания считается равным 1, если высказывание истинно, и равным 0, если высказывание ложно. Таким образом, таблица истинности есть таблица такой функции, которая сама принимает только два значения 0 и 1 и у которой каждый аргумент тоже принимает только два значения 0 и 1. Термины «утверждение», «предложение», «высказывание», «суждение» суть синонимы. Вообще в этой главе используются некоторые элементы современной формальной логики, называемой часто также *математической логикой*. — *Прим. ред.*

**** Условные обозначения в этой области пока еще не стандартизированы. — *Прим. авт.*

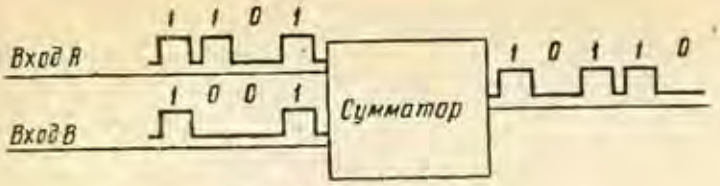


Рис. 15.5. Сумматор.

запретом: действительно, наличие импульса на входе B запрещает импульс на выходе*. В § 15.2 рассматриваются схемы, механизирующие эти функции. Однако сначала мы покажем, как строить арифметические операции из этих четырех основных логических функций.

Одной из простейших арифметических операций является сложение. Рассмотрим случай, показанный на рис. 15.5, когда в последовательный сумматор поступают числа 1101 и 1001, а на выходе образуется их сумма 10110. Каждое из слагаемых поступает в последовательной форме (это означает, что цифры числа поступают друг за другом в моменты времени, отстоящие друг от друга на одну единицу), начиная с самого младшего разряда. При поступлении первой пары цифр сумматор должен дать импульс на выходе в том и только том случае, если одна из этих цифр есть 1, а другая 0, а также выработать перенос 1, если обе эти цифры суть 1 (см. табл. 15.2).

Таблица 15.2

Таблица сложения для двоичной арифметики

A	B	$A+B$	
		перенос	сумма
0	0	0	0
1	0	0	1
0	1	0	1
1	1	1	0

Реализацию этой таблицы начнем с логической схемы, показанной на рис. 15.6. Верхняя часть этой схемы, содержащая два клапана НЕ и один клапан ИЛИ, составляет «исключающее или», и ее выход есть разряд суммы. Нижняя часть схемы, содержащая клапан И, дает на выходе цифру переноса. Сравнение рис. 15.6 с табл. 15.2 показывает, что в каждом из четырех возможных случаев

*Так как операция И называется в математической логике конъюнкцией, а операция ИЛИ — дизъюнкцией, то клапан И называется также конъюнктором, а клапан ИЛИ — дизъюнктором. Клапан НЕ называется тогда негатором, т. е. «отрицателем». — *Прим. ред.*

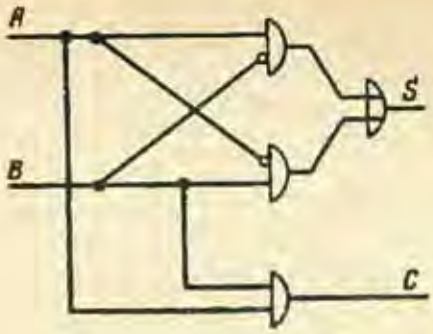


Рис. 15.6. Полусумматор.

схема дает правильный результат. Схема рис. 15.6 называется *полусумматором.*

Комбинируя два полусумматора, как показано на рис. 15.7, мы можем получить сумматор. На рис. 15.7 каждый из блоков с надписью «полусумматор» представляет собой комбинацию клапанов, показанную на рис. 15.6. Единичная задержка работает таким образом, что перенос C' от каждой пары цифр данного разряда слагаемых прибавляется к сумме S' двух цифр следующего разряда, как и должно быть. Если при этом образуется новый перенос C, то этот перенос прибавляется к следующей сумме, и т. д. Таким образом, схема на рис. 15.7 выполняет все функции, необходимые для сложения двух последовательных двоичных чисел, с одним только кажущимся затруднением: если бы импульсы переноса появились на выходах обоих полусумматоров в одно и то же время, то сумматор дал бы неверный ответ. К счастью, одновременные переносы невозможны, как это будет сейчас показано.

Заметим, кстати, что это обстоятельство, однако, отнюдь не очевидно и что данная схема сумматора является поэтому весьма изобретательной. Для логического проектирования каждой цифровой вычислительной машины требуется весьма много такой изобретательности.

Чтобы доказать, что обе цифры переноса не могут быть единицами одновременно, заметим, что при сложении двух двоичных цифр мы не можем получить одновременно цифру 1 и в сумме, и в переносе. Следовательно, если $C'=1$, то $S'=0$. Поэтому прибав-

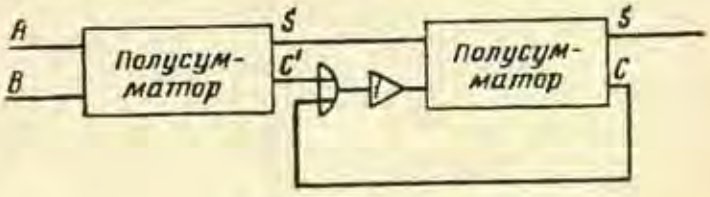


Рис. 15.7. Сумматор.

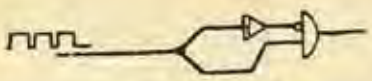


Рис. 15.8. Индикатор начала серии импульсов.

ление $S'=0$ к предыдущему переносу не может дать нового переноса 1. Следовательно, если $C'=1$, то $C=0$.

Подобным же образом может быть построен *вычитатель* (схема вычитания) (см. задачу 15.1). Ясно, что умножение и деление можно выполнить путем повторных сложений и вычитаний. Мы не будем строить этих сложных логических схем, но приведем пять простых примеров построения полезных функциональных схем из четырех основных операций (рис. 15.8—15.12). Другие примеры будут встречаться по всей главе.

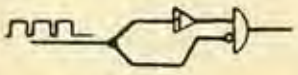


Рис. 15.9. Индикатор конца серии импульсов.

На рис. 15.8 и 15.9 изображены две функции, полезные для целей управления. Пусть у нас имеется серия импульсов (т. е. серия единиц, которой предшествуют и за которой следуют нули); тогда схема на рис. 15.8 выдает импульс только в момент поступления первого импульса серии, а схема рис. 15.9 выдает импульс спустя одну единицу времени после последнего импульса серии. Эта пара выходных импульсов может быть использована, например, для открывания и затем для закрывания какого-нибудь клапана, причем первый импульс может действовать через клапан И, а второй — через запрет (клапан НЕ).

На рис. 15.10 показан метод запоминания, весьма широко применяемый в цифровых вычислительных машинах. Задержка на 100 единиц времени теоретически может быть составлена из 100 последовательно соединенных единичных задержек; на практике задержкой на 100 единиц времени могла бы служить акустическая линия задержки (§ 15.3). Во всяком случае это устройство может хранить 100 битов (например, 10 десятиразрядных двоичных чисел). Хранящиеся в устройстве числа появляются на выходе линии задержки и при отсутствии управляющих

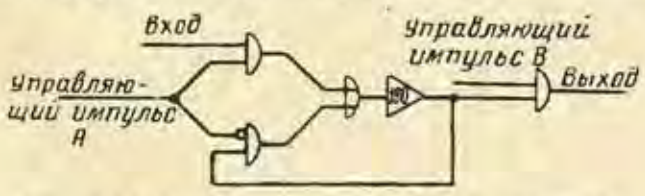


Рис. 15.10. Запоминающее устройство на линии задержки.

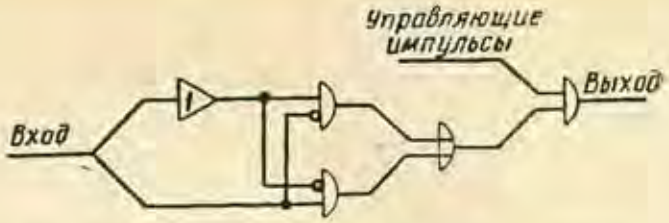


Рис. 15.11. Устройство перевода из двоичного кода в рефлексный.

импульсов возвращаются на ее вход через клапаны НЕ и ИЛИ.

Для записи одного или нескольких чисел в запоминающее устройство эти числа подаются в последовательной форме на общий вход устройства и, кроме того, в точку A подаются управляющие импульсы. Управляющие импульсы открывают клапан И и позволяют пройти входным импульсам; в то же время управляющие импульсы закрывают клапан НЕ и прекращают циркуляцию импульсов, выходящих из линии задержки. Для *считывания* хранимых чисел из запоминающего устройства, т. е. для получения их на выходе, нужно подать управляющие импульсы в точку B, чтобы открыть клапан И. Считывание не *стирает* чисел в запоминающем устройстве, так как они идут не только к общему выходу устройства, но также и ко входу линии задержки; запись же стирает числа в запоминающем устройстве.

На рис. 15.11 показана простая комбинация «исключающего или» и единичной задержки, позволяющая переводить двоичные числа в рефлексный код. Двоичное число подается на вход схемы начиная либо с самого старшего, либо с самого младшего разряда. В обоих случаях схема вырабатывает лишний разряд на младшем конце «рефлексного» числа; однако этот разряд не проходит через оконечный клапан И. На рис. 15.12 показана аналогичная схема, используемая для перевода из рефлексного кода в двоичную форму; самый старший разряд подается

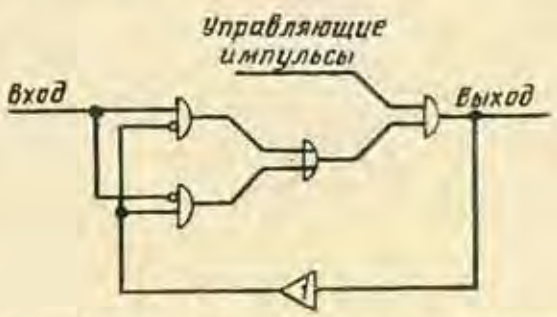


Рис. 15.12. Устройство перевода из рефлексного кода в двоичный.

здесь в схему первым. Чтобы проверить правильность работы этих схем, можно ввести в них какое-нибудь число из табл. 14.1.

15.2. Основные электронные схемы цифровых вычислительных машин

Для механизации любой вычислительной функции существует целый ряд методов. Как правило, мы будем описывать лишь один распространенный или простейший тип, чтобы показать возможность механизации и дать некоторое представление об используемой технике. В некоторых случаях (особенно в § 15.5) мы будем приводить по нескольку примеров, когда описываемые схемы могут встречаться и в других частях системы, кроме вычислительной машины.

Предположим, что в нашем распоряжении имеются хорошо сформированные прямоугольные импульсы, и пусть длительность этих импульсов равна $^1/_4$ *мксек*, расстояние между импульсами 1 *мксек*, амплитуда импульсов 10 *в*. Такие импульсы можно сформировать либо с помощью генератора синусоидальных колебаний частотой 1 *Мгц*, соединенного с подходящими схемами для придания этим колебаниям прямоугольной формы, либо с помощью свободно генерирующего мультивибратора.

В цифровой вычислительной технике используются различные виды мультивибраторов, в том числе: астабильный, или свободно генерирующий, мультивибратор; моностабильный мультивибратор, иначе называемый одновибратором, и бистабильный мультивибратор, иначе называемый статическим триггером. Статический триггер, показанный на рис. 15.13, характеризуется двумя устойчивыми состояниями, одно из которых может изображать цифру 1, а другое — цифру 0.

Предположим, что лампа V_1 проводит ток, а лампа V_2 нет. Это состояние устойчиво, пока нет импульсов на входе. Напряжение на аноде лампы V_1 (а следовательно, и на выходе *1*) отрицательно (т. е. более отрицательно, чем напряжение $B+$) из-за падения напряжения на сопротивлении R_1; напряжение же на аноде V_2 (а следовательно, и на выходе *2*) положительно (т. е. равно напряжению $B+$). Далее, отрицательное напряжение на аноде V_1 обусловливает отрицательное напряжение на сетке V_2 и удерживает эту лампу в запертом состоянии; напряжение же на сетке лампы V_1 положительно, и эта лампа поддерживается в проводящем состоянии.

Пусть теперь на вход *1* поступает отрицательный импульс. Этот импульс опускает катодный потенциал лампы V_2 ниже ее сеточного потенциала, и лампа V_2 начинает проводить. Ток, протекающий через сопротивление R_2, понижает напряжение на аноде лампы V_2, и благодаря конденсатору C_4 потенциал на сетке лампы V_1 становится весьма отрицательным. В результате лампа V_1 запирается, ток через R_1 прекращается и потенциал на сетке лампы V_2 повышается, увеличивая тем самым ее проводимость. Так достигается новое устойчивое состояние, в котором V_2 проводит, а V_1 заперта. На схему оказывает действие только передний фронт входного импульса (было бы достаточно даже простого изменения уровня на менее положительный). Задний фронт импульса, будучи повышением потенциала, не оказывает действия на схему.

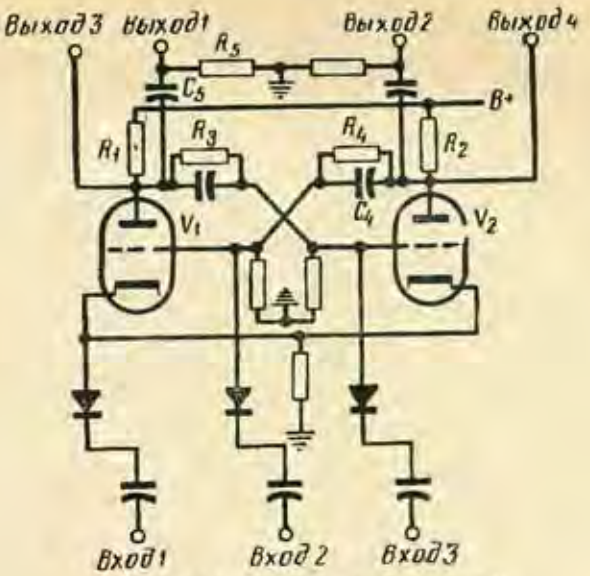


Рис. 15.13. Статический триггер.

Отрицательный импульс на входе *1* заставляет триггер изменить свое состояние независимо от первоначального состояния. Отрицательный импульс на входе *2* вызовет смену состояния («установку триггера») только в том случае, если первоначально лампа V_2 проводила; в противном случае импульс не окажет влияния. Аналогично, отрицательный импульс на входе *3* вызовет смену состояния («сброс триггера») только в том случае, если первоначально проводила лампа V_2. Выходным сигналом триггера служит уровень напряжения на выходах *3* и *4*, но его можно превратить в импульс на выходах *1* и *2* дифференцированием в цепи P_5—C_5. В зависимости от желательной полярности можно использовать либо тот, либо другой выход. Смещение в одном направлении от данного выходного уровня вызывает импульс желательной полярности, а смещение в другом

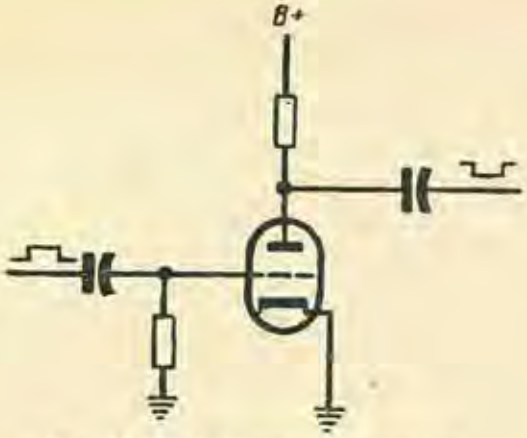


Рис. 15.14. Импульсный инвертор.

направлении — импульс противоположной полярности, который можно срезать и сделать тем самым эквивалентным отсутствию сигнала.

До сих пор предполагалось, что на триггер поступают отрицательные входные импульсы. Если в нашем распоряжении имеются положительные входные импульсы, то их можно *инвертировать* (обратить) с помощью схемы, показанной на рис. 15.14. Если в этой схеме лампа нормально имеет отрицательное сеточное смещение, достаточное для запирания, то положительный импульс на сетке отпирает лампу и вызывает падение напряжения на ее выходе. Эта же схема может использоваться для преобразования отрицательного импульса в положительный, если лампа нормально имеет положительное сеточное смещение, поддерживающее лампу в проводящем состоянии.

Схема на рис. 15.15 отличается от предыдущей схемы только наличием дополнительной управляющей сетки. Если обе сетки имеют положительные смещения или получают положительные импульсы, то лампа проводит и ее выходное напряжение является отрицательным относительно $B+$. Если одна

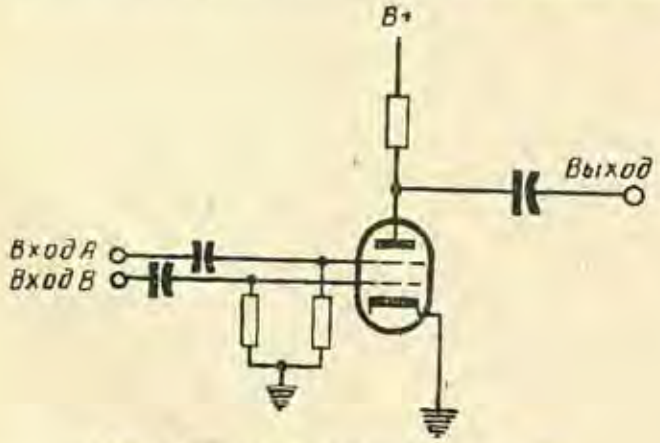


Рис. 15.15. Клапан (И, ИЛИ или НЕ).

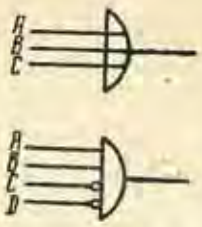


Рис. 15.16. Многовходовой клапан (логический символ).

из сеток имеет отрицательное смещение или получает отрицательный импульс, то лампа запирается и ее выходное напряжение поднимается до уровня $B+$. Эта схема служит клапаном И, ИЛИ и НЕ, в зависимости от смещений на сетках и полярности подаваемых импульсов.

Для получения клапана И обе сетки смещаются до запирания и подаются положительные входные импульсы. В этом случае выходной импульс (отрицательный) появляется тогда и только тогда, когда присутствуют оба входных импульса.

Для получения клапана ИЛИ обе сетки смещаются положительно и подаются отрицательные выходные импульсы. В этом случае выходной импульс (положительный) появляется всякий раз, когда присутствует либо один, либо оба входных импульса.

Для клапана НЕ одна из сеток (например A) смещается положительно и питается отрицательным импульсом, а другая сетка смещается отрицательно и питается положительным импульсом. В этом случае выходной импульс (отрицательный) появляется тогда и только тогда, когда присутствует импульс на входе B и отсутствует импульс на входе A.

В современных цифровых вычислительных машинах наблюдается тенденция избегать употребления электронных ламп, где только можно. Оказалось возможным применить кристаллические (полупроводниковые) диоды в качестве заменителей при выполнении всех логических и клапанирующих функций, сохраняя лампы и транзисторы только для усиления и формирования импульсов. Диодная схема, показанная на рис. 15.17, не только обладает всеми преимуществами кристаллических диодов над электронными лампами, но способна также к механизации таких логических схем, как показанные на рис. 15.16, которые изображают A ИЛИ B ИЛИ C и A И B И НЕ C И НЕ D. Такие конфигурации потребовали бы $n-1$ клапанов того типа, который был показан на рис. 15.15, где n — число входов.

Полупроводниковый диод обладает свойством оказывать очень малое сопротивление (менее 100 *ом*) электронам, движущимся от анода к катоду, и очень большое сопротивление (более 100 *ком*) электронам, движущимся в противоположном направлении. На рис. 15.17 сопротивление R выбрано около 10 *ком*. Поэтому если все входы имеют одн-

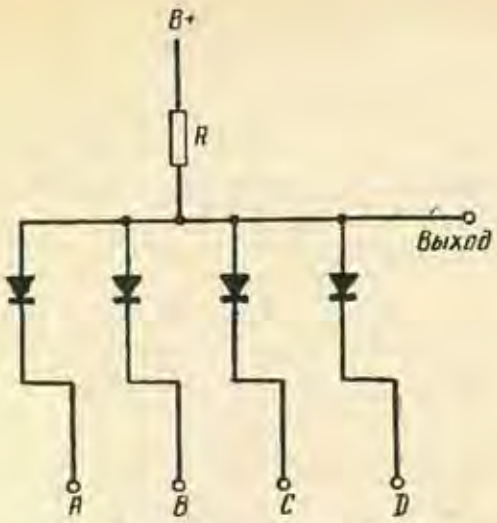


Рис. 15.17. Многовходовой клапан (диодная схема).

наковый потенциал (ниже $B+$) и способны подавать ток, то выходное напряжение будет примерно равно входному напряжению. Если входы имеют разные потенциалы, то выходной потенциал будет равен самому отрицательному из входных потенциалов, так как остальные входы будут отделены от выхода большими обратными сопротивлениями диодов. Иными словами, диоды служат для привязки выходного потенциала к самому отрицательному из входов.

Пусть теперь все входы имеют смещение, скажем, в 0 *в*; тогда отрицательный импульс, скажем, в 10 *в* на любом входе вызовет отрицательный импульс на выходе; следовательно, схема может служить клапаном ИЛИ. С другой стороны, положительный импульс на одном входе не окажет влияния на выход, если только все другие входы также не получат положительных импульсов; это значит, что схема представляет собой клапан И для положительных импульсов. Наконец, если A

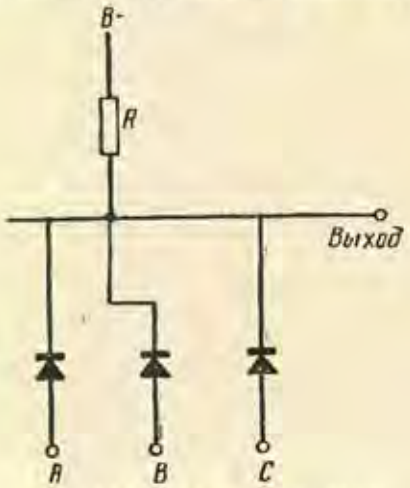


Рис. 15.18. Клапан (И, ИЛИ или НЕ).

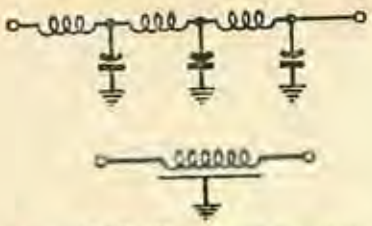


Рис. 15.19. Линии задержки.

и B имеют смещение в 0 *в*, а C и D — в 10 *в* и если на входы A и B поступают положительные импульсы, а на входы C и D — отрицательные импульсы, то схема работает как клапан НЕ, изображенный на рис. 15.16. На рис. 15.18 показана схема того же типа, но с обратным включением диодов. Эта схема может действовать как клапан И, как клапан ИЛИ или как клапан НЕ, совершенно так же, как схема на рис. 15.17, но полярность входных и выходных импульсов здесь другая.

Оставшаяся логическая функция — функция задержки — может быть механизирована с помощью распределенной индуктивности и емкости, как показано на рис. 15.19. Такие схемы называются *линиями задержки*; общая задержка на одну секцию линии выражается формулой $\sqrt{LC}$. Если, например, емкость равна 62,5 *пф* и индуктивность равна 1 *мгн*, то задержка составит $\sqrt{62{,}5\cdot 10^{-12}\cdot 1\cdot 10^{-3}}=2{,}5\cdot 10^{-7}$, т. е. 1/4 *мксек*.

С помощью триггеров и линий задержки можно построить важное устройство цифровых вычислительных машин — регистр сдвига.

Будем изображать триггер, как он изображен на рис. 15.20. Отрицательный импульс на сбросовом входе не оказывает влияния, если триггер хранит 0; если же хранится 1, то триггер сбрасывается на 0 и на выходе появляется импульс. Отрицательный импульс на установочном входе не оказывает влияния, если триггер хранит 1; если же хранится 0, то триггер устанавливается на 1 и импульса на выходе не появляется.

Регистр сдвига строится соединением любого числа триггеров по схеме, показанной на рис. 15.21. Он запоминает двоичное число, количество разрядов которого равно количеству триггеров. Если нужно сдвинуть это число вправо, то по общей шине подается импульс на все сбросовые входы. В тригге-

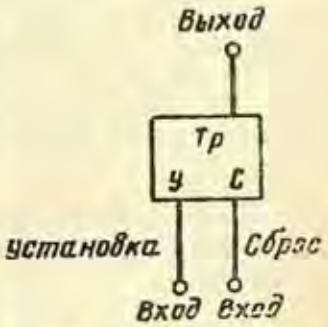


Рис. 15.20. Триггер.

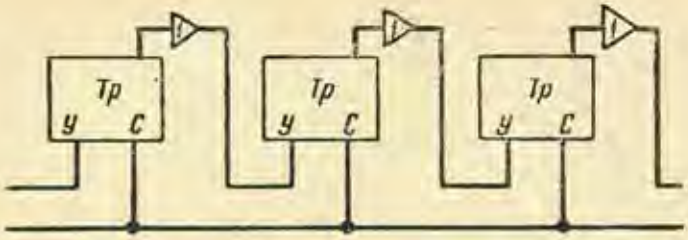


Рис. 15.21. Регистр сдвига.

рах, где хранится 0, ничего не происходит, но в каждом триггере, где хранится 1, эта единица стирается и в линию задержки выдается импульс. Через одну единицу времени этот импульс появляется на другом конце линии задержки (в это время все триггеры регистра должны хранить 0) и поступает на установочный вход ближайшего справа триггера, переводя его в состояние 1, но не вызывая импульса на его выходе. На практике задержка в схеме по рис. 15.21 должна равняться самое меньшее длительности сбросового импульса и могла бы составлять в случае нашей гипотетической вычислительной машины $^1/_4$ *мксек*.

Вводимое в регистр число нужно подать на установочный вход первого триггера в форме отрицательных импульсов (импульс для 1, отсутствие импульса для 0), начиная с самого младшего разряда, с расстоянием между разрядами 1 *мксек* и со сбросовыми импульсами спустя $^1/_2$ *мксек*. Сбросовые импульсы служат для сдвига числа вправо, пока оно не займет правильного положения. При помощи дополнительных схем можно обеспечить возможность сдвига в обе стороны.

Регистр сдвига чрезвычайно полезен как запоминающее устройство, что видно из ряда приводимых ниже примеров. Он используется также для умножения двоичных чисел, которое выполняется следующим образом: исследуется множитель, разряд за разрядом; если первый разряд есть 1, то множимое принимается за временное произведение; если же первый разряд множителя есть 0, то временное произведение остается равным нулю; затем в обоих случаях множимое сдвигается на один разряд вправо и исследуется следующий разряд множителя; если этот разряд есть 1, то новое (т. е. сдвинутое) множимое прибавляется к временному произведению; и т. д.

Таким образом, умножение может быть выполнено при помощи сумматора, регистра сдвига и некоторого вспомогательного запоминающего устройства. Регистр сдвига может быть также использован для нахождения самой старшей единицы в хранящемся числе; это позволяет снабжать цифровые вычислительные машины весьма полезной командой «определение порядка числа» (см. § 16.2). Умножение и деление на целые степени двойки осуществляются простым сдвигом числа влево или вправо.

Аналогичную цепочку триггеров можно использовать как двоичный счетчик.

15.3. Запоминающее устройство

Универсальная цифровая вычислительная машина требует от нескольких тысяч до 100 и более тысяч битов *внутреннего* запоминающего устройства, дополняемого в большинстве случаев вспомогательным запоминающим устройством с менее жесткими требованиями. Обычно двумя самыми важными характеристиками запоминающего устройства являются его время доступа* и стоимость, причем они связаны между собой обратной зависимостью. Одну крайность представляет триггер, где время доступа измеряется долями микросекунды, а стоимость — многими долларами на бит. Другую крайность представляют разные виды бумажной перфоленты или фотографической записи, где время доступа измеряется в секундах или минутах, а стоимость — малыми долями цента на бит. Существует и ряд промежуточных типов.

Обычно оказывается удобным снабдить вычислительную машину по крайней мере двумя различными видами памяти: одним — со сравнительно высокой скоростью работы и малой емкостью и другим — с низкой скоростью работы и большой емкостью. *Скорость* работы запоминающего устройства измеряется *временем доступа*, т. е. тем временем, которое требуется для доступа (записи или считывания) к первому требуемому биту. Доступ может быть *произвольным* (любой бит доступен в пределах фиксированного времени доступа) или *циклическим* (время доступа зависит от того, в каком месте цикла бит находится). В запоминающем устройстве с циклическим доступом последовательные биты, если их все записать вместе, доступны за гораздо меньшее время, чем первый бит, но это не принимается во внимание в условной классификации запоминающих устройств на *быстродействующие* (время доступа измеряется микросекундами), *среднедействующие* (время доступа измеряется миллисекундами) и *медленнодействующие* (время доступа — около секунды и более).

* Оно называется также «временем выборки» и «временем обращения» — *Прим. ред.*

Мы уже рассмотрели два запоминающих элемента (триггер и индуктивно-емкостную линию задержки) и методы использования их для запоминания в управляющей схеме из § 15.1 (рис. 15.10) и в регистре сдвига из § 15.2 (рис. 15.21). Любую из этих последних схем можно было бы использовать в качестве основного запоминающего механизма цифровой вычислительной машины, но их стоимость была бы чрезвычайно высока. В описанной нами линии задержки импульсы должны подформировываться и усиливаться по крайней мере после каждых 4 *мксек* задержки; положив 1 *мксек* задержки на бит помехи, мы видим, что для хранения каждых четырех битов необходима одна электронная лампа. В регистре сдвига потребуется как минимум один двойной триод (триггер) на бит памяти, и хотя такое запоминающее устройство весьма удовлетворительно, стоимость была бы непомерной.

Акустическая линия задержки. Как мы уже видели, задержка эквивалентна запоминанию. Если импульс (или отсутствие импульса) можно задержать и выход линии задержки соединен цепью обратной связи с ее же входом, то получается запоминающее устройство; нужно только добавить соответствующие цепи управления для записи и считывания запоминаемой информации. Такие цепи были уже описаны в § 15.1. Акустическая, или ультразвуковая, линия задержки дает возможность получить задержку достаточной длительности (например, значительную долю миллисекунды), чтобы запоминать большое число битов. Хотя время доступа в акустических линиях задержки может доходить до миллисекунды, акустические линии задержки часто относят к классу быстродействующих запоминающих устройств.

Акустическая линия работает следующим образом. Импульсы модулируют высокочастотную несущую, которая затем подается в кварцевый пьезокристаллический преобразователь. Возникающий ультразвуковой сигнал действует на столб кварца или ртути, на другом конце которого находится другой аналогичный пьезокристаллический преобразователь. Электрические сигналы, образующиеся во втором преобразователе, усиливаются, демодулируются и формируются, а затем, пройдя через надлежащие цепи управления, снова подаются на вход линии задержки.

Частота несущей составляет около 10 *Мгц*. Затухание в ртути на такой частоте равно примерно 5 *дб* на 1 *мсек*; это совсем немного (затухание возрастает пропорционально квадрату частоты). Скорость распространения ультразвука в ртути равна примерно 1 *см* в 6,9 *мксек*, с температурным коэффициентом примерно 1/3000 на 1° С. Иными словами, для запоминания тысячи битов с интервалами в 1 *мксек* друг от друга нужен столб ртути в 145 *см*; в этом случае повышение температуры на 3° С приведет к тому, что 999-й двоичный разряд появится на выходе в тот момент времени, в который при нормальной температуре появился бы тысячный разряд. Такой исход совершенно недопустим, так как хранящиеся биты опознаются исключительно по времени их появления на выходе линии задержки. Поэтому температура такой линии задержки должна поддерживаться постоянной с точностью примерно $\pm 1/2$° С.

Физическая длина столба ртути не составляет проблемы, так как при помощи отражателей ультразвук можно заставить пробежать несколько раз по линии туда и обратно и только после этого преобразоваться в электрический сигнал на выходе. Длина волны на этих частотах столь мала, что возникает остронаправленный луч, и поэтому грани пьезокристаллов должны быть весьма точно ориентированы.

Акустические сопротивления ртути и кварца очень хорошо согласуются на этих частотах, и именно это обстоятельство является главной причиной использования ртути (хотя кроме ртути и кварца иногда используются и другие материалы). При плохой передаче звуковой энергии между этими двумя веществами происходило бы большое затухание сигналов в линии; кроме того (и это более важно), возникали бы отражения (эхо), которые давали бы ложные сигналы. Благоприятным результатом хорошего согласования между ртутью и кварцем является также уменьшение добротности преобразователя за счет его эффективной нагрузки; малое Q обеспечивает широкую полосу частот, необходимую для пропускания достаточно прямоугольных сигналов с большой частотой.

Количество управляющих цепей для одной линии задержки не зависит от длины линии, поэтому линия должна быть настолько длинна, насколько это будет допускаться выбираемым нами временем доступа. Для запоминающих устройств цифровых вычислительных машин обычна ртутная линия длиной в 300—1000 битов при частоте 1 *Мгц*. Такое запоминающее устройство весьма надежно; если поддержание постоянной абсолютной температуры затруднительно, то можно поддерживать все линии задержки в машине при одной температуре (т. е. поддерживать только постоянную относительную температуру)

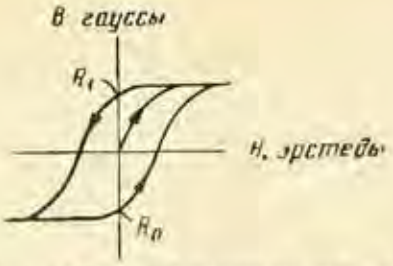


Рис. 15.22. Нормальная петля гистерезиса.

и синхронизировать главный задающий генератор машины от одной из этих линий задержки.

Отметим еще два свойства акустических линий задержки. Во-первых, такое запоминание является *кратковременным* («нестойким»): если электрическое питание на время прерывается, то хранящаяся в линии информация стирается. Во-вторых, доступ в такую память является циклическим, и потому устройство управления не знает заранее, сколь долгим окажется время доступа для данного конкретного числа.

Таким образом, когда требуется передать число из запоминающего устройства в арифметическое устройство, точное время доступа, вообще говоря, неизвестно, и потому, прежде чем приступать к следующей операции, устройство управления должно определить, что передача числа действительно уже произошла (это требует дополнительных цепей управления), или же каждый раз устройство управления должно ждать, пока не истечет максимальное время доступа. Обычно выбирается первая альтернатива.

Магнитные сердечники [103]. Время доступа для магнитных сердечников составляет всего несколько микросекунд; доступ является произвольным; запоминание *долговременное* («стойкое»), т. е. сохраняется после перерыва в электропитании. Стоимость была снижена до 1 доллара за бит (это весьма приблизительная цифра, так как стоимость запоминающего устройства не прямо пропорциональна его размерам). Вес и объем запоминающего устройства на магнитных сердечниках также невелики, причем в современных запоминающих устройствах самые сердечники занимают меньше места и весят меньше, чем цепи управления.

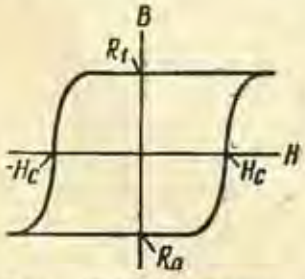


Рис. 15.23. „Прямоугольная" петля гистерезиса.

Петля гистерезиса обычного ферромагнитного материала показана на рис. 15.22. Если ненамагниченный образец такого материала (изображенный точкой в начале координат) постепенно намагничивать, увеличивая в положительном направлении магнитодвижущую силу H, то магнитная индукция образца будет изменяться по кривой в направлении, отмеченном стрелкой, и в конце концов достигнет насыщения. Если теперь поле уменьшено до нуля, то магнитное состояние материала возвращается в точку R_1 со значительной остаточной магнитной индукцией. Если поле меняет знак, то магнитная индукция изменяется по петле, как показано на рисунке стрелками.

Оказалось возможным изготовлять материалы с почти прямоугольной петлей гистерезиса, как на рис. 15.23. Остаточная магнитная индукция в такой прямоугольной петле в точности равна индукции насыщения. Кусок такого материала можно использовать как ячейку памяти; в состоянии R_1 он хранит цифру 1, а в состоянии R_0 — цифру 0. Тор из такого материала (рис. 15.24) называется магнитным сердечником. Как показано на рисунке, на сердечник намотаны две обмотки: одна для записи и одна для считывания.

Рис. 15.24. Магнитный сердечник.

Для того чтобы записать в сердечник единицу, на обмотку записи подают импульс, который создает магнитодвижущую силу несколько большую, чем критическое поле H_c. Для того чтобы записать в сердечник 0, на обмотку записи подают импульс такой же амплитуды, но противоположной полярности.

При считывании в обмотку записи подают такой же точно импульс, как при записи нуля. Если в сердечнике перед этим хранилась цифра 0, то магнитная индукция в сердечнике не изменяется и сигналов в обмотке считывания не наводится. Если в сердечнике хранится 1, то сердечник изменяет свое состояние от R_1 до R_0, пересекая линии потока индукции и наводя выходной импульс в обмотке считывания. При этом хранящаяся в сердечнике цифра 1 стирается, однако импульс из обмотки чтения поступает в клапан И вместе с положительным управляющим импульсом, а импульс из клапана И поступает в обмотку записи; следовательно, если в сердечнике хранилась цифра 1, то она опять регенерируется (восстанавливается) в нем. На

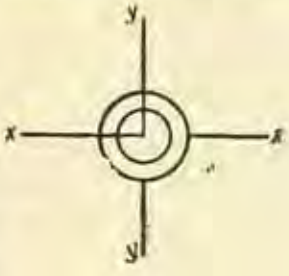


Рис. 15.25. Магнитный сердечник — условное обозначение.

практике вместо обмоток достаточны прямые проволоки (x и y на рис. 15.25).

Все описанное выше было бы возможно и в случае материала с такими магнитными свойствами, как на рис. 15.22. Однако запоминание с помощью магнитных сердечников оказывается практичным лишь благодаря применению таких сердечников в матричных схемах, требующих минимального количества цепей управления, а для применения сердечников в матричных схемах требуется прямоугольная петля гистерезиса.

Рассмотрим двумерную матрицу такого вида, как на рис. 15.26. Если мы хотим записать 1 в какой-либо сердечник, например в сердечник C_{11}, то по проводам X_1 и Y_1 нужно подать импульсы величиной приблизительно в 0,6 H_c. Тогда в сердечник C_{22} импульсы не поступают вообще; импульсы, поступающие в сердечники C_{12} и C_{21}, не обладают амплитудой, достаточной для изменения магнитного состояния; и только в сердечник C_{11} поступает импульс с амплитудой, достаточной для перевода этого сердечника в состояние R_1, если он уже не находился в этом состоянии. Для считывания того, что хранится в сердечнике C_{11}, импульсы той же самой амплитуды подаются на те же самые провода X_1 и Y_1, и опять только сердечник C_{11} получает импульс достаточной амплитуды. Провод считывания проходит через все сердечники матрицы и получает выходной импульс тогда и только тогда, когда выбранный сердечник хранит единицу.

Преимущество описываемой матрицы состоит в том, что в ней используется только один комплект цепей управления для всей матрицы плюс $2\sqrt{n}$ управляющих проводов (где n — число хранимых битов) и соответствующая коммутация для них. На практике иногда используют трехмерную матрицу. В такой матрице для считывания того, что хранится, например, в сердечнике C_{111}, подаются отрицательные импульсы на провода X_1 и Y_1 и положительные импульсы — на провода Z_2, Z_3 и т. д. Таким образом, другие сердечники получают положительный или отрицательный импульс 0,6 H_c или никаких импульсов, но только в один сердечник C_{111} поступит импульс 1,2 H_c. Такая матрица требует $3\sqrt[3]{n}$ управляющих проводов (например, 30 в 1000-битовой матрице) плюс один комплект цепей управления. Взяв, скажем, 40 таких матриц, мы можем хранить тысячу 40-битовых слов, и любое из них может быть считано параллельно по всем разрядам в течение нескольких микросекунд.

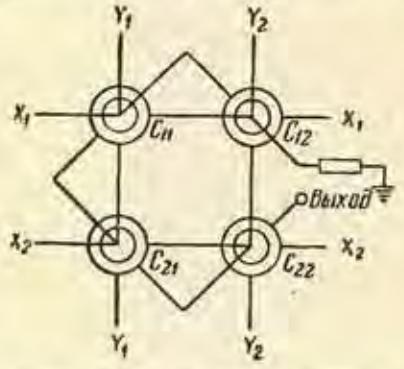


Рис. 15.26. Матрица магнитных сердечников.

Рис. 15.27. Малая петля гистерезиса.

Если петля гистерезиса сердечника не идеально прямоугольна, то под действием импульса 0,6 H_c магнитное состояние сердечника изменяется по малой петле гистерезиса, которая в сильно увеличенном виде показана на рис. 15.27. Большое число сердечников, изменяющих одновременно свое магнитное состояние по таким петлям, может дать ложное считывание единицы; более того, если сердечник неоднократно подвергается перемагничиванию по малой петле гистерезиса в одном и том же направлении без регенерации путем повторной записи или запроса, то хранящаяся в сердечнике цифра может быть разрушена. Поэтому прямоугольность петли является решающим фактором.

Может даже оказаться необходимым периодически прерывать использование запоминающего устройства в вычислительных процессах и проводить регенерацию всего содержимого матрицы, поочередно считывая один сердечник за другим. В практике работы с магнитными сердечниками обходятся без регенерации, однако в некоторых других типах быстродействующих запоминающих устройств регенерация необходима.

Сегнетоэлектрические запоминающие устройства. Некоторые кристаллы обнаруживают электрические свойства, которые можно изобразить таким же графиком, как на рис. 15.22, т. е. после приложения и снятия электрического напряжения они сохраняют остаточную электрическую поляризацию. Эти так называемые сегнетоэлектрические кристаллы могут быть использованы как быстродействующие запоминающие устройства. Типичным сегнетоэлектриком является титанат бария. Используются сегнетоэлектрические кристаллы аналогично магнитным сердечникам.

Криотрон. При температуре, близкой к

точке кипения гелия (4° К), некоторые металлы становятся «сверхпроводящими», т. е. их электрическое сопротивление падает до нуля. В магнитном поле сопротивление появляется опять, так что отношение сопротивлений в этих двух состояниях равно бесконечности. Это значит, что надлежащий металлический элемент при указанной температуре может действовать как переключатель или как элемент памяти, будучи переводим из одного состояния в другое сравнительно малым магнитным полем.

Другие типы быстродействующих запоминающих устройств. Любой элемент, способный принимать два различных состояния, может действовать как запоминающая ячейка, и если смена этих состояний может осуществляться электронной коммутацией, то данный элемент может оказаться пригодным для использования в быстродействующем запоминающем устройстве. Например, небольшой конденсатор мог бы хранить заряд той или другой полярности для регистрации единицы или нуля. Считывание и запись осуществлялись бы схемами на диодах. Запоминающее устройство на конденсаторах является сравнительно недорогим и исключительно быстрым. Однако вследствие утечки было бы необходимо время от времени считывать и регенерировать заряд. Во второй половине 40-х годов и в первой половине 50-х был исследован ряд таких элементов и по крайней мере один из них — электронно-лучевая трубка (называемая также *электростатическим запоминающим устройством*) — нашел широкое применение в цифровых вычислительных машинах.

В одном типе запоминающих электронно-лучевых трубок для получения двух устойчивых состояний используется вторичная эмиссия. Явление вторичной эмиссии состоит в том, что диэлектрическая мишень, бомбардируемая электронами, сама эмиттирует (испускает) электроны. Число *эмиттируемых* (вторичных) электронов может быть меньше и больше числа бомбардирующих (первичных) электронов, в зависимости от относительной разности потенциалов между мишенью и катодом трубки (источником первичных электронов). Экран трубки покрывается подходящим диэлектриком, и между катодом и экраном помещается положительно заряженный коллектор.

Если диэлектрик экрана имеет приблизительно тот же потенциал, что и катод, то состояние трубки устойчиво: столько же электронов испускается, сколько и поглощается, причем коллекторная пластинка забирает избыток. Если диэлектрик становится слегка положительным, то он поглощает электроны и возвращается к устойчивому состоянию. Это состояние используется для запоминания нуля. Если же диэлектрик становится весьма положительным, то в результате вторичной эмиссии он будет отдавать больше электронов коллектору, чем получать с катода, и будет становиться еще больше положительным, пока не достигнет нового устойчивого состояния, используемого для запоминания единицы.

Так как луч электронно-лучевой трубки может быть очень резко сфокусирован и так как каждая элементарная площадка диэлектрика электрически изолирована от соседних площадок, то любая из этих площадок может независимо от соседних площадок принимать то или иное состояние и, таким образом, хранить 1 или 0.

С другой (по отношению к катоду) стороны диэлектрика располагается металлическая пластина (обычно вне трубки), которая образует конденсатор с каждой из элементарных площадок диэлектрика. Когда записывается двоичная цифра (бит), металлической пластине придается некоторый определенный положительный потенциал. При этом все точки диэлектрика принимают тот же потенциал. Затем выбранная элементарная площадка диэлектрика облучается электронами с катода, а потенциал пластины снимается. В результате облучаемая площадка диэлектрика остается в новом устойчивом состоянии, а все остальные точки диэлектрика возвращаются к своим исходным потенциалам. При считывании на металлическую пластину подается отрицательный («нулевой») потенциал, электронный луч направляется на выбираемую точку и выходной сигнал считывается из цепи коллектора; импульс появится в том и только том случае, если в этой точке хранилась единица.

На практике часто используются две электронные пушки: одна — для считывания и записи, другая — для постоянного облучения мишени с целью поддержания каждой элементарной площадки в ее устойчивом состоянии. В тех случаях, когда используется только одна электронная пушка, хранимая цифра постепенно стирается, особенно если луч часто обегает соседние площадки. В запоминающем устройстве на таких электронно-лучевых трубках необходимо периодически считывать и регенерировать все хранимые цифры.

Существуют запоминающие устройства, в которых на каждой трубке запоминается до 1024 битов. Такие запоминающие устройства

имели то существенное преимущество перед другими устройствами, имевшимися около 1950 г., что они давали произвольный доступ к хранимой информации со столь малым временем доступа, как 10 *мксек*. Однако электронно-лучевые трубки даже в лучшем случае весьма капризные приборы, для управления которыми нужны точные и не создающие помех схемы; в современных цифровых вычислительных машинах электронно-лучевые трубки уже вытеснены магнитными сердечниками.

Реле. Электромеханическое реле, или переключатель, имеет два устойчивых состояния: разомкнутое и замкнутое. Существуют реле, которые срабатывают за 1 *мсек*, хотя в цифровых вычислительных машинах обычно использовались реле с временем срабатывания около 10 *мсек*. Электромеханические реле представляют значительный исторический интерес, так как в первой автоматической цифровой вычислительной машине запоминающие устройства строились почти исключительно на реле. В телефонной коммутации реле широко применяются до сих пор. Однако как запоминающие ячейки среднего быстродействия они не имеют особых преимуществ перед магнитным барабаном и обладают рядом недостатков. Поэтому реле применяют в настоящее время только в специальных целях.

Магнитный барабан. Техника, используемая для записи звука в магнитофоне, может быть приспособлена также для хранения импульсов, изображающих цифры. На поверхность барабана наносится постоянное покрытие из магнитного материала. Барабан вращается с большой скоростью мимо группы магнитных *головок*, с помощью которых производится запись и считывание. Магнитный барабан может хранить от нескольких тысяч до нескольких миллионов битов, со временем циклического доступа в несколько миллисекунд. Запоминание — долговременное («стойкое»).

Магнитный барабан представляет собой цилиндр диаметром от нескольких дюймов до нескольких футов (т. е. от дециметра до метра) и длиной от нескольких дюймов до нескольких футов. Каждый бит хранится на площадке размерами примерно в 0,254 *мм* по окружности и 2,54 *мм* по оси; другими словами, на барабане длиной в 91 *см* и диаметром в 91 *см* можно записать около 4 млн. двоичных цифр.

Группа таких элементов памяти, расположенных по всей окружности барабана, называется *дорожкой* или *каналом*. На одной магнитной дорожке может быть записано в зависимости от диаметра барабана от тысячи до 10 тысяч битов. Так как максимальная линейная скорость барабана равна приблизительно 50,8 *м/сек*, то барабан с бóльшим диаметром имеет соответственно большее время доступа. С другой стороны, количество цепей управления на одну дорожку более или менее не зависит от ее размеров, и потому барабаны большего диаметра стóят значительно дешевле в расчете на 1 бит.

Каждую дорожку обслуживает одна или несколько магнитных головок, которые производят запись, стирание и считывание информации. Хотя можно применить два или даже три разных типа головок для выполнения этих функций и расположить вдоль каждой дорожки несколько считывающих головок с целью уменьшения времени доступа, однако самое дешевое и самое распространенное решение состоит в применении одной головки для всех трех функций. Так как на одну дорожку обычно приходится по одной головке (или набору головок) и по одному комплекту цепей управления, то стоимость магнитного барабана в общем пропорциональна его длине. Головки обычно располагаются вдоль барабана по винтовой линии во избежание взаимных помех, как механических, так и магнитных. На некоторых магнитных барабанах устанавливаются подвижные головки, которые могут обслуживать две или более смежные дорожки.

Поверхность барабана покрывается тонким слоем какого-либо магнитного материала; обычно для этого используют порошкообразную окись железа, которая обладает малой электрической проводимостью. Магнитный материал должен иметь как можно более высокую коэрцитивную силу и как можно бóльшую остаточную индукцию, так как произведение этих величин определяет максимальную энергию, которая может быть запасена в заданном объеме материала, а следовательно, и мощность сигнала, которая может быть получена при считывании.

Поверхность барабана вращается под головкой с минимально возможным зазором, при котором еще отсутствует непосредственное физическое касание; касание стало бы разрушать как магнитное покрытие барабана, так и магнитную головку, и, кроме того, соскабливаемый магнитный материал начал бы набиваться в воздушный зазор магнитной головки и закорачивать его. Практически применяются зазоры между барабаном и головкой порядка 0,05 *мм*, фиксируемые механически или пневматически.

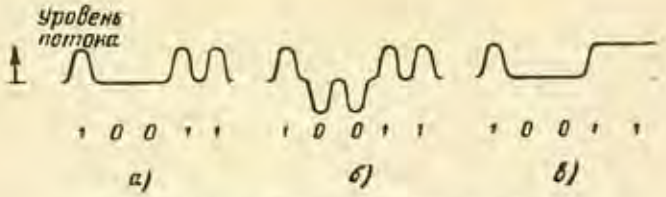


Рис. 15.28. Двоичные системы записи:

а) с двумя уровнями и возвратом к нулю; *б)* с тремя уровнями и возвратом к нулю; *в)* с двумя уровнями и без возврата к нулю.

Единица или нуль записывается в каждой элементарной площадке дорожки посредством намагничивания магнитного материала в разных направлениях. Обычно единицу изображает намагничивание в направлении вращения, а нуль — намагничивание в противоположном направлении. Можно также запоминать 0 посредством полного размагничивания материала переменным током.

Намагничивание производится одним из трех способов, показанных на рис. 15.28. В первых двух системах происходит возврат к уровню 0 с уровня 1, даже если записываются подряд единицы. Система с тремя уровнями является по существу троичной системой. Она может хранить 1, 0 или отсутствие бита. В системе без возврата к нулю более экономично используется место на магнитных дорожках благодаря тому, что при последовательной записи единиц уровень намагничивания не изменяется; однако эта система требует больше цепей управления, чем две другие.

Головки для записи и считывания представляют собой металлические кольца с воздушным зазором. С целью уменьшения вихревых токов кольцо набирается из отдельных изолированных пластин. Металл пластин должен обладать высокой магнитной проницаемостью и низкой коэрцитивной силой. Должны быть предусмотрены усилители тока как для считывания, так и для записи; обычно каждой головке придается по паре таких усилителей.

Барабан для запоминания 40-битовых слов содержал бы примерно 50 дорожек. Из них 40 дорожек можно было бы использовать для запоминания самих слов, так что каждое слово считывалось бы параллельно по всем разрядам. Остальные 10 дорожек можно было бы использовать для информации об адресах хранящихся чисел и для генерирования синхронизирующих импульсов. При желании увеличить емкость магнитного барабана без увеличения времени доступа можно увеличить число дорожек, например, до 90. При 90 дорожках потребовалось бы около 700 электронных ламп.

Среднедействующие запоминающие устройства чаще всего выполняются на магнитном барабане, хотя к этому же классу принадлежат и акустические линии задержки. Скорость работы запоминающего устройства на магнитном барабане значительно ниже, чем у описанных выше устройств, зато и стоимость хранения одного бита на магнитном барабане получается значительно меньшей; с другой стороны, магнитный барабан значительно быстрее и дороже устройств, которые описываются ниже. По всей видимости, магнитный барабан на некоторое время сохранит важное значение в цифровых системах.

Хотя большое время доступа у магнитного барабана ограничивает его использование в качестве первичного, главного запоминающего устройства среднедействующими (среднескоростными) системами (некоторые такие системы, называемые обычно *вычислительными машинами с магнитным барабаном*, имеются в продаже), магнитный барабан может быть эффективно использован в качестве вторичного запоминающего устройства в быстродействующих системах. Хотя для выборки первого слова с барабана может потребоваться от 10 до 50 *мсек*, соседние слова будут считываться по одному через каждые несколько микросекунд.

Решение задачи часто можно запрограммировать таким образом, что в течение нескольких минут вычислительная машина может производить вычисления, используя данные из своего внутреннего запоминающего устройства, затем на какую-то долю секунды вычисления прерываются и машина переписывает большую группу результатов на барабан и большую группу новых команд и констант с барабана в свое внутреннее запоминающее устройство. В некоторых машинах можно так кодировать задачу, что переносы между барабаном и внутренним запоминающим устройством происходят во время вычислений. Можно также путем так называемого *кодирования по минимальному времени доступа** достичь того, что требуемое слово будет проходить под магнитной головкой как раз в тот момент, когда оно должно быть считано. Однако такое кодирование представляет собой сложное и трудоемкое дело, кроме как в случае некоторых машин спе-

* Как будет видно дальше, авторы делят процесс, называемый у нас «программированием», на собственно программирование (выбор вычислительной процедуры) и кодирование (роспись ее в команды машины). Таким образом, в обычном смысле здесь идет речь о программировании. — *Прим. ред.*

циального назначения, например цифровых дифференциальных анализаторов.

Магнитный барабан может быть также использован в буферном запоминающем устройстве. Так, например, в системе наведения ЗУРС поступающие данные о положении ЗУРС и цели хранятся на магнитном барабане, пока они не потребуются вычислительной машине. Команды наведения, составленные вычислительной машиной, также могут некоторое время храниться на барабане в ожидании передачи или преобразования в аналоговые величины.

Магнитная лента и магнитная проволока. Магнитная лента состоит из тонкого слоя порошкообразной окиси железа на подложке из пластика. Она дает долговременное запоминание, но может стираться и затем использоваться снова. Количество битов, запоминаемых на 1 доллар или на 1 куб. фут, почти столь же велико, как у любого другого типа запоминающих устройств. Ввиду большого времени доступа магнитная лента обычно считается *медленнодействующим запоминающим средством*, однако в некоторых современных устройствах на магнитной ленте чтение и запись производятся со скоростями до 70 000 битов в секунду (500 битов на дюйм, 140 дюймов в секунду) на одной дорожке, причем на ленте обычно делается несколько дорожек.

Магнитная лента теоретически не отличается от магнитного барабана, так как последний можно рассматривать как цилиндр, обмотанный магнитной лентой (и, действительно, одно недавнее изобретение представляет собой барабан, в котором рабочая поверхность может изменяться именно этим способом, в целях увеличения емкости барабана без усложнения цепей управления*). Однако вследствие того, что лента используется для создания запоминающих устройств большей емкости и часто меньшей скорости работы, ее проблемы несколько иные. В частности, подобно многим другим видам медленнодействующих запоминающих устройств магнитная лента может быть использована во входных и выходных устройствах. Так как стоимость магнитной ленты невелика, то непосредственное соприкосновение между лентой и магнитной головкой вполне допустимо.

При использовании во входно-выходных устройствах лента должна двигаться с надлежащей скоростью, когда она подведена под головки, и здесь нужно учитывать, какая длина ленты проходится в процессе разгона. Для этого был разработан ряд остроумных устройств. При использовании ленты как медленнодействующего запоминающего устройства проблема состоит в быстром разгоне ленты до надлежащей скорости, чтобы можно было достичь нужного участка ленты. Для этого также были созданы исключительно эффективные устройства.

Иногда во входно-выходных устройствах используется магнитная проволока. Она занимает меньше места, может непосредственно проходить сквозь зазор магнитной головки, а не рядом с ним, и аппаратура с ней может стоить дешевле. Однако проволока менее гибка механически, менее крепка и ее труднее соединять при обрыве.

Фотографическое запоминающее устройство. Коммерческая запись аналоговых данных на фотографической пленке производится начиная с 20-х годов нашего века в виде звуковых дорожек на кинолентах. Существуют два различных способа записи: при одном из них аналогом звука служит оптическая плотность (непрозрачность), при другом — ширина изображения. Вполне естественно перенести применение фотопленки на цифровые запоминающие устройства. На 35-миллиметровой пленке можно непрерывно записывать 50 каналов данных со скоростью 10 000 битов в секунду на канал. Сама по себе фотопленка стоит очень дешево, методы обращения с ней хорошо разработаны. Очень большое количество данных может храниться в очень малом объеме (50 млн. битов на 1 $см^3$).

Из соображений надежности обычно применяются только два различных уровня оптической плотности изображения: совершенно темный и совершенно прозрачный. Для дальнейшего увеличения надежности каждый бит можно записывать на двух участках пленки: цифру 1 записывать как черное пятно слева и прозрачное пятно справа, а цифру 0 — наоборот. Оборудование для проявления пленки громоздко и неудобно, однако оно оправдывается в большой организации, занимающейся обработкой данных.

Запись данных на пленку производится от импульсных ламп или с электронно-луче-

* В американской литературе были опубликованы данные о так называемом «ленточном барабане». В этом устройстве информация записывается на бесконечной магнитной ленте шириной 25—30 *см* и длиной 2,4—135 *м*, охватывающей верхнюю половину барабана диаметром 30 *см*, вращающегося со скоростью 1 200 *об/мин*. Считывание и запись осуществляются головками, смонтированными внутри барабана и вращающимися вместе с ним. Во время считывания и записи лента останавливается. Характерной особенностью устройства является образование воздушной подушки между барабаном и лентой, что уменьшает износ последней и снижает требования к точности изготовления барабана [Д. 23]. — *Прим. ред.*

вых трубок при помощи надлежащих оптических приборов, направляющих свет на нужную часть пленки, и надлежащей аппаратуры для модуляции источника света. Для считывания используются фотоэлектрические элементы с подобной же оптической аппаратурой.

Хотя фотографическая пленка и использовалась таким образом для хранения цифровых данных, наибольшую ценность она представляет для запоминания чрезвычайно больших массивов данных (миллиардов битов, например) в форме микрофильмов, миникарт и т. п., о чем будет говориться в конце этого параграфа.

Перфокарты. Перфокарты, используемые в вычислительной технике, носят общее название голлеритовских карт, по имени изобретателя Голлерита, который впервые применил их в 1889 г. Хотя в деловом учете для разных целей используется много видов перфокарт (как, например, счета, посылаемые по почте клиентам), однако, если не оговорено противное, под термином «перфокарта» понимается стандартная карта размерами 82,5×180×0,17 *мм*, у которой один угол срезан для удобства работы. Эти перфокарты выпускаются различными компаниями, каждая из которых использует свой метод перфорации *буквенно-цифровой* (т. е. буквенной и/или цифровой) информации. В двух самых распространенных системах на карте имеется соответственно 80×12 и 90×6 позиций для пробивки отверстий. Если на колонку приходится 12 позиций, то достаточно одного отверстия для записи десятичной цифры и двух отверстий для записи буквы; если на колонку приходится 6 позиций, то обычно пробиваются два или три отверстия на одну цифру или букву *.

Каждое отверстие в голлеритовской карте пробивается механически. Перфорируемая буквенно-цифровая информация, иногда вместе с дополнительными данными, часто печатается на картах одновременно с перфорацией, так что карты легко опознаются и используются зрительно. Считывание с карты производится автоматически маленькими металлическими щупами, которые образуют электрический контакт через пробитые отверстия, а иногда — более быстрыми фотоэлектрическими методами.

Голлеритовские карты широко используются, и не только как средство запоминания медленного действия, но и для комплектования картотек, а также как внутреннее запоминающее устройство в большом классе среднедействующих электронных и медленнодействующих электромеханических вычислительных машин, в совокупности называемых *счетно-перфорационными машинами.* Эти машины существуют в большом количестве типов и размеров и могут сочетаться различными способами для образования вычислительных систем. В тех случаях, когда данные поступают в цифровой форме и когда при сравнительно большом объеме входа — выхода на каждую порцию информации приходится небольшой объем вычислений, счетно-перфорационные машины могут оказаться более эффективными, чем большие автоматические машины, особенно в небольших вычислительных бюро.

Основными типами счетно-перфорационных машин являются *сортировальная машина* («сортировка»), *табулятор* и *вычислительный перфоратор* («вычислитель»). Карты могут сортироваться по любой нужной колонке; например, чтобы снова расположить по порядку 10 000 карт, занумерованных в последовательном порядке и затем перемешанных совершенно случайно, их можно просортировать 4 раза, по каждому из четырех разрядов номера, в целом приблизительно за 1 час. Более совершенным типом сортировальной машины является *карторасскладочная машина,* которая, кроме того, производит сравнение карт друг с другом и с хранящимися в памяти значениями, например с таблицами функций.

Табулятор печатает искомые результаты или комбинации результатов с карт, и притом либо на другие карты (включая перфорацию), либо в табличной форме на листах бумаги (табулограммах); эти листы бумаги могут быть приспособлены для офсетной печати с них, благодаря чему исключается возможность внесения ошибки из-за переписывания данных вручную. Существует много разных типов вычислительных перфораторов, от простых механических (релейных) устройств до

---

* Карта 80×12 содержит слева направо 80 вертикальных колонок по 12 позиций в каждой, и в каждой колонке записывается (может быть записано) по цифре или букве. Такие карты широко распространены в СССР. В США их применяет Международная корпорация деловых машин (IBM). Карта 90×6 содержит слева направо только 45 вертикальных колонок по 12 позиций; но они разделены по всей длине на верхнюю и нижнюю секции по 6 позиций, что и дает в итоге 90 колонок по 6 позиций. В каждой из этих 90 колонок записывается по цифре или букве. Такие карты применяются, например, в США (фирма «Ремингтон Рэнд») и в Чехословакии (народное предприятие «Аритма»).

Карта 90×6 получилась в результате преобразования карты 45×12, применявшейся ранее фирмой «Пауэрс». Размеры всех этих карт совершенно одинаковы; отверстия в карте 90×12 пробиваются овальные и потому у́же, чем в картах 45×12 и 90×12, где они круглые. — *Прим. ред.*

электронных устройств с карточным программированием, которые почти так же быстры и гибки, как и большие машины, описываемые в следующей главе.

Перфолента. Существует большое количество различных видов перфолент. Обычно они делаются из бумаги, хотя иногда используются и более прочные материалы, и, как правило, имеют одну сплошную дорожку пробивок для протяжки и синхронизации, хотя их можно протягивать и при помощи фрикционных механизмов.

Число позиций для пробивки поперек ленты различно для разных типов ленты. Лента для телетайпа содержит 5 позиций, и любые из них или же все они вместе могут быть пробиты; это дает 32 различные комбинации, используемые для 26 букв * и 6 каких-либо других символов (могут быть также одна или две контрольные позиции). Система автоматического учета переговоров Белловской телефонной системы (см. § 2.2) использует 6-позиционную ленту, в которой всегда пробиваются в точности две позиции в колонке; это дает 15 возможных комбинаций и, кроме того, обеспечивает надежность, так как наличие в колонке более чем двух или менее чем двух пробитых отверстий свидетельствует об ошибке.

Перфолента перфорируется механически, хотя проводились опыты по перфорации отверстий с помощью электрической дуги. Считывание с ленты производится либо при помощи механических щупов, как для перфокарт, либо фотоэлектрическим путем. Фотоэлектрические устройства считывания могут считывать за 1 *мсек* по нескольку позиций в колонке. Перфоленты изумительно прочны и могут легко копироваться перед окончанием своего ожидаемого срока службы. Ленты стоят дешево и хранят довольно большое количество информации в небольшом объеме.

Другие типы медленнодействующих запоминающих устройств. До сих пор мы говорили о запоминающих устройствах, технические требования к которым изменялись от нескольких тысяч битов с доступом до микросекунды до многих миллионов битов с доступом до миллисекунды или секунды. Но существует другой класс запоминающих устройств для систем, которые могут нуждаться в хранении многих миллиардов битов со временем доступа порядка секунд и даже минут. Такие системы иногда бывает чрезвычайно трудно осуществить ввиду большого числа категорий классификаций или большой трудности классификации.

Системы хранения данных этого типа практически представляют собой системы картотек. Примеры нужды в них приводились в начале § 2.4: журналы с миллионами подписчиков; Управление материально-технического обеспечения ВВС США, имеющее 1 200 000 различных предметов снабжения в своем складском учете; система социального обеспечения, в которой производятся операции над большими суммами денег для 10^8 человек; страховые компании с 10^7 держателей полисов; проблемы кадров, закупок, складирования и производства в компании, выпускающей миллионы автомобилей в год; и т. д. Все это, конечно, системы большого масштаба со всем многообразием свойственных им проблем. Но в этом месте мы хотим рассмотреть только проблему хранения данных.

Для таких задач вполне возможно записывать по 10^{10} и даже по 10^{12} битов на магнитной ленте или перфоленте или даже на перфокартах или фотопленке. Можно также придумать много способов, обеспечивающих автоматический доступ к нужному рулону ленты или к нужной пачке карт, однако добиться малого времени доступа при этом трудно. Поскольку, однако, финансовая отдача от повышения эффективности таких систем весьма велика, на разработку их затрачиваются очень большие усилия. Читатель отсылается к текущей литературе за обсуждением этой проблемы, которая разрабатывается настолько быстро и всесторонне и в условиях такой коммерческой тайны, что адэкватное рассмотрение вопроса не представляется здесь возможным.

15.4. Типы цифровых вычислительных машин

Многие из рассмотренных нами схем применяются только в ограниченном классе цифровых вычислительных машин. В большинстве действительно построенных машин используется ряд остроумных методов усовершенствования этих схем для увеличения скорости работы и уменьшения требуемого оборудования.

Последовательные и параллельные машины. До сих пор мы описывали машины с последовательным представлением чисел, в которых каждый разряд следует во времени за предыдущим разрядом. Поэтому в 40-битовой машине, работающей на частоте в 1 *Мгц*, выполнение любой операции над числом, даже такой, как перенос числа из одного места в другое, требует минимум 40 *мксек*. Возмож-

* Латинского алфавита. — *Прим. ред.*

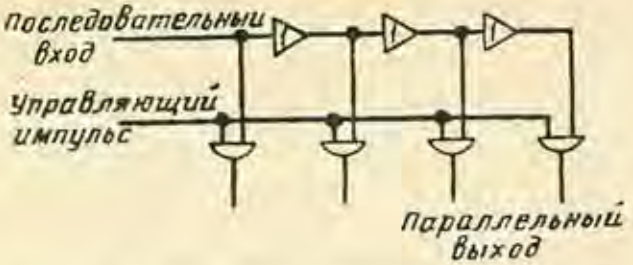


Рис. 15.29. Устройство перевода чисел из последовательной формы в параллельную.

но, однако, все операции над числом выполнять не последовательно по разрядам, а параллельно, так что перенос числа будет осуществляться за 1 *мксек* по 40 отдельным проводам.

Логически перевод числа из параллельного представления в последовательное и обратно выполняется, как показано на рис. 15.29 и 15.30. Так, для перевода четырехразрядного двоичного последовательного числа в параллельную форму нужно подать это число в схему на рис. 15.29, затем через три единицы времени направить в клапаны управляющий импульс, и в этот момент все разряды числа появятся параллельно на выходных линиях. На практике такое преобразование чисел из последовательной формы в параллельную проще всего выполняется с помощью регистра сдвига. Как рассказывалось в § 15.2, числа могут подаваться в такой регистр последовательно, храниться там сколь угодно долго и затем считываться параллельно. И обратно, числа могут записываться в триггерах параллельно и затем выдаваться в последовательной форме с помощью сдвигов.

Существуют полностью параллельные машины. Возможны и так называемые последовательно-параллельные машины. Например, 40-разрядное двоичное число может быть разделено на 4 группы по 10 битов в каждой и передано по четырем параллельным проводам за 10 *мксек*. Это уменьшает требуемое время на 75%, а дальнейшая экономия времени, которую можно было бы достичь, сделав машину полностью параллельной, не очень

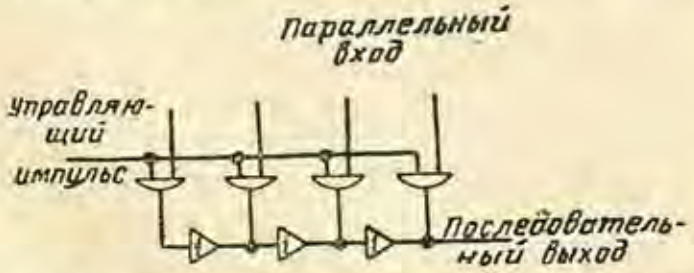


Рис. 15.30. Устройство перевода чисел из параллельной формы в последовательную.

существенна (по сравнению с другими элементами машинного времени, каковы, например, время доступа запоминающих устройств или время выполнения арифметических операций), а обходится сравнительно дорого. Кроме того, полностью параллельный динамический сумматор (см. следующий подпараграф) нельзя построить так, чтобы он выполнял сложение за один или два периода повторения импульса; для сложения двух n-разрядных двоичных чисел он должен затратить $n-1$ периодов повторения импульса *, чтобы дать возможность пройти импульсам переноса (например, при сложении 10-разрядных двоичных чисел 0 111 111 111 и 0 000 000 001 результатом будет число 1 111 111 111 и перенос должен пройти через 9 разрядов).

С другой стороны, упомянутая последовательно-параллельная машина может сложить два 40-разрядных двоичных числа за 10 *мксек* при помощи следующего трюка. Первая пара цифр вводится в один из четырех параллельных сумматоров; вторая пара цифр задерживается на $^1/_4$ *мксек* и вводится во второй сумматор вместе с цифрой переноса из первого сумматора, и т. д. Наконец, четвертая пара цифр задерживается на $^3/_4$ *мксек* и перенос из 4-го сумматора вводится также в 1-й сумматор вместе с 5-й парой цифр, которая теперь готова появиться. Для такого сложения необходимы четыре сумматора, на три входа каждый (см. задачу 15.3). В результате машина с основной частотой 1 *Мгц* может работать с такой скоростью, как если бы она была машиной на 4 *Мгц*.

Такой метод задержки на неполный период импульса типичен для цифровых машин; интервалы в $^1/_4$ *мксек* называются *фазами синхронизации*. Как отмечалось в § 15.1, при вычерчивании логических схем предполагается, что на прохождение импульса через клапан время совсем не затрачивается. На практике, конечно, для этого требуется какое-то время, и потому все схемы проектируются так, чтобы избежать бесконтрольного накопления задержек. Например, на прохождение импульса через каждый клапан можно использовать интервал в одну фазу синхронизации и после двух или трех клапанов можно провести формирование и усиление импульсов при помощи схемы рис. 15.15, получая один точно синхронизированный и хорошо сформированный импульс от задающего генератора. Таким путем возможен точный кон-

* На практике как минимум $n-1$ длительностей нарастающих фронтов импульса, за исключением специальных случаев [82]. — *Прим. авт.*

троль над синхронностью. При проектировании вычислительной машины необходимо следить самым пристальным образом за тем, чтобы все временные зависимости работали надлежащим образом.

Статические и динамические схемы. В статических схемах запоминающее устройство (например, триггер) дает на выходе уровень напряжения; такое устройство можно заставить давать при запросе импульс, если в нем хранится единица, и не давать импульса, если в нем хранится нуль. В динамических схемах соответствующее устройство выдавало бы серию импульсов, если в нем хранится цифра 1, и не выдавало бы ничего, если в нем хранится цифра 0.

Динамический триггер может быть осуществлен в виде свободно генерирующего мультивибратора, который запирается при хранении цифры 0, или же в виде одновибратора с петлей обратной связи, содержащей единичную задержку. Схема на рис. 15.13 будет свободно генерировать, если из нее удалить сопротивления R_3 и R_4.

В статических машинах каждая операция может производиться в любое время после того, как закончена предыдущая операция; различные операции часто не синхронизированы между собой.

В динамических машинах необходимо, чтобы устройство управления точно синхронизировало все операции.

Машины с фиксированным и с переменным циклом. При описании машин мы взяли в качестве примера 40-битовую машину и молчаливо предполагали, что все числа в машине были одинаковой длины. Если нам нужно сложить 1 и 1 (операция эта производится довольно часто), то нам придется послать в арифметическое устройство два 40-битовых числа, состоящих каждое из 39 нулей и одной единицы, а затем ждать 40 *мксек*, пока последовательный сумматор выполнит свою рутинную процедуру. Кроме того, так как машина построена для работы с 40-битовыми числами, то и каждая команда должна быть закодирована в виде 40-битового числа, которое будет обрабатываться внутри машины совершенно так же, как и настоящие числовые величины. (Команду или число можно обозначать общим термином *слово*).

Такая процедура, разумеется, весьма неэкономична. В машинах с переменным циклом новая операция начинается сразу же после окончания значащей части предыдущей операции; благодаря этому, например, экономится значительное время при умножениях. В последние годы были построены машины с произвольной длиной слов. Однако такие машины обходятся дорого, и поэтому иногда предпочитают компромисс, при котором машина может работать со словами определенной дробной длины, например со словами в 10 или 20 битов, наряду со словами основной длины в 40 битов.

Специализированные и универсальные вычислительные машины. Различие между специализированными и универсальными вычислительными машинами часто выражено недостаточно резко, особенно в случае цифровых вычислительных машин. Даже столь простое устройство, как пара регистров сдвига плюс сумматор, можно считать цифровой вычислительной машиной, которая в этом случае будет, конечно, специализированной. Машина, построенная для определенной цели, может быть спроектирована так, что она будет выполнять какую-нибудь одну операцию (например, извлечение квадратного корня) с особой эффективностью, так как ей придется выполнять эту операцию очень часто, или же эта машина может иметь необычайно короткое или длинное основное слово ввиду особых требований к ее точности. Такие машины часто также называются *специализированными*, хотя на них можно решать любые задачи. Термин «специализированная машина» часто относят только к машинам с жестко закоммутированной программой, так что они не способны решать эффективно или не способны решать вообще другие задачи. С другой стороны, под *универсальной вычислительной машиной* понимают обычно лабораторную вычислительную машину, предназначенную для научно-исследовательских работ, крайне гибкую в применении к широкому классу задач*.

Велись большие споры о сравнительных преимуществах и недостатках специализированных и универсальных вычислительных машин. Однако при проектировании систем нужно придерживаться простого правила — приспосабливать машину к нуждам задачи. Преимущества гибкости (универсальная машина) необходимо сопоставить с преимуществами низкой стоимости, малых размеров и высокой скорости (специализированная машина) в свете конкретной решаемой задачи, как это справедливо и для многих других проблем в системотехнике.

---

* Специализированные машины называются также *машинами специального назначения*, а универсальные машины — *машинами общего назначения*. — *Прим. ред.*

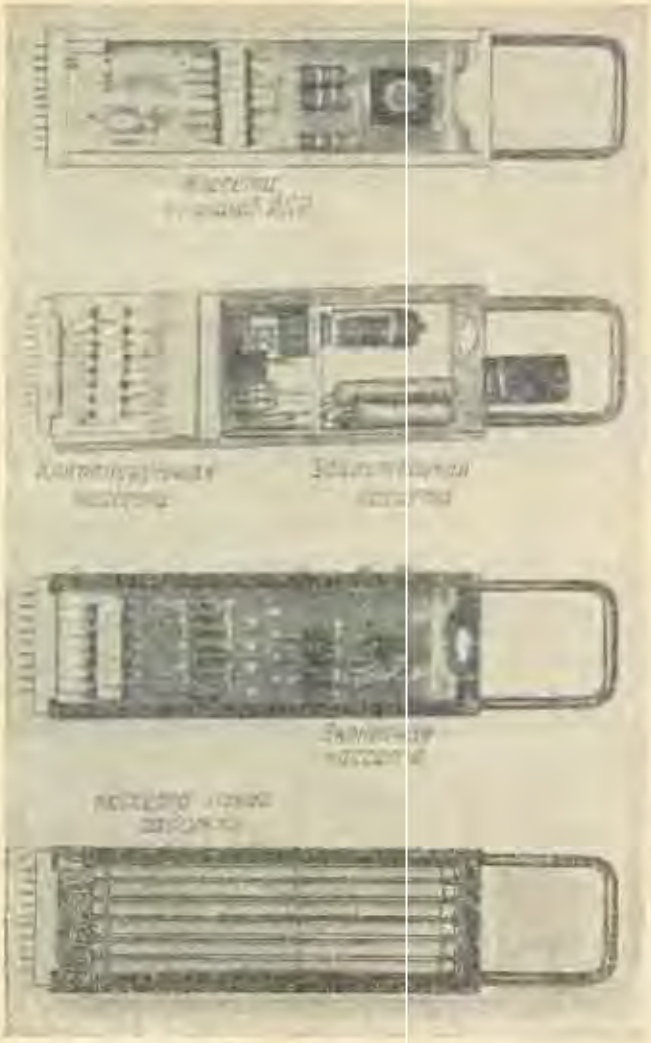


Рис. 15.31. Стандартные кассеты.

15.5. Управление

Функция управления в цифровой вычислительной машине обычно выполняется исключительно теми или иными подходящими комбинациями клапанов И, ИЛИ, НЕ и задержки, дополняемыми главным задающим генератором — источником синхронизации. Устройства управления различных машин сильно отличаются друг от друга, но все их можно описать, сказав, что устройство управления способно принять команду и выполнить ее. Команда содержит следующие пять основных элементов информации: 1) адрес первого слова, участвующего в операции; 2) адрес второго слова, участвующего в операции; 3) наименование операции; 4) адрес результата и 5) адрес следующей команды. Некоторые из этих элементов команд могут фигурировать неявно или содержаться более чем в одном «командном» слове.

Важно отметить, что в устройство управления не должны поступать и соответственно не поступают сами участвующие в операции числа — поступают только их адреса (номера ячеек запоминающего устройства). Таким образом, если машине нужно сложить числа 2 и 3, то команда этой операции будет иметь, например, следующий вид: «взять число, находящееся в ячейке памяти 132 (кодировщик знает, что в этой ячейке хранится число 2), прибавить его к числу, хранящемуся в ячейке памяти 213 (число 3), результат поместить в ячейку памяти 169 и затем получить следующую команду из ячейки памяти 035». Вся эта команда должна быть закодирована числами, например в виде следующего командного слова: 132 213 169 01 035 (где 01 есть код сложения); на практике команда была бы, по-видимому, закодирована в двоичной форме.

Итак, первая функция, которую должно осуществлять устройство управления, — это принять командное слово, разбить его на составные части и запомнить каждую часть там, где она может управлять соответствующими операциями. Запоминание производится при помощи триггеров, регистров сдвига и т. п. Управление какой-либо операцией осуществляется путем посылки разрешающих или за-

Рис. 15.32. Шасси из стандартных кассет (вид спереди).

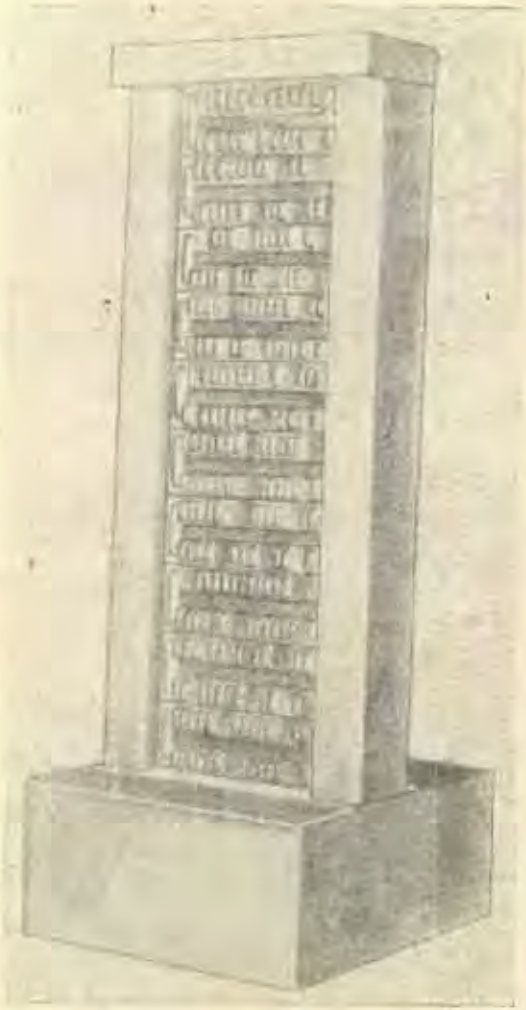


Рис. 15.33. Стойка из стандартных кассет.

Рис. 15.34. Общий вид вычислительной машины MIDSAC.*

1 — усилители считывания-записи магнитного барабана; *2* — цепи управления барабаном; *3* — устройство передачи чисел; *4* — устройство управления; *5* — арифметическое устройство; *6* — регистр сдвига; *7* — электростатическое запоминающее устройство; *8* — печатающее устройство „флексорайтер„; *9* — магнитный барабан; *10* — пульт управления.

* MIDSAC (Michigan Digital Special Automatic Computer, т. е. „Мичиганская цифровая специальная автоматическая вычислительная машина")—цифровая вычислительная машина специального назначения, находящаяся в вычислительном центре „Уиллоу Ран" при Мичиганском университете и служащая для управления режимами сложных систем в реальном времени. — *Прим. ред.*

прещающих импульсов в нужные моменты времени на соответствующие клапаны в различных устройствах машины.

Например, для выполнения операции умножения необходимо последовательно просмотреть цифры множителя и при каждом нуле производить сдвиг множимого на один разряд вправо, а при каждой единице производить сложение и сдвиг. Чтобы выполнить это сложение, устройство управления должно послать импульсы, открывающие входные клапаны сумматора, а также импульсы, соединяющие сумматор с регистром, в котором накапливается частичное произведение. Наконец, нужно уловить окончание операции и послать произведение в нужный адрес.

Кроме команд о выполнении арифметических операций, устройство управления будет получать команды о передаче данных из входного устройства в запоминающее и из запоминающего устройства в выходное, производить выбор нового шага вычислений на основании результата сравнения и выполнять некоторые другие операции, которые упоминаются в гл. 16. Понятно, сколь большая схема необходима для выполнения всех этих функций, однако каких-либо новых принципов уже не требуется.

15.6. Реальная вычислительная машина

На рис. 15.31—15.34 показана конструкция одной быстродействующей цифровой вычислительной машины. Так как эта машина — одна из первых построенных, различные функциональные части ее легко заметить и распознать. Машина использует последовательно-параллельную логику, построена на динамических схемах и работает на основной частоте 1 *Мгц*. Она складывает 31-разрядные двоичные числа за 40 *мксек* и умножает их за 88 *мксек*. Пять стоек в центре рис. 15.34 составлены из четырех типов кассет (сменных, вставных блоков с штепсельным разъемом), показанных на рис. 15.31.

В усилительно-клапанной кассете импульсы смешиваются, усиливаются, формируются и синхронизируются; клапанная часть кассеты содержит несколько многовходовых клапанов И. Кассета линий задержки содержит 16 *стержней* с задержками $^1/_4$ *мксек*. Оконечная кассета содержит сопротивления, уменьшающие отраженные импульсы, и диоды, срезающие нежелательные всплески противоположной полярности в задних фронтах импульсов. Кассета клапанов ИЛИ содержит клапаны ИЛИ и свободные диоды. Всего в вычислительной машине используется 1100 кассет, содержащих 686 электронных ламп и почти 20 000 диодов. Эти кассеты размещаются на шасси и стойках, как показано на рисунках.

ЛИТЕРАТУРА

См. гл. 16.

ЗАДАЧИ

а) Построить логическую схему для полувычитателя.

Указание. Сначала построить таблицу вычитания, подобную табл. 15.2; вместо цифр суммы и переноса в следующий разряд использовать цифры разности и займа из предыдущего разряда.

б) Построить логическую схему вычитателя.

в) Построить логическую схему блока, который будет работать как сумматор, если управляющие импульсы поступают на один провод, и как вычитатель, если импульсы поступают на другой провод.

15.2. Начертить логические схемы для триггера с одним входом (импульс на входе изменяет состояние триггера с 0 на 1 или с 1 на 0, импульс на выходе появляется только при переходе с 1 на 0) и для триггера с двумя входами (установочный импульс не оказывает влияния, если хранилась 1, и изменяет 0 на 1, если хранился 0; импульс сброса не оказывает влияния, если хранился 0, и изменяет 1 на 0, если хранилась 1, причем на выходе триггера вырабатывается импульс). Схема должна быть способна принимать импульсы, отстоящие друг от друга на единицу времени; задержки на дробную часть единицы времени не допускаются.

15.3. Построить логическую схему сумматора с тремя входами для последовательно-параллельной машины, описанной в § 15.4. Иметь в виду, что перенос не нужно подавать обратно на вход и потому задержек не требуется.

Указание. Использовать три трехвходовых клапана, в которых сигналы на выходе образуются, когда на входах имеются соответственно один, два и три сигнала.

ГЛАВА 16

РАБОТА ЦИФРОВЫХ ВЫЧИСЛИТЕЛЬНЫХ МАШИН

Мы уже говорили, что цифровая вычислительная машина решает задачи таким же по существу путем, как и человек. Принципиальная разница состоит в том, что автоматическая вычислительная машина работает гораздо быстрее, а также в том, что по сравнению с человеческим мозгом она почти невероятно глупа. В связи с этим последним обстоятельством все действия машины должны быть весьма тщательно запрограммированы до на-

чала вычислений. Должен быть подробно разработан каждый шаг вычислений, каждое логическое решение, которое машина должна делать на соответствующих этапах, каждый критерий, положенный в основу этих решений. Должны быть предусмотрены все возможные случайности. Например, если попытаться сложить 600 000 000 и 600 000 000 на настольной счетной машине, фиксирующей только 9 десятичных разрядов, то вместо правильного ответа 1 200 000 000 машина выдаст 200 000 000. Человек-оператор заметит ошибку и исправит ее при записи результата в рабочую таблицу. Автоматическая же вычислительная машина, сделав подобную ошибку, будет продолжать решение задачи, давая неверные ответы, быть может, при миллионах дальнейших вычислений.

Ошибка этого частного вида называется *переполнением*. Существуют три основных метода предотвращения ее (а также многих других ошибок). Во-первых, можно ввести сигнализацию об ошибке для оператора, который остановит машину и исправит ошибку; во-вторых, можно дополнительно усложнить машину, чтобы она сама автоматически предотвращала ошибку; и, в-третьих, можно избежать ошибки путем более тщательного программирования.

Первый метод (сигнализация), очевидно, нежелателен, так как для исправления ошибки необходимо прерывать решение задачи и проверять программу; тем не менее этот метод применяется по необходимости в большинстве машин, в которых не предусмотрено автоматическое предотвращение ошибки переполнения.

Метод автоматического предотвращения ошибки переполнения заключается в применении плавающей запятой (§ 14.3). При суммировании или вычитании с плавающей запятой одно из чисел сдвигается, при умножении или делении показатели чисел складываются или вычитаются. При этом остается возможность переполнения для показателя, однако эта проблема гораздо менее серьезна.

В машинах с фиксированной запятой методом предотвращения ошибки переполнения является *подбор масштабов*, т. е. умножение участвующих в операции чисел на подходящие масштабные множители согласно программе. Подбор масштабов обсуждается более подробно при описании аналоговых вычислительных машин.

В § 16.3 приведен пример программирования задачи, где показаны некоторые приемы предотвращения простых ошибок.

16.1. Число адресов

Как уже отмечалось в § 15.5, команда содержит пять основных элементов информации: четыре адреса и код операций *. Существуют машины, использующие команды такого вида; они называются *четырехадресными машинами*. При кодировании программ для таких машин обнаруживается, что в большинстве случаев адрес каждой последующей команды оказывается на единицу бóльшим, чем адрес предыдущей команды. Поэтому оказывается целесообразным вообще не указывать адрес следующей команды, а сделать так, чтобы машина выбирала адреса команд в последовательном порядке. Это осуществляется при помощи счетчика команд, содержимое которого автоматически увеличивается на единицу после каждой операции. Имеются также специальные операции для установки счетчика команд в нужное положение в том случае, когда желательно нарушить нормальный порядок возрастания адресов команд. Машины, работающие с командами такого вида, называются *трехадресными*.

Главным преимуществом трехадресной машины по сравнению с четырехадресной является возможность увеличения внутренней памяти. Рассмотрим, например, 40-битовую машину, т. е. машину, ячейки памяти которой имеют по 40 двоичных разрядов. Каждое числовое слово должно состоять из 39 двоичных цифр и знака и эквивалентно по разрядности приблизительно 12 десятичным цифрам.

Такая большая разрядность чисел обычно требуется в универсальных машинах вследствие накапливающейся ошибки округления при длинных вычислениях (например, при обращении матрицы 20-го порядка). Командное слово также должно содержать 40 двоичных цифр, так как всегда желательно, чтобы длина числовых и командных слов была одинаковой; увеличивать же длину используемых в машине слов более, чем это диктуется соображениями разрядности, мы не хотим, так как это потребовало бы лишнего оборудования и лишнего машинного времени.

Предположим, что требуется 16 различных операций; тогда для задания кода в командном слове будут нужны четыре двоичных разряда. На каждый из четырех адресов в команде остается по 9 двоичных разрядов,

---

* Предполагается, что только два числа вводятся в арифметическое устройство одновременно. Так обычно и бывает в цифровых вычислительных машинах, хотя, конечно, можно построить специализированную машину, которая будет складывать одновременно более двух чисел. — *Прим. авт.*

и возможное число адресов не превышает 2^9, т. е. 512. Таким образом, все *внутреннее запоминающее устройство* машины ограничивается 512 словами, включая все команды и константы, используемые при вычислениях, и промежуточные результаты. Часть информации можно хранить вне машины (например, на магнитном барабане или даже на магнитной ленте или перфолентах) и вводить по мере надобности во внутреннюю память (внутреннее запоминающее устройство), однако может оказаться желательным увеличить именно внутреннюю память. В машинах с более короткими словами ограничение внутренней памяти оказывается еще более серьезным.

В трехадресной 40-битовой машине каждый адрес может содержать 12 битов, допуская тем самым 4096 слов во внутренней памяти. Если такого количества слов не требуется, то можно использовать для каждого адреса 11 битов, а оставшиеся 3 бита использовать для дополнительных операций или для других целей управления машиной.

Существуют также *одноадресные* машины. В таких машинах сложение производится в три команды: 1) ввести первое число в арифметическое устройство; 2) ввести второе число в арифметическое устройство и сложить; 3) перенести число, стоящее теперь в арифметическом устройстве, в заданную ячейку памяти. В *двухадресной* машине второй адрес является адресом следующей команды. Используются и другие комбинации.

В трехадресной машине необходимые перестановки счетчика команд производятся обычно как результат операции логического решения. Полезность автоматической вычислительной машины при решении практических задач зависит от ее способности изменять свою программу посредством подобных логических решений, которые производятся обычно путем сравнения. При команде сравнения не возникает необходимости запоминать какой-либо числовой ответ и потому место, занимаемое третьим адресом, может быть использовано в этой команде для адреса второй возможной следующей команды. Команду сравнения можно выразить словами следующим образом: «Перенести в арифметическое устройство числа, определяемые двумя первыми адресами команды, и вычесть второе число из первого; если результат есть нуль или отрицательное число, то не предпринимать ничего (т. е. содержимое счетчика команд автоматически увеличивается на единицу и машина переходит к следующей по порядку команде); если же результат положителен, то установить счетчик команд на число, содержащееся по третьему адресу».

16.2. Коды операций

В табл. 16.1 показаны 16 операций гипотетической трехадресной 40-битовой машины, которая используется в приводимых ниже примерах. Символы α, β и γ обозначают три адреса, а символы α', β' и γ' — соответственно слова, хранимые по этим адресам. Если, например, число 1 хранится в ячейке памяти № 067, то $\alpha' = 1$ и $\alpha = 067$.

Таблица 16.1

Операции для гипотетической трехадресной 40-битовой машины

Номер операции	Название	Значение символов α, β, γ
01	Сложение	Записать $\alpha' + \beta'$ в γ
02	Вычитание	Записать $\alpha' - \beta'$ в γ
03	Умножение (нижнее)	Записать последние 40 битов произведения $\alpha' \times \beta'$ в γ
04	Умножение (верхнее)	Записать первые 40 битов произведения $\alpha' \times \beta'$ в γ
05	Умножение (округленное)	Записать первые 39 битов произведения $\alpha' \times \beta'$ в γ; 40-й бит числа γ' положить равным 1
06	Деление	Записать $\alpha' : \beta'$ в γ
07	Определение порядка	Сосчитать число нулей в α' слева от самой старшей единицы и записать результаты в γ
08	Сдвиг	Записать $\alpha' \times 2^{\beta}$ в γ
09	Выборка	Для каждого бита в β', равного 1, подставить соответствующий бит из α' на место соответствующего бита в γ'
10	Перевод из десятичной системы в двоичную	Перевести число в α и β в двоичную систему и записать в γ
11	Перевод из двоичной системы в десятичную	Перевести α' в двоично-десятичную систему и записать в β и γ
12	Сравнение (абсолютное)	Если $(\alpha') > (\beta')$, перебросить счетчики команд на γ
13	Ввод	Принять ближайшие α слов из входного устройства № γ и записать их в ячейках $\beta, \beta+1, \ldots, \beta+\alpha-1$ соответственно
14	Вывод	Вывести α слов, записанных в ячейках № $\beta, \beta+1, \ldots, \beta+\alpha-1$, через выходное устройство № γ
15	Сравнение (алгебраическое)	Если $\alpha' > \beta'$, перебросить счетчик команд на γ
16	Остановка	

Операции № 03 и 04 используются при вычислениях с двойной разрядностью, т. е. при операциях над 80-битовыми числами.

Возьмем некоторое 80-разрядное число x и обозначим его 40 старших разрядов через A, а 40 младших — через B. Тогда $x=A+B\cdot 2^{-40}$. Затем умножим x на другое число $y=C+D\times 2^{-40}$. Произведение имеет вид

$$xy=AC+(BC+AD)\cdot 2^{-40}+BD\cdot 2^{-80}.$$

Сначала образуем обе части произведения AC, затем — старшие части произведений BC и AD; складываем их и прибавляем сумму к младшей части произведения AC. Все переносы от этих двух сложений можно прибавлять к старшей части произведения AC; правда, при этом не будут учитываться возможные переносы округления в 80-й разряд из младших частей произведений BC и AD и старшей части произведения BD (при желании эти переносы можно было бы принять во внимание). Подобными методами можно достичь еще большей разрядности (например, 120 битов), однако эти методы требуют значительной памяти и машинного времени.

Операция № 05 используется при обычном умножении; приравнивание последнего разряда к 1 представляет собой метод случайного округления, так как в среднем единица встречается в последнем разряде в половине всех случаев, нуль — в другой половине и в среднем можно увеличивать последний разряд на единицу тоже в половине всех случаев. Обычно применяются более сложные методы случайного округления, так как данный простой метод может значительно способствовать росту кумулятивной (накопительной) ошибки округления.

Операция № 07 используется для нахождения абсолютной величины числа, например при автоматическом подборе масштаба. Часть β команды не используется.

Операция № 08 — это просто сдвиг двоичной запятой на число разрядов, указанное в β.

Операция № 09 применяется для изменения команд. Нули в β' используются для обозначения тех отрезков γ', которые мы хотим оставить без изменения.

Операции № 10 и 11 являются примерами дополнительных операций, которые могут использоваться в машине для упрощения кодирования и других аналогичных целей. Если при вводе числа в машину или выводе из машины требуется вся имеющаяся разрядность, то каждое двоично-десятичное число (см. § 14.2) должно занимать при хранении две ячейки памяти.

Операции № 12 и 15 суть операции сравнения. В обоих случаях, если второе число равно первому или больше него, устройство управления не предпринимает никаких особых действий, т. е. счетчик команд увеличивает свое содержимое на 1.

Операции № 13 и 14 используются для связи вычислительной машины с внешним миром. Так как входных устройств может быть два и более (например, электропишущая машинка и фотоэлектрическое считывающее устройство), то часть γ в командах 13 и 14 используется для указания, к какому именно входному устройству подключаться машине. Предполагается, что лента на входном устройстве находится в нужной позиции. Операцию № 13 нельзя применять для считывания слов с магнитного барабана во внутреннюю память, так как при этом нужно еще указать «барабанный» адрес; для работы с барабаном требуется видоизмененная команда.

Для практической машины нужна еще одна операция, а именно «пуск». Будем считать, что наша условная машина имеет кнопку пуска, при нажатии которой с ленты входного устройства считывается первое слово и направляется в ячейку памяти 000, а счетчик команд устанавливается на 000.

16.3. Пример программирования и кодирования

Нас интересует не столько решение чисто математических задач на цифровых вычислительных машинах, как использование их для логического управления системами большого масштаба. Однако мы научимся лучше понимать системные задачи, если сначала рассмотрим простую математическую задачу. Решение логической задачи рассматривается в гл. 22.

В качестве примера выбрана следующая задача: найти корни многочлена

$$y=Ax^4+Bx^3+Cx^2+Dx+E. \qquad (16.1)$$

График этого многочлена приведен на рис. 16.1. Предположим, нам известно, что все четыре корня действительны и лежат в интервале $0<x<10$ и любые два корня отличаются между собой не менее чем на 0,001; будем считать, что больше нам ничего не известно. Будем решать эту задачу методом «грубой силы», вычисляя значение многочлена для каж-

Рис. 16.1. Многочлен 4-й степени, все корни действительны и положительны.

дого последовательного значения аргумента x через каждые 0,001 начиная с нуля и до тех пор, пока мы не найдем все четыре корня. О прохождении через корень будет свидетельствовать изменение знака y.

Итак, наша вычислительная процедура будет иметь следующий вид: вычислить y; исследовать знак y и определить, изменился ли он по сравнению с прошлым разом; если не изменился, взять следующее по порядку значение x и снова вычислить y; если изменился, записать результат и отметить, что мы нашли один из корней; чтобы знать, когда остановиться, проверить, нашли ли мы уже все четыре корня; если нет, взять следующее значение x и вернуться к началу процедуры; если да, остановиться.

На практике такой метод не стали бы применять. Для решения подобных задач разработаны мощные аналитические методы, которые позволяют точно определить каждый корень уже после немногих итераций. Однако следует отметить, что описанный метод не является слишком нереалистическим. Машина работает столь быстро, что вычисление многочлена для рассматриваемых 10 000 значений переменной x занимает всего несколько секунд. Если только мы не хотим многократно повторять эти вычисления, то вполне можно пожертвовать несколькими секундами машинного времени ради того, чтобы избежать работы по составлению программы для более сложного метода.

При вычислении многочлена нам, разумеется, нецелесообразно сначала вычислять x^4, затем умножать x^4 на A, затем начинать с самого начала вычисление x^3 и т. д. Такой метод будет весьма неэкономичен. Вместо этого мы можем начать с вычисления x^2, x^3, x^4. Тогда для вычисления многочлена понадобится совершить семь умножений и четыре сложения. Так как умножения отнимают больше времени, чем сложения, мы попытаемся уменьшить их число. Одним из методов является деление уравнения (16.1) на A.

Благодаря четырем делениям, которые производятся лишь один раз, мы избавляемся от одного умножения, которое производится тысячи раз. В результате деления получаем уравнение

$$z=\frac{y}{A}=x^4+\frac{B}{A}x^3+\frac{C}{A}x^2+\frac{D}{A}x+\frac{E}{A}=$$
$$=x^4+Px^3+Qx^2+Rx+S. \qquad (16.2)$$

Для вычисления z необходимо шесть умножений и четыре сложения.

Число умножений можно еще уменьшить, перегруппировав (16.2) следующим образом:

$$z=S+x\{R+x[Q+x(P+x)]\}. \qquad (16.3)$$

Уравнение (16.3) требует только трех умножений и четырех сложений. Теперь мы можем составить блок-схему вычислений, которая показана на рис. 16.2. Этим кончается процесс *программирования*, если отличать его от *кодирования* *.

Рис. 16.2. Блок-схема вычисления корней многочлена.

Таблица 16.2

Номер команды	α	β	γ	Операция	Примечания
000	022	001	001	13	Ввод команд
001	011	101	001	13	Ввод констант
002	112	112	112	02	Начало подсчета корней
003	103	102	103	06	Вычисление P, Q, R, S
004	104	102	104	06	
005	105	102	105	06	
006	106	102	106	06	
007	101	109	101	01	Фиксация x_i
008	101	103	102	01	Вычисление z
009	102	101	102	05	
010	102	104	102	01	
011	102	101	102	05	
012	102	105	102	01	
013	102	101	102	05	
014	102	106	102	01	
015	102	107	007	15	Изменился ли знак?
016	015	107	102	01	Изменение команд
017	111	107	015	01	
018	102	107	111	01	
019	001	101	002	14	Печатание корня
020	112	108	112	01	Подсчет корней
021	110	112	007	15	Закончена ли задача?
022	000	000	000	16	Остановка

Самый код приведен в табл. 16.2 и 16.3. В табл. 16.2 в первом столбце помещены порядковые номера команд, во втором и в треть-

* См. примечание на стр. 172. — *Прим. ред.*

Таблица 16.3

Адрес	Содержимое	
	при вводе в машину	позже
101	x_0	x
102	A	Частный многочлен
103	B	P
104	C	Q
105	D	R
106	E	S
107	0	То же
108	1	"
109	0,001	"
110	4	"
111	107 102 007 15	102 107 007 15
112	—	Число корней

ем столбцах — адреса двух участвующих в операциях слов (α' и β'), в четвертом — адрес результата (γ') и в пятом—коды операций. В табл. 16.3 мы для удобства приводим список ячеек памяти, в которых хранятся константы и промежуточные результаты. На вводимой в машину ленте записаны данные из третьего, четвертого и пятого столбцов табл. 16.2 и второго столбца табл. 16.3.

Рассмотрим теперь подробно каждую команду. При нажатии кнопки пуска в ячейку памяти 000 вводится первое число с ленты. Это число представляет собой первую команду, согласно которой происходит считывание всех остальных команд. Первая команда содержит указание для машины считать с устройства 001 (лента) 22 команды и записать их в ячейки памяти, начиная с ячейки 001. Когда мы только начинаем кодирование, мы не знаем, сколько всего будет команд в программе, поэтому в команде 000 первый адрес α мы пока не заполняем. Впоследствии мы запишем в этом месте число 22.

Команда 001. Согласно этой команде одиннадцать следующих чисел на ленте записываются в соответствующие ячейки памяти, начиная (для удобства) с ячейки 101. Этим производится ввод констант.

Команда 002. Если мы хотим подсчитать число найденных корней, то нам нужно начать с нуля. Для этого в ячейку 112 вводится 0.

Команды с 003 по 006. После вычисления P число B нам не нужно. Поэтому мы стираем его, записывая P на место B. При кодировании практической задачи место в запоминающем устройстве часто ограничено, и такой метод экономии ячеек широко применяется.

Команда 007. На этом этапе мы не прибавляем на блок-схеме приращение (0,001) к аргументу x. Однако прибавлять приращение здесь оказывается удобным (см. конец параграфа), и никаких затруднений это не причиняет. Нужно только ввести в ячейку 101 константу — 0,001, если мы хотим, чтобы первое значение z было вычислено для $x=0$.

Команда 008. Число A нам больше не нужно, и мы используем занимаемое им место (ячейку 102) для первой частной суммы $x+P$.

Команды с 009 по 013. Ячейку 102 мы продолжаем использовать для хранения промежуточных результатов, так как каждый результат используется только один раз. Если бы он был нужен и в дальнейшем, нам пришлось бы отвести для его хранения отдельную ячейку.

Команда 014. В результате этой операции получаем z.

Команда 015. Это критическая операция всего вычисления. Мы хотим определить, нашли ли мы уже корень или нет. С этой целью мы задаем машине логический вопрос. В логической блок-схеме в этом месте имеется разветвление, и дальнейший ход решения задачи зависит от результата решения этого логического вопроса. В использованном методе учитывается то обстоятельство, что z должен быть положительным в (16.2) для больших по абсолютной величине значений x и что тем самым при прохождении первого корня знак z должен измениться с плюса на минус. Таким образом, мы можем сравнивать алгебраическое значение z с нулем. Если z больше нуля, то мы еще не нашли первого корня и должны вернуться к команде 007. Если z меньше нуля (т. е. значение z отрицательно), то нам нужно сделать ряд новых операций. Первой из них, как мы увидим ниже, будет изменение команды 015.

Для обнаружения перемены знака можно было бы использовать и другие методы. Согласно одному из них первое вычисленное значение z запоминается в некоторой ячейке памяти, второе вычисленное значение z запоминается в другой ячейке памяти. Затем производится сравнение знаков этих двух значений z. Если знаки обоих z оказываются одинаковыми, то содержимое второй ячейки памяти переносится в первую ячейку и машина переходит к вычислению следующего значения z. Сравнение знаков может производиться путем перемножения двух значений z и сравнения алгебраического значения произведения с нулем или путем выделения знаковых разрядов при помощи команды выборки (по этой команде в некоторую ячейку записывается цифра 1, если знак числа есть плюс, и цифра 0, если знак — минус) и последующего вычитания

этих знаковых разрядов одного из другого и сравнения абсолютной величины разности с нулем.

Обычно, как и в данном случае, имеется целый ряд способов кодирования логического решения и опытный кодировщик * должен выбрать среди них наиболее подходящий для решаемой задачи, учитывая при этом относительную стоимость времени кодирования, машинного времени, емкости памяти и гибкости (выбранный в нашем примере метод не гибкий, так как в нем предполагается знание знака первого вычисленного значения z).

Команды с 016 по 018. Новая команда, хранимая в ячейке 111, отличается от прежней команды перестановкой α и β. Заметим, что над командой (т. е. над содержимым ячейки 015) совершаются операции точно так же, как над числом. В нашем случае к команде с целью переноса в другую ячейку прибавляется число 0. Во многих программах изменение какой-либо команды производят путем последовательного вычитания из нее числа 1. Это делают для того, чтобы не записывать в программе серию почти одинаковых команд. Вычислительная машина может производить над командой любую арифметическую операцию таким же образом, как над числом. После того как последовательность команд 016, 017, 018 выполнена, прежняя команда 015 оказывается в ячейке 111, откуда она после нахождения второго корня будет возвращена в ячейку 015.

Команда 019. По этой команде машина печатает в устройстве печати число x_r, которое служит признаком того, что корень x лежит в интервале $x_r - 0{,}01 \leqslant x \leqslant x_r$. До окончания вычислений будут отпечатаны четыре таких значения.

Разумеется, в этом месте вычислений можно ввести подпрограмму, по которой машина вычтет 0,001 из рассматриваемого значения x и затем начнет вычислять значения многочлена через интервалы приращения x, равные, скажем, 0,000 001. Закончив подпрограмму, машина вернется к исходной программе. Это небольшое усложнение дает некоторое представление о силе вычислительных методов, которые используются на практике.

Команды 020 и 021. Так как машина не обладает собственным разумом, она будет работать без конца, если мы не прикажем ей остановиться. По команде 020 мы подсчитываем число уже отпечатанных корней, а по команде 021 сравниваем это число с числом 4. Это сравнение может быть как алгебраическим, так и абсолютным. В командах сравнения необходимо обращать внимание на то, входит ли понятие «равно» в понятие «больше» или в понятие «меньше». Наша гипотетическая машина в операции 015 включает понятие «равно» в понятие «меньше», т. е. считается, что $\alpha' < \beta'$ даже при $\alpha' = \beta'$. Если бы считалось, что $\alpha' > \beta'$ даже при $\alpha' = \beta'$, то в ячейку 110 нам нужно было бы поместить вместо константы 4 константу 3. Аналогично, если бы установка счетчика команд в положение 007 производилась при $\alpha' < \beta'$, а не наоборот, то нам пришлось бы в команде 021 поменять местами α и β.

Тонкости такого рода весьма затрудняют кодирование и приводят к тому, что каждую программу перед пуском в машину требуется «отлаживать». Подбор маштабов также затрудняет практическое кодирование, но мы здесь не будем углубляться в эту проблему. Мы не затронули также вопросов перевода чисел из двоичной системы в десятичную и обратно. При кодировании следует также предусмотреть многие другие возможные затруднения (деление на 0, мнимые корни, корни, лежащие друг к другу ближе чем на 0,001, и т. д.).

Предположим теперь, что мы составили программу в точности согласно табл. 16.2, но задание нового x_i (команда 007) производим только после операции сравнения. В этом случае команда сравнения будет иметь номер 014. В результате сравнения машина в одном случае будет переходить к прежней команде 016, а в другом случае — к новой команде 015 (прежняя команда 007). Трудность заключается в том, что мы теперь не можем просто перейти к команде 008. В четырехадресной машине переход к этой команде не представлял бы трудности, но в трехадресной машине для этого перехода нам нужно ввести искусственную операцию сравнения (например, 110 108 008 15, что означает: если 4 больше, чем 1, — а это, конечно, так, — то перейти к операции 008). Таким образом, в программе оказывается на одну команду больше; требуется лишняя ячейка памяти для ее запоминания и несколько микросекунд для ее исполнения. Поэтому выбранный нами первоначально метод программирования более удобен.

Если в программе обнаруживается ошибка и должна быть добавлена новая команда, то все следующие за ней команды должны быть перенумерованы заново. Кодировщик должен при этом проследить, чтобы соответственно были изменены все ссылки на номера этих команд (например, должно быть измене-

* «Кодировщик» в смысле авторов — это «программист» в обычном, широком смысле. — *Прим. ред.*

но число, хранящееся в ячейке 111). В четырехадресной машине перенумерацию команд производить не нужно. Легкость исправления программ и является некоторым преимуществом четырехадресной машины.

16.4. Подпрограммы

Подпрограммой можно называть любую группу команд, пригодную для работы с различными значениями аргумента, например группу команд от 007 до 015 в табл. 16.2. Однако этот термин чаще относят к таким группам команд, как вся группа команд табл. 16.2, которая могла бы служить подпрограммой для нахождения корней многочлена 4-й степени. Вычислительные центры составляют «библиотеки» часто используемых подпрограмм, к которым можно обращаться всякий раз, когда нужно.

Подпрограммы могут использовать разложения в ряд, например для функций $\sin x$ и e^x. Они могут также использовать быстро сходящиеся итерационные процедуры. Например, для нахождения $y=1/x$ (в машине без команды деления) мы можем воспользоваться следующей итерационной формулой:

$$y_{k+1}=y_k(2-xy_k). \qquad (16.4)$$

Пусть, например, мы хотим найти по этой подпрограмме обратную величину числа 9 и в качестве первого приближения y_1 взяли число 0,1. После двух итераций получим $y_2=0{,}11$ и $y_3=0{,}1111$, верные соответственно в двух и четырех десятичных знаках.

Для извлечения корня $y=\sqrt{x}$ существует аналогичная формула

$$y_{k+1}=\frac{y_k+x/y_k}{2}. \qquad (16.5)$$

При применении этой формулы в качестве первого приближения можно взять $y_1=x$. Более сложные методы потребовали бы больше команд, но зато и меньше машинного времени; например, для нахождения y_1 можно было бы применить команду определения порядка, чтобы найти количество разрядов между двоичной запятой и первой единицей в числе, и затем сдвинуть x на половину этого количества разрядов. Формула (16.5) сходится очень быстро, в особенности когда y_k уже близко к правильному ответу. Так, если $x=100$ и $y_1=100$, то последовательные значения y_k будут: 50,5; 26,24; 15,02; 10,84; 10,03; 10,00005.

16.5. Автоматическое программирование [106]

Хотя кодирование программы представляет собой, по крайней мере теоретически, сравнительно несложную процедуру, на практике оно требует больших затрат труда опытных специалистов. Пока код программы перфорируется, отлаживается и будет, наконец, готов ко вводу в машину, проходит несколько недель и затрачивается около 5 долл. на одну команду. Методы, позволяющие машине делать самой часть этой работы, называются обычно *автоматическим программированием*, хотя при нашем различении между программированием и кодированием эти методы относятся скорее к кодированию. Некоторые шаги к автоматизации, например два первых из перечисленных ниже, довольно тривиальны, в то время как другие приводят к существенному упрощению работы кодировщика и позволяют перфорировать коды за минуты вместо недель.

Мнемонические знаки. Вместо номеров операций применяются двухбуквенные обозначения например AD для сложения (addition) и DV для деления (division).

Ввод данных. Наша гипотетическая цифровая машина требовала довольно простой пусковой процедуры. Во многих настоящих вычислительных машинах ввод и запуск программы представляет собой сложную процедуру, которая может быть устранена при помощи автоматического программирования.

Анализ ошибок. Хотя термин «*анализ ошибок*» относят обычно к математическим ошибкам (в смысле, обсуждавшемся в гл. 11 и 12), которые могут накапливаться до значительных размеров, здесь этот термин означает отыскание ошибок в программе, а также ошибок, сделанных машиной. Анализ может быть поверхностным или же почти полным. В одном часто используемом методе, называемом *проверкой по четности*, контролируется число знаков 1 в каждом слове; этот метод по своему принципу подобен алгорифму делимости на 9, изучаемому в начальной школе.

Перевод чисел. Все вводимые в машину команды и все выдаваемые ею результаты даются в десятичной системе, и машина автоматически производит перевод из десятичной системы в двоичную и из двоичной в десятичную. Это не только упрощает кодирование, но также значительно облегчает отладку программы и наблюдение за ходом вычислений.

Расчленение задачи. Часто задача оказывается настолько сложной, что внутренняя память не может вместить сразу все команды и константы. Часть информации должна хра-

ниться на магнитном барабане или во входном устройстве до тех пор, пока не будет закончена часть программы. Автоматическое программирование позволяет машине решать самой, когда принять в себя новые данные и сколько именно.

Изменение задачи. Простейшим типом изменения задачи является изменение одного из параметров. Более сложные типы изменений включают изменения некоторой функции, изменение числа членов в ряде, изменение числа итераций в процессе приближения и т. п. Автоматическое программирование делает все эти перемены более легкими для кодировщика.

Подпрограммы. Использование подпрограммы может быть довольно сложным. Последняя команда подпрограммы должна быть командой возврата к основной программе; в подпрограмму должны быть введены необходимые константы; ответ или ответы должны быть помещены там, откуда их затем можно получить. Все это связано с видоизменениями подпрограммы, которые должны быть произведены еще до ухода из основной программы. Может оказаться необходимым вызывать подпрограмму с магнитного барабана и знать, сколько это потребует команд.

Все это можно делать гораздо легче при помощи автоматического программирования. При одном методе для перехода к подпрограмме требуется только указать номер подпрограммы и дать команду «разрешение»; последняя команда подпрограммы является командой «возврат», при которой счетчик команд автоматически устанавливается на нужное число.

Плавающий адрес. Вместо того чтобы каждой команде и константе назначать определенный адрес в виде номера ячейки памяти, кодировщик может приписывать им произвольные адреса не на машинном языке. При этом кодировщику легче следить за тем, что он делает, а также добавлять или вычеркивать команды, не изменяя при этом остальной программы, чего в трехадресной машине нельзя достичь другим путем. Машина сама назначает каждому «плавающему» адресу конкретный номер ячейки внутренней памяти и ведет на магнитном барабане справочник по взаимному переводу между адресами на машинном языке и адресами на внешнем языке.

Плавающая запятая. Подбор масштабов — одна из самых трудоемких задач для программиста и один из главных источников ошибок — может быть совершенно устранен путем использования плавающей запятой (§ 14.3). Если машина не имеет плавающей запятой, то все же можно закодировать вычисления таким образом, чтобы они производились так же, как и при наличии плавающей запятой. Такое кодирование может быть выполнено автоматически. Например, с помощью команды определения порядка и команды сдвига можно привести число к нормальному виду ($1 > x \geqslant 1/2$), причем команда определения порядка даст и соответствующую степень двойки. Сдвиги, необходимые для сложения и для сложения порядков, при умножении также могут программироваться самой машиной.

Во всех перечисленных выше случаях экономия во времени программирования достигается за счет затраты дополнительного машинного времени и дополнительной емкости запоминающего устройства. Затраты машинного времени могут быть весьма значительными; так, программа с плавающей запятой может занять в 20 раз больше времени, чем при расчете масштабов для той же самой задачи самим программистом. При *интерпретативных* программах (когда машина постоянно делает перевод с внешнего языка на внутренний и обратно) и при использовании всех перечисленных выше методов машина может затратить в 500 раз больше времени, чем если бы задача была закодирована сразу на машинном языке. При *компилятивных* программах (когда машина приготовляет свою собственную программу из хранящихся в ней подпрограмм) затрачивается меньше машинного времени, но может потребоваться бо́льшая память.

Кроме автоматического программирования (с его перерасходом памяти и машинного времени) и ручного программирования (с его перерасходом времени кодирования), существует еще третья альтернатива. А именно, можно усложнить конструкцию машины, расплачиваясь расходом оборудования. Так, некоторые новые большие машины построены специально для работы с плавающей запятой, благодаря чему операции в них занимают не на много больше времени, чем операции с фиксированной запятой. Подпрограммы и некоторые другие задачи программирования создают гораздо меньше трудностей, когда в машине имеются некоторые дополнительные команды.

В частности, в некоторых современных машинах имеется один или несколько так называемых «*B*-блоков», каждый из которых представляет собой вспомогательный счетчик (подобный счетчику команд в трехадресной машине), фиксирующий без затраты лишних

команд такие вещи, как число итераций. Некоторые машины в каждой ячейке памяти имеют специальный разряд контроля четности для автоматической проверки правильности вычислений. Многие современные машины, и в особенности машины, приспособленные для обработки деловых данных, имеют входные устройства, способные непосредственно воспринимать буквенно-цифровую информацию, так что применение мнемонических кодов операций не составляет проблемы.

Говоря об автоматическом программировании, мы должны помнить, что оно было разработано для машин образца 1950 г., которые, в конечном счете, являются примитивными устройствами, первыми представителями своего семейства, когда-либо построенными. Автоматическое программирование может поэтому оказаться лишь временным средством, применяемым только до появления более усовершенствованных машин, спроектированных с гораздо большим учетом требований потребителей. С другой стороны, автоматическое программирование может оказаться новым шагом к использованию машин для выполнения более высокой функции, чем любая выполняемая ныне автоматами, — функции планирования или организации, которая в настоящее время считается недоступной для машины.

ЛИТЕРАТУРА

Книга [56], подготовленная Инженерной исследовательской ассоциацией, была первой книгой по цифровым вычислительным машинам, и хотя она была превосходной, но теперь уже совсем устарела. Хорошим пособием является также выпуск журнала «Proceedings of the Institute of Radio Engineers», посвященный вычислительной технике [31]). По отдельным вопросам вычислительной техники хорошим пособием может служить Ричардс [57], хотя он уделяет слишком много места машинам одной фирмы. В настоящее время готовится несколько книг по цифровым вычислительным машинам.

ЗАДАЧИ

16.1. Закодировать деление для вычислительной машины, которая отличается от гипотетической машины, описанной в этой главе, только тем, что у нее нет операции 06.

16.2. Закодировать извлечение квадратного корня по алгоритму, изучаемому в средней школе.

16.3. Закодировать подпрограмму для уравнения (16.5) с точностью до 10^{-9} (2^{-30}).

ГЛАВА 17

КОМПОНЕНТЫ АНАЛОГОВЫХ ВЫЧИСЛИТЕЛЬНЫХ МАШИН

17.1. Механические устройства

В механической аналоговой машине аналогами участвующих в вычислениях величин являются углы (например, повороты валов) или расстояния (например, линейные перемещения).

Дифференциал автомобиля (рис. 17.1) устроен таким образом, что мощность передается с карданного вала на полуоси с очень малыми потерями, даже когда одно колесо автомобиля вращается быстрее другого (как, например, на повороте). И наоборот, если поднять заднюю часть автомобиля домкратом и вращать оба колеса (или вращать одно, а другое закрепить неподвижно), то во вращение придет карданный вал. Во всех этих случаях (и при соответственно выбранных передаточных отношениях) угловые повороты трех валов связаны уравнением

$$z = x + y, \tag{17.1}$$

где каждая переменная изображает число радиан поворота вала. Такой дифференциал, вместе с надлежащими средствами управления поворотами x и y и измерения поворота z, используется в аналоговых машинах для выполнения одной основной операции — сложения.

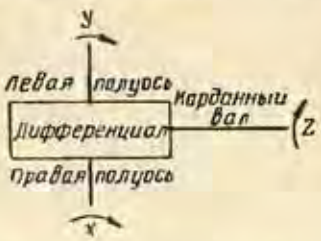


Рис. 17.1. Дифференциал.

Вторая основная операция — умножение на постоянную величину — выполняется посредством еще более простого устройства — зубчатого механизма, показанного на рис. 17.2. Уравнение этого зубчатого механизма имеет вид

$$z = kx. \tag{17.2}$$

Интегрирование. Третьей основной операцией в механических аналоговых машинах является интегрирование. Показанный на

рис. 17.3 дисковый интегратор изобрели в 1875 г. лорд Кельвин и его брат, однако технология продвинулась достаточно для изготовления точных интеграторов только к 20-м годам нашего столетия.

В дисковом интеграторе одной входной величине x соответствует угол поворота диска. На диске, на расстоянии y от его центра, рас-

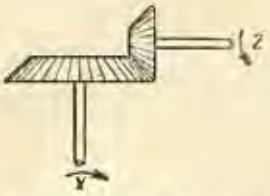


Рис. 17.2. Зубчатый механизм для $z = kx$.

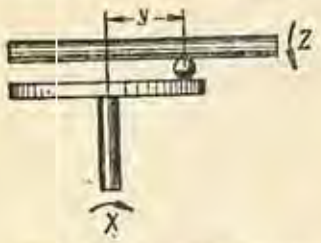


Рис. 17.3. Шариково-дисковый интегратор.

полагается шарик (на практике обычно два шарика или один ролик). Диск вращает шарик, который в свою очередь вращает выходной вал z. Выходная переменная выражается формулой

$$z = \int y dx. \qquad (17.3)$$

Входная переменная y, конечно, может быть любой функцией от x, какую можно получить путем комбинации основных операций. Эта переменная вводится как линейное перемещение, а не как поворот вала; преобразование между этими двумя аналогами производится посредством червячной передачи или посредством соединения зубчатого колеса с зубчатой рейкой (рис. 17.4), где линейное перемещение y преобразуется в поворот вала x и обратно.

Погрешности. Главными причинами ошибок в таких вычислительных машинах являются мертвый ход в зубчатых передачах и отсутствие точечного контакта в интеграторе. Очевидно, что уравнение (17.3) строго справедливо только в том случае, если контакты между диском и шариком и между шариком и валом представляют собой точки с нулевой площадью; однако для предотвращения проскальзывания к шарику должна быть приложена значительная сила. Эта сила превращает зоны соприкосновения в маленькие кружки и тем самым создает погрешность.

Рис. 17.4. Зубчатое колесо и зубчатая рейка.

Дифференциальные уравнения. Можно показать [52], что все математические операции, необходимые для решения обыкновенных дифференциальных уравнений, могут быть построены из сложений, умножения на константу и интегрирования. Так, например, умножение переменных можно выполнять по формуле

$$xy = \int y dx + \int x dy. \qquad (17.4)$$

Функции могут быть получены с помощью разложений в ряды или другими специальными способами. Экспоненциальные и тригонометрические функции могут быть образованы путем задания соответствующего дифференциального уравнения, решением которого служит требуемая функция.

Первой большой автоматической вычислительной машиной, когда-либо построенной, было механическое аналоговое устройство, состоящее из компонентов, подобных тем, которые мы описали. Эта машина получила название *дифференциального анализатора** и была создана в Массачусетском технологическом институте в 1933 г. под руководством Ванневара Буша**. Такие устройства еще применяются в специализированных вычислительных машинах, например в авиационных пушечных прицелах (прицелах для воздушной стрельбы), где программа и нужные константы могут быть встроены постоянно, но в качестве универсальных вычислительных машин они сегодня мало полезны.

Программирование механического дифференциального анализатора происходит путем сложной настройки зубчатых передач, и эти машины тяжелы, громоздки, дороги и действуют медленно. При особой тщательности изготовления точность этих машин может быть доведена до 0,01%, однако ту же самую точность можно получить при помощи гораздо более быстрых и гибких электромеханических устройств.

17.2. Операционный усилитель

В электронных аналоговых вычислительных машинах аналогами участвующих в вычислениях величин являются электрические

---

* Дифференциальные анализаторы называются также в русской технической литературе «машинами для решения (или интегрирования) дифференциальных уравнений». — *Прим. ред.*

** Русский академик А. Н. Крылов построил аналоговую машину для дифференциальных уравнений в Морском опытовом бассейне в Петербурге в 1912 г. При постройке этой машины были впервые осуществлены многие идеи, использованные затем нашими и зарубежными конструкторами таких машин. Машина Крылова была разобрана во время I мировой войны 1914 г. и более не восстанавливалась. В 1938 г. механическая машина для интегрирования дифференциальных уравнений была построена в Академии наук СССР И. С. Бруком. — *Прим. ред.*

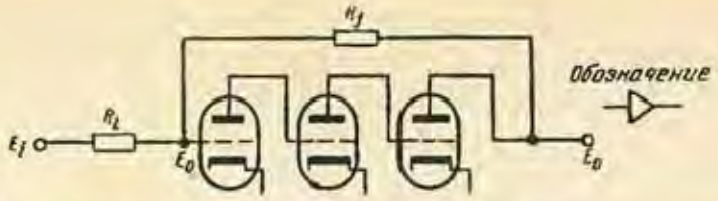


Рис. 17.5. Операционный усилитель с сопротивлением в цепи обратной связи.

напряжения. В электромеханических аналоговых машинах в качестве аналогов применяются как напряжения, так и повороты валов. Преобразования поворотов валов в электрические напряжения и обратно осуществляются при помощи сервомеханизмов (см. § 17.4).

Как в электромеханических, так и в электронных аналоговых машинах наиболее важным устройством является операционный усилитель, показанный на рис. 17.5 и 17.6. При помощи операционных усилителей производится сложение, интегрирование и другие операции. Операционный усилитель состоит из трехкаскадного усилителя постоянного тока с большим усилением, снабженного надлежащими цепями RC на входе и в цепи обратной связи. На рис. 17.5 и 17.6 показаны две наиболее простые и наиболее часто применяемые схемы. Отмеченные на рисунках напряжения измеряются по отношению к земле; E_0 и E_i изображают аналоги переменных из задач и обычно имеют номинальный диапазон от —100 до +100 *в*, хотя E_0 может быть и значительно больше.

Передаточная функция. Анализ такого устройства, как операционный усилитель, обычно производится путем исследования его передаточной функции, которая определяется как отношение напряжения на выходе устройства к напряжению на входе (в § 29.2 дается другое, более употребительное определение передаточной функции).

Для нахождения передаточной функции схемы на рис. 17.5 отметим, что входное сопротивление сетки первой лампы обычно велико и практически весь ток, который проходит через сопротивление R_i, проходит через R_f. Отсюда

$$\frac{E_a - E_i}{R_i} = \frac{E_0 - E_a}{R_f}. \qquad (17.5)$$

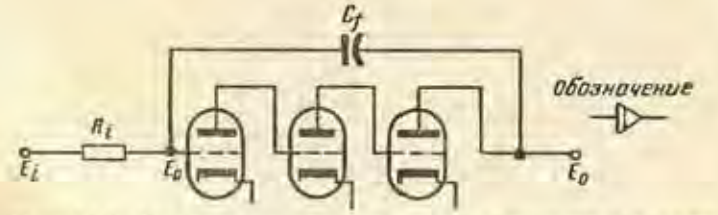


Рис. 17.6. Операционный усилитель с емкостью в цепи обратной связи.

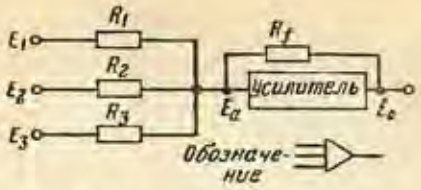


Рис. 17.7. Операционный усилитель как сумматор.

Далее, отношение напряжения на выходе к напряжению на сетке первой лампы определяется коэффициентом усиления μ:

$$E_0 = \mu E_a. \qquad (17.6)$$

Используя (17.6), исключаем E_a из уравнения (17.5) и находим:

$$\frac{\frac{E_0}{\mu} - E_i}{R_i} = \frac{E_0 - \frac{E_0}{\mu}}{R_f},$$

$$-\frac{R_f}{R_i} E_i = E_0 - \frac{E_0}{\mu} - \frac{R_f}{R_i}\frac{E_0}{\mu} \approx E_0. \qquad (17.7)$$

Ошибка в этом приближении пропорциональна отношению $\frac{1 + R_f/R_i}{\mu}$. При $R_f/R_i \leqslant 10$ (что на практике всегда выполняется) и при $\mu = -50\,000$ (а коэффициент усиления обычно больше) ошибка будет меньше $0{,}02\%$.

Утверждение, что уравнение (17.7) является хорошим приближением, эквивалентно утверждению, что E_a приближенно равно нулю. Вообще говоря, E_a равно E_0/μ, но так как E_0 всегда меньше 100 *в*, а μ обычно порядка 1 000 000, то приближение $E_a \approx 0$ является очень точным. Дальнейший анализ передаточных функций мы будем производить в предположении, что напряжение на сетке первой лампы равно нулю.

Сложение. Если вход на рис. 17.5 заменить входом с рис. 17.7, уравнение (17.7) примет вид

$$E_0 = -\sum \frac{R_f}{R_i} E_i \qquad (17.8)$$

с ошибкой того же порядка, как и раньше. Таким образом, операционный усилитель может работать как сумматор, причем слагаемые при суммировании могут умножаться на константы. На практике эти константы обычно больше единицы, так как для умножения на константу меньше единицы существуют более удобные способы (§ 17.3). Величины констант часто указываются в условном обо-

а) б)

Рис. 17.8. Обозначение операционных усилителей с цепью обратной связи:
а—интегратор; б—сумматор.

значении операционного усилителя, как на рис. 17.8.

Интегрирование. Проводя такой же анализ схемы, показанной на рис. 17.6, получаем

$$\frac{E_a - E_i}{R_i} = \frac{dQ}{dt} = \frac{d}{dt}[C_f(E_0 - E_a)],$$

или ввиду $E_a = 0$

$$-\frac{E_i}{R_i} = C_f \frac{dE_0}{dt}.$$

Интегрирование дает

$$E_0 = -\frac{1}{R_i C_f}\int_0^t E_i dt + E_{ic}, \qquad (17.9)$$

где постоянная интегрирования E_{ic} изображает напряжение на конденсаторе C_f в момент $t=0$. Таким образом, операционный усилитель может также интегрировать; одновременно с интегрированием он может складывать несколько входных переменных, умножая каждую из них на определенную константу.

Можно показать таким же анализом, что операционный усилитель с конденсатором на входе и с сопротивлением в цепи обратной связи осуществляет дифференцирование. Эта схема редко применяется в аналоговых машинах, так как всякие ошибки (шум) на входе испытывают при дифференцировании тенденцию возрастать, в то время как при интегрировании они испытывают тенденцию сглаживаться. Кроме того, дифференцирующие усилители обнаруживают тенденцию к неустойчивости. Наконец, интегральное уравнение может быть всегда аналитически продифференцировано и преобразовано в дифференциальное уравнение, которое может быть решено на электронных интеграторах.

Применение. Более сложные передаточные функции операционного усилителя можно получить, если применить более сложные цепи RC во входной цепи, или в цепи обратной связи, или в обеих этих цепях одновременно.

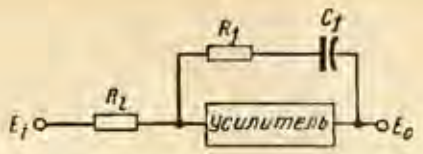


Рис. 17.9. Операционный усилитель со специальной цепью обратной связи.

Например, схема на рис. 17.9 описывается уравнением

$$E_0 = -\frac{R_f}{R_i}E_i - \frac{1}{R_i C_f}\int E_i dt + E_{ic}. \qquad (17.10)$$

Такие схемы могут уменьшить число требуемых усилителей за счет добавления нескольких сопротивлений и конденсаторов, и поэтому они заслуживают подробного рассмотрения при проектировании специализированных аналоговых вычислительных машин. Однако синтезировать такие схемы трудно, и для универсальных аналоговых машин они применяются мало.

В обычных операционных усилителях сопротивление R_f (если оно в схеме имеется) выбирается равным 1 *Мом*, емкость конденсатора C_f (если он в схеме имеется) — равной 1 *мкф*, а входные сопротивления берутся в диапазоне от 0,1 до 1 *Мом*. Усилитель обычно делается трехкаскадным с гальваническими связями. Для достижения стабильности применяется отрицательная обратная связь (заметим, что в схемах на рис. 17.7 и 17.9 напряжения на входе и выходе имеют противоположные знаки). Усиление усилителя на нулевой частоте может быть от —10 000 до —600 000 000.

Усиление усилителя без обратной связи должно оставаться большим в рабочем диапазоне частот, ибо только в этом случае уравнение (17.7) будет хорошим приближением и передаточная функция (при наличии обратной связи) будет линейной относительно амплитуды входного сигнала. Однако усиление не обязательно должно оставатся постоянным в рабочем диапазоне частот, и во многих усилителях при изменении частоты от нуля до 5 *гц* усиление уменьшается в 1000 раз. Рабочий диапазон частот в электромеханических вычислительных машинах простирается от нулевой частоты до нескольких герц, а в электронных вычислительных машинах — до нескольких килогерц, однако в целях уменьшения фазовых искажений операционные усилители часто делаются плоскими в диапазоне до нескольких сот килогерц. Передаточная функция усилителя также должна быть линейной по напряжению в выбранном диапазоне, обычно в диапазоне ± 100 *в* на выходе и на сумме всех входов.

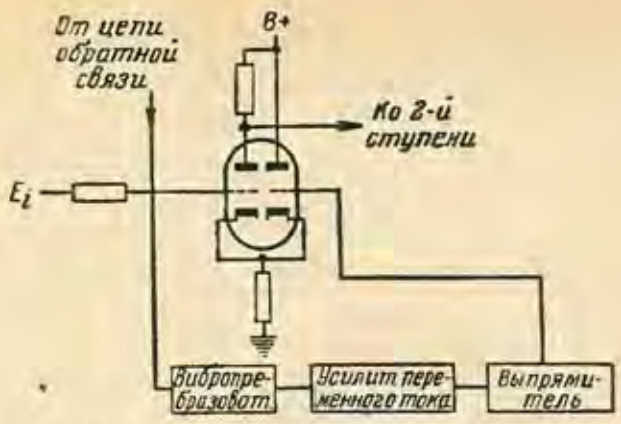


Рис. 17.10. Стабилизатор дрейфа, предназначенный для первой ступени операционного усилителя.

Усилители постоянного тока подвержены дрейфу. Каждые несколько минут необходимо компенсировать этот дрейф или же использовать специальные схемы автоматической компенсации дрейфа. В наиболее обычной схеме компенсации дрейфа входное напряжение преобразуется на вибропреобразователе в переменное напряжение, затем это переменное напряжение усиливается, выпрямляется и подается на сетку правой лампы входного каскада (рис. 17.10). Напряжение на сетке правой лампы изменяет эффективный коэффициент усиления μ операционного усилителя.

Если операционный усилитель используется как интегратор, то при установке начальных значений интегратора конденсатор обратной связи соединяется посредством переключателя с источником напряжения. Этим источником напряжения обычно служит потенциометр (§ 17.3).

Кроме интегрирования, сложения и умножения на константу, операционный усилитель можно использовать для инвертирования знака (т. е. для умножения на -1), а также для развязки одной схемы от другой. В хорошо спроектированной аналоговой машине операционные усилители для изменения знака и развязки схем применяются очень редко.

17.3. Потенциометры

Применение потенциометра для умножения напряжения на произвольную константу k, $0 \leqslant k \leqslant 1$, показано на рис. 17.11,*а*. Движок потенциометра обычно устанавливается вручную до начала вычислений. Если желательно изменять константу умножения в обе стороны от нуля, то можно применить схему рис. 17.11,*б*, в которой операционный усилитель используется как инвертор знака. В этом случае константа умножения k может иметь

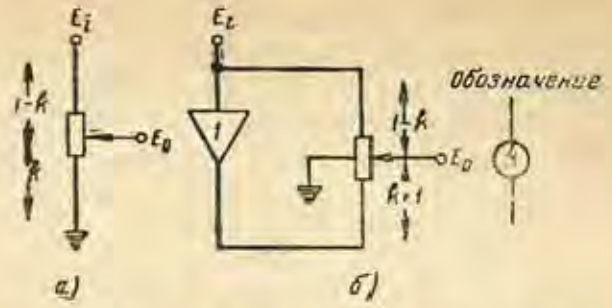


Рис. 17.11. Линейные потенциометры.

любое значение в диапазоне $-1 \leqslant k \leqslant 1$. Передаточная функция обеих этих схем равна $E_0/E_i = k$.

При помощи нелинейного потенциометра теоретически можно образовать любую функцию. На схеме рис. 17.12,*а* выходное напряжение пропорционально сопротивлению между движком потенциометра и землей, которое, в свою очередь, пропорционально площади треугольника слева от движка, а эта площадь пропорциональна квадрату расстояния x. Следовательно, при помощи такого потенциометра можно образовать функцию $y = x^2$, если обеспечить соответствующее линейное или угловое перемещение x. На практике эта схема обычно не применяется, так как квадратическая функция получается проще путем умножения (см. § 17.4).

Зато схема, изображенная упрощенно на рис. 17.12,*б*, часто применяется на практике. Сопротивление, намотанное с переменным шагом по окружности, получается обычно путем намотки проволоки на синусоиду (рис. 17.12,*в*), которая затем изгибается так, чтобы образовался цилиндр. На рис. 17.12,*б* движки потенциометров установлены в угловых положениях $x = \pm 135°$. С одного движка считывается функция $\sin x$, с другого — функция $\cos x$. На практике движки могут быть помещены также под углами $\pm 45°$ для считывания функций $-\sin x$ и $-\cos x$. Вход ± 100 *в* можно заменить произвольной переменной, как на

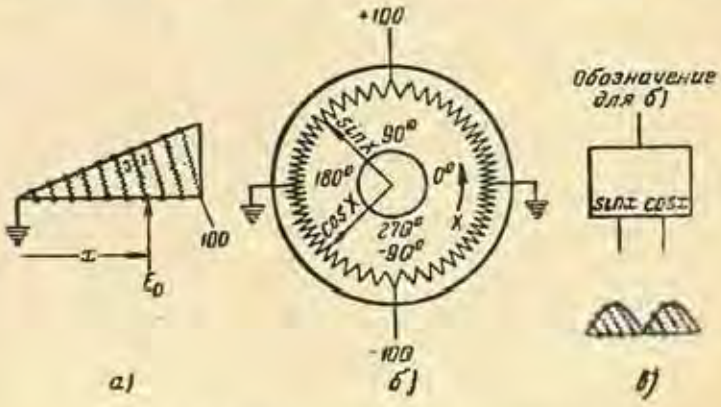


Рис. 17.12. Нелинейные потенциометры.

рис. 17.11,*б*. В этом случае вся схема, включая операционный усилитель, называется *синусно-косинусным устройством* (потому что выполняет операцию, эквивалентную разложению вектора R на его составляющие $x=R\cos\theta$ и $y=R\sin\theta$)

17.4. Сервомеханизмы

Как показали предыдущие примеры, переменные в электромеханической аналоговой машине могут быть представлены в виде электрических напряжений и в виде поворотов валов. Преобразования переменных из одной формы в другую производятся при помощи сервомеханизмов, или следящих устройств (следящих систем). Сервомеханизм по существу изображает математическую «операцию» равенства. Сервомеханизм изменяет переменную на одном входе так, чтобы она стала равна переменной на втором входе. Это изменение производится (как это описано в гл. 29) путем измерения разности напряжений на входах (*напряжение ошибки*), которая затем усиливается и используется для приведения во вращение сервомотора. Сервомотор передвигает движок потенциометра в направлении, при котором напряжение ошибки уменьшается.

Умножение. Функция равенства вместе со схемой на рис. 17.11,*а* позволяет производить умножение двух переменных. Соответствующая коммутация показана на рис. 17.13. Одна из перемножаемых величин (y) подается на один вход сервомеханизма (следящего устройства); напряжение на другом его входе автоматически устанавливается равным этой величине благодаря перемещению движка. Движок первого потенциометра жестко связан с движком второго потенциометра, на концы которого подается вторая переменная величина (x). На движке второго потенциометра и образуется искомое произведение xy (точнее говоря, напряжение, снимаемое с движка, равно $xy/100$).

Деление. Тот, кто умеет умножать, умеет и делить. Показанная на рис. 17.14 схема производит деление $z=y/x$ путем приравнивания $y=xz$. Часть схемы деления показана на рис. 17.14,*а*. Если к этой схеме применить уравнение (17.5) и заменить E_0 на E_p, то мы получим

$$E_p=-(R_f/R_i)E_i.$$

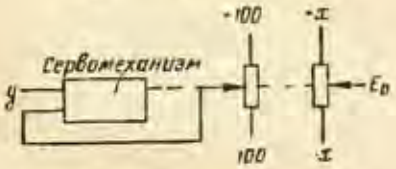


Рис. 17.13. Сервоумножитель.

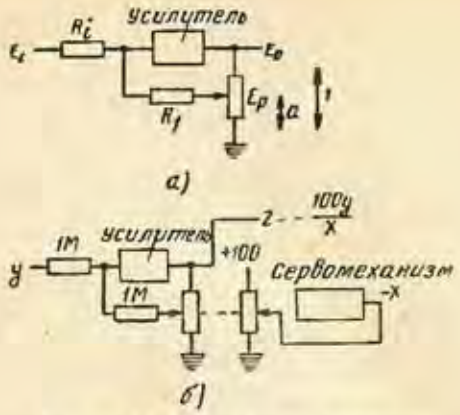


Рис. 17.14 Устройство деления:
а — операционный усилитель с потенциометром в цепи обратной связи;
б — полная схема.

Если движок потенциометра установлен на относительном значении a ($0\leqslant a\leqslant 1$), то напряжение на выходе схемы становится равным

$$E_0=-\frac{1}{a}\frac{R_f}{R_i}E_i.$$

На рис. 17.14,*б* напряжение E_i представляет делимое y, E_0 представляет частное z, а движок потенциометра устанавливается сервомеханизмом на $a=x/100$, где x — делитель. Тогда

$$z=-\frac{100}{x}\cdot\frac{1}{1}\cdot y=-\frac{100y}{x}.$$

Если выход схемы связан с входом x, то уравнение принимает вид

$$z=-100\frac{y}{z},$$

откуда $z=10\sqrt{-y}$. Если $y>0$, то $z=-10\sqrt{y}$, с соответствующими поправками.

Так как сервомеханизм вырабатывает равенство, то с помощью сервомеханизма и генератора надлежащей функции можно образовать любую обратную функцию. Так, например, если движок потенциометра, намотанного по закону синуса, приводить в движение от сервомотора, то можно образовать обратные тригонометрические функции (см., например, рис. 18.11).

17.5. Генераторы функций

Кроме таких отдельных функций, как синус и косинус, которые могут быть образованы специальными методами, в форме напряжений могут быть образованы вообще любые функции. На рис. 17.15 показано электромеха-

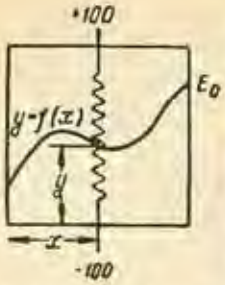


Рис. 17.15. Сервотаблица.

ническое устройство для генерирования функций, называемое *сервотаблицей*. Проволока изгибается в форме требуемой функции $f(x)$, и сервомеханизм перемещает в горизонтальном направлении по таблице линейный потенциометр. Расстояние x пропорционально входному напряжению сервомеханизма. Выходное напряжение снимается с одного конца проволоки (другой конец изолирован). Это напряжение пропорционально сопротивлению от нижнего конца потенциометра до точки касания проволоки и потенциометра, т. е. пропорционально расстоянию y.

Фотоформирователь. На рис. 17.16 показано сходное электронное устройство, называемое *фотоформирователем*. Нижняя часть экрана электронно-лучевой трубки закрывается непрозрачной маской, профиль которой вырезан по кривой $y=f(x)$. Независимая переменная x подается в виде напряжения на горизонтальные отклоняющие пластины, а зависимая переменная снимается с вертикальных отклоняющих пластин. Если x уменьшается, то яркое пятно на экране смещается влево и, как показано на рис. 17.16, уходит под маску. Это приводит к уменьшению напряжения на выходе фотоэлемента и к уменьшению напряжения на выходе усилителя. При уменьшении напряжения на пластинах пятно движется вверх. Если x увеличивается, то изменения происходят в обратном направлении.

Усиление и смещение таковы, что пятно всегда частично закрывается маской. Существуют специальные трубки, в которых функция $f(x)$ изображается на экране при помощи материала, вторичные эмиссионные свойства которого отличаются от вторичных эмиссионных свойств остальной части экрана. В трубке имеется коллектор. Напряжение сигнала снимается не с фотоэлемента, а с мишени. У этого устройства время установления выходного напряжения при заданном входном напряжении меньше миллисекунды, а при введении соответствующей обратной связи это время может быть уменьшено до микросекунды.

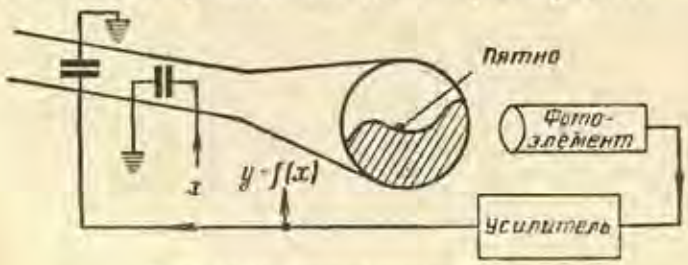


Рис. 17.16 Фотоформирователь.

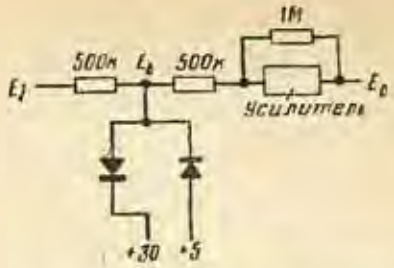


Рис. 17.17. Ограничительная схема.

Схемы ограничения. При моделировании часто требуется ограничение тех или иных величин. Например, элерон крыла самолета не может отклониться более чем на $\pm 30°$. Другой пример: механические ограничители не позволяют снаряду из авиационной пушки попасть при качании в хвост того самого самолета, на котором она установлена. Электронный аналог этого явления выполняется на диоде. Если после подбора масштабов установлено, что напряжение на входе некоторого усилителя не должно выходить за пределы $10 \leqslant E \leqslant 60$ *в*, то на входе этого усилителя включается схема на диодах, показанная на рис. 17.17.

Когда входное напряжение изменяется в диапазоне $10 \leqslant E_i \leqslant 60$, то напряжение E_b равно половине входного напряжения и изменяется в диапазоне $5 \leqslant E_b \leqslant 30$. Если E_i становится больше 60 *в*, то первый диод начинает проводить и E_b не превышает 30 *в*. Если E_i становится меньше 5 *в*, то проводит второй диод и E_b остается на уровне 5 *в*.

17.6. Компоненты полноэлектронных аналоговых вычислительных машин

Если построить аналоговую вычислительную машину из описанных выше компонентов, то скорость ее работы будет ограничиваться скоростью работы устройств с сервомоторами (устройства умножения, сервотаблицы, синусно-косинусные устройства и т. д.). Максимальная частота работы этих устройств лежит в диапазоне от 2 до 10 *гц*. Если бы можно было выполнить все операции на электронных устройствах, то скорость работы увеличилась бы в 1000 раз. Труднее всего осуществить электронным путем операцию умножения, однако в настоящее время разработан ряд чисто электронных умножителей.

Выполнять умножение по формуле $xy = \int y\,dx + \int x\,dy$, к сожалению, нельзя, так как операционный усилитель может производить

интегрирование только по времени*. Но можно применить так называемый принцип *„четверти квадратов"*

$$xy = \frac{1}{4}[(x+y)^2 - (x-y)^2].$$

Сложение, оба вычитания и умножение на $^1/_4$ легко выполняются на электронных устройствах; остается только построить устройство возведения в квадрат. Для этой цели можно использовать фотоформирователь с параболической кривой x^2 на маске или на мишени; использовались также нелинейные сопротивления, диодные схемы и другие устройства.

В одном чисто электронном умножителе генерируется непрерывно и с неизменной частотой последовательность прямоугольных импульсов, амплитуда которых пропорциональна одной переменной умножения, а ширина — другой. Последовательность импульсов пропускается через фильтр. Постоянная слагающая на выходе фильтра будет пропорциональна произведению сомножителей. Очевидно, что такое устройство может быть осуществлено, однако для построения такого устройства, обеспечивающего высокую точность, требуются сложные схемы. Поэтому полноэлектронная аналоговая машина получается довольно дорогой, и применение таких машин по сравнению с электромеханическими аналоговыми машинами ограничено.

17.7. Аналоговые вычислительные системы

Наиболее распространенной формой аналоговой вычислительной системы является универсальная аналоговая вычислительная машина, называемая часто *дифференциальным анализатором* (потому что она лучше всего приспособлена для решения дифференциальных уравнений). Она состоит из совокупности компонентов, описанных в этой главе, вместе с соответствующими источниками питания и входно-выходными устройствами (гл. 19). Для того чтобы отдельные компоненты могли соединяться между собой любым желательным способом, применяются удобные вставные разъемы. Прежде чем пустить машину в ход, на потенциометрах устанавливаются константы, а на конденсаторах — начальные условия задачи. Установка этих данных производится вручную, а в более совершенных машинах — автоматически.

Другой класс аналоговых вычислительных машин — специализированные устройства, например прицелы для воздушной стрельбы.

* Можно производить интегрирование по формуле $\int f(x)\left(\frac{dx}{dt}\right)dt$, однако при этом нужно устройство умножения. — *Прим. авт.*

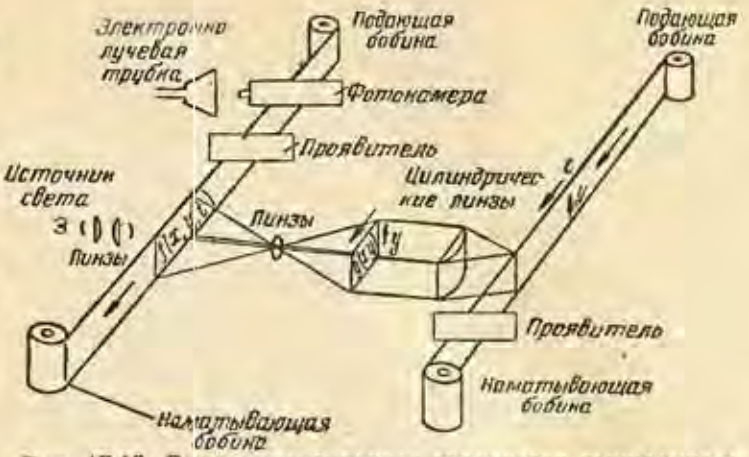


Рис. 17.18 Специализированная аналоговая вычислительная машина для нахождения $\int_{x_1}^{x_2} f(x, y, t)g(x, y)dx$.

В устройствах этого класса все соединения и часть констант закоммутированы постоянно.

Кроме того, существует не рассмотренный нами большой класс аналоговых вычислительных машин, отличающихся от дифференциальных анализаторов. В машинах этого класса могут использоваться как электронные, так и другие аналоговые устройства, упомянутые в § 14.2. Сюда относится, например, обширная группа *анализаторов сетей*, в которых течение электричества или какой-либо жидкости сделано аналогичным течению воды в трубах, электричества в линиях передачи, нефти под землей, нейтронов в ядерном реакторе и т. п.

Иногда бывает так, что специфическая задача эффективно решается на специальной, необычной аналоговой машине. Примером остроумного устройства, применяемого для решения специфической задачи, может служить аналоговая вычислительная машина (рис. 17.18), на которой непрерывно вычисляется функция $\int f(x, y, t)g(x, y)dx$. Функция f непрерывно преобразуется в оптический сигнал при помощи электронно-лучевой трубки и проецируется на фотопленку в координатах x и y. Пленка быстро проявляется в виде транспаранта. Величина функции g наносится постоянно на другой транспарант.

Умножение производится пропусканием света через оба транспаранта при помощи подходящей системы линз, интегрирование — фокусированием света с двумерной площадки на одномерную линию при помощи цилиндрической линзы. Интеграл как функция от y и t непрерывно записывается фотографически или же преобразуется фотоэлектрически в группу электрических сигналов для дальнейших вычислений.

ЛИТЕРАТУРА

См. гл. 18.

ГЛАВА 18

РАБОТА ЭЛЕКТРОМЕХАНИЧЕСКИХ АНАЛОГОВЫХ ВЫЧИСЛИТЕЛЬНЫХ МАШИН

В устройствах, описываемых в этой главе, применяются: операционные усилители — для выполнения операций сложения, интегрирования, умножения на константы, большие чем 1, и для изменения знака; линейные потенциометры — для умножения на константы, меньшие чем 1; потенциометры с сервомеханизмами — для умножения переменных; комбинации усилителей с сервомеханизмами — для деления одной переменной на другую; синусно-косинусные потенциометры — для генерирования тригонометрических функций; специальные генераторы функций — для образования других функций. Такие устройства часто называются *электронными аналоговыми вычислительными машинами или электронными дифференциальными анализаторами*, хотя такие названия правильнее относить к полностью электронным устройствам.

18.1. Решение дифференциальных уравнений

Мы начнем с простого примера — с обычного линейного дифференциального уравнения 2-го порядка

$$A\frac{d^2x}{dt^2}+B\frac{dx}{dt}+Cx+D=0. \qquad (18.1а)$$

Производную по времени будем обозначать точкой над функцией (такое обозначение весьма удобно в нашем случае, когда дифференцирование производится только по времени). Тогда уравнение (18.1а) запишется в виде

$$A\ddot{x}+B\dot{x}+Cx+D=0. \qquad (18.1б)$$

При решении дифференциального уравнения на аналоговой вычислительной машине первый шаг обычно состоит в решении уравнения относительно высшей производной:

$$\ddot{x}=-\frac{B}{A}\dot{x}-\frac{C}{A}x-\frac{D}{A}=-P\dot{x}-Qx-R. \qquad (18.2)$$

Аналогично блок-схеме вычислений, которую мы составляли при программировании решения на цифровой вычислительной машине, составим блок-схему вычислений на аналоговой машине, в которой наметим порядок выполнения математических операций.

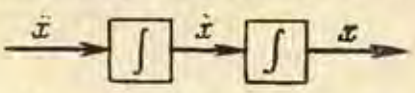


Рис. 18.1 Последовательное интегрирование.

На рис. 18.1 показано, что, зная $\ddot{x}$, мы можем найти $\dot{x}$ и x путем последовательного интегрирования. Если произвести теперь операции сложения и умножения согласно правой

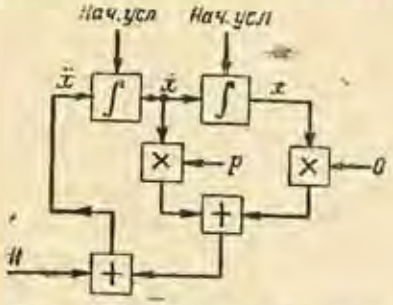


Рис. 18.2. Блок-схема для уравнения (18.2).

части уравнения (18.2), то мы можем получить $\ddot{x}$ и подать его обратно на вход. Эти операции показаны на рис. 18.2 (знаки мы пока не учитываем). Петли обратной связи такого вида являются основным методом решения уравнений на аналоговых вычислительных машинах. После того как сделаны нужные соединения, все выходные напряжения вынуждаются удовлетворять заданному уравнению.

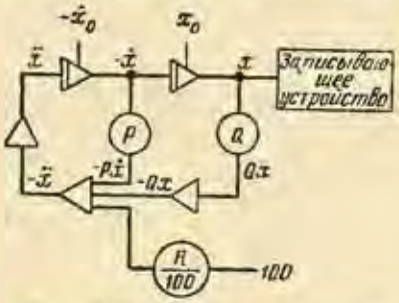


Рис. 18.3. Путевая карта для уравнения (18.2).

Кодированию в цифровых вычислительных машинах соответствует *путевое картирование* в аналоговых вычислительных машинах. На рис. 18.3 показана путевая карта, соответствующая блок-схеме на рис. 18.2*. Знаки интегрирования заменены здесь интеграторами, знаки умножения — потенциометрами и знак сложения — сумматором. Для коррекции изменения знака, которая в данном случае происходит в интеграторе и в сумматоре, преду-

* В русской технической литературе такие путевые карты называются иногда картами набора. — *Прим. ред.*

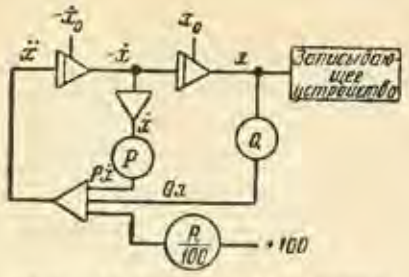


Рис. 18.4. Путевая карта для уравнения (18.2).

смотрено использование двух дополнительных операционных усилителей в качестве инверторов знака.

Мы предполагаем, что P, Q и R — положительны. Вопрос о подборе масштабов будет рассматриваться в § 18.3, а пока мы считаем, что каждая переменная моделируется в масштабе l *в* на единицу ее измерения. Начальные значения $\dot{x}_0$ и x_0 предполагаются известными и установленными на конденсаторах обратной связи интеграторов еще до начала решения задачи; при этом условливаемся, что каждое начальное значение имеет тот же знак, что и напряжение на выходе. Интересующей нас выходной переменной будет, очевидно, x (как функция времени). Однако любая другая из фигурирующих здесь величин также может быть при желании подана на записывающее устройство.

На рис. 18.4 показана несколько улучшенная путевая карта. Она дает тот же результат, но в ней на один операционный усилитель меньше. Но путевую карту можно еще более упростить, ибо мы видим сумматор на входе интегратора, а это является излишеством (если только мы не желаем выводить на записывающее устройство именно функцию $\ddot{x}$), так как суммирование можно производить на интеграторе одновременно с интегрированием. Нужно учесть также, что при устранении одного сумматора устраняется одно инвертирование знака. Полученная карта показана на рис. 18.5.

Путевые карты с рис. 18.3—18.5 настолько похожи на блок-схему с рис. 18.2, что обычно блок-схема не составляется вообще. При

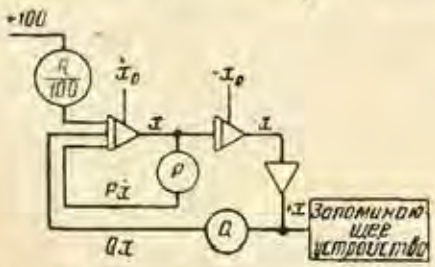


Рис. 18.5. Путевая карта для уравнения (18.2).

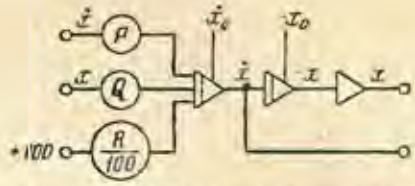


Рис. 18.6. Путевая карта для уравнения (18.2).

вычерчивании предварительной путевой карты часто делается другое упрощение. Так как петли обратной связи могут стать очень сложными, то их указывают просто надлежащей маркировкой входов и выходов; например, рис. 18.5 заменяется рисунком 18.6.

Нелинейные уравнения. Разумеется, уравнение (18.2) гораздо легче решается аналитически. Однако те же самые принципы с небольшими усложнениями применяются в аналоговых вычислительных машинах к уравнениям, которые аналитически решать весьма трудно. Пусть, например, нам дана крайне нелинейная формула, вроде

$$\dddot{x} - 0{,}4\ddot{x}^2\dot{x} - x^{3/2}x = 6. \qquad (18.3)$$

Решим сначала это уравнение относительно $\dddot{x}$. Затем, предполагая, что $\dddot{x}$ нам известно, проинтегрируем его три раза, чтобы получить $-\ddot{x}$, $\dot{x}$ и $-x$. Затем при помощи подходящих схем умножения и извлечения квадратного корня построим отдельные слагаемые и, наконец, сложим их с нужными знаками для того, чтобы получить $\dddot{x}$. См. задачу 18.1.

Уравнения в частных производных как таковые не могут решаться на аналоговых машинах, однако существуют разнообразные методы преобразования этих уравнений (с бóльшим или меньшим приближением) в такие формы, которые могут решаться на аналоговых машинах.

Системы уравнений. Аналоговые машины используются также для решения систем дифференциальных уравнений. Рассмотрим следующую систему (в которой для простоты опущены все константы):

$$\left.\begin{aligned} \ddot{x} + \dot{x} + y &= 0 \\ \ddot{y} + \dot{y} - x &= 0 \end{aligned}\right\}. \qquad (18.4)$$

Сначала решим эту систему относительно высших производных:

$$\left.\begin{aligned} \ddot{x} &= -\dot{x} - y \\ \ddot{y} &= -\dot{y} + x \end{aligned}\right\}. \qquad (18.5)$$

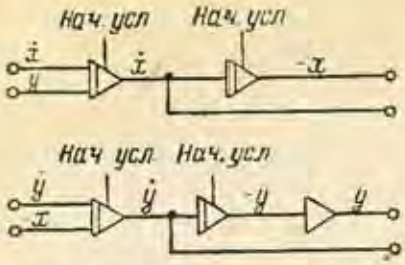


Рис. 18.7. Путевая карта для уравнения (18.5).

Тогда решение получается по карте, представленной на рис. 18.7. Вход y для верхней части карты берется с выхода нижней части, и аналогично вход $-x$ для нижней части схемы берется с выхода верхней части.

18.2. Моделирование

Любая физическая система может моделироваться (имитироваться) на электромеханической аналоговой вычислительной машине, если известны математические уравнения, описывающие ее поведение, и если имеется в наличии необходимое оборудование. Уравнения могут задаваться неявно.

Моделирование массы, подвешенной на пружине. Рассмотрим физическую систему, показанную на рис. 18.8. Масса m_1 прикреплена пружиной к жесткой опоре, и к массе приложена синусоидальная вынуждающая сила. Меньшая масса m_2 подвешена к массе m_1 на пружине и демпфере. Упругая сила, создаваемая каждой пружиной, пропорциональна ее растяжению или сжатию, тормозящая сила, создаваемая демпфером, пропорциональна скорости поршня относительно стенок цилиндра.

При составлении уравнений мы будем отсчитывать все смещения от положения равновесия в двух разных координатных системах: y_1 для массы m_1 и y_2 для массы m_2. Положительным направлением смещений, скоростей, ускорений и сил будем считать направление вниз. Уравнение (18.6) выражает тот факт, что сумма всех сил, приложенных к m_1, равна нулю, а уравнение (18.7) выражает тот факт, что сумма всех сил, приложенных к m_2, равна нулю. Важно отметить, что каждое из уравнений составлено независимо от другого.

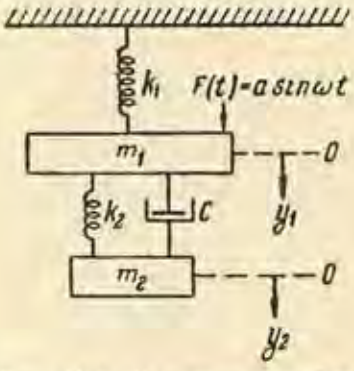


Рис. 18.8. Условное изображение масс, подвешенных на пружинах.

Ценнейшим свойством аналоговой машины является то, что при решении сложной системы дифференциальных уравнений аналоговая машина решает фактически каждое уравнение в отдельности, обходя при этом дополнительные усложнения, обусловленные наличием связей между уравнениями.

$$-m_1\ddot{y}_1 - c(\dot{y}_1 - \dot{y}_2) - k_1 y_1 - k_2(y_1 - y_2) + F(t) = 0. \qquad (18.6)$$

Очень важно, разумеется, расставить всюду правильные знаки. Заметим, что в уравнении (18.6) знаки четырех слагаемых отрицательны по следующей причине: когда $\ddot{y}_1$ положительно, то ускорение направлено вниз, а сила инерции, противодействующая ускорению, направлена вверх и, значит, отрицательна; когда разность $\dot{y}_1 - \dot{y}_2$ положительна (например, когда $\dot{y}_1 > 0$, а $\dot{y}_2 = 0$), то относительная скорость направлена вниз и сила торможения демпфера, приложенная к m_1, отрицательна; когда разность $y_1 - y_2$ положительна (например, когда m_2 находится в положении равновесия, а m_1 смещена вниз), то пружина сжата и сила пружины, приложенная к m_1, отрицательна. Соответствующий член в (18.7) имеет противоположный знак.

$$-m_2\ddot{y}_2 + k_2(y_1 - y_2) + c(\dot{y}_1 - \dot{y}_2) = 0. \qquad (18.7)$$

Следующий шаг — решение уравнений (18.6) и (18.7) относительно высших производных:

$$\left.\begin{aligned} \ddot{y}_1 &= -\frac{c}{m_1}(\dot{y}_1 - \dot{y}_2) - \frac{k_1 + k_2}{m_1} y_1 + \\ &\quad + \frac{k_2}{m_1} y_2 + \frac{1}{m_1} F(t) \\ \ddot{y}_2 &= \frac{c}{m_2}(\dot{y}_1 - \dot{y}_2) + \frac{k_2}{m_2}(y_1 - y_2) \end{aligned}\right\}. \qquad (18.8)$$

Теперь мы можем начертить путевую карту (рис. 18.9).

Функцию вынуждающей силы получим, решая дифференциальное уравнение $\ddot{x} + \omega^2 x = 0$ при начальных условиях

$$x_0 = a\sin\omega t_0 = 0 \quad \text{и} \quad \dot{x}_0 = a\cos\omega t_0 = 0.$$

Другими начальными условиями служат смещения и скорости каждой массы в момент

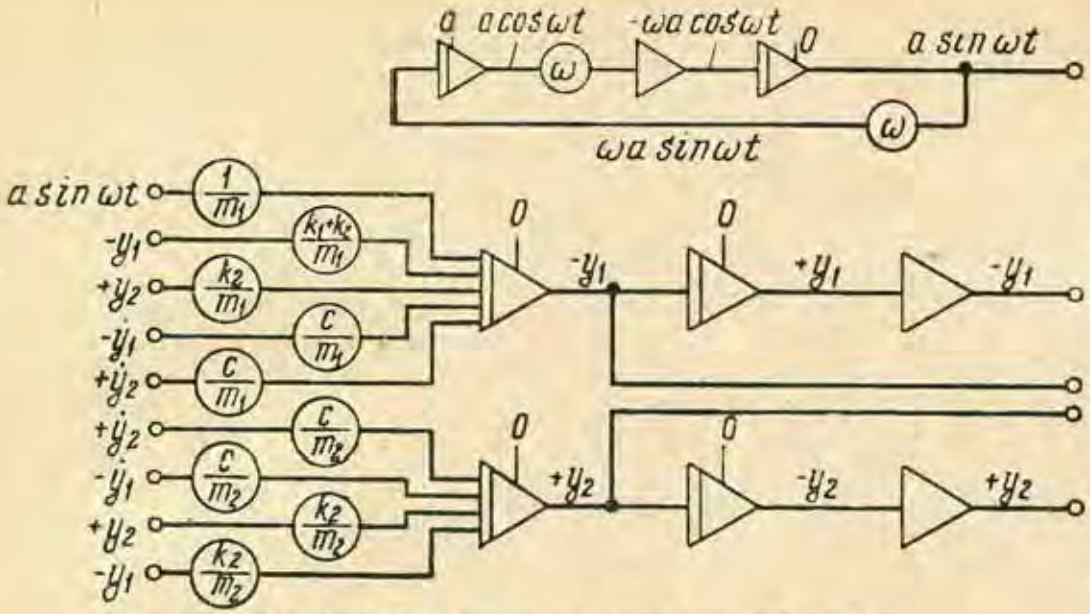


Рис. 18.9. Путевая карта для уравнений (18.8).

стоит в том, что если мы желаем провести моделирование какого-то ряда значений параметра, то при изменении одной из величин m_1, m_2, k_2 или c нам приходится одновременно менять уставки на четырех потенциометрах. Это очень неудобно и является дополнительным источником ошибок.

На рис. 18.10 показана одна из многих других возможных путевых карт. Эта карта составлена путем *непосредственного моделирования* [125] и

времени $t=0$. Мы предположим, что все они равны нулю; во многих случаях переходный процесс нас не интересует, и тогда не важно, какие значения имеют начальные условия.

В начерченной путевой карте много недостатков. Наиболее серьезный недостаток состоит совсем не требует формулировки уравнений. Мы замечаем, что пружина представляет собой элемент с двумя входами (смещения двух концов пружины) и одним выходом (сила, пропорциональная разности входов). Один из входов пружины k_1 фиксирован, и его аналог поэтому равен 0 *в*. Демпфер также имеет два входа (скорости цилиндра и поршня) и один выход (сила, пропорциональная разности входов). Масса изображается в виде элемента с несколькими входами (действующие на нее силы)и тремя выходами (ускорение, пропорциональное сумме сил; скорость, пропорциональная интегралу ускорения; и смещение, пропорциональное интегралу скорости). Карта на рис. 18.10 начерчена без учета знаков и содержит, очевидно, несколько лишних усилителей. На рис. 18.13 эта карта перечерчена в пригодном для практики виде.

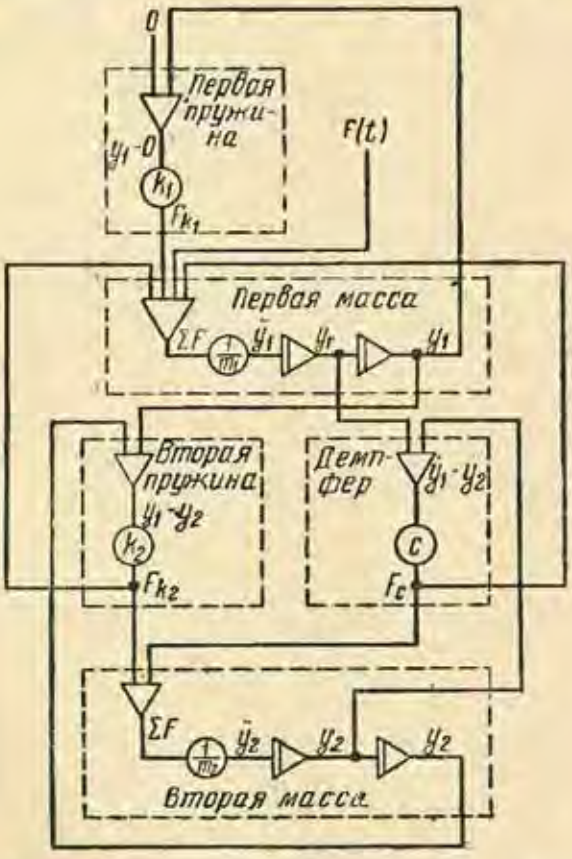


Рис. 18.10. Путевая карта для рис. 18.8, начерченная по методу прямого моделирования. Знаки, масштабные коэффициенты и лишние суммирующие усилители не учитываются. Схема для $F(t)$ не показана.

Путевая карта, начерченная по принципу непосредственного моделирования, не может содержать лишних интеграторов, которые иногда вносятся при составлении уравнений и могут привести к неверным решениям. Однако она может содержать лишние инверторы (в нашем случае путевая карта на рис. 18.10 содержит на один инвертор больше, чем карта на рис. 18.9). Другим недостатком путевой карты на рис. 18.10 является то, что в ней величины y_2, $\ddot{y}_1$ и $\ddot{y}_2$ не образуются непосредственно и отсутствуют на выходах. Ускорения $\ddot{y}_1$ и $\ddot{y}_2$ сами по себе обычно не представляют интереса, а смещение y_2 можно получить при необходимости, добавив в схему суммирующий усилитель. Путевая карта на рис. 18.13 весьма удобна для подбора масштабов, о чем будет говориться далее, в § 18.3.

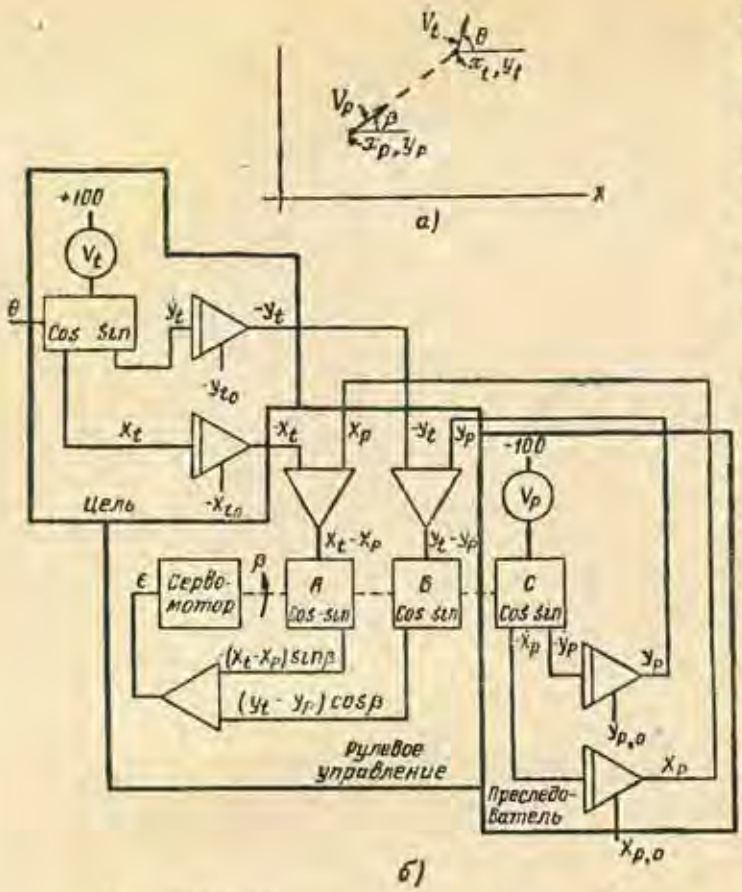


Рис. 18.11. Моделирование курса погони:
а) геометрия; б) аппаратура.

Моделирование преследования. Проведем теперь моделирование более сложной системы. Мы хотим исследовать процесс, при котором преследователь (например, самолет-перехватчик, или управляемый реактивный снаряд, или морская торпеда) пытается перехватить цель, руководствуясь тем критерием, что в каждый момент времени он должен держать курс прямо на цель. (Хотя во многих случаях более эффективны другие критерии, однако этот так называемый *критерий курса погони* используется весьма часто.) Входами системы являются мгновенное положение и скорость цели; требуется определить курс преследователя. Геометрический чертеж приводится на рис. 18.11,*а*, а путевая карта — на рис. 18.11,*б*. Заметим, что нам не приходится даже писать уравнений движения, а мы прямо переходим к моделированию физической ситуации.

На рис. 18.11,*б* вверху слева синусно-косинусное устройство вырабатывает составляющие $\dot{x}_t$ и $\dot{y}_t$ скорости V_t цели. Интегрирование этих величин дает координаты x_t и y_t цели, которые затем вычитаются из координат x_p и y_p преследователя. Полученные разности поступают в синусно-косинусные устройства A и B, где они умножаются соответственно на $\sin\beta$ и $\cos\beta$. Из геометрического чертежа видно, что только при правильном курсе β эти произведения будут равны, а потому сервомеханизм должен вращать вал вплоть до достижения угла β, при котором вход сервомеханизма (величина погрешности ε) сводится к нулю. Заметим, что с синусно-косинусного устройства A снимается — sin, а не sin, для того чтобы лишний раз не инвертировать знак и сэкономить один операционный усилитель.

Выход сервомеханизма связан с синусно-косинусным устройством C, в котором происходит умножение скорости V_p преследователя на соответствующие тригонометрические функции угла β и вырабатываются составляющие скорости для преследователя. После интегрирования этих составляющих получаются координаты преследователя, которые используются при вычитании координат цели из координат преследователя.

Какие выходы записывать при моделировании, зависит от назначения моделирования.

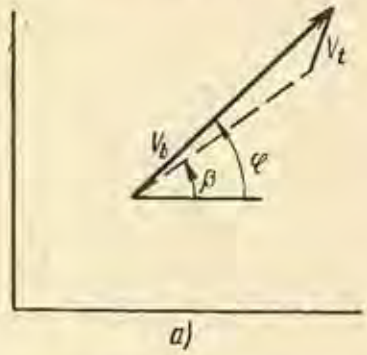
Очень интересно, например, записывать одним пером самописца кривую движения преследователя* (x_p, y_p), а другим пером на том же листе — кривую движения цели (x_t, y_t). В этом случае два пера могут начертить действительный ход перехвата. Движение цели можно задавать вручную или автоматически. При моделировании можно учесть и другие детали, как, например, предел ускорения преследователя, шумы (ошибки) во входной информации о положении цели и т. д.

Моделирование испытания авиационного пушечного прицела. Описанную модель можно использовать в качестве элемента более крупной системы. Предположим, например, что разработан новый автоматический прицел для воздушной стрельбы и должен быть испытан с участием человека — оператора. Ситуация показана на рис. 18.12. Самолет летит тем же курсом погони, как и в предыдущем примере, но теперь в подходящий момент времени он выпускает в цель снаряд со скоростью V_b и углом опережения $\varphi - \beta$ (т. е. курс снаряда относительно оси x равен φ).

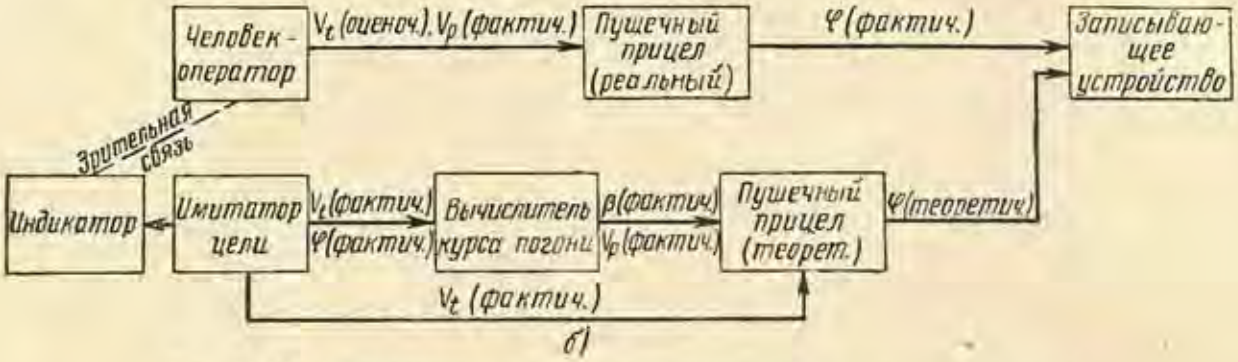
Вся модель, приведенная на рис. 18.11, изображается теперь одним блоком с надписью «Вычислитель курса погони». Имитатор цели моделирует движение цели, которая может выполнять маневры уклонения. Выходы этих двух блоков подаются в «теоретический» прицел, который вычисляет идеальный угол опережения. В то же время выход ими-

* Эта кривая движения преследователя курсом погони называется кривой погони или кривой гончих. — *Прим. ред.*

татора цели подается визуально реальному стрелку, применяющему реальный прицел. Стрелок вводит в прицел скорость своего самолета и свою оценку скорости цели и производит прицеливание. Прицел автоматически вычисляет угол опережения, который затем сравнивается с идеальным углом.



а)



б)

Рис. 18.12. Моделирование испытания авиационного пушечного прицела:
а) геометрия, *б*) аппаратура.

В упрощенной схеме на рис. 18.12 рассматривается только двумерная задача и, кроме того, не учитывается целый ряд тонкостей, таких, как предел возможного ускорения, пределы возможного направления прицела, точная баллистика движения снаряда, помехи и т. д., однако все это может быть учтено при реальном моделировании.

18.3. Подбор масштабов

Обычно каждая переменная в аналоговой машине должна оставаться в диапазоне ± 100 *в*. На входе или выходе усилителя эти пределы иногда могут быть расширены, но в такой, например, схеме, как на рис. 17.13, увеличение входного напряжения y сервомеханизма за пределы ± 100 *в* приводит к большой погрешности. Так как погрешности в машине обычно не зависят от значений переменных величин, то относительная погрешность будет минимальной в том случае, когда переменные поддерживаются возможно ближе к их максимальным значениям. Поэтому почти после каждой операции производится подбор масштабов при помощи потенциометров или мультипликативных констант на входах усилителей.

На практике мы не всегда знаем точные диапазоны изменения переменных, поэтому приходится вводить некоторые допуски. При выборе слишком малого масштаба возрастает относительная погрешность, но эта погрешность возрастает еще больше при выборе слишком большого масштаба, при котором переменные в машине превосходят 100 *в*. Очень часто оценки диапазонов могут быть получены из физики ситуации или же назначены произвольно. Например, при моделировании преследования возможное расстояние от перехватчика до цели, очевидно, не должно превышать дальности радиолокации. Некоторое представление о диапазонах изменения переменных часто можно получить из аналитических решений вырожденных случаев (например, таких случаев, когда некоторые из переменных равны нулю). И наконец, можно провести пробное моделирование, замерить напряжения и изменить при необходимости масштабы, что не слишком трудно.

Особые проблемы связаны с масштабом времени. Если масштаб времени равен единице, то говорят, что вычисление происходит в *реальном или натуральном времени*. Как мы уже отмечали, моделирование, в котором участвуют реальные компоненты (в том числе и люди), должно производиться в реальном времени. Если моделирование процесса на машине протекает быстрее, чем сам физический процесс, то говорят, что работа производится в *ускоренном времени*. Преимущество такой быстрой работы очевидно — экономия машинного времени. Эта экономия может быть весьма значительной, если моделируемый физический процесс занимает несколько часов. Возможность работы в ускоренном времени обычно ограничивается частотными ха-

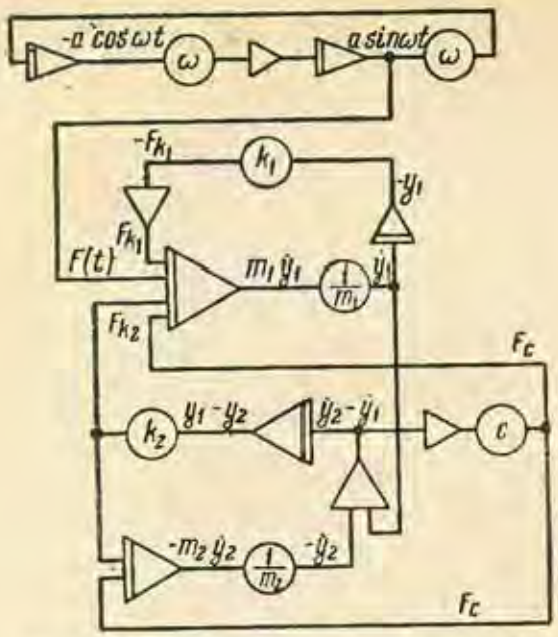


Рис. 18.13. Путевая карта для рис. 18.10, перечерченная в целях удаления лишних суммирующих усилителей и учета знаков.

рактеристиками механических элементов (см. § 17.6). Во многих практических случаях такое нижнее ограничение рабочего времени не особенно существенно ввиду большого времени набора (настройки) модели и необходимости наблюдения результатов. Там же, где оно существенно, необходимо применять чисто электронные устройства.

Возможность работы в *замедленном времени* ограничивается не только расходами, но и накоплением погрешности. Эта погрешность зависит от формы моделируемой функции, от коэффициента усиления интеграторов и от сопротивления утечки конденсаторов в петлях

Таблица 18.1*

Значения параметров

$a = 10$ футов
$m_1 = 64{,}4$ фунта $= 2$ слага
$m_2 = 3{,}22$ фунта $= 0{,}1$ слага
$k_1 = 2\,000$ фунтов на фут
$k_2 = 100$ фунтов на фут
$c = 2$ фунта на *фут/сек*
При $t = 0$ $y_1 = y_2 = \dot{y}_1 = \dot{y}_2 = 0$

Таблица 18.2

Значения переменных

$|y_1| < 0{,}05$ фута
$|y_2| < 0{,}15$ фута
$|\dot{y}_1| < 1{,}5$ *фут/сек*
$|\dot{y}_2| < 3{,}0$ *фут/сек*
$|\ddot{y}_1| < 50$ *фут/сек²*
$|\ddot{y}_2| < 150$ *фут/сек²*
Критическое значение $\omega = 31{,}7$ *рад/сек* $= 5{,}05$ *гц*.

обратной связи. Обычно считается нежелательным проводить моделирование какой-либо задачи в течение более чем 1—5 *мин*, если не прибегают к каким-либо специальным мерам предосторожности.

Другая проблема при выборе масштаба времени состоит в том, что каждое интегрирование эквивалентно умножению на время. При подборе временной шкалы это необходимо учитывать.

В качестве примера выберем масштабы для путевой карты на рис. 18.10, перечерченной (рис. 18.13) с удалением лишних усилителей и с правильной расстановкой знаков. В табл. 18.1 приведены принятые нами значения параметров; сильно различающиеся значения масс и констант пружин взяты преднамеренно, чтобы усложнить подбор масштаба. В табл. 18.2 приведены максимальные абсолютные значения переменных, найденные из грубого аналитического решения уравнений [59].

Для переменных y_1 и y_2 удобно иметь один и тот же масштаб (как и для других пар), и мы начнем с того, что смещение в 0,2 фута приравняем к напряжению в 100 *в*. Такой масштаб обеспечивает запас надежности для y_1 и y_2 в отдельности, но является предельным для разности $y_1 - y_2$. Однако напряжение, немного превышающее 100 *в*, в этой схеме не опасно, потому что сервомеханизмов в схеме нет и после первого пробного моделирования мы всегда можем изменить масштабы, если это окажется необходимым. Таким образом, наш масштабный коэффициент в этой точке равен 500, и мы пишем на рис. 18.14 не y_1, а 500 y_1 (рис. 18.14 отличается от рис. 18.13 только наличием масштабных коэффициентов). Запись этого коэффициента означает, что любое значение напряжения, замеренное в этой точке, численно равно 500-кратному значению переменной y_1. Аналогично мы пишем 500 y_2 вместо y_2.

Для скоростей мы приравниваем 4 фута в секунду 100 вольтам и приписываем ко всем $\dot{y}$ масштабный коэффициент 25. Для того чтобы от масштаба 25 : 1 для скорости перейти к масштабу 500 : 1 для смещения, нам нужно взять коэффициент усиления интегратора 20.

* Слово „фунт", как и слово „килограмм", может обозначать либо единицу массы, либо единицу силы (веса). Соответственно на основе американской традиционной системы мер и весов были созданы две системы механических единиц: фут—фунт-сила—секунда и фут—фунт-масса—секунда (аналогично технической и практической системам MKS). В первой системе (используемой в табл. 18.1) единицей силы служит 1 фунт-сила $= 453{,}6$ *г*, а единицей массы служит 1 слаг (slug) $= 1\ \frac{\text{фунт-сила}}{\text{фут/сек}^2} = g$ фунтов-масс, где g — ускорение силы тяжести. Так как $g \approx 9{,}81$ *м/сек²* $\approx 32{,}5$ *фут/сек²*, то 1 слаг $\approx 32{,}2$ фунта-массы $\approx 14{,}6$ *кг*. [В системе же фут—фунт-масса—секунда единицей массы служит 1 фунт-масса $= 453{,}6$ *г*, а единицей силы служит 1 фунтал (poundal) $= 1$ *фунт-масса·фут/сек²* $= \frac{1}{9}$ фунта-силы $\approx 14{,}1$ *г*.] — *Прим. ред.*

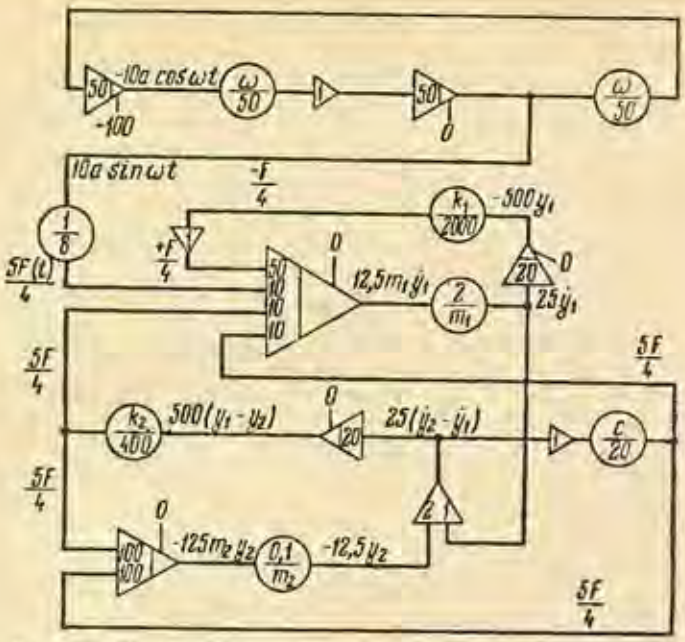


Рис. 18.14. Подбор масштабов в реальном времени для рис. 18.13.

На практике такие большие коэффициенты усиления не встречаются. Мы увидим, однако, что при введении масштаба времени коэффициент усиления интегратора будет значительно уменьшен.

Заметим, что критическая частота моделируемого процесса равна примерно 5 *гц* (5 циклам в секунду). Хотя наша аналоговая машина может оказаться в состоянии работать на таких частотах, применяемое нами записывающее устройство этого не допускает. Поэтому проведем моделирование в масштабе 0,1 к реальному времени *, так что полуцикл (полное отклонение в одном направлении) будет занимать 1 *сек*. Так как каждое интегрирование эквивалентно умножению на время, то коэффициент усиления интегратора уменьшается с 20 до 2. И даже если бы мы не подсчитали критическую частоту, то все равно мы могли бы прийти к мысли провести моделирование значительно медленнее реального времени ввиду необходимости уменьшить коэффициент усиления интегратора. Таким образом, процедура подбора масштабов часто позволяет гораздо лучше войти в задачу.

Дальнейший подбор масштабов будем пока вести в реальном времени, а переход к одной десятой реального времени произведем в конце. Рассмотрим теперь потенциометры, изображающие массы. Удобно установить движки этих потенциометров на значения $a=1$. Для этого запишем на путевой карте соответственно ** $2/m_1$ и $0{,}1/m_2$. Если при моделировании нам нужно увеличить какую-либо из масс, то это легко сделать путем изменения уставки соответствующего потенциометра. Если же мы хотим уменьшить какую-либо из масс, то сначала нужно увеличить вдвое усиление соответствующего интегратора и уменьшить вдвое уставку потенциометра, так как усиление потенциометра не может превысить единицу. Масштабный коэффициент 2 определяет выражение для количества движения, стоящее перед потенциометром массы m_1; так как масштабный коэффициент скорости равен 25, то масштабный коэффициент количества движения будет равен $25/2=12{,}5$. Таким образом, мы пишем $12{,}5\ m_1\dot{y}_1$. Мы должны были бы написать также $250\ m_2\dot{y}_2$, однако, как будет видно из дальнейшего, этот член будет записан иначе.

Точно так же выбирается масштаб для потенциометра, изображающего пружину k_1. Движок этого потенциометра устанавливается на значении $a=1$, и на путевой карте записывается $k_1/2000$. Выходное напряжение потенциометра равно тогда одной четверти действительной силы, и мы записываем $F/4$. Это в свою очередь определяет усиление интегратора: входной масштабный коэффициент $^1/_4$ и выходной масштабный коэффициент 12,5, определяют при моделировании в реальном времени усиление интегратора, равное 50.

Нам следовало бы взять для потенциометра k_2 тот же масштаб, что и для потенциометра k_1, но тогда напряжение на выходе k_2 будет небольшим. Поскольку это касается входов интегратора m_1, причина не в том, что масштабы выбраны плохо, а в том, что пружина k_2 действует на массу m_1 со сравнительно малой силой (по сравнению с силой от пружины k_1). В нашем случае целесообразно взять масштаб для потенциометра в 5 раз больше, а усиление интегратора по соответствующему входу — в 5 раз меньше. Таким образом, мы запишем $k_2/400$, а для интегратора возьмем коэффициент усиления 10. Рассуждая таким же образом, запишем $c/20$ (на потенциометре уставка 0,1) и снимем с этого потенциометра $^5/_4F$.

Но теперь мы видим, что выходные напряжения с интегратора m_2 очень малы. Поэтому мы применяем на входах усиления, равные 100; тем самым масштабный коэффициент количества движения будет равен 125, а масштабный коэффициент скорости 12,5. Это выражение скорости при суммировании с $25\ \dot{y}_1$

* Т. е. в 10 раз медленнее, чем в реальном времени. — *Прим. ред.*

** Здесь масса измеряется в слагах; см. примечание на стр. 203. — *Прим. ред.*

должно быть умножено на 2 (на входе суммирующего усилителя).

И наконец, мы применяем тот же самый масштабный коэффициент 5/4 к функции вынуждающей силы, которая поступает в интегратор с коэффициентом усиления 10. Такой масштаб легче всего осуществить устанавливая на интеграторе функции силы начальное условие 100 *в* для изображения силы в 10 фунтов и применяя потенциометр с уставкой $a=0{,}125$. Так как нас интересуют угловые частоты порядка 30 *рад/сек*, мы установим потенциометры на $\omega/50$, а усиление интегратора возьмем равным 50. Начальные условия на всех других интеграторах равны нулю.

На этом заканчивается для задачи подбор масштабов в реальном времени. Для того чтобы перейти к одной десятой реального времени, нам нужно просто уменьшить в 10 раз усиление на каждом из 10 входов шести интеграторов.

18.4. Дополнительные соображения

Для практического решения задачи на аналоговой вычислительной машине требуется большой опыт, так как при решении могут возникнуть многочисленные трудности. Для преодоления этих трудностей существуют различные методы. Приведем несколько примеров.

$$\left.\begin{aligned} 2x+y+5&=0\\ x+2y+3&=0 \end{aligned}\right\}. \qquad (18.9)$$

Решением этой системы уравнений являются числа $x=-7/3$, $y=-1/3$. Предположим, что мы пытаемся решить эту систему на аналоговой вычислительной машине по методике, описанной в § 18.1. Путевая карта приводится на рис. 18.15. Петля обратной связи имеет положительный коэффициент усиления 4, поэтому схема не стабильна. Если напряжения случайно окажутся равными $x=-7/3$ и $y=-1/3$, то схема будет находиться в статическом равновесии, но если напряжения будут хотя бы слегка отличаться от этих значений, то обратная связь приведет ко все возрастающему расхождению напряжений.

Таким образом, еще недостаточно, что каждая цепь дает математически правильный

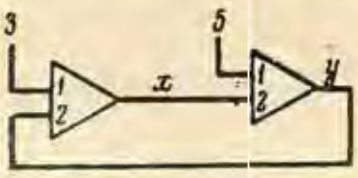


Рис. 18.15. Путевая карта для уравнений (18.9).

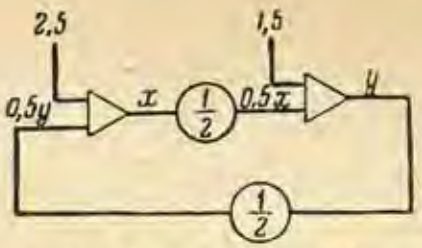


Рис. 18.16. Путевая карта для уравнений (18.10).

результат; необходимо также, чтобы каждая цепь и каждая петля обратной связи были стабильны. Существует ряд способов добиться этого на практике в конкретных случаях. В нашем примере для этого достаточно разделить каждое из уравнений на 2, получив

$$\left.\begin{aligned} x+0{,}5y+2{,}5&=0\\ 0{,}5x+y+1{,}5&=0 \end{aligned}\right\}. \qquad (18.10)$$

Путевая карта для этой системы уравнений (рис. 18.16) является стабильной.

Решая систему уравнений

$$\left.\begin{aligned} \ddot{x}+3\ddot{y}+x&=7\\ \dot{x}+2\dot{y}-3y&=5 \end{aligned}\right\}, \qquad (18.11)$$

мы сталкиваемся с затруднениями другого рода: мы не можем получить на выходах одновременно $\ddot{x}$ и $\ddot{y}$, так как первое уравнение может быть решено только относительно одной из этих переменных, а из второго уравнения нельзя получить ни одну из них. Из этого затруднения можно легко выйти, продифференцировав второе уравнение и получив

$$\ddot{x}+2\ddot{y}-3\dot{y}=0.$$

Это уравнение может быть теперь решено относительно одной из вторых производных, а первое уравнение — относительно другой второй производной, после чего обычным путем составляется путевая карта. Однако эта процедура приводит к одному лишнему интегратору, который может ввести ложные ре-

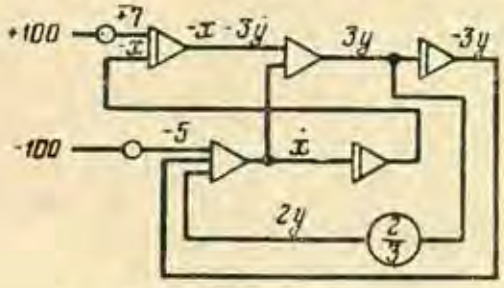


Рис. 18.17. Путевая карта для уравнений (18.11).

шения. В нашем случае лучше решить первое уравнение относительно $\ddot{x}+3\ddot{y}$, а второе уравнение относительно $-\dot{x}$. Соответствующая путевая карта показана на рис. 18.17. Начальным условием для интегратора в верхнем левом углу будет $-\dot{x}_0-3\ddot{y}_0$.

ЛИТЕРАТУРА

По аналоговым машинам имеется ряд хороших книг. Корн и Корн [50] ограничиваются почти целиком электромеханическими дифференциальными анализаторами. Сорока [51] весьма подробно рассматривает большое количество аналоговых устройств, но без оценки. Хорошая подборка статей, весьма отличных от указанных учебников, дается в книге [132].

ЗАДАЧА

18.1. Начертить путевую карту для уравнения (18.3).

ГЛАВА 19

ВХОДНО-ВЫХОДНЫЕ УСТРОЙСТВА

За исключением цифро-аналоговых и аналого-цифровых преобразователей, описание которых занимает основную часть этой главы, приводимое ниже рассмотрение входных и выходных устройств для вычислительных машин не затрагивает никаких новых принципов. Однако важность этой части систем нельзя недооценивать. При проектировании систем большого масштаба нередко бывает так, что входно-выходные устройства создают более серьезные проблемы, чем сама вычислительная машина. Даже в тех случаях, когда рассмотрение входно-выходных устройств не откладывается на конец (а оно часто откладывается), они ввиду своего специального характера требуют значительно большего времени на реализацию и, как показывает опыт, часто создают значительную долю затруднений во время рабочих испытаний систем большого масштаба.

Цифровые вычислительные машины. Любое из медленнодействующих запоминающих устройств, описанных в § 15.3, может быть использовано в качестве входного оборудования, выходного оборудования или же их обоих для цифровых вычислительных машин. Различные преобразователи между этими формами памяти (например, устройства перезаписи данных с перфокарт на перфоленту) представляют собой обычнейшие компоненты входно-выходного оборудования.

Часто требуется еще одна важная форма памяти — обычное печатание буквенно-цифровых знаков на бумаге. Почти всегда это окончательная форма выхода. Выходную информацию печатают табуляторы (см. § 15.3), а также различные автоматические печатающие устройства. Эти устройства могут работать от медленнодействующего запоминающего устройства любого типа, например от бумажной перфоленты; кроме того, они могут работать от электрических импульсов внутреннего языка машин.

Между вычислительной машиной и печатающим устройством необходимо специальное буферное запоминающее устройство, так как печатающее устройство работает намного медленнее, чем вычислительная машина, и не синхронно с ней. Кроме того, числа на печатающем устройстве обычно нужно подавать начиная со старшего разряда, в то время как в вычислительных машинах последовательного действия операции производятся над числами, начиная с младшего разряда.

В настоящее время печатная информация всегда вводится вручную в какое-либо медленнодействующее запоминающее устройство, например перфорируется на бумажную ленту, и только оттуда передается в вычислительную машину. Однако ведутся исследования по автоматическому чтению печатной информации. Некоторые результаты этих исследований описаны в § 25.3.

В качестве выходного устройства в цифровых вычислительных машинах часто применяется цифро-аналоговый преобразователь. Напряжение с выхода преобразователя можно подать в какое-либо исполнительное или индикаторное устройство; в качестве индикаторного устройства часто применяется электронно-лучевой осциллограф.

Аналоговые вычислительные машины. Входы аналоговых вычислительных машин, кроме функциональных связей, указываемых путевой картой, включают в свой состав специальные функции, константы умножения и начальные условия. Специальные функции вырабатываются генераторами функций, два типа которых (сервотаблица и фотоформирователь) были описаны выше. Константы умножения устанавливаются на потенциометрах или механических передачах, начальные усло-

вия устанавливаются на конденсаторах. Эти уставки делаются обычно вручную, однако они могут производиться и автоматически на основании команд, записанных на перфоленте, перфокартах и т. п. Даже внутренние соединения в аналоговой машине могут изменяться автоматически в соответствии с заранее составленной программой.

В качестве выходных устройств в аналоговых машинах обычно применяются сервомоторы (которые являются исполнительными устройствами этих машин), самописцы, измерительные приборы, осциллоскопы, аналогово-цифровые преобразователи и т. д. В большинстве случаев результаты записываются на движущейся бумажной ленте самописца, при этом каждый след пера на ленте представляет собой график одного из напряжений (т. е. одной из переменных) в зависимости от времени. На ленте могут записываться по такому методу одновременно несколько переменных. Для записи функции двух переменных может использоваться сервотаблица.

Ценными приборами для индикации результатов решений на аналоговой машине являются электронно-лучевые осциллоскопы. Их показания могут записываться фотографически, хотя при фотозаписи невозможно получить такую же точность, как в случае записывающих вольтметров. В полноэлектронных аналоговых вычислительных машинах частотное содержание выходов превышает возможности механических записывающих устройств, и в этом случае обычно требуются осциллоскопы того или иного типа.

19.1. Аналого-цифровые преобразователи

Как уже отмечалось выше, аналоговые величины («аналоги») в вычислительных машинах обычно имеют форму напряжений, однако они могут иметь также большое количество других форм. Наиболее важными из этих форм являются поворот вала, длительность (например, длительность импульса или длительность паузы между импульсами), частота, положение (например, положение луча в фотоформирователе). Так как аналоговые величины могут быть преобразованы из одной

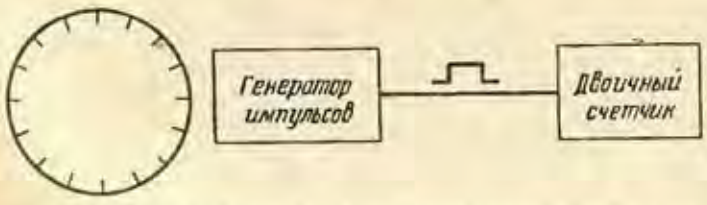


Рис. 19.1. Преобразователь поворотов вала в цифры по принципу счета импульсов.

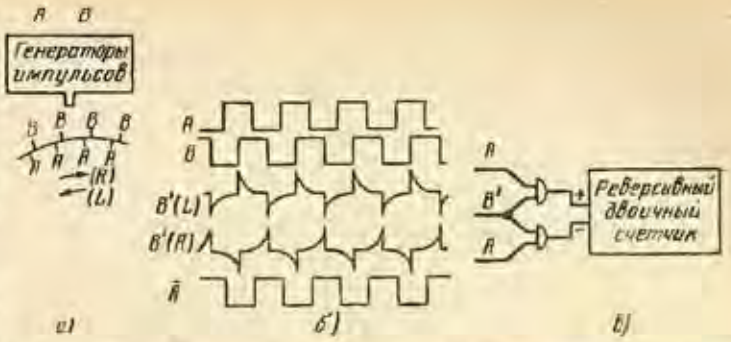


Рис. 19.2. Реверсируемый преобразователь поворотов вала в цифры:
а) диск и импульсный генератор; б) формы импульсов; в) управление счетчиком.

формы в другую, то обычно считают, что аналоговая величина, подлежащая преобразованию в цифровую величину, представлена в форме электрического напряжения. Однако существует целый ряд специальных преобразователей аналоговых величин в цифровые, которые непосредственно воспринимают аналоговые величины, представленные не в форме напряжения.

Преобразование поворотов вала в цифровые величины. В качестве примера таких устройств рассмотрим преобразование поворотов вала в цифровые величины. На рис. 19.1 показан преобразователь поворотов вала в цифровые величины, работающий по принципу счета импульсов. По окружности изображенного слева диска расположены какие-нибудь метки, способные давать сигналы. Этими метками могут быть электрические контакты, отверстия, сквозь которые может проходить луч света, и т. д. Генератор импульсов воспринимает сигналы от этих меток и вырабатывает один импульс каждый раз, когда метка проходит мимо генератора. Импульсы от генератора поступают в двоичный счетчик, который производит подсчет общего числа поступивших в него импульсов.

На одном диске может быть размещено до 1024 меток; для увеличения числа оборотов применяют зубчатые передачи (однако такие передачи часто вводят погрешность из-за люфта). Двоичный счетчик на триггерах легко может считать импульсы на частотах 1 *Мгц* и выше; при меньшей частоте поступления импульсов можно использовать более простые счетчики.

Если требуется преобразовать повороты вала, который периодически изменяет направление вращения, то нужно применить реверсивный двоичный счетчик. Кроме того, генератор импульсов должен иметь способность воспринимать направление вращения вала и

в зависимости от этого направления вырабатывать импульсы на том или другом из двух имеющихся выходов. Как показано на рис. 19.2,*а* генераторы импульсов реагируют на две серии меток A и B, нанесенные по окружности диска. Расстояние между метками A и B равно четверти расстояния между одинаковыми метками. Выходные напряжения генераторов, после подлежащего прямоугольного формирования, показаны как сигналы A и B на рис. 19.2,*б*.

Если продифференцировать импульсный сигнал B, то при вращении вала по часовой стрелке продифференцированные импульсы будут иметь форму $B'(R)$; если же вал вращается против часовой стрелки, то продифференцированные импульсы будут иметь форму $B'(L)$. Сигнал $\bar{A}$ представляет собой просто дополнение (инверсию) сигнала A. На рис. 19.2,*в* показано, что если B' совпадает с A, то счетчик считает со знаком плюс, а если B' совпадает с $\bar{A}$, то счетчик считает со знаком минус.

Другой тип преобразователя поворотов вала в цифровые величины показан на рис. 19.3,*а*. Здесь изображен 5-битовый кодовый диск. Один сектор диска освещен; выходная величина считывается фотоэлементами последовательно или параллельно после того,

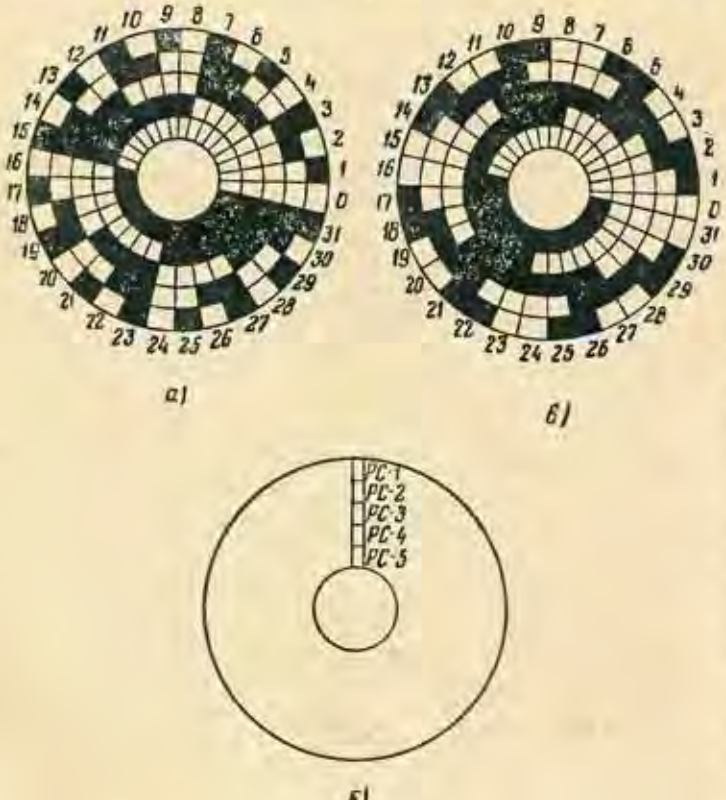


Рис. 19. 3. Преобразователь поворотов вала в цифры с прямым отсчетом:
а) двоичный кодовый диск; *б*) метод отсчета; *в*) рефлексный кодовый диск.

как вал повернулся менее чем на один полный оборот. На рис. 19.3,*б* показана схема параллельного считывания; *PC-1* — это фотоэлемент, который считывает самый младший разряд числа, и т. д. Число разрядов может быть и больше пяти; разрядность ограничивается только разрешающей способностью фотоэлементов, неточностями мертвого хода и расцентровкой меток. Вследствие возможности расцентровки (угловых ошибок) меток и возможности остановки диска на границе между секторами может возникнуть большая ошибка из-за того, что часть цифр будет считана с одного сектора и часть — с другого.

Рис. 19.4. Кодирующая трубка для преобразования напряжения в цифры.

По этой причине на практике следует использовать запись чисел на диске в рефлексном коде (коде Грея), а не в обычном двоичном коде. При использовании рефлексного кода максимальная ошибка расцентровки не превышает единицы самого младшего разряда. Преобразование чисел из кода Грея в обычный двоичный код легко выполняется при помощи схемы последовательного действия на рис. 15.12.

Преобразование напряжения в цифровые величины. Для преобразования напряжения в цифровые величины может быть использована кодирующая трубка, по своему действию во многом похожая на описанный выше кодовый диск. На рис. 19.4 показана 3-битовая кодирующая трубка для перевода напряжения в обычный двоичный код. На практике возможно использовать до 8 битов, причем лучше применять не обычный двоичный код, а рефлексный. Изображенная на рисунке фигура наносится краской на экран электронно-лучевой трубки в качестве постоянной маски. Преобразуемое напряжение прикладывается к отклоняющим пластинам y, а к отклоняющим пластинам x прикладывается пилообразное напряжение развертки. Фотоэлемент вырабатывает последовательно выходные импульсы. Устройство описанного типа может работать на мегагерцевых частотах.*

Вторым типом преобразователя напряжения в цифровые величины является так называемый *аналого-время-цифровой преобра-*

* Аналого-цифровые преобразователи типа кодирующей трубки и кодового колеса называются преобразователями с пространственным кодированием. — *Прим. ред.*

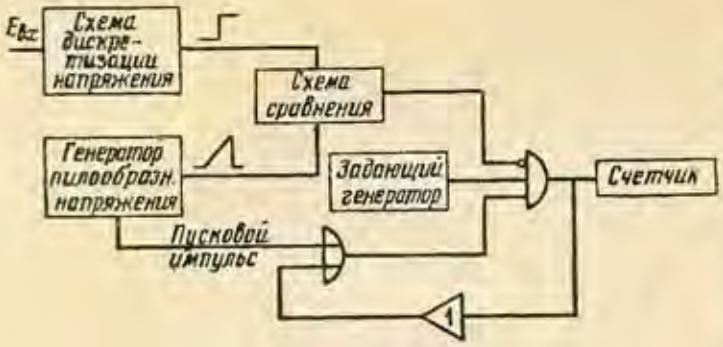


Рис. 19.5. Аналого-время-цифровой преобразователь.

зователь, в котором аналоговое напряжение преобразуется во время, а время — в цифровую величину *. Преобразование напряжения во время производится посредством генерирования наклонного (пилообразного) напряжения и сравнения его с дискретизированным аналоговым напряжением **; когда оба напряжения станут равны, протекшее время будет пропорционально преобразуемому напряжению. Это время преобразуется в цифровую величину посредством запуска генератора импульсов в начальный момент генерации пилообразного напряжения и запирания этого генератора в момент сравнения.

Блок-схема преобразователя показана на рис. 19.5. Преобразуемое напряжение дискретизируется и подается на один вход схемы сравнения, наклонное напряжение подается на другой вход. С началом генерации наклонного напряжения отпирается клапан ИЛИ и клапан И, разрешающий импульсам с генератора импульсов проходить на счетчик. В момент равенства двух напряжений схема сравнения выдает запрещающий импульс, который запирает клапан И и останавливает счетчик. Полученное на счетчике число затем может быть считано в параллельном или последовательном коде. Устройства такого типа работают сравнительно медленно: n-битовое преобразование может занять 2^n периодов повторения импульса, против n периодов в случае описанной выше кодирующей трубки.

Третий тип преобразователей напряжения в цифровые величины основан на применении обратной связи, с использованием цифро-аналогового преобразователя в системе с замкнутой петлей ***. Этот принцип поясняется рис. 19.6,*а*, где изображен n-битовый преобразователь, требующий до 2^{n-1} периодов повторения импульса. Двоичный счетчик здесь устанавливается первоначально на 2^{n-1} и после этого приводится в действие цифрово-аналоговой преобразователь. Затем схема сравнения и обнаружитель знака ошибок запускают генератор импульсов и двоичный счетчик считает импульсы в том или ином направлении до тех пор, пока схема сравнения не обнаружит равенства. В этот момент счетчик содержит число, являющееся цифровым значением заданной аналоговой величины.

На рис. 19.6,*б* показано устройство, которое работает по тому же принципу, но благодаря более сложной схеме сокращает время преобразования до $n+1$ периодов повторения импульса. Двоичный счетчик, детально показанный на схеме, состоит из n триггеров, каждый из которых может быть *установлен* (на 1) и *сброшен* (на 0). В момент T_0 все триггеры хранят 0. В момент T_1 первый триггер устанавливается на 1 и цифро-аналоговый преобразователь преобразует это двоичное число в эквивалентное напряжение, которое сравнивается с дискретизированным напряжением. Если ошибка (разность этих напряжений) отрицательна, то ничего не происходит (и единица в первом триггере остается единицей); если же ошибка положительна, то в момент T_2 первый триггер сбрасывается на 0 посредством клапана И. В тот же момент времени T_2 2-й триггер переводится в состояние 1. Затем весь процесс повторяется.

Преобразователи только что описанного

---

* Аналого-время-цифровые преобразователи называются иначе преобразователями с время-импульсным кодированием. — *Прим. ред.*

** Дискретизация непрерывной переменной состоит в замере ее значений в фиксированные дискретные моменты времени, отстоящие друг от друга на определенные интервалы; в итоге, если запоминать каждое замеренное значение вплоть до следующего замера, непрерывная кривая заменяется ступенчатой кривой (ср. § 28.3). Эти замеренные дискретные значения переменной иногда называют ее «дискретами» (единственное число — «дискрета»).

В данном случае дискретизация применяется для фиксации преобразуемого напряжения на неизменном уровне в течение одного цикла преобразования (от замера напряжения до выдачи цифровых импульсов); это облегчает условия работы преобразователя. Дискретизирующая схема на рисунке не показана. Дискретизация преобразуемого напряжения широко применяется и в других аналого-цифровых преобразователях. Однако, вообще говоря, можно обойтись и без дискретизации преобразуемого напряжения, если выполнять преобразование с такой скоростью, при которой напряжение изменяется за цикл преобразования весьма незначительно, скажем не более чем на единицу младшего разряда своего численного представления. — *Прим. ред.*

*** Такие преобразователи напряжения называются преобразователями со ступенчатой компенсацией напряжения. Метод пространственного кодирования, метод время-импульсного кодирования и метод ступенчатой компенсации напряжений суть три основных метода преобразования аналоговых величин в цифровые. — *Прим. ред.*

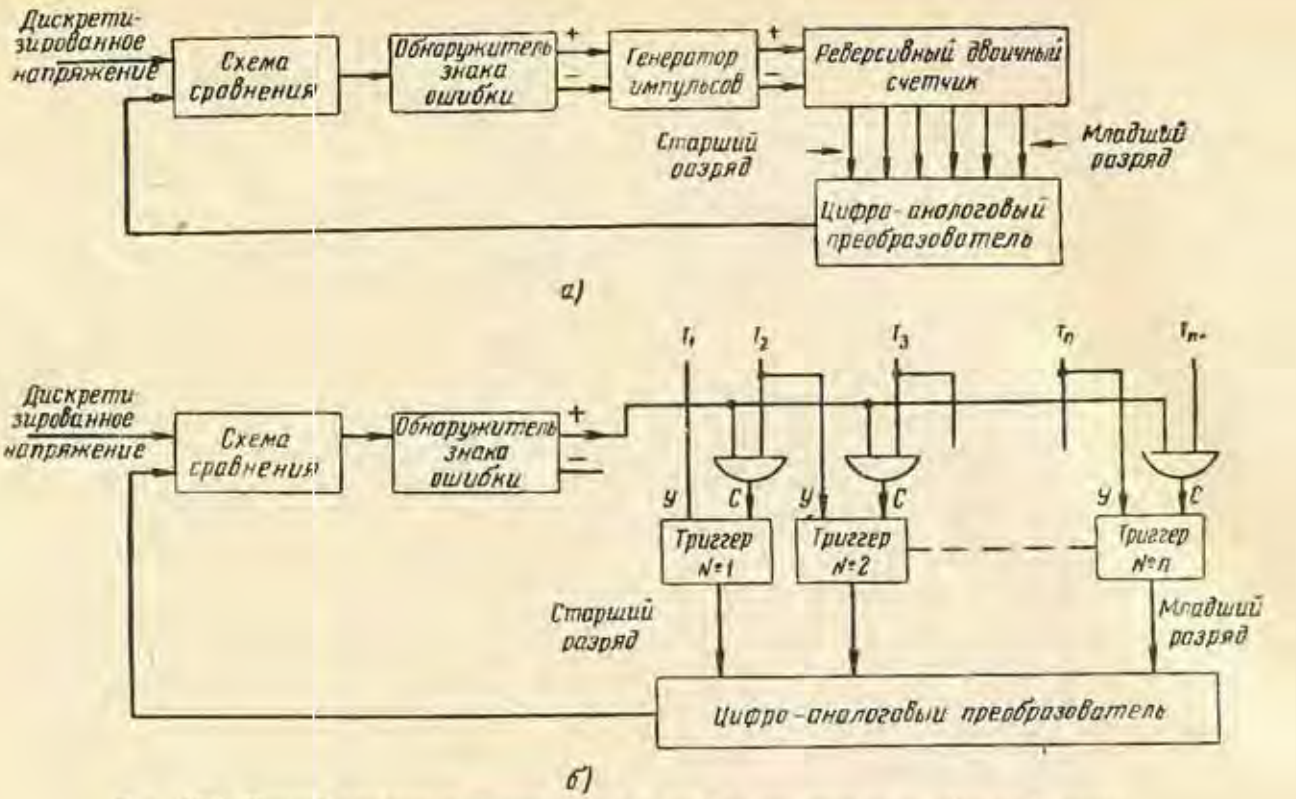


Рис. 19.6. Преобразователь напряжения в цифры по принципу обратной связи.

типа дороги, но точны и быстры. Схема сравнения обычно работает на переменном токе, и знак ошибки находится по фазе сигнала ошибки.

19.2. Цифро-аналоговые преобразователи

Цифровые величины обычно преобразуются в напряжение, потому что напряжение, если нужно, легко может быть преобразовано в любую другую аналоговую величину. В частности, если требуется преобразовать число в угол поворота, то можно использовать схему, в которой отсчет с кодового диска сравнивается с заданным числом, пока их разность (получаемая с помощью цифрового вычитателя) не станет равна нулю. Однако проще и удобнее преобразовать цифровую величину в напряжение, а затем при помощи сервомотора преобразовать напряжение в угол поворота вала.

Разработан целый ряд очень изящных и сложных схем преобразования цифровых величин в аналоговые. Приводимые здесь два примера схем преобразования отличаются, однако, в первую очередь своей простотой. Одна из схем предназначается для чисел, заданных в последовательной форме, а другая — для чисел, заданных в параллельной форме.

Преобразование последовательных чисел в напряжение. На рис. 19.7 показана исключительно простая схема, которую разработали Шеннон и Рэк [135]. Переключатель замыкается на один период импульса, когда в очередном разряде преобразуемого числа находится цифра 1, и размыкается, когда в очередном разряде находится цифра 0. Преобразуемое число поступает в схему в последовательной форме младшими разрядами вперед. Цепь RC подобрана так, что накопленный в ней заряд стекает ровно наполовину за один период повторения импульса. После того как на схему подан последний, самый старший разряд числа, на выходе образуется искомое эквивалентное напряжение; оно должно быть считано немедленно, потому что заряд продолжает стекать.

Работу схемы можно пояснить следующим образом. Предположим, что первая (младшая) цифра числа есть 1. Пока переключатель замкнут, на конденсаторе накапливается единичный заряд от источника постоянного

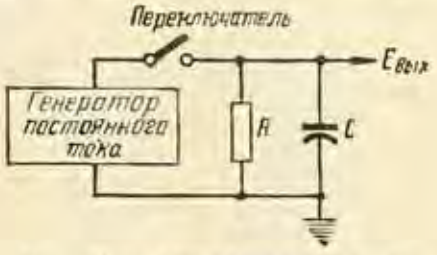


Рис. 19.7. Преобразователь последовательных цифр в напряжение.

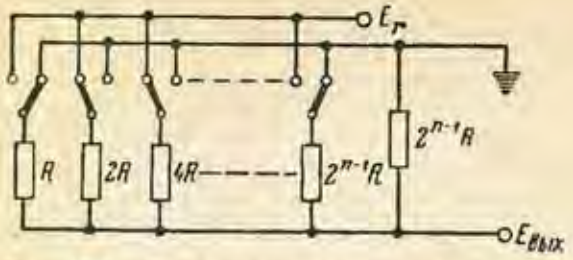


Рис. 19.8. Преобразователь параллельных цифр в напряжение.

тока. В конце следующего периода импульса этот заряд уменьшается вдвое, в конце еще одного периода импульса он уменьшается вчетверо и т. д. Таким образом, независимо от того, будет ли переключатель разомкнут или замкнут в течение последующих импульсов, самая младшая цифра числа создает на конденсаторе в момент окончания последнего импульса напряжение, равное 2^{-n} от основного напряжения, если эта цифра была 1, и равное 0, если она была 0. Аналогично, вторая цифра числа создает долю $2^{-(n-1)}$ или 0 и т. д. Все эти доли надо сложить.

Описанное устройство успешно работало с 7 битами (т. е. с точностью 1/128), хотя использованные схемы коммутации и синхронизации несколько сложны. В качестве источника постоянного тока использовался пентод.

Преобразование параллельных чисел в напряжение. На рис. 19.8 показана схема преобразования параллельного двоичного числа в эквивалентное аналоговое напряжение. В схеме на каждый двоичный разряд имеется по одному переключателю, который, если данный разряд равен 1, соединяется с опорным напряжением E_r (левое положение), а если разряд равен 0, — с землей (правое положение). Старший разряд подается на самый левый переключатель. Показанное на рисунке положение переключателей соответствует числу 011...0. Сопротивления действуют как делители напряжения, приводя выходное напряжение к надлежащей величине между землей и опорным напряжением E_r. Выходное сопротивление схемы равно $^1/_2R$, оно постоянно и не зависит от положения переключателей.

ЛИТЕРАТУРА

Материал этой главы заимствован в основном из статьи Х. Гарнера [126].

ГЛАВА 20

СРАВНЕНИЕ АНАЛОГОВЫХ И ЦИФРОВЫХ МЕТОДОВ

Вопрос о том, какой воспользоваться аппаратурой — цифровой или аналоговой, постоянно возникает при проектировании систем большого масштаба. Собственно говоря, вопрос касается не только относительных преимуществ двух типов вычислительных машин, но и использования того или иного класса методов. Иногда должны быть использованы методы обоих классов, и тогда возникает вопрос о выборе тех точек внутри системы, в которых должны производиться преобразования.

Кроме того, когда система становится более сложной и более автоматической, существует тенденция к сдвигу в сторону цифровой техники, и в связи с этим может возникнуть вопрос о наилучшем моменте в истории разработки для такого сдвига. Предметом спора тогда становятся лучшие рабочие характеристики в течение промежуточного периода, полученные благодаря сохранению или введению аналоговой техники, по сравнению с ранним получением еще лучших рабочих характеристик при предпочтении цифровой системы; преимущества, вытекающие из введения промежуточной (временной) аналоговой системы, следует также сопоставить со стоимостью такой быстро устаревающей разработки.

Трудность сравнения аналоговых и цифровых систем хорошо иллюстрируется спорами, которые часто велись при выборе между ними. Внутренние или потенциальные преимущества одних и других могут почти совсем утрачиваться на практике. Кроме того, выбор между двумя предложениями: применить аналоговую или цифровую технику — часто весьма сильно зависит от других факторов в этих предложениях. Такие факторы, хотя они и не вытекают по существу из различий между аналоговой и цифровой техникой, часто бывают практически связаны с ними; например, предлагаемая цифровая система часто оказывается более централизованной и более автоматизированной, чем предлагаемая аналоговая система. Окончательное решение о выборе может основываться не только на эффективности и стоимости, как это мы рассматриваем здесь, но и на сроках разработки, легкости приобретения, гибкости в случае из-

менения цели системы и даже на догадках о том, насколько точны оценки проектировщиком всего предыдущего.

20.1. Разрядность и точность

Разрядность можно определить для наших целей как количество значащих цифр, с которым записываются результаты вычислений; эти цифры могут быть как верными, так и неверными. Точность можно определить для наших целей как количество верных значащих цифр. Таким образом, точность числа никогда не превышает его разрядности, но разрядность может значительно превышать его точность.

Например, мы можем измерить сторону квадрата линейкой и определить, что ее длина равна 8,63 дюйма. Возможно, что мы ошибаемся на одну или две единицы в последнем разряде, однако разрядность ответа равна 3 десятичным цифрам, и если линейка правильно прокалибрована, то точность также равна 3 десятичным цифрам. Мы можем попытаться невооруженным глазом считывать 0,001 дюйма, однако вряд ли будем получать одинаковые результаты при повторных измерениях и, таким образом, не сможем добиться разрядности более чем в 3 десятичные цифры. Если теперь мы будем вычислять площадь нашего квадрата, то в результате может получиться 74,4769 кв. дюйма. Разрядность этого результата равна 6 десятичным цифрам, а точность — только 3 десятичным цифрам.

В цифровых устройствах разрядность может быть получена любая. Длину слов в универсальных цифровых вычислительных машинах часто делают равной примерно 12 десятичным разрядам. Увеличить разрядность можно, увеличив длину слов. При этом возрастает стоимость вычислительной машины. Можно также увеличить время выполнения операций и вести вычисления с удвоенным числом разрядов. Первая из электронных цифровых вычислительных машин, ENIAC, работавшая сравнительно медленно и обладавшая сравнительно небольшой памятью, вычислила математическую константу π с разрядностью (и точностью) свыше 2000 десятичных знаков *.

В пределах своей разрядности цифровая вычислительная машина не имеет другого внутреннего источника ошибок, кроме кумулятивной ошибки округления. Хотя при некоторых длинных вычислениях погрешность округления может оказать влияние на много десятичных разрядов, однако в вычислениях, используемых для логического управления системами, ошибка округления вряд ли опасна. Цифровому процессу, который следует отличать как таковой от цифровой вычислительной машины, свойственны еще дополнительные источники ошибок, например ошибка обрыва («ошибка формулы»), связанная с применением конечного числового (дискретного) процесса вместо бесконечного дискретного процесса, эквивалентного непрерывному математическому процессу. Эту ошибку, однако, всегда можно удерживать в заданных границах, увеличивая число членов при вычислениях (и, следовательно, длительность вычислений).

Разрядность, достижимая при помощи аналоговых устройств, может быть различной. В качестве грубого эмпирического правила можно сказать, что в лаборатории можно получить точность в 5 десятичных разрядов, при разработке — в 4 десятичных разряда и в реальных эксплуатационных условиях — в 3 разряда. Ошибки (неточности) возникают из-за дрейфа в усилителях, люфта в механических устройствах, изменения величин сопротивлений и других бедствий, угрожающих любому точному прибору. Однако в отличие от ошибок округления в цифровой вычислительной машине, которые имеют тенденцию суммироваться статистически (т. е. как квадратный корень из суммы квадратов), ошибки отдельных компонентов аналоговой машины на практике почти не суммируются, потому что большинство элементов охвачено петлями обратной связи.

20.2 Многоцелевость и гибкость

Под многоцелевостью здесь понимается способность решать различные задачи, под гибкостью — возможность легко и быстро перестраиваться с решения одной задачи на решение другой (о гибкости в смысле расширяемости см. § 20.4).

Сравнение аналоговых и цифровых машин по многоцелевости провести трудно, так как машины обоих типов можно приспособить для решения любой разрешимой математической задачи. В некотором смысле аналоговая машина обладает внутренней, натуральной многоцелевостью благодаря своей способности

* ENIAC (Electronic Numerical Integrator and Automatic Calculator, т. е. «Электронный численный интегратор и автоматический вычислитель») — первая американская электронная вычислительная машина; строилась во время II мировой войны в Пенсильванском университете для Управления вооружения армии США. Впервые публично демонстрировалась в феврале 1946 г, содержала 18 000 электронных ламп и 1500 электромеханических реле. — *Прим. ред.*

к непосредственному выполнению столь различных операций, как интегрирование, образование синуса и умножение, в то время как цифровая машина должна составить эти операции из простых математических процессов сложения, вычитания и умножения на число, служащее основанием системы счисления. По этой причине аналоговая машина может оказаться более подходящей как вычислительная машина для частных ситуаций, как, например, для решения системы обыкновенных дифференциальных уравнений. Но с точки зрения всей системы большого масштаба, это преимущество может оказаться тривиальным.

По сравнению с аналоговой цифровая вычислительная машина является более гибкой, так как для настройки ее на решение новой задачи достаточно ввести в машину новую ленту. На аналоговой машине в этом случае нужно производить перекоммутацию, переставлять движки потенциометров и задавать новые начальные условия на конденсаторах. Хотя все необходимые коммутации могут быть подготовлены заранее на сменных коммутационных досках, а потенциометры и конденсаторы могут настраиваться автоматически с помощью телеуправления (возможно по программе на ленте), однако на практике настройка аналоговой машины на решение новой задачи представляет собой длительный процесс.

Наиболее важный вопрос, связанный с гибкостью при проектировании систем, состоит в том, выбрать ли универсальную или специализированную вычислительную машину. Если мы выбираем специализированную машину (а в ныне действующих системах центральная вычислительная машина почти всегда является несколько специализированной), то вопрос о внутренней многоцелевости и гибкости аналоговых или цифровых машин будет представлять сравнительно малый интерес.

20.3. Скорость

Здесь опять аналоговая вычислительная машина обладает внутренним преимуществом, которое, однако, на практике может не иметь значения. Аналоговая машина выполняет такие операции, как интегрирование, непрерывно и быстро, и 10 последовательно соединенных интеграторов совершает 10 последовательных интегрирований примерно так же быстро, как один интегратор совершает одно интегрирование. В цифровой вычислительной машине интегрирование должно осуществляться суммированием приращений по методам численного анализа; последовательные интегрирования должны выполняться по очереди друг за другом.

Сказанное относится к одному частному решению; по мере того как вход изменяется во времени, решение должно перевычисляться. Рассмотрим вычислительное устройство пушечного прицела. По мере изменения входа (информация от радиолокатора) должен изменяться и выход (угол опережения). Аналоговое вычислительное устройство совершает изменение непрерывно, в то время как цифровое устройство должно производить повторные вычисления через дискретный интервал времени на основе новых входных данных; на практике, однако, входная информация должна сглаживаться, чтобы предотвратить случайные колебания наводки прицела из-за помех, и поэтому реального различия между непрерывным решением и дискретным решением не существует, если только дискретное решение повторяется достаточно часто (по сравнению с постоянной времени при сглаживании).

Если мы нашли, что повторять вычисление чаще чем один раз в три секунды нецелесообразно, и если цифровая машина может произвести все дискретные вычисления за 300 *мсек*, то цифровая машина может обрабатывать 10 различных входов по методу разделения времени. Хотя разделение времени теоретически возможно также и для аналоговых машин, однако для аналоговых машин оно неудобно и вместо него, вероятно, применили бы 10 параллельных аналоговых вычислительных машин. Таким образом, для некоторых типов задач (например, для задач, связанных с решением систем дифференциальных уравнений) аналоговая машина обычно может получить решение быстрее, чем цифровая машина; однако на практике почти всегда возможно построить цифровую машину, которая может получить решение достаточно быстро (если это решение вообще может быть произведено на какой-либо вычислительной машине). Когда входы множественны, нередко одна цифровая машина может выполнить работу нескольких аналоговых машин.

20.4. Стоимость

Стоимость аналоговой машины примерно пропорциональна ее размерам, тогда как стоимость цифровой вычислительной машины, по-видимому, никогда не бывает меньше некоторой довольно значительной суммы. Механические цифровые устройства, например счетно-перфорационные машины, обычно гораздо

дешевле, чем электронные устройства, но часто совершенно не сопоставимы с ними. Сравнение стоимости обычно делается между одной большой цифровой установкой и комплектом аналоговых установок.

С вопросом стоимости связан вопрос о способности к расширению системы. Цифровая вычислительная машина должна строиться в расчете на максимальную предвидимую загрузку, так как после постройки цифровой машины ее размеры относительно неизменяемы. Некоторые изменения, конечно, можно произвести; например, иногда возможно заменить запоминающее устройство машины каким-либо более быстрым запоминающим устройством и таким образом выиграть машинное время для обработки большего количества входов. Но такие усовершенствования не могут продолжаться бесконечно, и если цифровая машина перегружена, то наступает время, когда она должна быть полностью дублирована. В аналоговой машине, наоборот, можно всегда добавить новое устройство, если это нужно. Таким образом, сравнение по стоимости связано с точностью предсказания нагрузки.

Разрядность гораздо дешевле получить на цифровой машине, чем на аналоговой. Стоимость цифровой машины примерно пропорциональна разрядности, в то время как повышение разрядности в аналоговой машине приводит к очень быстрому увеличению стоимости, если только повышенная разрядность вообще в этом случае достижима. Поэтому относительная предпочтительность аналоговых вычислительных машин, по-видимому, возрастает намного там, где не требуется большой точности и где нас не ограничивает разрядность.

20.5. Надежность

Под надежностью понимается многое. Если речь идет о том, что устройство должно работать, когда это требуется, а не выходить из строя, то между аналоговыми и цифровыми устройствами нет особой разницы. Современные электронные вычислительные машины, несмотря на свою сложность, имеют в настоящее время такую степень надежности, которая казалась почти невозможной 10 лет назад. Серийная вычислительная машина (в отличие от лабораторного образца) при обычном обслуживании будет с 90-процентной вероятностью производить расчеты в течение многих часов без ошибок или поломок. Там, где действительно требуется 100-процентная вероятность безошибочной работы, вычислительную машину можно поставить в параллель с одной или даже двумя резервными машинами [см. уравнение (5.14)].

Если случится ошибка, то цифровая вычислительная машина обладает, по крайней мере теоретически, способностью поставить диагноз своим неисправностям и тем самым локализировать источник ошибки; закладывается ли действительно такая способность в машину, зависит от относительной ценности и стоимости дополнительного оборудования. После локализации неисправность должна быстро устраняться, поскольку и цифровые, и аналоговые машины могут и должны конструироваться так, чтобы неисправный блок можно было легко удалить и целиком его заменить новым. Ремонт этого блока может оказаться сложным, но это уже не сказывается на надежности машины в целом.

Надежность связывается также с вероятностью того, что полученный ответ является правильным. Аналоговые устройства имеют тенденцию делать маленькие ошибки (§ 20.2), в то время как цифровые устройства имеют тенденцию делать большие ошибки (потому что ошибка в самом старшем разряде так же вероятна, как ошибка в самом младшем разряде). Для того чтобы избежать грубых ошибок в цифровых системах, процесс решения в них иногда непрерывно и автоматически контролируется. Такой контроль может состоять в добавлении к каждому слову особых битов для проверки по четности или в периодической приостановке программы для пропуска контрольной («тестовой») задачи, ответ на которую известен.

Для еще большей надежности можно применить две полные вычислительные машины параллельно и пропускать их ответы через устройство сравнения. Во всех таких случаях при появлении какого-либо сомнения в правильности результата автоматически вырабатывается сигнал тревоги. Так как две вычислительные машины имеют в два раза большую вероятность выхода из строя, чем одна машина, и так как всегда существует возможность того, что обе вычислительные машины работают правильно, а устройство сравнения ошибается, то этот метод проверки оказывает отрицательное влияние на надежность (в первом из двух рассматриваемых смыслов), а также на стоимость. Если, однако, кто-нибудь захочет перейти к четырем вычислительным машинам — трем работающим параллельно и одной резервной, признавая правильным любой результат, совпадающий у двух машин, то у него могут быть и волки сыты и овцы целы.

20.6. Другие соображения

Многое писалось [31а, 32] и говорилось об относительной целесообразности применения аналоговых и цифровых машин для решения отдельных классов математических задач. В частности, от аналоговых машин можно получить больше интегрирования на доллар, а от цифровых — больше арифметики на доллар. Однако нас интересует в первую очередь не это. Там, где решение, принимаемое управляющим органом системы, является более логическим решением, чем вычислением, обычно более целесообразна цифровая техника. Это особенно справедливо, когда входы системы дискретны и относительно каждого входа должно быть принято отдельное решение. Пример системы такого рода приведен в гл. 22, где описана машина, играющая в игру «ним». Ясно, что эта машина должна быть цифровой.

Нужно помнить, что хотя универсальная вычислительная машина может решить любую разрешимую математическую задачу, но она не может решить никакой системной задачи без входного и выходного оборудования, а последнее всегда является специализированным. Трудности проектирования и реализации входных и выходных устройств особенно велики в том случае, когда вычислительная машина является цифровой, а основные входы и выходы — аналоговыми. В таких случаях, если только требования к системе не являются действительно крайне сложными, аналоговая система часто оказывается более удовлетворительной и более дешевой, чем цифровая.

Там, где данные должны передаваться на значительные расстояния, цифровые методы обычно оказываются более эффективными. Правда, телефон — аналоговое устройство, но та же самая информация (минус узнаваемость голоса, интонация и т. п.) может быть передана столь же быстро по гораздо более узкому и шумному каналу при помощи телетайпа (цифровое устройство). Абсолютные значения аналоговых величин, таких, как напряжение, не могут передаваться точно на большие расстояния, в то время как кодированные цифровые данные могут передаваться в форме импульсов на сотни миль с весьма большой скоростью и практически без потери информации.

Человек, по крайней мере по своему входу и выходу, является натурально аналоговой системой, и поэтому машина, с которой он имеет дело, также должна быть в большинстве случаев аналоговой. Конечно, возможна ситуация, когда оператор связан с машиной по принципу «да — нет», т. е. цифровой зависимостью, но в этом случае полоса частот оператора очень узка. Возьмем, например, экран радиолокатора, на котором световое пятно указывает положение самолета. Предположим, что оператор должен оценить положение светового пятна на экране в координатах x и y с точностью до 1/100 диаметра экрана.

По аналоговому методу эта операция выполняется путем установки на световом пятне двух взаимно перпендикулярных визирных нитей. Оператор может сделать это примерно за 3 сек. По цифровому методу нужно было бы предъявлять оператору одно за другим 10 000 квадратиков сетки экрана, с тем чтобы он сказал «да», когда появится квадратик, в котором имеется световое пятно. Очевидно, что эта операция займет очень большое время. Этот, на первый взгляд явно неэффективный, метод работы может, однако, выполняться автоматическим цифровым устройством, и автоматическое устройство может определять положения многих световых пятен на экране за небольшую долю секунды. Вот пример, показывающий целесообразность перехода к цифровой технике с ростом степени автоматизации системы.

Этот пример показывает нам также систему, где вход и выход должны быть аналоговыми (из-за присутствия человека), но где данные затем должны быть преобразованы в цифровую форму (для передачи). Если между этими двумя стадиями должны производиться вычисления, то выбор между аналоговой и цифровой вычислительной машиной становится реальным вопросом. Общего ответа на этот вопрос не существует. За исключением тех случаев, когда один метод явно лучше другого, необходимо всегда произвести предварительное проектирование по обоим методам, чтобы получить грубую сравнительную оценку расходов.

ЧАСТЬ 5

ВНУТРЕННЕЕ ПРОЕКТИРОВАНИЕ СИСТЕМ

ГЛАВА 21

РЕШЕНИЕ ЗАДАЧИ — ЭТАПЫ И ОРУДИЯ

В гл. 3 мы установили две стороны проектирования систем: внешнее проектирование, имеющее целью формулировку задачи системы, и внутреннее проектирование, имеющее целью ее решение. Рассматривая внешнее проектирование систем в ч. 3, мы не могли указать достаточные условия для формулировки задачи. В дополнение к необходимым условиям мы рассмотрели этапы формулировки и применяемые при этом орудия и привели несколько примеров неправильных и правильных формулировок. Точно так же в ч. 5, касающейся внутреннего проектирования систем, мы не в состоянии полностью описать, как решить задачу. Мы рассматриваем этапы решения и применяемые при этом орудия; мы рассказываем об имеющемся опыте, даем примеры, выводим некоторые обобщения и излагаем методы, но функция действительного решения еще не решенных задач пока остается неуловимой человеческой способностью.

21.1. Входы

В § 21.2 мы классифицируем системы по типам входов *, которые должны действовать на них. Но перед этим мы хотим перечислить и классифицировать входы нескольких типичных систем и отметить некоторые важные входные характеристики.

В табл. 21.1 приведены входы и некоторые входные характеристики и характеристики окружения для десяти систем. Мы определяем входы и окружение системы как такие действующие на систему элементы, которыми проектировщик системы не может управлять. Мы различаем входные характеристики, изменяющие нагрузку системы (и потому имеющие особое значение при проектировании большой нагрузки), и характеристики окружения, влияющие на работу системы (и потому имеющие особое значение при проектировании единичной нити). Различия эти не очень резки, но тем не менее их наличие полезно. Такие классификации не принесут вреда, пока мы не будем говорить: то-то и то-то было определено как характеристика окружения и поэтому не может влиять на нагрузку.

Следует заметить, что *условия погоды* (метеорологические условия) иногда указываются как входная характеристика, иногда как характеристика окружения, а иногда как и то и другое. Так, в транспортной системе дождливая погода оказывает очевидное влияние на окружение (уменьшая видимость, делая дороги скользкими), но она влияет также на нагрузку, как мы отметили в случае транспортной системы г. Денвера, когда погода вызвала густой поток автомашин в город около 5 часов дня. Вероятность механической поломки автомобиля отнесена к характеристике окружения, так как поломка автомобиля не влияет на общее число автомобилей, въезжающих в город, но, подобно дереву, лежащему поперек дороги, создает препятствие движению.

Составление такого перечня входов часто составляет существенную часть первоначальной работы по изучению системы.

Стандартизация входов. С какого момента нужно начинать стандартизацию и как далеко должна заходить стандартизация входов — эти существенные вопросы следует решить на первых шагах проектирования системы. Телефония сделала большой шаг вперед, когда входы были стандартизированы в самом исходном пункте тем, что на абонентском аппарате был установлен диск для набора номера. Теперь тот же принцип применен и в дальней телефонной связи.

В системе предупреждения службы гражданской обороны существенным является во-

* См. примечание на стр. 18. — *Прим. ред.*

Таблица 21.1

Входы. Частные случаи

Система	Входы	Входные характеристики	Характеристики окружения
Автомобильный транспорт (внутригородской или междугородный)	Автомобили	Распределение во времени: а) среднее значение б) функция распределения в) суточные колебания г) недельные колебания д) сезонные колебания е) долговременная тенденция Географическое распределение: а) начальные пункты выезда б) пункты отправления в) распределение длительностей поездок Тип машины (грузовик, легковая и т. д.) Условия погоды	Местоположение города Вероятность поломки Условия погоды
Управление воздушным движением („Телеран")	Самолеты	Распределение вылетов во времени Пункты а—е, как выше Распределение прибытий во времени Пункты а—е, как выше Распределение транзитных самолетов во времени Тип (скорость, размер, электронное оборудование) Условия погоды	Местоположение аэропорта. Количество и местоположение соседних аэродромов Условия погоды
Автоматическая слепая посадка	Самолеты	Те же, что в системе управления воздушным движением, но имеют значение лишь прибытия самолетов.	Окружающие строения
Координатная система № 5	Разговоры (вызовы)	Распределение во времени: Пункты а—е, как выше Распределение длительностей разговоров Тип (внутристанционный, междустанционный, междугородный) Привычки абонентов Чрезвычайные происшествия и случаи паники	Плотность населения: Степень автоматизации ближайших станций
Автоматический ремонт	Неисправности реле	Частота неисправностей	Нет (управляемое окружение)
Автоматический учет разговоров	Платные разговоры	Распределение во времени Распределение длительностей Число разговоров с разовой оплатой	Законы (федеральные, штата и т. д.)
Автоматический завод	Сырье	Нет (управляемая нагрузка)	Физические свойства входных факторов Изменения характеристик входных факторов.
Деловая система	Продажи	Распределение во времени Географическое распределение Тактика конкурентов	Различные — смотря по бизнесу, например: законы (страхование); местоположение складов; дефицит сырья и т. д.
Радиозонд	Давление, температура и т. д.	Предельные значения Скорость изменения во времени Скорость изменения в пространстве	Местоположение станций
Военная система	Цели	Распределение во времени Распределение в пространстве. Физические характеристики Применение контрмер Пути следования, тактика и т. д.	Национальные ресурсы: люди, материалы и т. д. Географический рельеф Среда (воздух, вода) Условия погоды и т. д.

прос: следует ли предусмотреть несколько стандартных входных сигналов (например: красный сигнал тревоги, желтый сигнал тревоги) или один стандартный входной сигнал (тревога). В военной системе необходимо стандартизировать информацию о цели (тип, количество, положение, скорость и т. д.); однако может оказаться удобным ввести очень

сложный код для логического управления в центре системы при значительно меньшей стандартизации сообщений наблюдателей с передней линии обороны. Обычно желательно допустить некоторые сообщения нестандартного вида для извещения о непредвиденных или маловероятных событиях. В телефонной системе это обеспечивается возможностью позвонить дежурным по станции.

Фирма «Юнайтед Стейтс Стил Корпорейшен» и Американская административная ассоциация предложили [2] оборудовать каждую конторскую машину, от пишущих машинок и арифмометров до телетайпов и быстродействующих вычислительных машин, дополнительным устройством для набивки и считывания перфолент, с применением кода телетайпного типа, который был бы единым для всех Соединенных Штатов. Это заманчивое предложение было разработано довольно подробно; читатель может получить дальнейшие сведения в упомянутом отчете.

21.2. Классификация систем

В этой главе мы классифицируем системы по таким признакам: 1) принадлежит ли ее вход всегда к одному или к различным типам; 2) появляется ли он периодически (или очень редко) или распределен во времени случайно и 3) стремится ли он разрушить систему или нет. Из этой классификации мы заключаем, что решение задачи проектирования большой или сложной системы распадается на три этапа, которые мы будем называть соответственно *проектированием единичной нити*, *проектированием большой нагрузки* и *состязательным проектированием*. К каждому этапу можно подходить по существу независимо от других; иначе говоря, решение вопросов, что делать со множественными входами, как поступать с входами, статистически распределенными во времени, и как изменить реакцию системы при состязательных входах, можно в значительной мере вести в разных группах проектировщиков.

Однотипные или многотипные входы. Если система всегда получает один и тот же вход, то для определения вызываемой им реакции не нужно никакого решающего механизма. Проектировщик определяет реакцию раз навсегда, и система отвечает на появление входа всегда одинаково. Наличие же нескольких типов возможных входов сразу приводит к необходимости в решающем устройстве для выбора правильного ответа. Это вносит в систему качественно новый вид сложности. Чем больше возможных типов входов, тем более слежен выбор ответа и тем сложнее система. Подход к решению этой задачи мы называем *проектированием единичной нити*.

Распределение входов во времени. Если ответ системы на любой вход короче, чем кратчайшее время между входами, то их можно обрабатывать по мере появления. Если вход может придти настолько быстро после предыдущего, что система еще отвечает на этот предыдущий вход, то может образоваться очередь (линия ожидания). Если средний промежуток времени между входами мал по сравнению со временем ответа, то образуется очередь бесконечной длины и система будет перегружена. Это положение можно исправить, ускорив ответ или введя большее число каналов. Однако большей частью входы интересующих нас систем распределены во времени вероятностным образом. В этом случае можно допустить, что ответ системы будет продолжительнее кратчайшего возможного интервала между входами (так что через некоторое время образуется очередь), и уменьшать длину очереди в какой-нибудь другой отрезок времени, когда интервал между входами станет больше.

Таким образом, мы можем выбирать одно из трех: большее число обслуживающих каналов, более быстрый ответ или промежуточное хранение. Статистическое распределение входов характерно для весьма многих систем. Эту сторону проектирования мы называем *проектированием большой массовой нагрузки*.

Состязание. В военных системах вход поступает от разумной силы, стремящейся полностью разрушить систему. В большинстве других систем также присутствует некоторый соперничающий фактор, хотя обычно он не является столь сильным и проникающим и может представлять собой просто тенденцию к несогласованности. Во всех случаях, связанных с деньгами (система учета социального страхования, система учета телефонных разговоров, система частного страхования и т. д.), можно предположить, что некоторые люди будут пытаться обмануть систему, если, по их мнению, это может принести им выгоду. Во многих промышленных системах нужно учитывать действия конкурентов. В системах автомобильного транспорта необходимо предусмотреть меры против нарушителей.

Любопытный пример состязательных факторов имел место в одной системе управления уличным движением, где для пешеходов была установлена кнопка, которой они могли включать красный свет при переходе улицы.

Как и следовало ожидать, когда окрестные юнцы обнаружили кнопку, они вызвали страшный беспорядок в уличном движении.

Отличительным свойством состязательных систем является то, что они могут действовать различным образом при одних и тех же входах и что нельзя предсказать ответ на данный вход. Как мы увидим, переменный ответ может также появляться в других случаях (см. ниже «Примеры» и § 25. 3), но случайный ответ в хорошо спроектированной системе появляется только в ответ на состязательные входы.

При проектировании всех состязательных систем нужно специально учитывать конфликтные стороны — как при выборе ответа (единичная нить), так и при расчете времени ответа (большая нагрузка). Эту сторону проектирования систем мы называем *состязательным проектированием.*

Разделение сторон проектирования. В системе «Телеран» (§ 2.1) единичная нить включает в себя данные единичного самолета (радиус поворота, скорость, торможение), требуемую длину взлетно-посадочной дорожки, необходимые средства речевой связи в диспетчерской вышке, работу телевизионного передатчика и приемника и т. д. Проблемы большой нагрузки включают необходимые интервалы между самолетами, возможность перепутывания сигналов на телевизионном экране, метод обслуживания нагрузки на диспетчерской вышке и т. д. Кодирование и разделение сигналов высоты составляют задачу проектирования единичной нити, хотя они необходимы лишь в системе с большой нагрузкой; поэтому объем нагрузки влияет на метод обслуживания единичной нити. Это служит одной из причин того, что различные этапы проектирования систем должны осуществляться совместно и между ними должна быть постоянная взаимосвязь.

В телефонной системе единичная нить включает потребление мощности телефоном, сигналы абоненту (ответ готовности станции, сигнал вызова, сигнал занятости), работу номеронабирателя и характеристики линии (полоса частот, шум, затухание). Проблемы большой нагрузки включают коммутационное и станционное оборудование. Вопрос о том, что делать с абонентами, которые набирают номер не дождавшись ответа готовности либо не начинают (или не заканчивают) набор номера после сигнала готовности, относится к состязательному проектированию независимо от того, злонамерен абонент или нет. В монетных телефонах-автоматах состязательные факторы очевидны.

В системе управляемых реактивных снарядов для противовоздушной обороны все эти стороны проектирования чрезвычайно сложны. Единичная нить включает сопровождение, предпусковые команды, двигатели, навигацию, действие взрывателя и т. д.; проблемы большой нагрузки включают тыловое снабжение, опознавание, оценку числа атакующих самолетов, износ, координацию и т. д.; состязательные факторы включают электронные контрмеры (противорадиолокационные средства) и контрконтрмеры, возможность существования слабейшего звена, уязвимого для противника, возможную тактику противника с целью избежать обнаружения или уничтожения и т. д.

Примеры. При прочих равных условиях система с несколькими типами входов сложнее системы с входами одного типа; система с входами, распределенными во времени, сложнее системы с периодическими входами; и система, противостоящая неприятелю, сложнее системы, в которой можно ожидать сотрудничества (или, самое большее, безразличия) входов. Мы приводим примеры всех сочетаний этих пар, приблизительно в порядке возрастающей сложности.

На заводе-автомате промежуток времени между появлениями входов определяется инженером-проектировщиком. Этот промежуток времени фиксирован и периодически повторяется, и этот период больше (или равен) времени, необходимого для того, чтобы закончился ответ системы на вход. Далее, в такой системе входы всегда одного типа, например сырье, применяемое при изготовлении изделия.

Застава для сбора пошлин с автомобилей, железные дороги с единой платой за проезд и электроэнергетические распределительные системы суть примеры систем, у которых имеется только один вход, т. е. от системы всегда требуется одно и то же. С другой стороны, моменты появления этих входов распределены статистически, и в системе необходимо предусмотреть возможность появления очереди. Вход в основном способствует работе системы и не направлен на ее разрушение.

В цифровой вычислительной машине или в сортировальной машине имеется много типов возможных входов, но фиксированная частота, или же имеется группа входов, разделенных большими интервалами времени, причем входные факторы не состязательны (не направлены на разрушение системы). Здесь надо предусмотреть зависимость ответа системы от входа, но обычно не требуется пропускать входы по различным каналам. Кроме

того, вход данного типа вызывает всегда один и тот же ответ.

Телефонная система представляет собой пример системы, у которой входы соединяют свойства множественности и статистического распределения во времени. Так, телефонная система в данном районе должна отвечать на любой из 10^7 возможных входов (10^{10}, если учесть дальнюю автоматическую связь). Кроме того, не известно в точности, когда именно произойдет следующий вызов на станцию. Известны лишь такие статистические характеристики, как средняя частота вызовов и распределение.

Рассматривать ли входы такой системы как состязательные или как несостязательные, зависит от точки зрения проектировщика. Мы указали состязательные явления при выписывании счетов (учете разговоров) и сборе монет в телефонных автоматах и несколько менее состязательные явления, вызванные неправильным набором номера абонентами. Как указано выше, к состязательным сторонам относится также возможность разного ответа на одинаковые входы; в телефонной системе это имеет место при дальней связи, где соединение зависит от наличия свободных линий в данный момент.

В «игровой» машине, для которой решение задано заранее (например, в машине для игры «тик-так-ту»* или для игры «ним», см. гл. 22), налицо множественность входов, которые состязательны вследствие самой природы игры. Но распределение во времени не существенно, так как игра проводится настолько медленно, что входы можно пропустить через один канал.

При поисках подводных лодок мы встречаем много возможных входов в виде целей разных типов с их распределением в пространстве. От системы требуется также состязательный ответ. Но время между входами велико сравнительно с ответом системы, т. е. в каждый момент система должна определить местоположение лишь одной цели.

Наконец, в некоторых военных системах, как система противовоздушной обороны, входы обладают всеми сложными свойствами, какие только могут быть у систем: зависимостью работы системы от входа, статистическим распределением входов и состязательностью ответов, которые должна давать система.

* Тик-так-ту (tick-tack-too) — игра в «крестики и нолики»: два игрока по очереди вписывают кресты или кружки в клетки некоторой матрицы и каждый из них старается получить ряд из трех кружков или трех крестов раньше, чем противник. — *Прим. ред.*

21.3. Единичная нить

Проектирование единичной нити системы завершается тогда, когда можно точно описать, что произойдет с каждым возможным входом на каждом этапе его прохождения через систему, или описать каждый ответ, который он вызовет в системе. При полном завершении нужно также описать все возможные выходы системы. Часто такое решение системной задачи можно описать, не указывая, что нужно сделать в том случае, когда два входа появляются одновременно или когда в одном и том же звене системы одновременно возникают два ответа, т. е. не решая полностью проблемы большой нагрузки.

Первый шаг к решению этой части задачи — начертить функциональную блок-схему. Если система уже существует, то ее функциональная блок-схема является хорошим исходным пунктом. Если системы еще нет (хотя что-нибудь похожее всегда имеется), то функциональную блок-схему нужно придумать. Чтобы вычертить эту схему, необходимо, очевидно, полностью понять на основании внешнего проектирования работу данной системы и связь каждой ее части со всей системой.

Следующий шаг — изменение функциональной блок-схемы в соответствии с возможным решением, пришедшим на ум проектировщику, и с требованиями, установленными ранее при внешнем проектировании системы. Затем эти функции осуществляются с помощью оборудования, известного бригаде проектирования системы, при содействии соответствующих специалистов (проектировщиков компонентов). В результате появляется блок-схема оборудования системы.

Следующая затем детализировка составляющих блоков должна быть достаточно подробной, чтобы в случае необходимости систему можно было анализировать посредством описанных в следующих главах орудий: системной логики, теории информации, технической психологии и т. д. Этот анализ покажет, насколько выполняются требования, поставленные перед системой. Тем самым выяснится, в каких пунктах требования не выполняются, и весь цикл повторяется.

Результаты этой части решения системной задачи таковы: а) функциональная блок-схема, приводящая к достаточно полному пониманию системы; б) блок-схема оборудования, позволяющая описать входы и выходы и способ работы каждого блока оборудования, и в) блок-схема компонентов. Успешно сделав эти вещи, проектировщик близко подойдет к составлению функционального задания, ко-

торое, как было сказано в гл. 3, составляет основной результат проектирования системы.

Функциональная блок-схема. Для материально-технического обеспечения военно-воздушных сил США требуется вести текущий учет свыше 1 200 000 типов предметов снабжения при большом количестве предметов каждого типа и обеспечить надлежащий ответ системы на заявки на материал любого типа. Надлежащий ответ означает доставку нужного предмета в достаточно короткое время. На рис. 21.1 изображена часть схемы единичной нити (см. вклейку в конце книги).

Основную роль в этой системе выполняет цифровая вычислительная машина, так что эта функциональная схема представляет собой по существу блок-схему работы вычислительной машины при поступлении заявок. Однако функциональные блок-схемы выглядят одинаково — независимо от того, происходит ли работа во многих местах системы или в одном месте, как показано на рисунке. Этот пример ясно показывает сходство между блок-схемой программы для вычислительной машины и функциональной блок-схемой системы.

Блок-схема оборудования. После того как функции определены, нужно указать, как они осуществляются реальным оборудованием. На рис. 2.3 и 2.4 представлены типичные примеры такого осуществления для координатной системы № 5 и связанной с ней системы автоматического учета разговоров. Такие схемы должны быть простыми и ясными. В них не обязательно указывать все мелкие соединения. Основное назначение блок-схемы оборудования — наглядно представить поток информации через систему, чтобы можно было составить четкое, логически последовательное описание входов и выходов и способа действия системы на входы. Там, где участвуют люди, иногда желательно указать их присутствие в системе, используя разные обозначения для линий, соединяющих между собой оборудование, и для линий, соединяющих людей с оборудованием (инженеры-психологи обозначают в таких блок-схемах оборудование квадратами, а людей — кружками).

Блок-схема компонентов. Каждой единице оборудования будет соответствовать весьма полное описание нужного для него компонента. Типичная схема с таким описанием представлена на рис. 2.2, где изображена система «Резервизор» вместе с запоминающим устройством на магнитном барабане и соответствующими цепями. Здесь дается как поток информации, так и наименования устройств. Указана также связь с другими единицами оборудования. Эта схема дает примерно ту степень деталировки, какая необходима для составления функционального задания на систему. Конечно, она недостаточна для изготовления самого элемента, но дает прекрасный исходный пункт для проектировщика компонентов.

Анализ и испытание предложенного компонента лучше всего предоставить группам оборудования, но составление задания для такого анализа и испытания является по существу делом проектировщика-системника.

Выбор подсистем. При решении единичной нити нужно указать различные подсистемы, которые будут составлять более широкую систему. Выбор таких подсистем зависит от многих факторов, часть из которых здесь указана. Один очевидный фактор — желательность размещения подсистемы в одном географическом пункте. Другой очевидный фактор — требование, чтобы подсистема имела возможно меньше входов и выходов. Особенно важно предусмотреть, чтобы соединительные звенья, пересекаемые границами подсистемы, не оказались теми, которые вызывают наибольшие нарушения при соединениях между блоками.

Можно указать хорошее практическое правило при выборе подсистемы: группа оборудования должна быть в состоянии проводить проектирование без больших обсуждений в течение примерно 3 месяцев, не отклоняясь слишком далеко от основных линий построения системы. Если группе оборудования требуются направляющие указания через более короткие промежутки, то это указывает, что в качестве подсистемы был взят слишком малый блок, требующий слишком частой координации, или что к проектированию подсистемы приступили до начала фазы основного проектирования (гл. 3). Ведь эти 3 месяца приблизительно соответствуют периодическому оглашению новых вариантов функционального задания. Тогда замкнутая петля отчетности от проектировщика компонентов к группе системы и обратно через функциональные задания синхронизирована более или менее правильно.

В телефонной системе нельзя найти границу подсистемы, которая не пересекала бы многих соединительных звеньев, ибо телефонная система сделана из столь многих соединительных звеньев! Однако было бы плохо выбрать в качестве подсистемы маркер, так как в этом случае граница пересекала бы все критические соединительные звенья. Зато телефонная станция образует идеальную подсистему. Она находится географически в од-

ном месте, и число линий между станциями и особенно число их типов сравнительно малы по сравнению с линиями внутри станции.

Мы уже говорили, что такие вещи, как автоматический учет разговоров, являются подсистемами. Но удобнее считать подсистемой лишь центр учета разговоров (рис. 2.4,*б*) и рассматривать те элементы учета разговоров, которые относятся к центральной станции (рис. 2.4,*а*), как дополнительные требования к подсистеме координатной системы № 5.

В цифровой вычислительной машине, как уже было сказано, подсистемами являются устройство управления, арифметическое устройство, память и устройства входа-выхода. Последние обычно объединяются, потому что они имеют сходные задачи и часто в них применяется одно и то же оборудование.

В большинстве воздушных бортовых систем («Телеран», «Ника») при первоначальном делении на подсистемы желательно отделить бортовые элементы от наземных. Каждую из этих подсистем можно делить дальше. В управляемом реактивном снаряде корпус и двигатель проектируется отдельно от аппаратуры наведения. В системе «Телеран» телевизионный приемник проектируется отдельно от радиолокационного ответчика; радиолокационная станция, телевизионный передатчик и пункт управления образуют отдельные подсистемы. Конечно, эти подсистемы отнюдь не независимы одна от другой, но после решений об их взаимных соединениях, хотя эти решения и пересматриваются по мере необходимости в порядке периодического согласования, разработку подсистем можно проводить более или менее независимо.

21.4. Большая нагрузка

Проблема большой нагрузки связана с одновременным обслуживанием многих входов или с большим числом ответов на входы. Так, мы можем ясно представлять себе, как посылать речевые сообщения по проводам, но задача одновременного удовлетворения большого числа требований на обслуживание вызывает осложнения нового вида. В военных системах мы можем ясно представлять, как атаковать цель одним оружием. Но атака нескольких целей оружием нескольких разных классов представляет собой чрезвычайно трудную задачу. Вполне понятно, как вести автомобиль между двумя пунктами, но решение задачи пропускания тысяч машин через лабиринт городских улиц — это совсем другая и чрезвычайно трудная проблема. Необходимость решать эти задачи и приводит к изучению вопросов большой нагрузки в системе большого масштаба. Напомним, что система большого масштаба характеризуется именно множественностью, многократностью.

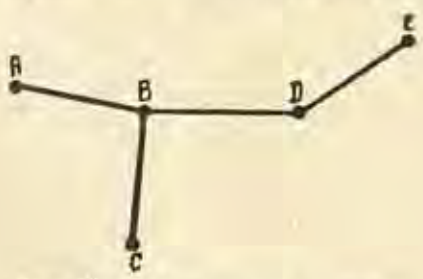


Рис. 21.2. Упрощенная географическая блок-схема.

Подход к проблеме большой нагрузки подобен подходу к проблеме единичной нити. Начинают с функциональной схемы, составленной на основе изучения старой системы в процессе внешнего проектирования. Предполагается, что виды оборудования выбраны и нужно решить, где будет выполняться каждая функция. «Где» — это следует понимать в смысле географического местоположения, физического размещения и функции оборудования, как описано ниже. Решение проблемы большой нагрузки приводит к географической блок-схеме, схеме физического размещения аппаратуры и множественной функциональной схеме. Орудиями анализа служат теория массового обслуживания (гл. 23) и моделирование (гл. 26), позволяющие оценить сделанный выбор. После анализа будет ясно, какие изменения еще необходимы в приготовленных схемах и, может быть, в проекте единичной нити, и цикл повторяется.

Географическая блок-схема. В § 2.3 мы спрашивали: выбирать ли стационарную или подвижную установку? посылать ли информацию во всех направлениях и предоставить ей действовать на людей (циркулярное управление), или адресовать информацию особо для каждого участника (коллективная линия), или предусмотреть отдельный канал связи для каждого участника (индивидуальная линия)? Ответы на эти вопросы воплощаются в географической блок-схеме.

На рис. 21.2 мы приводим весьма упрощенную географическую блок-схему телефонных кабелей, соединяющих пять городов. Легко показать, что для соединения между собой n городов необходимо и достаточно $n-1$ кабелей. Если мы намерены использовать именно это минимальное число кабелей, то при помощи линейного программирования (§ 25.1) можно определить, какая из комбинаций дает минимальную общую длину кабеля.

Однако на практике потребуется большее число кабелей, если учесть экономические факторы, необходимость разделения разных маршрутов и т. д. Нужно будет не только ввести несколько каналов для каждого звена, но , вероятно, нужно будет добавить некоторые (но не все) из возможных диагональных звеньев (как $C—D$, $C—E$ и т. д.). Если ввести все возможные диагонали, то общее число соединительных звеньев будет равно $C_2^n = \frac{n(n-1)}{2}$. Географическая блок-схема как раз и показывает, какие узлы мы соединяем.

Если мы начертим блок-схему единичной нити для системы радиозонда, то нам нужно будет посылать данные телеизмерительными средствами на землю, декодировать их там и, наконец, обобщать данные от отдельных станций в вычислительной машине. Ясно, что на каждой станции должен быть телеизмерительный приемник и должны быть линии связи с центральной вычислительной машиной, но не ясно, где декодировать данные: на отдельных станциях или на центральной станции.

Другими словами, один из основных вопросов при проектировании системы такой: передавать ли по каналу связи сырые телеизмерительные данные или проводить некоторую их обработку на отдельных станциях? Выбирая ту или другую из этих альтернатив, получим две системы совершенно разного вида. Решение этого вопроса тесно связано с другими вопросами проектирования системы: будут ли телеизмерительная линия, линия связи и вычислительное устройство непрерывные или дискретные? если они разного типа, то где производить преобразование данных? какая разрядность на каждой ступени? сколько нужно станций? Ясно, что решение этих вопросов, составляющее существо проектирования системы, совсем не простое дело.

Очевидно также, что от решения этих вопросов зависит как проектирование большой нагрузки, так и проектирование единичной нити, хотя их влияние на последнее носит вторичный характер. Во всяком случае, географическую блок-схему нельзя составить без решения этих вопросов. Попытка начертить географическую блок-схему не только укажет на необходимость их решения, но может также пролить некоторый свет на способы решения.

Схема физического размещения. Такая схема должна подробно указывать, как нужно расположить оборудование в центре управления, или в подсистеме, или на оперативной базе. На рис. 21.3 показан пример такой схемы для центра управления системы «Телеран». Художественное оформление не обязательно. В § 30.7 кратко разобраны методы, позволяющие определить размещение аппаратуры.

Множественная функциональная схема. На рис. 21.2 изображена часть схемы единичной нити для предложенной системы материально-технического обеспечения военно-воздушных сил, где производится обслуживание заявок. Но система должна иметь дело со многими типами информации и материалов: инвентарями, отчетностью, самолетными деталями, транспортными средствами и т. д. В проекте должно быть учтено, на каком уровне должна выполняться каждая функция, и должны быть предусмотрены соединения между всеми элементами системы, дубли-

Рис. 21.3. Физическое размещение центра управления системы „Телеран" (по Юингу и др. [4]).

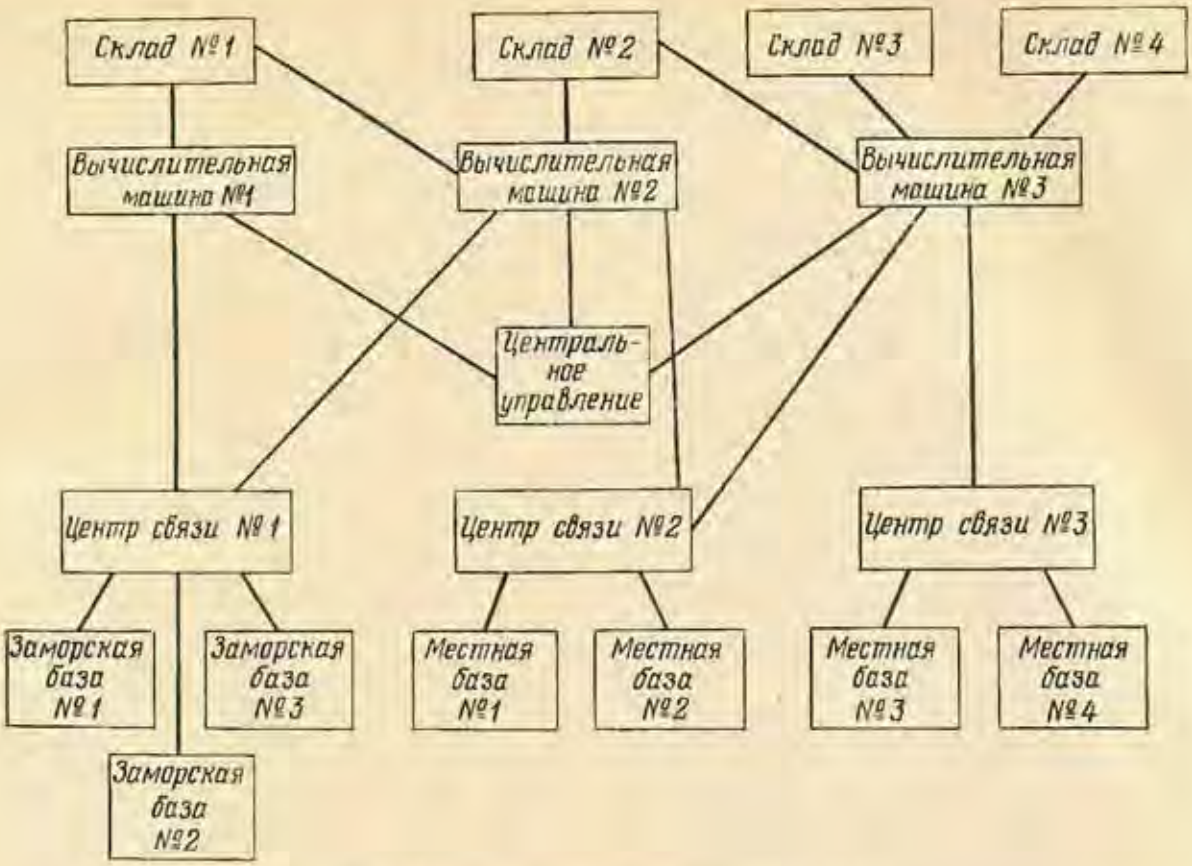


Рис. 21.4. Множественная функциональная схема.

рующими каждую функцию. На рис. 21.4 показана часть множественной функциональной схемы, какая может при этом получиться.

Центральными при проектировании большой нагрузки являются следующие задачи: сколько ставить блоков каждого типа (таких, как на рис. 21.4)? как часто при статистически меняющихся входах это количество окажется недостаточным? к каким последствиям это приведет (в смысле задержки входов, плохого их обслуживания или полного перерыва в обслуживании)? Как было указано выше, существуют три метода приспособления к распределению входов во времени: ускорение ответа системы, введение добавочных каналов и промежуточное (буферное) хранение. Во всякой большой системе применяется какое-нибудь сочетание этих методов. Скорость ответа системы определяется при проектировании единичной нити. Число каналов и емкость и тип промежуточных хранилищ определяются при проектировании большой нагрузки.

Сравнение буферного хранения и добавочных каналов. В аэропортах добавочными каналами служат дополнительные взлетно-посадочные дорожки; более эффективные методы посадки могут привести к увеличению числа посадок в единицу времени; буферное хранение выражается в том, что самолет кружит над аэродромом. Неудобства буферного хранения выражаются здесь в недовольстве пассажиров ожиданием и в некотором увеличении расходов на горючее; предел буферного хранения определяется возможностью того, что самолет израсходует горючее до посадки.

В телефонной системе добавочные каналы могут потребоваться на любой ступени: в абонентских регистрах, линейных искателях, маркерах; буферное хранение выражается в том, что абонента заставляют ждать. Ожидание приводит к недовольству абонентов и отказу от вызовов; кроме того, если сигнал готовности станции задерживается, абонент может все равно начать набор номера и в итоге получить не тот номер; это плохо в любом случае и особенно плохо в часы наибольшей загрузки.

В коммерческой системе увеличенный спрос можно удовлетворить пустив в работу запасные машины для производства товаров (добавочные каналы) или при помощи больших складов товаров (буферное хранение). В системах связи и в вычислительных установках нужно иметь буферные устройства хранения, где можно сохранять подлежащую обработке информацию. В транспортных си-

стемах, системах резервирования и т. п. буфер — это просто очередь. Во всех этих случаях аналитическим орудием для определения величины буферного хранилища служит теория массового обслуживания.

21.5. Состязательность

Существо состязательного проектирования состоит в том, чтобы «рандомизировать» ответ — сделать его случайным *. Рассмотрим какую-либо правительственную систему обработки данных, например систему подоходного налога. Проверять каждую налоговую декларацию было бы слишком дорого и ненужно. Но некоторые декларации, во избежание обмана, необходимо проверять. Число деклараций, которые нужно проверять, сравнительно невелико: оно должно лишь заставить налогоплательщиков бояться обманывать.

Допустим, правительство решает проверять ежегодно 5% деклараций. Если предположить, что всякий обман будет наверняка обнаружен при проверке и никак не обнаружится иначе, и оставить в стороне всякие моральные соображения, то тот, кто обманывает, имеет один шанс из двадцати быть пойманным. По-видимому, штраф за обман нужно сделать достаточно большим, чтобы выгода от обмана не оправдывала этого риска. Но 5% деклараций нужно выбирать каждый год случайно. Если их выбирать по плану, то кто-нибудь сможет изучить этот план и будет знать, что его декларация в этом году не будет проверяться. Если же само правительство не знает, какие декларации оно будет проверять, то, очевидно, и никто не сможет это узнать.

В действительности некоторые, вероятно, будут обманывать в любом случае, и число обманывающих будет возрастать с уменьшением процента проверяемых деклараций. Процент деклараций, проверяемых правительством, должен быть таким, чтобы ожидаемая от этого выгода была наибольшей. Дополнительные проверки требуют дополнительных расходов, но приводят к увеличению доходов по двум причинам: пойманных станет больше и возрастет число отказавшихся от обмана. Пока предельное увеличение доходов превосходит предельное увеличение расходов, нужно увеличивать процент проверяемых деклараций.

В более сложных случаях не всегда легко определить, как именно можно получить наибольшую ожидаемую выгоду и как сделать действия случайными. Соответствующим аналитическим орудием является теория игр, разбираемая в гл. 24.

21.6. Некоторые принципы системного проектирования

Хотя проектирование систем не есть некая твердо установленная наука, тем не менее можно сделать некоторые обобщения, заслуживающие названия принципов. Они имеют различную важность и допускают исключения, но все же их стоит изложить, потому что задача, кажущаяся чрезвычайно трудной, нередко значительно проясняется, если исследовать ее в свете надлежащих общих положений.

Основной принцип проектирования систем. Основной принцип можно изложить очень кратко так: «Добиваться, в некотором смысле, максимальной ожидаемой выгоды». Пусть, например, ожидаемая выгода определяется уравнением (5.3); тогда мы должны рассчитать для каждой альтернативы вероятность каждого возможного исхода и получающийся при этом платеж. Перемножая их и суммируя произведения по всем возможным исходам, получим ожидаемую выгоду альтернативы. После этого выбирается альтернатива с наибольшей ожидаемой выгодой **.

Примеры этого мы видели по всей книге. В предшествующем параграфе мы определили этим методом одно из явлений состязательности. Ранее мы рекомендовали определять ошибки I и II рода максимизацией минимального ожидаемого выигрыша. По существу это правило входит почти во всякое решение, которые мы принимаем при проектировании системы.

Маловероятные события. Почти во всякой системе возникает вопрос: как учитывать входы, появляющиеся сравнительно редко? Ответить на него можно при помощи изложенного выше основного принципа, но вопрос заслуживает особого разбора, так как часто на него склонны давать совершенно другие ответы.

Пусть, например, проектируется система обработки данных для военных применений. Возникает вопрос: какие типы информации нужно хранить? или, точнее, можно ли не

---

* По-английски random — «случай». — *Прим. ред.*

** Мы не рассматривали трудные и важные проблемы *теории статистических решений* (см. например, [141]), с которыми связано получение наибольшей ожидаемой выгоды при неполной информации. Трудность состоит, очевидно, в том, чтобы найти надлежащие численные значения вероятностей и платежей. Мы здесь не рассматриваем критерии максимизации наибольшей ожидаемой выгоды, но в гл. 25 мы разбираем один из таких критериев, а именно минимаксный критерий. — *Прим. авт.*

учитывать такие-то события, наступающие очень редко? На этот вопрос почти всегда отвечают: «Храните все». Обоснование такой непреклонности похоже на старую сказку: «Не было гвоздя — потерялась подкова, потерялась подкова — пропал конь, пропал конь — пропал всадник, пропал всадник — погибло царство». Ошибка этого рассуждения, очевидно, в том, что вероятность потери подковы из-за отсутствия одного гвоздя очень мала; вероятность пропажи коня и всадника из-за потери одной подковы очень мала; и вероятность гибели царства из-за пропажи одного солдата очень мала. Вероятность всей цепи событий равна произведению этих отдельных вероятностей и, разумеется, ничтожна.

Часто при проектировании большой автоматической системы склонны забывать, что прежняя ручная система, которую она заменяет, не могла бы обслужить ни обычный вход, ни необычный; поэтому требовать от новой системы, чтобы она обслужила и то и другое, значит требовать слишком многого.

Предположим, что мы создаем автоматическую систему для регистрации, сортировки и опознания отпечатков пальцев. Нам предлагают, чтобы мы давали на каждой карточке полное описание человека, закодированное (при помощи перфорации) так, чтобы карточки можно было сортировать автоматически по любому ключу. В обоснование этого говорят: «Может случиться, что относительно какого-нибудь опасного преступника имеется лишь показание свидетеля, что у него рыжие волосы, зеленые глаза и шрам на левой щеке. Без предлагаемой системы кодирования мы, возможно, никогда его не найдем». Действительно, может оказаться желательным создать систему перфокарт для сортировки по ключу такого рода. Это будет определяться сравнительными стоимостями — создания автоматической системы, выполнения работы вручную и отказа от ее выполнения вообще. Суть дела в том, что последнюю альтернативу нельзя просто отбросить. До того как была установлена система, такой преступник не был бы пойман; даже после того, как система установлена, не все преступники будут пойманы немедленно. И мы не должны подвергать опасности главную цель (составление картотеки отпечатков пальцев), делая излишний упор на побочной системе, ценность которой может быть в действительности очень мала.

Меры, предусматривающие все возможные случаи, приводят к колоссальному удорожанию системы. Телефонные компании допускают несколько напряженных дней, когда телефонная станция будет перегружена. Теория указывает, что такая система обработает должным образом почти все вызовы и что стоимость кратковременной перегрузки (в виде недовольства абонентов, вмешательства правительства и т. д.) мала сравнительно со стоимостью обслуживания пиковой нагрузки.

Можно предусмотреть участие людей в тех случаях, когда задача превосходит возможности автоматической системы. Так, в канадской системе почтовой связи (§ 2.4) все сомнительные письма обрабатываются вручную. Общее количество обрабатываемых таким образом писем составляет лишь небольшую долю общего количества писем, но экономия в аппаратуре по сравнению с полностью автоматической системой получается очень большая. Обычно можно построить систему так, что при наступлении маловероятного события появляется сигнал о передаче функций человеку. Такое решение предусмотрено в телефонной системе в виде полуавтоматической регулировки линейной нагрузки. В системах учета наличия и движения материалов, как упомянутая выше система материально-технического обеспечения военно-воздушных сил, вычислительная машина может быть снабжена инструкцией дать сигнал тревоги при возникновении непредусмотренной ситуации. Тогда человек может решить задачу на бумаге, как он делал это до установки системы.

В военной (или любой другой состязательной) ситуации, прежде чем отбросить некоторое событие как имеющее ничтожно малую вероятность, нужно убедиться, что его вероятность не во власти противника. Если у нас имеются три радиолокатора, каждый с надежностью 99%, то вероятность того, что все три одновременно откажут, равна 10^{-6}, т. е. ею можно пренебречь. Если противник располагает противорадиолокационными средствами и наши радиолокаторы не имеют достаточной защиты от этих средств, то вероятность отказа каждого радиолокатора близка к 1; кроме того, вероятности отказа отдельных радиолокаторов уже не будут независимы, так что если один не работает, то и другие почти наверное тоже не работают.

Информационные и материальные системы. Нужно различать поток информации (число проданных мест в системе «Резервизор», вызываемый телефонный номер в телефонной системе, положение самолета в системе управления воздушным движением) и поток материала (электронные приборы на автоматическом заводе, автомобили

в системе регулирования уличного движения, самолеты в системе управления воздушным движением). Мы предполагаем здесь, что проектирование системы можно полностью провести на основе информационной стороны системы и что впоследствии можно приспособить систему для потока материала, причем не потребуется очень больших изменений в первоначально намеченной информационной системе.

В системе управления воздушным движением главная задача проектировщика системы — не столько управление положением самолета, сколько отыскание того места, где он находится. Мы различаем в любой момент для каждого самолета два положения. Одно положение — это та точка пространства, в которой самолет находится по данным системы; второе положение — точка пространства, в которой самолет действительно находится. Согласно нашему предположению все проектирование системы можно провести по положению самолета, определяемому системой. Так, точки, представляющие положение самолета по данным системы, могут понадобиться для поддержания определенных интервалов, для регулировки скоростей на данные условия, для выполнения поворотов, которые требуются для опознавания, и т. д.

После того как такие точки были выбраны согласно желанию проектировщика системы, может возникнуть вопрос, с какой точностью самолет может «держать место». Из-за навигационных ошибок, ошибок в системе связи и неточностей в управлении самолетом появляются расхождения между действительным положением самолета и данными системы о его положении. Взаимосвязь между материальной системой и информационной системой можно рассматривать как небольшую задачу по теории следящих систем, где регулирующим положением является положение, запоминаемое системой, а следящим механизмом является самолет, стремящийся сохранить это положение. Если ошибки слишком велики, то может потребоваться разработка новых навигационных приборов или средств связи. Наше предположение, кроме того, устанавливает, что необходимость в изменении проекта информационной системы будет возникать относительно редко.

В этом смысле всякая большая система есть информационная система. При проектировании автоматического завода, системы автомобильного движения, любой военной системы основное внимание должно быть обращено на поток информации через элементы системы.

Обратная связь. В этой книге мы неоднократно подчеркивали важность обратной связи. Этот вопрос рассматривается далее в гл. 29 и упоминается в § 25.3 как один из принципов проектирования систем.

Оптимум групповой и оптимум локальный. Необходимость в системном подходе при проектировании возникает потому, что субоптимизация (оптимизация компонента или части системы) не обязательно улучшает работу системы. Оптимизация всех элементов также не обеспечивает достижения общего оптимума. Например, инженеры-транспортники заметили, что если поток через некоторые перекрестки увеличивается, то получающийся вследствие этого приток машин в систему вызывает затор на других перекрестках, который в конце концов отражается на первых перекрестках и совершенно останавливает поток. В этом случае нужно замедлить поток машин, проходящих через упомянутые перекрестки. Проектирование системы есть процесс достижения группового оптимума.

Сетевые системы. Большинство систем содержит какую-либо сеть. В случае систем связи, транспортных систем и систем распределения энергии вся система основана на сети. Но и во многих других системах имеются сети, как это видно по диаграммам такого типа, как на рис. 21.4. В этой диаграмме надо выбирать для обслуживания данной ситуации одно из нескольких вычислительных устройств.

По поводу сетей разработана довольно большая математическая теория. Здесь мы просто укажем, что в общем имеются два способа изучения движения через сеть: а) найти точный путь каждого входа через систему и б) исследовать лишь пересечения. Второй способ гораздо легче и часто является достаточным. Так, в системе моделирования уличного движения Калифорнийского университета (§ 10.3) «индивидуальность» автомобиля теряется, как только он входит в систему, но процесс, развертывающийся на каждом перекрестке, остается таким же, как если бы все эти индивидуальности сохранялись.

Централизация и децентрализация. В большинстве систем большого масштаба возникает вопрос, в какой мере система должна быть централизована. Нужно различать централизацию материальной системы и централизацию информационной системы. Примером первой может служить телефонная система, где важное значение имеет выбор размера телефонной станции. Крайние возможные случаи — вырожденный случай полной централизации (одна станция для всей си-

стемы, собирающая линии всех абонентов) и вырожденный случай полной децентрализации (каждый абонент есть телефонная станция и должен быть соединен проводной линией с каждым другим абонентом). В действительности применяется компромиссное решение: телефонные станции примерно на 10 000 абонентов в больших городах и, конечно, несколько меньшие станции в сельских местностях.

В соответствии с нашим предположением, что все системы суть информационные системы, более существенным вопросом является централизация информации. В централизованной системе информация о входах, окружении и состоянии системы притекает к некоторому центральному штабу, где принимаются все решения; штаб посылает приказы о соответствующем воздействии на входы. В децентрализованной системе каждая низшая ступень имеет право воздействовать на проходящий к ней вход; затем она посылает сообщение высшим ступеням, которые сохраняют лишь право вето. Предполагается, что низшая ступень будет обслуживать все обычные входы. Необычные ситуации докладываются высшему органу: чем важнее решение, тем более высокая ступень отвечает за решение. Если вход необычен и вместе с тем требует неотложного решения, то низшая ступень должна взвесить стоимость неправильного решения и сравнить со стоимостью задержки действия. Так как низшая ступень может взвешивать эти стоимости неправильно или эти стоимости могут оказаться чрезмерными, в обоих случаях централизованные системы в этом отношении имеют явные преимущества.

Следует отметить, что реальное различие между централизованной системой и децентрализованной заключается прежде всего в скорости сообщений между низшими и высшими ступенями. Даже в крайне слабо организованной системе предполагается, что центральный штаб сохраняет некоторый контроль над местными операциями, основывая свои решения на сообщениях, получаемых им за большие промежутки времени. Итак, способность централизованной системы распознавать ситуацию во всей системе при определении оптимального ответа на данный вход означает реально способность сделать это за короткое время.

Недостатки централизованного управления могут быть четырех видов: система может быть нечуткой к входам или отвечать на них недостаточно быстро; информация может ослабляться при прохождении через промежуточные ступени, что приводит к ошибкам; центральный орган управления может оказаться не в состоянии справиться с огромным количеством информации, собранной большой системой, и может выйти из строя вследствие перегрузки; система является негибкой, так как низшие ступени становятся беспомощными, если разрываются их связи с высшим органом управления. Однако в настоящее время имеются способы устранения этих недостатков при сохранении всех преимуществ централизованной системы.

Современные средства связи и устройства обработки данных позволяют работать с централизованной системой по крайней мере столь же быстро, как и с децентрализованной. В системе «Резервизор» кассир может запросить магнитный барабан, находящийся в аэропорте Ла Гуардия *, и получить ответ с расстояния во много миль значительно быстрее, чем если бы он прочитал цифры в книге, лежащей перед ним на столе. Хотя автоматические системы иногда бывают негибкими и нечуткими, это не обязательно должно быть так. Если предусмотрен сигнал тревоги и ручное вмешательство при появлении необычных входов, то система может сохранить гибкость, характеризующую поведение человека.

Ослабления информации можно избежать двумя способами: использовать надежные средства связи с применением таких кодов, как импульсная модуляция, и применять множественные адреса, при которых информация, проходящая по ступеням, поступает к ближайшему контрольному органу и одновременно к центральному штабу. Таким образом, каждая ступень контроля сохраняет право вето над своей организацией, как и должно быть, а высшие ступени получают полные, быстрые и точные сообщения о работе всей системы.

Конечно, такая система множественных адресов приводит к тому, что высший штаб переполняется еще большим количеством информации, чем он получает теперь, а оно и так больше того, что он может переварить. Но в наши дни уже появляется аппаратура, способная справиться с этим. Центральный штаб может получать и эффективно обрабатывать на несколько порядков величины больше информации, чем это было возможно лет десять назад. Если логика системы хорошо разработана, то простое дублирование

---

* Ла Гуардия — крупный аэропорт (227 *га*) в восточной части Нью-Йорка (район Квинс на о. Лонг-Айленд), открыт в 1939 г. — *Прим. ред.*

входов не должно привести к неразрешимым задачам при обработке данных.

Чтобы учесть уязвимость системы по отношению к неисправностям высших ступеней или линий связи, можно сохранить зачаточную форму управления на низших ступенях, которые должны быть обучены принимать на себя управление, но нормально не управляют. Тогда при появлении неисправности центр зачаточного управления на каждой ступени принимает управление на себя. Он действует несколько менее эффективно, чем полностью централизованная система, но если проектирование было проведено должным образом, зачаточное управление не менее эффективно, чем если бы система была децентрализованной с самого начала. Другими словами, эта система управления не будет менее эффективной от того, что она вводится при проектировании в централизованную систему.

Во всех видах человеческой деятельности были разработаны ступени ответственности и единицы различной величины. В военном деле иерархия корпуса, дивизии, полка и т. д. оправдала себя в огне многочисленных сражений. Завод в промышленности, автономное отделение в торговле выросли до определенного размера, дающего наибольшую эффективность. Но при наличии новых средств в виде вычислительных машин, аппаратуры связи, аппаратуры индикации и теории статистических решений представляется целесообразным поставить под вопрос выбор размеров во всех этих операциях, хотя до сих пор они оказывались удовлетворительными. При проектировании больших систем сильная централизация обычно дает преимущества.

ГЛАВА 22

ЕДИНИЧНАЯ НИТЬ. СИСТЕМНАЯ ЛОГИКА

Мы неоднократно упоминали понятие единичной нити через систему большого масштаба. В гл. 21 мы утверждали, что во всякой большой системе, какой бы сложной она ни была, можно выделить то, что мы называем *единичной нитью*, т. е. способ действия системы на один вход (каждого типа). Эта единичная нить представляет собой, однако, не способ действия системы на изолированный вход, а нить, проходящую через систему с большой нагрузкой, с состязательными аспектами. Так, в гл. 2 мы упомянули, что была построена единичная нить системы «Телеран» и что она содержала ряд компонентов (например, аппаратуру обработки данных), необходимых только в системе с большой нагрузкой. Единичная нить связана с другими аспектами системы, но эта связь слабая. В гл. 3 мы включили единичную нить в надлежащую хронологическую перспективу как первый аспект, с которого надо начинать внутреннее проектирование системы, хотя единичную нить нельзя полностью решить, пока не уточнены нагрузочные и состязательные аспекты. Так, буферное хранение, необходимое при большой нагрузке, будет видоизменять единичную нить, но в большинстве случаев не окажет на нее существенного влияния.

До полного выяснения единичной нити проектировщику системы придется глубоко вникать в рассмотрение компонентов системы. Однако первым шагом является составление логических правил работы системы. Это выполняется методом, весьма родственным программированию задачи для цифровой вычислительной машины и составляющим предмет этой главы. На рис. 21.1, представляющем небольшую часть логической схемы для системы материально-технического обеспечения ВВС, показан пример конечного результата этой работы*. Ясно, что в практических задачах составление такой схемы может оказаться сложным делом.

Рассмотрим автоматическую телефонную систему. Абонент снимает трубку с рычага, и система, т. е. коммутационное оборудование на телефонной станции, получает соответствующий сигнал. Теперь система должна иметь набор логических правил, предписывающих искать линию, с которой она может соединить этот телефон, и искать абонентский регистр; выполнить соединения, когда соответствующие линии найдены; и если все эти

---

* Таким образом, системная логика, которую авторы трактуют как особую научную дисциплину, сложившуюся в 40-х годах (см. § 3.5), по существу совпадают с тем, что другие авторы у нас и за рубежом называют *теорией алгорифмов*. Как известно, теория алгорифмов действительно сложилась около 1940 г. и получила сейчас значительное развитие (работы А. Тьюринга, А. А. Маркова и др.), в современной автоматизированной технике она играет все более важную роль. Составление логических правил работы системы, о котором идет речь в этом месте книги, есть не что иное, как составление алгорифма управления системой. — *Прим. ред.*

действия успешны, послать абоненту сигнал готовности станции, извещая его, что он может набирать номер.

Теперь система ожидает, что абонент наберет семь цифр; она ждет набора седьмой цифры, затем исследует цифры и снова, согласно другому набору логических правил, приступает к выполнению надлежащих коммутационных соединений.

Однако может случиться, что абонент совсем не будет набирать номер или наберет менее семи цифр и остановится. Такие ситуации маловероятны, но все же будут наступать достаточно часто для того, чтобы система имела логическое правило, учитывающее эти ситуации. Такое правило могло бы гласить: «Если по истечении 60 *сек* полное число цифр не набрано, а трубка абонента все еще снята с рычага, отключить его линию от абонентского регистра и послать в его аппарат специальный сигнал неисправности; затем ждать 5 *мин*; если трубка все еще будет снята с рычага, оповестить о неисправности контрольный аппарат; если в течение 5 *мин* трубка будет положена на рычаг, выключить сигнал о неисправности и восстановить нормальное положение».

Конечно, изложенное представляет собой лишь небольшую часть логических правил, необходимых для работы автоматической телефонной системы, и даже в этом случае для каждого частного правила потребовались бы довольно сложные обоснования (некоторые из этих обоснований и правил рассматривались в гл. 13). Нужно также отметить, что эти правила связаны с особыми требованиями к оборудованию. Так, если принять вышеуказанное правило, то нужно предусмотреть надлежащие устройства отметки времени. Однако эти соображения можно на время отложить, ограничившись общим утверждением, что требования к аппаратуре осуществимы.

Рассмотрим другой пример, относящийся к системе наблюдения самолетов, которая может входить в систему ПВО или в систему управления воздушным движением. Рассмотрим лишь случай одиночного самолета, поскольку мы разбираем единичную нить, проходящую через систему. Имеется несколько радиолокационных станций в разных пунктах, и мы хотим, чтобы по меньшей мере одна из них следила за самолетом все время и сообщала его положение органам логического управления системы. Будут моменты, когда за самолетом могут одновременно следить несколько радиолокационных станций; мы хотим найти логические правила, гарантирующие, что система не будет поступать так, как если бы эти многократные сообщения относились к разным самолетам. Мы можем решить усреднять все данные для одного самолета или даже применять какую-нибудь сложную схему усреднения с весовыми коэффициентами. Или можем решить отбрасывать в центре логического управления все данные, кроме данных от одной станции, или можем решить прекратить подачу сообщений всеми станциями, кроме одной. В любом случае нам нужен специальный набор логических правил и мы должны быть уверены, что в этих правилах не будет упущений.

Пусть, например, мы решили, что сообщать о цели будет только ближайшая к ней радиолокационная станция. Предположим, что две радиолокационные станции отстоят друг от друга на 100 миль и самолет находится на высоте 10 миль и точно посередине между ними. Тогда каждая станция, измерив наклонную дальность до самолета, установит, что она равна 51 миле; каждая станция вычислит, что самолет находится в 49 милях от другой станции и в 51 миле от нее самой, и поэтому не будет сообщать о нем. Мы видим, что для учета подобных ситуаций требуется довольно сложный набор логических правил.

22.1. Пример. Игра «ним»

Вывод и обоснование всех логических правил для любой системы большого масштаба занял бы книгу такого же объема, как наша (см., например, [142]). Поэтому выведем лишь логические правила для простой игры «ним».

Предположим, что наша «система» представляет собой цифровую вычислительную машину с подходящим кодом для игры в ним, и начертим блок-схему действий (как на рис. 16.2). Мы не будем стремиться составить подробный код, который был бы слишком длинен, хотя разработать его довольно легко. Поскольку сама игра очень проста, мы будем в состоянии обосновать подробно каждый шаг. Логическая схема, которую мы строим (рис. 22.1), будет сравнительно простой, но в других отношениях она будет вполне аналогична логической схеме единичной нити большой системы, с такими же типами элементарных выборов.

В игру «ним» играют два игрока (которых мы будем называть «вычислительная машина» и «противник»). В начале игры перед ними имеются несколько кучек фишек. Игроки ходят поочередно, причем при каждом ходе из какой-нибудь кучки удаляется по крайней мере одна фишка; игрок может взять

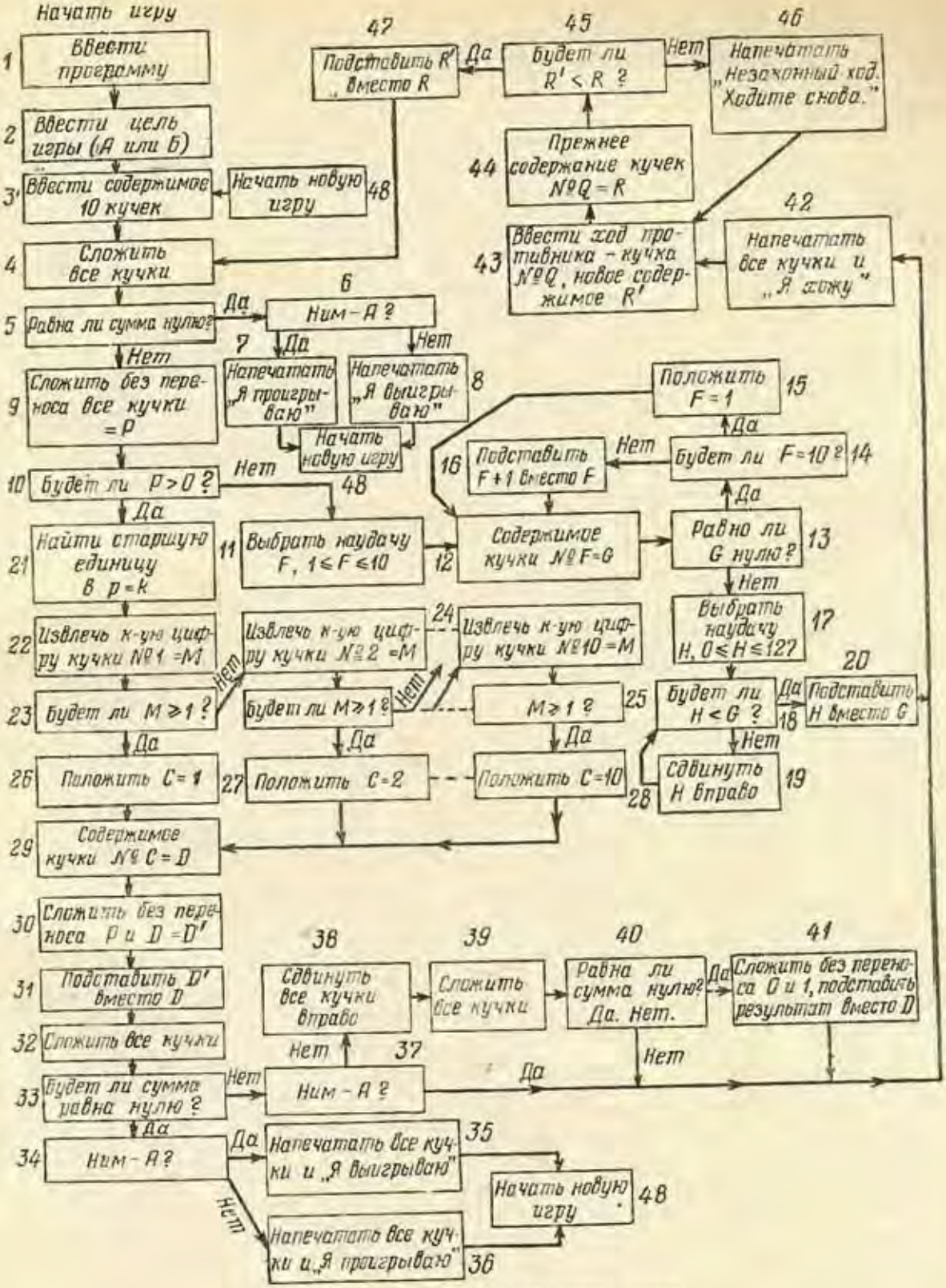


Рис. 22.1. Логическая схема игры „ним".

за один ход сколько угодно фишек из одной кучки, но не может за один ход брать фишки из двух разных кучек. Выигрывает тот, кто возьмет последнюю фишку*.

Например, пусть имеются три кучки: с тремя фишками в первой кучке, двумя во второй и одной фишкой в третьей; предположим, что сейчас ход противника. Тогда машина всегда может выиграть, так как у противника лишь шесть возможных ходов, а именно:

1. Противник берет три фишки из первой кучки. Машина тогда берет одну фишку из

* Игра «ним» происходит из Китая. — *Прим. ред.*

второй кучки, оставляя одну фишку во второй кучке и одну в третьей. Противник тогда берет одну из них, а машина берет другую и выигрывает.

2. Противник берет две фишки из первой кучки. Машина тогда берет две фишки из второй кучки, оставляя одну фишку в первой кучке и одну в третьей. Машина выигрывает, как и раньше.

3. Противник берет одну фишку из первой кучки. Машина тогда берет одну фишку из третьей кучки, оставляя две фишки в первой кучке и две во второй. Тогда у противника две принципиальные возможности:

а) противник берет две фишки из одной кучки; машина тогда берет две фишки из другой кучки и выигрывает;

б) противник берет одну фишку из одной кучки; машина тогда берет одну фишку из другой кучки; после этого машина выигрывает, как при ходе 1.

4. Противник берет две фишки из второй кучки. Машина тогда берет две фишки из первой кучки и выигрывает, как при ходе 1.

5. Противник берет одну фишку из второй кучки. Машина тогда берет три фишки из первой кучки и выигрывает, как при ходе 1.

6. Противник берет одну фишку из третьей кучки. Машина тогда берет одну фишку из первой кучки, оставляя по две фишки, и выигрывает, как при ходе 3.

Мы видим, таким образом, что существуют определенные основные *выигрышные комбинации* и что если один из игроков встречается с одной из этих комбинаций, он должен проиграть, если другой игрок играет правильно*. Однако если число кучек и число фишек в каждой кучке велико, то отнюдь не сразу видно, как прийти к такой выигрышной комбинации. Например, пусть имеется 53 фишки в первой кучке, 67 во второй, 13 в третьей, 2 в четвертой и 19 в пятой. Правильный ход — взять 26 фишек из второй кучки. Если машина сделает это и после этого будет играть правильно, она должна выиграть; если же она сделает какой-либо другой ход и противник после этого будет играть правильно, то должен выиграть он. Теперь мы опишем логическое правило, определяющее, какие существуют выигрышные комбинации и как прийти от невыигрышной комбинации к выигрышной.

Логика игры. Запишем сначала число фишек в каждой кучке, как в табл. 22.1, и против каждого числа напишем соответствующее

Таблица 22.1

Вычисления для игры „ним“

53	0110101	
67	1000011	D
13	0001101	
2	0000010	
19	0010011	
	1101010	P
67	1000011	D
41	0101001	D'

двоичное обозначение. Сложим эти двоичные числа без переноса.

Например, сумма цифр в первом столбце (считая, как всегда, справа) равна 4, или в двоичном обозначении 100, но поскольку мы не учитываем переносов, пишем лишь цифру 0. Продолжая складывать таким же образом по всем столбцам, мы получим ряд двоичных цифр, представляющий сумму без переноса, которую мы обозначим символом P.

Теперь найдем самую левую (самую старшую) единицу в P и заметим ее позицию, обозначив ее буквой k; в данном случае самая левая единица находится в седьмом столбце, так что $k=7$. Затем найдем строку, содержащую единицу на k-м месте, и обозначим эту строку буквой D; в данном случае единица на седьмом месте имеется во второй строке, так что $D=67$ (или, вернее, двоичному эквиваленту от 67). Наконец, мы запишем D под P и сложим их без переноса; эту сумму без переноса мы обозначим буквой D'.

Правильная игра состоит в том, чтобы вместо D получить D'. В нашем случае из 67 мы делает 41, т. е. на первом ходе нам нужно взять 26 фишек из второй кучки.

Доказательство логики. Мы докажем, что: 1) если заменить D на D' и повторить сложение без переноса, то новое P будет состоять целиком из нулей; 2) если $P\neq 0$ при ходе машины, то машина всегда может найти такое D', при котором $P=0$; 3) если машина сделает P равным 0, то ход противника приведет к $P\neq 0$; 4) если игра продолжается именно так: с $P=0$ после хода машины и $P\neq 0$ после хода противника,— то машина в конце концов должна выиграть.

Доказательство первого предложения таково. Обозначив знаком «плюс» сложение без переноса и буквой A сумму без переноса всех столбцов, кроме D, мы найдем P из

$$A+D=P$$

и D' из

$$P+D=D'.$$

* Это изложение логики выигрыша не нужно смешивать с теорией игр (гл. 24), с точки зрения которой игра «ним» имеет чистые стратегии и тривиально проста. — *Прим. авт.*

Отсюда

$$D' = A + D + D.$$

Если мы теперь вычислим новое P, то оно будет равно

$$P = A + D' = A + A + D + D.$$

Но всякое число при сложении без переноса с самим собой дает одни нули.

Чтобы доказать второе предложение, мы должны показать, что всегда можем найти кучку D, содержащую единицу на k-м месте, и что D' всегда меньше D, так что ход является законным. Первая часть вытекает из того, что при сложении нет переноса; если все кучки имеют на k-м месте нули, то и P на k-м месте будет иметь нуль. Вторая часть вытекает из известного нам факта, что P содержит единицу на k-м месте, но ни одной единицы левее. Так как и P, и D содержат единицу на k-м месте, то D' будет содержать на k-м месте нуль, а поскольку в P нет единиц левее k-го места, то D и D' левее k-го места будут одинаковы. Следовательно, $D' < D$.

Чтобы доказать третье предложение, заметим, что противник может изменить лишь одну кучку. Так как двоичное обозначение всякого целого числа единственно, то противник должен изменить по крайней мере одну цифру из нуля в единицу или наоборот; поскольку все другие кучки не меняются, это должно привести к изменению соответствующей цифры или цифр числа P.

Чтобы доказать четвертое предложение, заметим, что общее число остающихся фишек всегда уменьшается. Поэтому в конце концов оно должно дойти до нуля. Но когда это произойдет, P будет равно 0 и, согласно нашему предположению, машина в этот момент как раз закончит играть. Следовательно, машина выиграет.

Может оказаться, что несколько кучек содержат единицу на k-м месте; в этом случае будет несколько «правильных» ходов, и не имеет значения, какой из них мы делаем; мы показали выше, что всегда существует хотя бы один правильный ход (при условии, что у противника не было возможности первым применить выигрышную стратегию).

Другая форма игры «ним». Описанную выше игру мы будем называть «ним-А». В игре, которую мы назовем «ним-Б», ставится другая цель — заставить противника взять последнюю фишку. Для этого в стратегии в какой-то точке должно быть предусмотрено *логическое переключение*, посредством которого P придается некоторое ненулевое значение. Но это логическое переключение нельзя делать в начале игры; его нужно сделать лишь в тот момент, когда ни в одной кучке не останется больше одной фишки. Таким образом, мы вычисляем P и D как прежде, но это D' мы называем пробным. Прежде чем сделать этот ход, мы исследуем ситуацию, чтобы посмотреть, не останется ли после него в какой-нибудь кучке две или более фишки. Если такая кучка есть, мы делаем ход; если же нет, то найденное нами D' должно быть равно 0 или 1, и в обоих случаях мы заменяем единицу нулем или наоборот и затем делаем ход.

Например, если перед игрой было четыре кучки, содержащие три, одну, одну и одну фишку, то прежнее правило рекомендовало бы нам взять две фишки из первой кучки, сделав $D' = 1$; новое правило говорит, что в этой точке нужно проделать логическое переключение и взять три фишки из первой кучки ($D' = 0$). Если было три кучки, с тремя, одной и одной фишкой, то прежнее правило рекомендовало бы взять три фишки из первой кучки ($D' = 0$); новое правило рекомендует взять две фишки из первой кучки ($D' = 1$).

Докажем теперь, что эта стратегия дает выигрыш в игре «ним-Б». Если перед логическим переключением делает ход противник, то $P = 0$; поэтому, если имеется единица в каком-либо другом столбце, кроме первого, в нем должны быть по меньшей мере две единицы, и противник не может изменить обе в нули. Следовательно, противник не имеет никакой возможности выполнить логическое переключение. Однако в конце концов в других столбцах, кроме первого, будут в точности две единицы, и противник в конце концов должен будет переделать одну из них в 0. В этот момент машина имеет возможность сделать логическое переключение и затем оставить противника с нечетным числом кучек, содержащих каждая по одной фишке. После этого оба игрока должны будут брать по одной фишке, причем противнику придется взять последнюю.

Логическая схема. Прежде чем начертить логическую схему, мы должны сделать некоторые дополнительные допущения. Мы примем, что имеется 10 кучек (если их меньше, машина будет считать, что некоторые кучки содержат нуль фишек) и что максимальное число фишек в кучке равно 127. При своем ходе машина будет печатать все 10 кучек с их содержанием; при ходе противника он будет вводить в машину на перфорированной ленте номер Q кучки, из которой он решает

брать фишки, и число фишек R', которое он решает оставить в этой кучке. Мы допускаем, что противник, по неведению или с целью плутовства, может попытаться сделать незаконный ход.

Блок-схема изображена на рис. 22.1. В пояснение ее сделаем следующие замечания:

Команды 4—8. Противник, возможно, взял последнюю фишку и закончил игру. С помощью команд 4 и 5 мы обнаруживаем, произошло ли это; с помощью команд 6, 7 и 8 мы находим и печатаем, кто из игроков выиграл. Затем мы переходим к команде «начать новую игру». Для выполнения команды «сложить все кучки» машине, описанной в гл. 16, нужно будет девять команд. Более совершенной машине для этого потребуется также девять операций, но только лишь две команды: первая заставит машину произвести сложение и укажет, куда поместить сумму, а вторая укажет, как изменить команду и когда нужно остановиться.

Команда 9. Машина должна иметь специальную команду «сложить без переноса». Эту операцию машина, описанная в гл. 6, может выполнить путем трех операций выборки и нескольких передач, но это требует большого времени и большого объема памяти. Осуществить аппаратурно сложение без переноса, конечно, проще, чем обычное сложение.

Команды 10—20. Может случиться, что противник благодаря удаче или искусству выбрал выигрышную стратегию на первом ходе. Если не предусмотреть такой возможности в этой точке программы, то машина впоследствии окажется не в состоянии сделать законный ход в соответствии со своими командами. Поэтому составляется довольно сложная подпрограмма для случайного выбора кучки и хода в этой кучке (в надежде, что это, может быть, собьет противника с правильного пути). В реальной состязательной системе стратегия могла бы быть определена по теории игр, и в этом случае применялись бы подпрограммы со случайными числами, аналогичные рассмотренным.

Команда 13. Мы должны убедиться, что в кучке, из которой мы хотим брать фишки, они имеются.

Команды 14—16. Если выбранная кучка пуста, то нам нужно выбрать другую кучку; мы не можем применить простой способ (16) для этого нового выбора без того, чтобы не возник вопрос (14). Этот, на первый взгляд, тривиальный шаг может оказаться очень существенным в логической схеме реальной системы.

Команда 18. Мы должны убедиться, что выбранный нами ход законен. (При желании мы можем ввести сюда специальную подпрограмму, по которой машина время от времени, скажем один раз из десяти с выбором по таблице случайных чисел, будет пытаться плутовать, если она обнаружит, что проигрывает.)

Команда 19. Эта команда состоит в делении на 2 с отбрасыванием остатка.

Команды 22—29. В этой подпрограмме мы определяем местоположение («адрес») кучки D. Применяя команду выборки из табл. 16.1, найдем содержание этой кучки α'; β' будет равна 2^{k-1}, а γ' будет сначала равна нулю. После того как команда выполнена, γ' по-прежнему будет равна нулю, если в кучке на k-м месте был нуль, и будет равна 2^{k-1}, если в кучке на k-м месте была единица. Следующая команда — сравнение — говорит нам, какая эта кучка. Если машина не имеет команды выборки, тот же результат можно получить с помощью сравнения, но нужно будет убедиться, что γ' меньше 2^k, чтобы быть уверенным, что кучка содержит 1 на k-м месте.

Команда 25. На этот вопрос может быть лишь один ответ. Поэтому для машины это лишняя инструкция, но она не приносит вреда и облегчает кодирование.

Команда 31. Хотя мы еще не определили ход D' (так как нам может потребоваться проделать логическое переключение), число D нам больше не понадобится.

Команды 32 и 33. Теперь мы должны определить, не взяли ли мы последнюю фишку, закончив тем самым игру.

Команда 33. Независимо от ответа на этот вопрос, следующая команда будет одна и та же. Однако эта избыточность более кажущаяся, чем действительная, и ее нельзя избежать.

Команда 37. Если мы играем в «ним-Б», нам нужно будет определить, делать ли логическое переключение.

Команда 38. При сдвиге (делении на 2 и отбрасывании остатка) мы должны запомнить первоначальные числа фишек в кучках, так как они понадобятся нам в команде 42.

Команда 40. При ответе «нет» мы делаем выбранный пробно ход D'; при ответе «да» мы делаем логическое переключение.

Команда 41. Эта операция есть логическое переключение. В результате сложения без переноса D' превращается из 0 в 1 или наоборот. Нельзя делать логическое переключение больше одного раза. Мы можем за-

страховаться от этого специальной подпрограммой, но она оказывается ненужной. После логического переключения мы получаем «нет» в команде 10 и проблема больше не возникает.

Команды 44—46. Противник может попытаться плутовать.

Эта блок-схема поясняет, как вся логика системы разбивается на крайне простые вопросы «да—нет», на которые может ответить вычислительная машина. Решение (команда 23 и следующие) о выборе кучки может быть аналогичным задаче о типе запускаемого реактивного снаряда в системе ПВО. Первым вопросом может быть такой: составляет ли высота цели больше 30 000 футов? При ответе «да» мы можем выбрать высотный реактивный снаряд; при ответе «нет» мы можем спросить: составляет ли высота цели больше 10 000 футов? При ответе «нет» мы можем взять маловысотный реактивный снаряд. При промежуточных высотах мы можем задать другие вопросы, например относительно скорости. В каждом случае нужно определить логику системы и разбить эту логику на простые элементарные выборы, подобные показанным на рассмотренной логической схеме.

ГЛАВА 23

БОЛЬШАЯ НАГРУЗКА. ТЕОРИЯ МАССОВОГО ОБСЛУЖИВАНИЯ

Предмет настоящей главы называли поразному: *теорией массового обслуживания* [Д. 14] *, *теорией линий ожидания* [78], *теорией очередей* [80], *проблемами скученности* [33] и *теорией группообразования* ** [102]. В гл. 21 было сказано, что вторым этапом внутреннего проектирования систем является изучение и определение нагрузочных аспектов системы. Если в какой-либо системе с большой нагрузкой предусмотрено слишком мало каналов для какого-либо обслуживания, то время от времени будут образовываться линии ожидания чрезмерной длины; поэтому определение числа каналов в разных пунктах, являющееся первым шагом при проектировании большой нагрузки, требует решения соответствующих задач об очередях.

Далее, в гл. 21 было указано, что требования к большой нагрузке системы можно исследовать, не определяя подробно единичных нитей системы. Для того чтобы применить теорию массового обслуживания, нужно задать три фактора: частоту (или распределение во времени) появления входов в каждом «блоке» системы; время (или распределение времен), которое нужно блоку для выполнения своих функций, и набор правил, указывающих, что будет делать система с входом, если блок еще обслуживает предыдущий вход (например, образуется ли линия ожидания или вход теряется). Первый фактор определяется при внешнем проектировании системы, а два последних — при внутреннем проектировании (в частности, при проектировании единичной нити). При этом правила определяются логикой системы, а распределение времен занятия — предварительными соображениями относительно оборудования (т. е. компонентов).

23.1. Входные источники, очереди и каналы

На рис. 23.1 изображена принципиальная схема рассматриваемой системы. Несколько источников могут питать одну очередь, или один источник — одну очередь, или один источник — несколько очередей; несколько очередей могут поступать в один канал, или

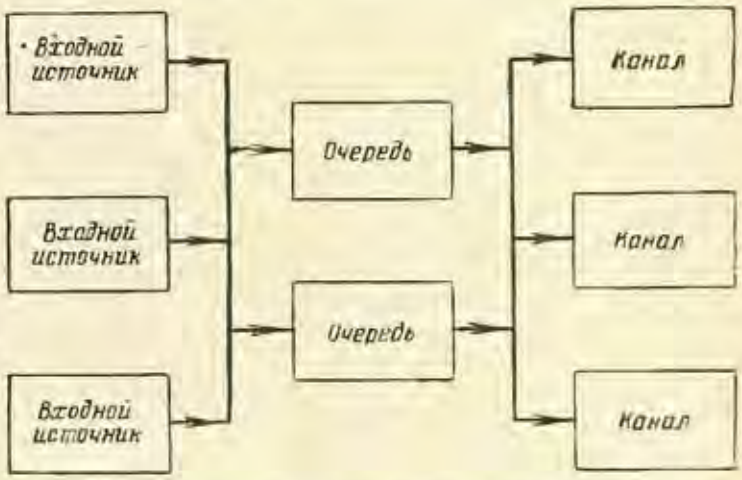


Рис. 23.1. Принципиальная схема входных источников, очередей и каналов.

* Термин «теория массового обслуживания» наиболее распространен в русской технической литературе; им пользовался русский математик А. Я. Хинчин, внесший большой вклад в развитие этой теории [Д. 14]. По этим причинам мы выбрали этот термин при переводе и ввели его здесь в текст книги. В оригинале авторы пользуются из упомянутых далее терминов преимущественно термином «теория очередей» — queueing theory, theory of queues. — *Прим. ред.*

** Термин «теория группообразования» является общепринятым русским эквивалентом термина «trunking theory» (буквально — «теория соединительных линий). — *Прим. ред.*

одна очередь — в один канал, или одна очередь — в несколько каналов.

Это, на первый взгляд, приводит к девяти различным типам систем; в действительности они сводятся лишь к двум или трем принципиально различным типам. Если несколько источников питают одну очередь, то обычно их можно рассматривать как один источник; конечно, распределение во времени входов от объединенного источника, вообще говоря, не совпадает с распределением входов от какого-нибудь одного источника. Если один источник питает несколько очередей, то какое-нибудь правило должно определять, в какую очередь направляется каждый вход. После того как такое правило задано, можно считать, что каждая очередь имеет один источник. Подобно этому, если несколько очередей поступают в один канал, их можно рассматривать как одну очередь.

Таким образом, обычно можно считать, что один источник питает одну очередь, которая затем поступает в один или несколько каналов. Если несколько очередей поступают в несколько каналов, то также должно быть какое-то правило. Если клиент в очереди может идти только к одному определенному каналу, то эта очередь действительно поступает в один канал; если же клиент в очереди может идти к любому свободному каналу, то все очереди представляют по существу одну очередь, поступающую в несколько каналов. Эти два случая, которые мы будем подробно разбирать, мы обозначим соответственно как *случай одиночного канала* и *случай множественных каналов*. Другие авторы дали им другие названия; например, в [33] они обозначаются соответственно как «некооперативные» и «кооперативные» каналы.

Входные источники. Мы различаем входной источник, вход и клиента. *Входной источник* производит входы; как только вход вливается в очередь, мы будем называть его *клиентом*. Входной источник характеризуется распределением во времени производимых им входов. Клиенты иногда различаются один от другого, как описано ниже в разделе об очередях.

Проще всего описать такой источник, из которого входы выходят в заданные равные промежутки времени. Если, кроме того, время, необходимое для действия канала, постоянно, как в автоматической производственной линии, то легко обеспечить, чтобы линия ожидания не образовывалась, и никакой проблемы не возникает. Если же времена задержки в канале распределены случайно, то может образоваться линия ожидания. Такая ситуация возникает в приемной врача, где расписание приема составляется заранее, но длительность осмотра распределена случайно.

Обратное положение, когда промежутки времени между входами распределены случайно, а времена задержки в канале постоянны, имеет место у турникетов метро и может приводить к линиям ожидания. Оказывается, что среди распределений входов легче всего исследовать распределение Пуассона — в некотором смысле самое случайное из всех дискретных распределений *. В настоящей главе мы будем иметь дело лишь с этим распределением входов.

Существенная характеристика входного источника состоит в том, конечен ли он или бесконечен. Например, в телефонной системе группа соединительных линий может обслуживать лишь ограниченное число абонентов. Хотя мы можем вполне допустить, что все абоненты имеют одинаковые вероятности произвести вызов (и что, следовательно, входы распределены по закону Пуассона), занятый абонент не может произвести вызов. Мы указали раньше, что параллельные входные источники можно вообще рассматривать как один источник. Но если число источников не намного больше, чем число каналов, и источник не может произвести новый вход, пока канал не пропустит предыдущий вход, то для гипотетического одиночного источника, изображающего группу из нескольких источников, вероятность произвести вход не будет постоянна: эта вероятность пропорциональна числу свободных источников.

В § 23.5 рассмотрен подобный случай, т. е.

---

* В гл. 28 мы дадим количественное определение понятия «наиболее случайное из всех распределений», а именно будем понимать под ним распределение ансамбля функций, имеющего максимальную энтропию. Мы покажем, что из непрерывных распределений таким является нормальное распределение. Из тех же рассуждений будет ясно, что среди дискретных распределений таким является распределение Пуассона, потому что дискретный входной источник с максимальной энтропией есть такой источник, у которого все члены его генеральной совокупности независимы и равновероятны; как показано в § 5.6, такой входной источник вызывает события, распределенные по закону Пуассона.

Обратным распределению Пуассона является экспоненциальное распределение, которое нужно также рассматривать как «наиболее случайное», но в другом смысле. А именно, если мы наблюдаем процесс из ансамбля с экспоненциальным распределением и узнаем, сколько времени он продолжается, то тем самым мы еще не получаем новых сведений о вероятном времени его окончания, как было показано в § 6.7. Этого нельзя сказать ни о каком другом распределении, и в этом смысле экспоненциальное распределение является наиболее случайным. — *Прим. авт.*

когда имеется несколько входных источников и несколько каналов; входами в этом случае служат поломки машин, а сломавшаяся машина не может образовать другой вход, пока ее не починили. Но в нашем исследовании мы большей частью будем полагать, что частота входов не зависит от длины очереди. Ввиду того что теория массового обслуживания была разработана в значительной мере специально для телефонии, имеется много литературы об очередях, происходящих от ограниченного числа входных источников.

Очереди. Хотя термин *очередь* и *линия ожидания* обычно применялись как синонимы, мы будем их различать. Удобно обозначать символом n длину очереди, включая клиента, обслуживаемого в данный момент (или — в случае нескольких каналов — клиентов, обслуживаемых в данный момент); таким образом, очередь, по нашему определению, включает клиента или клиентов, находящихся в канале. Линия ожидания, по определению, включает клиентов, не находящихся в канале. Следовательно, длина очереди больше, чем длина ожидания, на столько единиц, сколько имеется каналов, за исключением тех случаев, когда один или несколько каналов не заняты.

Существуют по крайней мере две ситуации, при которых линия ожидания не образуется и которые тем не менее представляют интерес и изучаются в теории очередей. Первая ситуация — когда клиент отказывается от запроса, если его немедленно не обслужили. Такое положение возникает, например, в телефонной системе при междустанционном вызове, когда в вызываемой станции все входящие соединительные линии или входящие регистры заняты. В этом случае абонент получает сигнал занятости и вынужден повесить трубку. Он может начать все снова, но прежний вызов «теряется» и исключается из системы. В таких случаях желательно, чтобы число потерянных клиентов было возможно меньше, и теория массового обслуживания исследует, какое число клиентов будет потеряно.

Другой крайний случай — когда самолет прибывает к аэропорту и все взлетно-посадочные дорожки заняты, так что ему остается лишь ждать очереди. Промежуточный случай — когда абонент снимает трубку и обнаруживает, что все приборы на станции заняты; тогда он не получает сигнала готовности станции и может ждать или не ждать, как ему угодно. Обстановка, когда клиенты, не получившие немедленного обслуживания, теряются, математически несколько проще, чем ситуация, когда они образуют очередь. Другая математически простая обстановка — это когда необслуживаемый клиент ждет (в линии ожидания и в канале, если он до него доходит) в течение времени, равного тому времени, которое он потратил бы в канале, если бы был немедленно обслужен. Оба эти случая подробно рассмотрены в [33]. Они представляют простые обобщения случаев, разобранных в этой главе, и дальше не рассматриваются.

Вторая из двух упомянутых выше ситуаций, при которых линия ожидания не образуется, — это когда имеется бесконечное число каналов. Конечно, такое допущение — идеализация. Тем не менее оно не только математически просто, но и представляет значительный интерес, ибо мы можем определить относительное время (т. е. вероятность), в течение которого произвольное количество из бесконечного множества каналов будет занято; если вероятность того, что, скажем, 10 каналов будут заняты, составляет лишь 0,001, мы можем построить только 10 каналов и будем знать, что линия ожидания ненулевой длины будет возникать лишь в течение одной тысячной всего времени. Математически это не совсем точно, но если вероятности быстро падают (как обычно бывает на практике), это очень близко к действительности. Случай бесконечного числа каналов мы разбираем первым в разделе о многоканальных системах.

Существует еще один тип очереди, который не охватывается схемой рис. 23.1, а именно очередь с двумя концами. В этом случае очередь можно также рассматривать как канал, а канал — как очередь. Например, на стоянке такси может быть очередь такси, если такси больше, чем людей, и очередь людей, если людей больше, чем такси. В некотором смысле все очереди можно считать двухконечными, так как бездействующие каналы можно рассматривать как линию ожидания. Возьмем, например, упомянутые выше машины и механиков, которые их чинят. Если поломанных машин слишком много, некоторые из них будут ждать в очереди, пока механики чинят другие; с другой стороны, будут моменты, когда некоторые или даже все механики не заняты, а все машины работают; в этом случае образуют очередь механики. Но эта очередь не имеет другого входного источника, кроме освобождения механиков по выполнении ими своих обязанностей, и поэтому ее длина строго ограничена.

Встает также вопрос о *порядке очереди.* Мы обычно имеем в виду тот порядок очереди, при котором каждый клиент сохраняет

свое место в линии. Так обстоит дело, например, в очереди такси, ожидающих пассажиров. Другая крайность — случай занятой телефонной станции, когда абонент снимает трубку с рычага и не получает гудка станции. Другие абоненты, также ждущие гудка станции, находятся в той же линии ожидания, но в ней нет определенных мест. Когда освобождается канал, один абонент выбирается случайно из линии ожидания, и с одинаковой вероятностью это может быть как первый, так и последний абонент. Возможны и промежуточные случаи, когда существуют предпочтения (приоритеты), т. е. когда каждый клиент имеет свое место в линии, но некоторые, более предпочитаемые клиенты могут быть обслужены раньше других.

Порядок очереди влияет на некоторые из факторов, которые мы хотим оценить, и не влияет на другие. Например, он не влияет на распределение длины очередей, но влияет на распределение времени ожидания. В этой главе мы будем предполагать, что в тех случаях, когда порядок очереди существен, места фиксированы. Но во многих реальных системах будет существовать система предпочтений. Например, в очереди над аэродромом самолету, у которого почти вышло горючее, очевидно, нужно дать предпочтение перед другими самолетами, у которых полный запас горючего. Во многих случаях проектировщик системы захочет дать некоторым входам предпочтение в очереди.

Существует один принцип взаимосвязи предпочтений, который кажется столь очевидным, что его не стоило бы упоминать, если бы его часто не упускали из вида. Если одного клиента продвигают в линии вперед, то в действительности все другие клиенты передвигаются назад. Если много клиентов получают большие предпочтения, то остальным, по-видимому, придется ждать чрезмерно долго.

Каналы. В большинстве случаев клиенты, дойдя до конца очереди, получают какое-то обслуживание, занимающее конечное время. Лицо или аппарат, который производит это обслуживание, называется *каналом* или *обслуживателем*; в этой книге употребляется лишь первый термин. Время, которое клиент затрачивает в канале, называется *временем занятия* или *временем обслуживания;* здесь мы употребляем лишь первый термин. Времена занятия могут быть постоянны или могут быть как-то распределены.

Рассмотрим упомянутый выше случай такси и пассажиров. Если такси образуют канал, то промежуток времени между прибытиями такси есть время занятия, а человека, занимающего первое место в линии, нужно считать находящимся в канале; когда такси являются клиентами, то промежуток времени между прибытиями такси задает распределение входов. Поэтому неудивительно, что имеется тесная связь между распределениями входов и распределениями времен занятия. Если моменты окончания времен занятия в занятом канале распределены по закону Пуассона, то сами времена задержки распределены экспоненциально (и наоборот).

Таким образом, экспоненциальное распределение времен занятия аналогично пуассонову распределению входных факторов, как наиболее случайное (см. предыдущее примечание) и как математически наиболее удобное. В этой главе мы будем заниматься только этим распределением времен занятия.

Хранение и буферное хранение. По определению буферное хранение есть хранение, предназначенное специально для удержания очереди. Так, при постройке театра нужно предусмотреть обширное запасное место (в виде вестибюля) для ожидаемых очередей, чтобы клиенты не были потеряны в плохую погоду. Это место, требующее больших расходов, можно рассматривать как буферное хранилище (накопитель).

Более типичным примером при проектировании систем является система кодирования для связи, подобная рассмотренной в § 28.1, в которой мы используем то обстоятельство, что некоторые символы (например, e и t в английском языке) встречаются чаще других (например, q и z). При кодировании применяются короткие коды для распространенных символов и более длинные коды — для менее распространенных. Однако если входной и выходной сигналы кодирующего устройства появляются с постоянной частотой, то будут случаи, когда в течение малого времени появится много редких символов, канал (т. е. кодирующее устройство) будет не в состоянии обработать их и образуется линия ожидания. Для удержания этих «клиентов» должно быть предусмотрено буферное хранение.

Теория массового обслуживания может нам сказать, как велико должно быть буферное хранилище (т. е. до какой длины может вырасти очередь), чтобы буфер не переполнялся чаще, чем, скажем, раз в неделю. Системная логика определяет, что нужно делать, если буфер переполняется; здесь возможны такие альтернативы: закрыть источник, допустить потерю некоторых сообщений или допустить вмешательство человека.

Не всегда легко провести границу между буферным и обычным хранением. Рассмот-

рим, например, склад товаров, представляющий собой очередь, у которой входами являются изделия завода, а каналом — продажи, причем времена занятия (т. е. промежутки времени между требованиями) распределены более или менее случайно.

Другой, более важный пример (подтверждающий наше положение, что все системы суть информационные системы) — это хранение данных в системе материально-технического обеспечения ВВС. Здесь имеются миллионы элементов информации, разбитых на 55 общих классов, и вопрос состоит в том, какой величины взять каждую из 55 секций запоминающего устройства; желательно иметь быстродействующее запоминающее устройство, поэтому дополнительная емкость дорого стоит. Время занятия равно промежутку времени от момента, когда элемент информации записывается в запоминающее устройство, до момента, когда он стирается. В этом случае, вероятно, удобнее считать, что запоминающее устройство состоит из множественных каналов (где каждый двоичный разряд является отдельным каналом) и задача состоит в определении числа каналов, потребных, чтобы вероятность образования линии ожидания в течение некоторого времени, скажем года, стала меньше заданного значения.

Во всех подобных задачах после того, как мы подобрали формулы из теории массового обслуживания, можно определить число каналов, размер буфера и т. п. с помощью основного принципа проектирования систем (§ 21.6), а именно: вероятность ошибок I и II рода можно вычислить из теории массового обслуживания, стоимость ошибок I и II рода можно вычислить из других соображений, а затем можно вычислить ожидаемое значение стоимости при различных выбранных величинах и свести ее к минимуму.

Искомые величины. Из этого разбора видно, что в теории массового обслуживания нам нужно вычислить две основные величины: длину очереди и время ожидания. Следует отметить, что одна из них дискретна, а другая — непрерывна; поэтому математические выводы различны: в одном случае они основаны на разностном уравнении, а в другом — на дифференциальном.

Иногда нужно определить лишь среднюю длину очереди и/или среднее время ожидания, и их часто легко бывает вычислить, особенно если допустить некоторую аппроксимацию *. В других случаях мы хотим знать вероятность того, что длина очереди (и/или время ожидания) превысит некоторое значение, и для таких величин часто бывает необходимо вычислить функцию распределения и определить интересующие нас параметры распределения. Часто бывает необходимо найти функцию зависимости всех или некоторых из этих величин от числа предусмотренных каналов. Ниже приведены примеры всех этих задач.

В отдельных случаях может потребоваться определить какую-нибудь специальную характеристику распределения. Например, если канал представляет собой (или содержит) электрический двигатель, то может потребоваться знать не только относительное время, когда двигатель работает, но и сколько времени он работает между двумя простоями, т. е. распределение времени работы двигателя, эквивалентное распределению того времени, когда очередь имеет ненулевую длину. Распределение рабочего времени выражается через бесселеву функцию [80] и здесь не рассматривается.

Если канал стоит дорого, то часто вычисляется *степень занятости* его, т. е. среднее относительное время, когда он занят. Однако не нужно впадать в ошибку, приписывая излишнее значение низкой степени занятости. Как мы увидим из численных примеров в конце этой главы, для уменьшения степени занятости до 50% и менее нередко приходится вводить большое число добавочных каналов и все же линии ожидания будут образовываться довольно часто. Если дорогостоящий канал используется лишь в течение небольшой доли времени, то целесообразно исследовать ситуацию, чтобы найти, какое улучшение можно ввести, но это не обязательно является признаком неэффективности. В одном аэропорте вполне могут потребоваться две полные установки наземного управления посадкой, хотя они будут бездействовать больше 99% всего времени.

23.2. Одиночный канал

Как указано выше, когда промежуток времени между входами постоянен и время занятия также постоянно, никакой задачи по теории массового обслуживания не возникает.

---

* Под средними величинами в этой главе всюду понимаются математические ожидания величин, в смысле гл. 6. Там слово «среднее значение» (average) не рекомендовалось, но при переводе этой главы мы были вынуждены им пользоваться (за счет термина «математическое ожидание») во избежание путаницы, так как здесь часто идет речь об ожидании в другом смысле — в смысле линии ожидания и времен ожидания. — *Прим. ред.*

Когда одна величина постоянна, а другая распределена случайно или когда обе величины распределены случайно, существуют линии ожидания, по крайней мере в течение какой-то доли времени. Нам интересно прежде всего определить, что можно сделать с очередями при произвольных распределениях.

Во всей этой главе мы будем применять следующие обозначения:

t — промежуток времени между появлением последовательных входов (t применяется также как переменная, обозначающая время вообще, но из контекста будет ясно, о чем идет речь, так что путаницы не возникнет);

r — число входов, появляющихся в интересующий нас отрезок времени;

m — среднее число входов в единицу времени;

T — время занятия;

R — число выходов, появляющихся из канала в интересующий нас отрезок времени;

M — среднее число выходов, появляющихся из *занятого* канала в единицу времени; $1/M$ равно среднему значению от T;

$$\rho=\frac{m}{M};$$

w — время ожидания, равное времени между появлением входа и его поступлением в канал;

n — длина очереди (включая клиента или клиентов, находящихся в канале или в каналах);

ν — число каналов;

j — длина линии ожидания ($j=n-\nu$, если $n>\nu$ и $j=0$, если $n\leqslant\nu$);

β — число источников (всюду равное единице, за исключением § 23.5).

Отношение ρ имеет большое значение в теории массового обслуживания *. Легко видеть, что если $\rho>1$ для случая одиночного канала, то очередь будет безгранично расти; вероятность любой заданной конечной длины стремится с течением времени к нулю, так же как вероятность любого заданного конечного времени ожидания. Отсюда ясно, что мы не можем найти явные выражения для этих интересующих нас вероятностей. Поэтому первое требование к проектировщику — определить условия *устойчивости* линии.

Устойчивость имеет место, если процесс является *стационарным*; для наших целей достаточно определить *стационарность* как состояние, которое имеет место, если

$$\frac{dp(n)}{dt}=0 \qquad (23.1)$$

для всех n. В случае одиночного канала это равенство не может иметь места при $\rho>1$, а также при $\rho=1$, как указано ниже. Во всех практически интересных случаях оно наступает после достаточно долгого времени, если m и M постоянны и $\rho<1$ (или, вообще говоря, если $\rho<\nu$). Во всем последующем изложении, если не оговаривается противное, мы будем считать, что $\rho<\nu$ и что процесс является стационарным.

Произвольные распределения входов и времен занятия. Мы покажем сначала, что вероятность занятия одиночного канала равна ρ. Интуитивно это в некотором смысле очевидно. Предположим, например, что входы появляются в среднем один раз в 2 *мин* и что среднее время занятия равно $^1/_4$ *мин*. Тогда $M=4$, $m=1/2$ и $\rho=1/8$. В течение 10 *час* появится около 300 входов и канал будет занят $^1/_4\cdot 300=75$ *мин*, или 1,25 *час*, т. е. одну восьмую всего времени. Следующий метод принадлежит Кендаллу [80].

Рассмотрим момент, когда какой-нибудь клиент C_0 уходит из канала. Длину очереди после его ухода (быть может, равную нулю) мы обозначим n_0. Если $n_0\neq 0$, то в линии находятся входы C_1, C_2, ..., причем C_1 как раз начинает свое время занятия T_1. Если $n_0=0$, то первый пришедший вход будет обслужен немедленно. В обоих случаях в течение T_1 (пока обслуживается C_1) придет некоторое число входов (быть может, нуль), которое мы обозначим r_1. Когда C_1 уходит из канала, новая длина очереди будет равна $n_1=n_0+r_1-1$ в первом случае и $n_1=r_1$ во втором (поскольку мы рассматриваем бесконечно краткий момент ухода, вероятность одновременного появления другого входа равна 0).

Эти два случая можно представить следующим одним уравнением:

$$n_1=n_0+r_1-1+\delta_0, \qquad (23.2)$$

где δ_0 по определению равно единице при $n_0=0$ и равно нулю при $n_0\neq 0$. Величина δ_0 принимает лишь значения 0 и 1, и ее математическое ожидание находится между этими двумя значениями. Заметим, что $n_0\delta_0=0$ и $\delta_0^2=\delta_0$.

* Эта величина измеряется в эрлангах, названных так по имени одного из основоположников теории массового обслуживания, который опубликовал ряд важных работ до 1920 г. Это безразмерное отношение подобно, как отмечает Кендалл [80], децибелам, октавам и звездным величинам. Но последние суть логарифмы отношения, тогда как эрланги выражают само отношение. — *Прим. авт.*

Теперь усредним уравнение (23.2), сначала по r_1 и затем по T_1. Из предположения о стационарности следует $E(n_0)=E(n_1)$. Поэтому при усреднении по r_1

$$E(\delta_0)=1-E(r_1)=1-mT_1,$$

так как ожидаемое число входов во время нахождения C_1 в канале равно среднему числу прибытий за единицу времени, умноженному на время занятия. При усреднении по T_1 математическое ожидание величины δ равно

$$E(\delta)=1-mE(T_1)=1-m\frac{1}{M}=1-\rho,$$

так как $E(T_1)=1/M$ по определению. Но $E(\delta)$ есть вероятность того, что $n=0$, т. е. вероятность того, что канал свободен. Следовательно,

$$P(n=0)\equiv p(0)=1-\rho, \qquad (23.3а)$$

$$P(n>0)\equiv p(>0)=\rho. \qquad (23.3б)$$

Мы можем провести дальнейшие рассуждения, не уточняя характера распределений времени между входами или времени занятия. Возводя в квадрат обе части равенства (23.2), получаем

$$n_1^2=n_0^2+(r_1-1)^2+\delta_0^2+2n_0(r_1-1)+$$
$$+2\delta_0(r_1-1)+2n_0\delta_0.$$

Последний член правой части равен нулю. Подставляя $\delta_0^2=\delta_0$ и вычисляя математические ожидания, находим

$$E(n_1^2)=E(n_0^2)+E[(r_1-1)^2]+E(\delta_0)+$$
$$+2E[n_0(r_1-1)]+2E[\delta_0(r_1-1)].$$

Ввиду стационарности первый член в правой части равен, как и раньше, левой части. На основании (6.25) мы можем перенести оператор E на члены произведений в правой части, так как r_1, число приходящих входов, по предположению независимо от числа входов, находящихся в линии, и должно быть независимо также от δ_0, зависящего только от n_0:

$$0=E[(r_1-1)^2]+E(\delta_0)+2E(n_0)E(r_1-1)+$$
$$+2E(\delta_0)E(r_1-1).$$

Усредняя по r_1 и T_1 и подставляя найденные выше значения $E(\delta)=1-\rho$ и $E(r)=\rho$, получаем:

$$0=E(r^2)-2\rho+1+(1-\rho)+2E(n)(\rho-1)+$$
$$+2(1-\rho)(\rho-1),$$

$$2E(n)(1-\rho)=E(r^2)-3\rho+2-2+$$
$$+4\rho-2\rho^2=E(r^2)-2\rho^2+\rho,$$

$$E(n)=\frac{E(r^2)-2\rho^2+\rho}{2(1-\rho)}=\rho+\frac{E(r^2)-\rho}{2(1-\rho)},\quad \rho\neq 1. \qquad (23.4)$$

Равенство (23.4) справедливо для произвольных распределений входов и времен занятия — при условии, что эти распределения не зависят от n и не изменяются с течением времени и что $\rho<1$. Следует отметить, что хотя $E(r)=\rho$, однако $E(r^2)$, вообще говоря, не равно ρ, и, следовательно, (23.4) означает, что $E(n)\to\infty$ при $\rho\to 1$. Мы не можем оценить $E(r^2)$, ничего не зная о распределении промежутков времени между появлениями входов.

Произвольное распределение времен занятия, пуассоново распределение входов. В остальной части этой главы мы будем предполагать, если не оговорено противное, что входы распределены по закону Пуассона. Поскольку распределение Пуассона будет очень часто применяться на следующих страницах, мы воспроизводим здесь две формулы из гл. 5:

$$p(k)=\frac{e^{-\mu}\mu^k}{k!}, \qquad (5.20)$$

$$p(k)=\frac{e^{-\mu t}(\mu t)^k}{k!}. \qquad (5.27)$$

В (5.20) $p(k)$ есть вероятность наступления точно k событий в единицу времени; в (5.27) $p(k)$ есть вероятность наступления точно k событий в течение времени t. В обоих случаях ожидаемое число событий в единицу времени равно μ, но среднее и дисперсия равны μ в (5.20) и μt в (5.27).

При вычислении (23.4) мы применяем вторую формулу:

$$p(r_1)=\frac{e^{-mT_1}(mT_1)^{r_1}}{r_1!},$$

поскольку r_1 есть число входов в течение времени занятия T_1. Математическое ожидание этого распределения равно mT_1, т. е. постоянному числу, и его дисперсия также равна mT_1. Но из (5.10)

$$\sigma_{r_1}^2=E(r_1^2)-[E(r_1)]^2.$$

Следовательно,

$$E(r_1^2)=\sigma^2_{r_1}+[E(r_1)]^2=mT_1+(mT_1)^2.$$

Теперь усредняем r_1 по всем T_1 и получаем

$$E(r^2)=mE(T)+m^2E(T^2).$$

Из (5.10)

$$E(T^2)=\sigma_T^2+[E(T)]^2=\sigma_T^2+\frac{1}{M^2}.$$

Следовательно,

$$E(r^2)=m\frac{1}{M}+m^2\left(\sigma_T^2+\frac{1}{M^2}\right)=\rho+m^2\sigma_T^2+\rho^2.$$

Наконец, подставляя это в (23.4), находим

$$E(n)=\rho+\frac{\rho^2+m^2\sigma_T^2}{2(1-\rho)}. \qquad (23.5)$$

Из (23.5) можно видеть, что поскольку числитель дроби должен быть всегда положителен, математическое ожидание числа n становится бесконечно большим, когда ρ стремится к единице, независимо от распределения времен занятия. Это равенство справедливо для произвольного распределения времен занятия. Для какого-нибудь распределения мы можем подставить σ_T и найдем $E(n)$. При экспоненциальном распределении $\sigma_T^2=1/M^2$ ввиду (6.36), написанное выше выражение приводится к следующему выражению:

$$E(n)=\frac{\rho}{1-\rho}. \qquad (23.6)$$

Чтобы найти среднее время ожидания, рассмотрим промежуток времени между моментом, когда появляется вход C_1, и моментом, когда он уходит из канала. Это общее *время очереди* равно времени ожидания w_1 плюс время занятия T_1. В момент, когда C_1 уходит из канала, очередь имеет длину n_1 (которая может равняться нулю). Среднее число входов в течение промежутка времени w_1+T_1 равно $m(w_1+T_1)$. Отсюда средняя длина очереди равна:

$$E(n_1)=m[E(w_1)+E(T_1)],$$

$$E(n)=mE(w)+\rho.$$

Подставляя вместо $E(n)$ его значение из (23.5) и решая уравнение, находим:

$$\rho+\frac{\rho^2+m^2\sigma_T^2}{2(1-\rho)}=mE(w)+\rho,$$

$$E(w)=\frac{\rho^2+m^2\sigma_T^2}{2m(1-\rho)}. \qquad (23.7)$$

Разделив (23.7) на $E(T)=1/M$, получаем следующую полезную формулу:

$$\frac{E(w)}{E(T)}=\frac{\rho^2+m^2\sigma_T^2}{2\rho(1-\rho)}=\frac{\rho}{2(1-\rho)}(1+M^2\sigma_T^2). \qquad (23.8)$$

Отношение (23.8) есть безразмерный показатель качества произвольной системы с очередями; оно выражает среднее время ожидания через число времен занятия. Эта величина представлена, например, на графике рис. 13.2.

Из общей формулы (23.8) мы можем вывести несколько качественных заключений. Поскольку величина в скобках всегда положительна, среднее время ожидания становится бесконечно большим, когда ρ стремится к 1. Далее, среднее время ожидания имеет наименьшее значение, когда $\sigma_T=0$, а это справедливо лишь тогда, когда времена занятия постоянны (т. е. не распределены случайным образом). Чтобы найти экономию, которую можно получить таким путем, предположим, что времена занятия распределены экспоненциально; тогда $\sigma_{2T}=1/M^2$ и среднее время ожидания равно

$$\frac{E(w)}{E(T)}=\frac{\rho}{1-\rho}, \qquad (23.9)$$

т. е. как раз вдвое больше, чем при постоянных временах занятия с тем же самым средним значением.

Можно показать [81], что, регулируя появление входов (а не времена занятия) так, чтобы интервал между ними был постоянным, можно получить примерно такое же уменьшение времени ожидания. Если нельзя применить ни один из этих двух способов, то остаются лишь два метода уменьшения времени ожидания: уменьшить ρ (уменьшив m и/или увеличив M) или ввести добавочные каналы.

Экспоненциальное распределение времен занятия, произвольное распределение входов. Были также выведены формулы для случая произвольного распределения входов и экспоненциального распределения времен ожидания. Тогда как в предыдущем случае мы нашли лишь средние значения, в этом случае были выведены формулы для распределений [130]. Эти формулы, вообще говоря, аналогичны выводимым ниже для пуассонова распределения входов, например (23.17), но вместо параметра ρ подставляется сложная функция.

Все предыдущее (и большая часть последующего) основано на некоторых допущениях,

которые должны быть проверены на практике, а именно что m, M и n независимы. Эти допущения разбираются в § 23.6.

Экспоненциальное распределение времен занятия, пуассоново распределение входов. В остальной части этой главы мы будем предполагать, если не оговорено противное, что распределение входов описывается формулами (5.20) и (5.27), а распределение времен занятия описывается, согласно изложенному в гл. 6, формулой

$$P(T)\,dT = Me^{-MT}dT. \qquad (6.34)$$

Длина очереди. Средняя длина определяется формулой (23.6), но мы хотим также знать распределение длин очереди. Для этого мы используем разностное уравнение, связывающее вероятности того, что очередь будет иметь определенную длину в момент t, и вероятность, что она будет иметь определенную длину в момент $t+dt$. В частности, чтобы очередь имела длину $n\,(n\neq 0)$ в момент $t+dt$, должно осуществляться одно из трех условий:

1) очередь имела длину $n-1$ в момент t, в течение интервала dt появился один вход и ни в одном канале не было закончено обслуживание;

2) очередь имела длину n в момент t, в течение интервала dt не появился ни один вход и ни в одном канале не было закончено обслуживание;

3) очередь имела длину $n+1$ в момент t, в течение интервала dt не появился ни один входной фактор, но в одном канале было закончено обслуживание.

Всякая другая мыслимая последовательность событий, приводящая к тому, что очередь имеет длину n в момент $t+dt$, включала бы по меньшей мере два события (например: два входа, или два выполненных обслуживания, или один вход и одно обслуживание) в течение бесконечно малого интервала, а согласно нашим допущениям при выводе распределения Пуассона, такие комбинации имеют нулевую вероятность. Вероятность появления одного входа в течение dt равна mdt; вероятность непоявления входов в течение dt равна $1-mdt$; вероятность выполнения одного обслуживания в течение dt равна Mdt; и вероятность невыполнения ни одного обслуживания в течение dt равна $1-Mdt$. Следовательно,

$$p(n;t+dt) = p(n-1;t)\,mdt\,(1-Mdt) + p(n;t)\,(1-mdt)\,(1-Mdt) + p(n+1;t)\,(1-mdt)\,Mdt, \quad n\neq 0, \qquad (23.10)$$

где мы использовали обозначение $p(n;t)$ для вероятности того, что линия имела длину n в момент t.

Раскрывая скобки и отбрасывая все члены с dt^2, получаем

$$p(n;t+dt) = p(n-1;t)\,mdt + p(n;t) - p(n;t)\,mdt - p(n;t)\,Mdt + p(n+1;t)\,Mdt, \quad n\neq 0.$$

Вычитая $p(n;t)$ и деля на dt, находим

$$\frac{p(n;t+dt)-p(n;t)}{dt} = mp(n-1;t) - (m+M)\,p(n;t) + Mp(n+1;t), \quad n\neq 0.$$

Если устремить dt к нулю, то левая сторона этого уравнения переходит в производную $dp(n)/dt$, согласно определению производной. Но эта производная ввиду (23.1) равна нулю. Решая уравнение относительно

$$p(n+1;t) = p(n+1),$$

получаем

$$p(n+1) = \frac{m+M}{M}\,p(n) - \frac{m}{M}\,p(n-1), \quad n\neq 0. \qquad (23.11а)$$

При $n=0$ первоначальное разностное уравнение переходит в следующее:

$$p(0;t+dt) = p(0;t)\,(1-mdt) + p(1;t)\,(1-mdt)\,Mdt,$$

откуда

$$\frac{p(0;t+dt)-p(0;t)}{dt} = -mp(0;t) + Mp(1;t) = 0$$

и

$$p(1) = \frac{m}{M}\,p(0). \qquad (23.11б)$$

Уравнения (23.11) можно также написать в виде:

$$p(n+1) = (\rho+1)\,p(n) - \rho p(n-1), \quad n\neq 0;$$
$$p(1) = \rho p(0).$$

Решение этого разностного уравнения мы находим из рассмотрения первых членов:

$$p(2) = (\rho+1)\,p(1) - \rho\,p(0) = (\rho^2+\rho)\,p(0) - \rho p(0) = \rho^2 p(0),$$
$$p(3) = (\rho+1)\,p(2) - \rho p(1) = (\rho^3+\rho^2)\,p(0) - \rho^2 p(0) = \rho^3 p(0),$$
$$p(n) = \rho^n\,p(n).$$

Значение $p(0)$ ввиду (23.3) равно $1-\rho$. Отсюда

$$p(n)=(1-\rho)\rho^n. \qquad (23.12)$$

Среднее этого выражения на основании (23.6) равно

$$\frac{\rho}{1-\rho}=\frac{m}{M-m}.$$

Дисперсия равна $\rho/(1-\rho)^2$.

Время ожидания. Как было указано выше, если времена занятия имеют экспоненциальное распределение, то окончания времен занятия в занятом канале распределены по закону Пуассона; иначе говоря, если канал занят, то вероятность того, что обслуживание закончится в течение интервала dt, равна Mdt. Следовательно, распределение числа выходов из непрерывно занятого канала определяется равенством (5.27):

$$p(R)=\frac{e^{-Mt}(Mt)^R}{R!}.$$

Распределение выходов из незанятого канала, очевидно, равно

$$p(R)=0, \quad R\neq 0.$$

Мы хотим теперь найти функцию плотности вероятностей $p_n(w)dw$ — условную вероятность того, что если при появлении какого-либо входа длина очереди равна n, то время ожидания этого входа будет больше w и меньше $w+dw$. Она равна произведению двух вероятностей: вероятности того, что в течение интервала w канал произведет точно $n-1$ выходов [эта вероятность определяется написанным выше выражением для $p(R)$, при $R=n-1$] и вероятности того, что в течение интервала dw канал произведет еще один выход (эта вероятность равна Mdw). Отсюда

$$p_n(w)\,dw=\frac{e^{-Mw}(Mw)^{n-1}}{(n-1)!}Mdw, \quad n\neq 0.$$

Вероятность того, что при появлении входа очередь имеет длину n, определяется из (23.12). Совместная вероятность двух событий, состоящих в том, что очередь будет иметь длину n и вход будет ожидать больше w и меньше $w+dw$, в силу (4.11) равна произведению этих вероятностей:

$$p(n, w)\,dw=(1-\rho)\rho^n\frac{e^{-Mw}(Mw)^{n-1}}{(n-1)!}Mdw, \quad n\neq 0.$$

Чтобы исключить из этого выражения n и получить плотность для w, мы суммируем его по всем значениям n от нуля до бесконечности. Очевидно, при $n=0$ время ожидания равно тождественно нулю.

$$p(w)\,dw=\sum_{n=1}^{\infty}(1-\rho)\rho^n\frac{e^{-Mw}(Mw)^{n-1}}{(n-1)!}Mdw=$$

$$=(1-\rho)\rho e^{-Mw}Mdw\sum_{n=1}^{\infty}\frac{\rho^{n-1}(Mw)^{n-1}}{(n-1)!}=$$

$$=(1-\rho)\rho Me^{-Mw}dw\sum_{h=0}^{\infty}\frac{(\rho Mw)^h}{h!},$$

где $h=n-1$. Сумма равна $e^{\rho Mw}$.

Следовательно,

$$p(w)\,dw=(1-\rho)\rho Me^{-Mw(1-\rho)}dw. \qquad (23.13)$$

Среднее этого выражения в силу (23.9) равно $\rho/(1-\rho)$ средним временам занятия (из которых каждое есть $1/M$), т. е. оно равно $\rho/M(1-\rho)=m/M(M-m)$.

Вероятность того, что вход будет ожидать по меньшей мере время W перед поступлением в канал, находим путем интегрирования (23.13) от W до бесконечности:

$$\int_W^{\infty}p(w)\,dw=\rho\int_{MW(1-\rho)}^{\infty}e^{-Mw(1-\rho)}\,d[Mw(1-\rho)]=$$

$$=\rho e^{-MW(1-\rho)}.$$

Так как (23.13) есть плотность вероятностей и w не может принимать отрицательные значения, то интеграл для (23.13) от нуля до бесконечности должен быть равен единице. Но если мы подставим в последнее выражение $W=0$, то получим ρ. Это вполне понятно. Вероятность того, что время ожидания равно нулю, в силу (23.3) равна $1-\rho$; сумма всех вероятностей, что время ожидания будет заключено между нулем и бесконечностью, должна быть поэтому равна ρ. Функция плотности вероятностей для времен ожидания имеет так называемую дельта-функцию* порядка

* Дельта-функцией $\delta(t-t_0)$ порядка γ называется функция, равная нулю при всех значениях t, кроме значения $t=t_0$, при котором она обращается в бесконечность:

$$\delta(t-t_0)=\left\{\begin{array}{l}0 \text{ при } t\neq t_0\\ \infty \text{ при } t=t_0\end{array}\right\},$$

причем выполняется следующее соотношение:

$$\int_{-\infty}^{\infty}\delta(t-t_0)\,dt=\gamma.$$

Прим. ред.

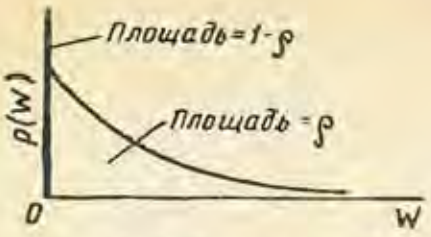


Рис. 23.2. Распределение времен ожидания.

$1-\rho$ при $w=0$, как показано на рис. 23.2. Сама функция плотности вероятностей в этой точке бесконечна, а площадь под этим бесконечно большим пиком нулевой ширины составляет $1-\rho$.

23.3. Множественные (параллельные) каналы

Во многих интересных случаях проектирования систем основным вопросом является число потребных каналов. Мы рассматриваем одиночную очередь, поступающую в группу каналов, причем передний клиент в линии ожидания перемещается в первый свободный канал. Как и раньше, мы считаем, что входы, поступающие в очередь, распределены по закону Пуассона, со средним количеством m входов в единицу времени; время занятия в каждом канале не зависит от других каналов (и от длины очереди) и распределено экспоненциально со средним, равным $1/M$. Длина очереди n включает всех клиентов в каналах и в линии ожидания. Если длина линии ожидания неотрицательна, мы будем обозначать ее через $j=n-\nu$, где ν — число каналов.

Как мы увидим, большинство уравнений, выведенных в § 23.2, можно прямо применить и в этом случае, сделав в них соответствующие изменения, обусловленные тем, что вероятность появления выхода в течение интервала dt зависит от длины очереди. Чтобы отметить это обстоятельство, мы всюду вместо M в (23.10) пишем M_n (или соответственно M_{n-1} и M_{n+1}). Впоследствии мы разберем случай, когда число входов в интервале dt также зависит от длины очереди, и поэтому вместо m в (23.10) мы пишем m_n (или m_{n-1} и m_{n+1}). Решая это уравнение, как и раньше, а именно, находя производную $dp(n)/dt$ и полагая ее равной нулю, мы получаем вместо (23.11):

$$p(n+1)=\frac{m_n+M_n}{M_{n+1}}\,p(n)-\frac{m_{n-1}}{M_{n+1}}\,p(n-1),\qquad n\neq 0; \tag{23.14а}$$

$$p(1)=\frac{m_0}{M_1}\,p(0). \tag{2.14б}$$

Эта более общая формула включает (23.11) как частный случай, который получается при $m_n=m$ и $M_n=M$.

Бесконечное число каналов. Допустим сначала, что $\nu=\infty$. Конечно, это не отражает никакой реальной ситуации, и линия ожидания здесь никогда не образуется, но задачу легко решить и получаются полезные выводы. Например, пусть мы нашли, что в некоторой ситуации с бесконечным числом каналов вероятность того, что будет применяться больше шести каналов, составляет только 0,01. Если мы теперь построим систему в точности с шестью каналами, то мы знаем, что: 1) если клиенты, не обслуженные немедленно, теряются, мы потеряем около 1% клиентов, и 2) если клиенты, не обслуженные немедленно, образуют очередь, линия ожидания будет существовать лишь около 1% времени.

В этом случае вероятность появления одного выхода в течение интервала dt равна $M_n dt=nMdt$, так как вероятность того, что в интервале dt освободятся два или больше каналов, состоит из бесконечно малых величин высшего порядка. Подставив $m_n=m_{n-1}=m$, $M_n=nM$ и $M_{n+1}=(n+1)M$ в (23.14), получим:

$$p(n+1)=\frac{m+nM}{(n+1)M}\,p(n)-\frac{m}{(n+1)M}\,p(n-1),\qquad n\neq 0;$$

$$p(1)=\frac{m}{M}\,p(0).$$

Эту систему можно решить итерационным методом, как в предыдущем случае:

$$p(2)=\frac{m+M}{2M}\,p(1)-\frac{m}{2M}\,p(0)=$$

$$=\frac{m+M}{2M}\,\frac{m}{M}\,p(0)-\frac{mM}{2M^2}\,p(0)=\frac{m^2}{2M^2}\,p(0),$$

$$p(3)=\frac{m+2M}{3M}\,p(2)-\frac{m}{3M}\,p(1)=$$

$$=\frac{m+2M}{3M}\,\frac{m^2}{2M^2}\,p(0)-\frac{m}{3M}\,\frac{m}{M}\,p(0)=$$

$$=\frac{m^3+2Mm^2}{6M^3}\,p(0)-\frac{2Mm^2}{6M^3}\,p(0)=\frac{m^3}{6M^3}\,p(0),$$

$$p(n)=\frac{m^n}{n!\,M^n}\,p(0)=\frac{\rho^n}{n!}\,p(0).$$

Для отыскания $p(0)$ суммируем по n от нуля до бесконечности:

$$\sum_{n=0}^{\infty}p(n)=1=p(0)\sum_{n=0}^{\infty}\frac{\rho^n}{n!}=e^{\rho}p(0).$$

Отсюда

$$p(0)=e^{-\rho}$$

и

$$p(n)=\frac{e^{-\rho}\rho^n}{n!}. \qquad (23.15)$$

Итак, длина очереди распределена по закону Пуассона, со средним значением $\rho=m/M$. Длина очереди в этом случае равна просто числу используемых каналов. Линия ожидания всегда имеет нулевую длину. Вероятность того, что потребуется больше чем данное число каналов (т. е. что они будут использоваться, если имеется бесконечное число каналов), можно определить из таблицы кумулятивных значений пуассонова распределения.

Конечное число каналов, длина очереди. Вероятность появления выхода из некоторого канала в течение интервала dt равна теперь $nMdt$, пока $n\leqslant\nu$, но становится νMdt, когда $n\geqslant\nu$. Следовательно, уравнения (23.16а) и (23.16б) тождественны с написанными раньше, но (23.16в) изменится:

$$p(1)=\frac{m}{M}p(0); \qquad (23.16а)$$

$$p(n+1)=\frac{m+nM}{(n+1)M}p(n)-\frac{m}{(n+1)M}p(n-1),$$

$$0<n<\nu; \qquad (23.16б)$$

$$p(n+1)=\frac{m+\nu M}{\nu M}p(n)-\frac{m}{\nu M}p(n-1),$$

$$n\geqslant\nu. \qquad (23.16в)$$

Уравнение (23.16б) имеет такое же решение, как и в предыдущем случае, а именно

$$p(n)=\frac{\rho^n}{n!}p(0), \quad 0\leqslant n\leqslant\nu. \qquad (23.17а)$$

Чтобы решить (23.16в), мы делим числители и знаменатели на M:

$$p(n+1)=\frac{\rho+\nu}{\nu}p(n)-\frac{\rho}{\nu}p(n-1), \quad n\geqslant\nu.$$

При $n=\nu$

$$p(\nu+1)=\frac{\rho+\nu}{\nu}p(\nu)-\frac{\rho}{\nu}\frac{\rho^{\nu-1}}{(\nu-1)!}p(0)=$$

$$=\frac{\rho+\nu}{\nu}p(\nu)-\frac{\rho^n}{\nu!}p(0)=\frac{\rho+\nu}{\nu}p(\nu)-p(\nu)=$$

$$=\frac{\rho}{\nu}p(\nu).$$

При $n=\nu+1$

$$p(\nu+2)=\frac{\rho+\nu}{\nu}p(\nu+1)-\frac{\rho}{\nu}p(\nu)=$$

$$=\frac{\rho+\nu}{\nu}\frac{\rho}{\nu}p(\nu)-\frac{\rho\nu}{\nu^2}p(\nu)=\frac{\rho^2}{\nu^2}p(\nu),$$

откуда

$$p(\nu+j)=\left(\frac{\rho}{\nu}\right)^j p(\nu)=\left(\frac{\rho}{\nu}\right)^{n-\nu}\frac{\rho^\nu}{\nu!}p(0)$$

и

$$p(n)=\frac{\rho^n}{\nu^{n-\nu}\nu!}p(0), \quad n\geqslant\nu. \qquad (23.17б)$$

Сумма вероятностей для всех значений n от нуля до бесконечности должна быть равна единице:

$$\sum_{n=1}^{\nu-1}\frac{\rho^n}{n!}p(0)+\sum_{j=0}^{\infty}\left(\frac{\rho}{\nu}\right)^j\frac{\rho^\nu}{\nu!}p(0)=1.$$

Вторая сумма бесконечна, если $\rho\geqslant\nu$ и $p(0)$ не равно нулю; таким образом, если число входов в единицу времени больше или равно числу обслуживаний в единицу времени, то средняя (ожидаемая) длина очереди бесконечно велика и вероятность того, что очередь будет иметь конечную длину, становится бесконечно малой. Но мы принимаем, как и раньше, что $\rho<\nu$:

$$p(0)\sum_{n=0}^{\nu-1}\frac{\rho^n}{n!}+p(0)\frac{\rho^\nu}{\nu!}\sum_{j=0}^{\infty}\left(\frac{\rho}{\nu}\right)^j=1.$$

Вторая сумма представляет теперь бесконечную геометрическую прогрессию, у которой первый член единица, а знаменатель ρ/ν. Напомним, что ее сумма равна $1/(1-\rho/\nu)$.

$$p(0)\left(\sum_{n=0}^{\nu-1}\frac{\rho^n}{n!}+\frac{\rho^\nu}{\nu!}\frac{1}{1-\rho/\nu}\right)=1,$$

$$p(0)=\frac{1}{\sum_{n=0}^{\nu-1}\rho^n/n!+(\rho^\nu/\nu!)[\nu/(\nu-\rho)]}. \qquad (23.18)$$

В любом частном случае, когда ρ и ν известны, можно найти численное значение (23.18) и после этого, применив (23.17), найти вероятность очереди любой данной длины, как в приведенном ниже примере. Чаще всего требуется найти вероятность того, что $n>\nu$ (т. е. вероятность того, что вход будет ждать). Вероятность того или иного времени ожидания нельзя определить из этих формул.

Время ожидания. Мы находим время ожидания по существу таким же образом, как и в предшествующих вычислениях времени ожидания, с учетом того, что у нас будет ν каналов вместо одного.

Вероятность того, что за время w закончится R обслуживаний, равна

$$p(R)=\frac{e^{-M\nu w}(M\nu w)^R}{R!},\quad n\geqslant\nu.$$

Нас не интересует случай, когда $n<\nu$, так как в этом случае время ожидания всегда равно нулю. Условная плотность вероятностей будет при этом равна

$$p_n(w)\,dw=\frac{e^{-M\nu w}(M\nu w)^{n-\nu}}{(n-\nu)!}M\nu dw.$$

Вероятность того, что линия имеет длину n, определяется из (23.17б):

$$p(n)=\left(\frac{\rho}{\nu}\right)^{n-\nu}\frac{\rho^\nu}{\nu!}p(0).$$

Совместная вероятность равна произведению этих двух выражений:

$$p(n,\ w)\,dw=$$
$$=p(0)\frac{\rho^\nu}{\nu!}\left(\frac{\rho}{\nu}\right)^{n-\nu}e^{-M\nu w}M\nu\frac{(M\nu w)^{n-\nu}}{(n-\nu)!}dw.$$

Суммируя по всем длинам линий $\geqslant\nu$, получаем

$$p(w)\,dw=$$
$$=p(0)\frac{\rho^\nu}{\nu!}e^{-M\nu w}M\nu dw\sum_{n=\nu}^{\infty}\left(\frac{\rho}{\nu}\right)^{n-\nu}\frac{(M\nu w)^{n-\nu}}{(n-\nu)!}.$$

Для вычисления суммы положим $n-\nu=h$ и получим

$$\sum_{h=0}^{\infty}\frac{(M\rho w)^h}{h!}=e^{M\rho w}.$$

Отсюда

$$p(w)\,dw=p(0)\frac{\rho^\nu}{\nu!}M\nu e^{-Mw(\nu-\rho)}dw,\qquad(23.19)$$

а среднее время ожидания равно

$$E(w)=\int_0^\infty wp(w)\,dw=$$
$$=p(0)\frac{\rho^\nu}{(\nu-1)!}M\int_0^\infty we^{-Mw(\nu-\rho)}dw=$$
$$=p(0)\frac{\rho^\nu}{(\nu-1)!}\frac{1}{M(\nu-\rho)^2}\int_0^\infty ze^{-z}dz,$$

где $z=Mw(\nu-\rho)$. Интеграл равен $\Gamma(2)=1$.

Итак,

$$E(w)=\frac{\rho^\nu p(0)}{(\nu-1)!\,M(\nu-\rho)^2}.\qquad(23.20)$$

Уравнение (23.19), как и (23.13), имеет дельта-функцию при $w=0$, а его интеграл от нуля до бесконечности не равен единице. Уравнение (23.20), как и (23.7), можно привести к безразмерному виду, разделив его на $E(T)$. В частных случаях оба эти выражения можно вычислить.

Пример. Прибытия и вылеты самолетов в аэропорту распределены по закону Пуассона с математическим ожиданием 60 самолетов в час. Время, в течение которого взлетно-посадочная дорожка занята, распределено экспоненциально с математическим ожиданием 2 *мин*. Сколько требуется взлетно-посадочных дорожек?

Решение. В этом случае $m=60$, $M=30$ и $\rho=2$. Следовательно, две взлетно-посадочные дорожки явно недостаточны, и мы вычислим распределения длин линии ожидания и среднее время ожидания для трех и четырех дорожек. Из (23.18) находим $p(0)=0{,}11$ в первом случае и $p(0)=0{,}13$ во втором. Это вероятности того, что ни одна дорожка не используется, и они мало отличаются. Из (23.17) мы вычисляем различные значения $p(n)$, приведенные в табл. 23.1.

Таблица 23.1

Вероятности числа ожидающих самолетов

Три дорожки:										
В очереди (включая самолеты на дорожках)	0	1	2	3	4	5	6	7	8	9
В линии ожидания	0	0	0	0	1	2	3	4	5	6
Вероятность	0,110	0,222	0,222	0,146	0,098	0,066	0,044	0,029	0,020	0,013
Четыре дорожки:										
В очереди (включая самолеты на дорожках)	0	1	2	3	4	5	6	7	8	9
В линии ожидания	0	0	0	0	0	1	2	3	4	5
Вероятность	0,130	0,260	0,260	0,174	0,087	0,044	0,022	0,011	0,005	0,003

Мы видим, что значения $p(n)$ в этих двух случаях весьма различны; в частности, вероятность очень длинной линии ожидания в одном случае во много раз больше, чем в другом. Среднее время ожидания, вычисленное из (23.20), равно соответственно 0,89 и 0,173 *мин*.

23.4. Каскадные (последовательные) каналы

Во многих случаях в сложных системах первичный вход должен пройти через целую серию устройств, каждое из которых действует как канал, в смысле настоящей главы. Перед большей частью каналов или перед всеми каналами образуются время от времени очереди, и нужно предусмотреть соответствующее хранение. Нас может интересовать длина очередей, время ожидания и другие характеристики очередей. Некоторые из каналов будут параллельны и могут иметь весьма различные распределения времен ожидания. Как указано ниже, такие задачи могут оказаться слишком трудными для аналитического решения. Но в одном случае целую цепочку последовательных каналов можно описать простыми выражениями в явном виде. Это случай, когда входы распределены по закону Пуассона, времена ожидания распределены экспоненциально и нет параллельных каналов.

Рассмотрим одиночный источник, одиночную очередь и одиночный канал, как в § 23.2. В силу (23.3) вероятность того, что канал не занят, равна $1-\rho$, а вероятность того, что канал занят, равна ρ. В первом случае вероятность появления выхода в течение интервала dt равна 0, а во втором случае вероятность появления выхода равна Mdt. Следовательно, общая вероятность появления выхода в течение интервала dt равна

$$p(R=1)=\rho Mdt+(1-\rho)\times 0=mdt.$$

Но если вероятность события одна и та же для любого интервала dt и эти вероятности для двух последовательных моментов независимы, то выходы распределены по закону Пуассона. Независимость выходов при входах, распределенных по закону Пуассона, обеспечивается экспоненциальным распределением времен занятия; если в некоторый момент появляется выход, нет оснований полагать, что в следующий момент не будет другого выхода (действительно, как отмечено в § 6.7, если в канале имеется клиент, его появление на выходе в следующий момент весьма вероятно). С другой стороны, если долго не было выхода, мы не имеем оснований ожидать его появления в следующий момент (как было также отмечено в § 6.7).

Это означает, что если у нас имеется источник входов, распределенных по закону Пуассона, со средним, равным m, и цепочка последовательных каналов, имеющих экспоненциально распределенные времена занятия, со средними, равными $1/M_1$, $1/M_2$, ..., $1/M_n$, причем каждое M больше m, то выходы последнего канала распределены подобно входам, а именно по закону Пуассона, со средним, равным m. Это утверждение может казаться удивительным лишь с первого взгляда. После установления стационарного состояния средняя частота выходов должна быть равна средней частоте входов; если моменты появления входов совершенно случайны и на них воздействует серия совершенно случайных процессов, то следует ожидать, что на выходе они будут совершенно случайны.

При этих условиях распределение длин очереди перед каждым каналом вычисляют из (23.12). Общее среднее время прохождения через систему равно сумме средних времен занятия $[\Sigma(1/M_i)]$ плюс сумма средних времен ожидания перед каждым каналом, из которых каждое равно (23.9). В длинной цепочке таких каналов может возникнуть сильный «эффект трубопровода», т. е. установление стационарного состояния может потребовать значительного времени.

23.5. Множественные входные источники

До сих пор мы рассматривали частоту входов, независимую от длины очереди. Это означает, что мы рассматривали либо одиночный пуассонов источник, либо бесконечное число независимых входных источников. Но во многих случаях число входных источников не намного больше числа каналов и источник не может создать новый вход, пока последний созданный им вход еще находится в очереди.

В качестве примера рассмотрим ремонт машин. Имеется фиксированное число машин, скажем β, из которых каждая при работе может поломаться с вероятностью mdt в течение интервала времени dt; эти поломки представляют входы. Имеется ν механиков, которые могут их чинить, и время, необходимое для починки, распределено экспоненциально, со средним, равным $1/M$; механики представляют каналы. Общее число работающих машин равно $\beta-n$, и вероятность появления нового входа в течение интервала dt прямо пропорциональна $\beta-n$ и равна $(\beta-n)mdt$. Вероятность того, что машина будет вновь пущена в ход в течение интервала dt, прямо пропорциональна числу ремонтируемых машин и равна меньшему из чисел $nMdt$ или νMdt.

Следовательно, в разностном уравнении (23.14) мы должны заменить m_{n-1}, m_n и m_{n+1} соответственно на $(\beta-n+1)m$, $(\beta-n)m$ и $(\beta-n-1)m$; M имеет то же значение, как и в случае многих каналов. Вместо уравнения (23.16) получается:

$$p(1)=\frac{\beta m}{M}p(0);$$

$$p(n+1)=\frac{(\beta-n)m+nM}{(n+1)M}p(n)-$$

$$-\frac{(\beta-n+1)m}{(n+1)M}p(n-1),\ 1\leqslant n<\nu;$$

$$p(n+1)=\frac{(\beta-n)m+\nu M}{\nu M}p(n)-$$

$$-\frac{(\beta-n+1)m}{\nu M}p(n-1),\ \nu\leqslant n<\beta.$$

Решая уравнение, как и раньше, итерационным методом, находим

$$p(2)=\frac{(\beta-1)m+M}{2M}p(1)-\frac{\beta m}{2M}p(0).$$

Но в этом случае удобнее подставить значение $p(0)$, а не $p(1)$, как мы делали раньше:

$$p(2)=\frac{(\beta-1)m+M}{2M}p(1)-\frac{\beta m}{2M}\frac{M}{\beta m}p(1)=$$

$$=\frac{(\beta-1)m}{2M}p(1),$$

$$p(3)=\frac{(\beta-2)m+2M}{3M}p(2)-\frac{(\beta-1)m}{3M}p(1)=$$

$$=\frac{(\beta-2)m+2M}{3M}p(2)-\frac{(\beta-1)m}{3M}\frac{2m}{(\beta-1)m}p(2)=$$

$$=\frac{(\beta-2)m}{3M}p(2),$$

$$p(n+1)=\frac{(\beta-n)m}{(n+1)M}p(n)=$$

$$=\frac{\beta!\,\rho^{n+1}}{(n+1)!\,(\beta-n-1)!}p(0),\ 0\leqslant n<\nu.$$

Продолжая итерацию получаем:

$$p(\nu+1)=\frac{(\beta-\nu)m+\nu M}{\nu M}p(\nu)-$$

$$-\frac{(\beta-\nu+1)m}{\nu M}p(\nu-1)=\frac{(\beta-\nu)m+\nu M}{\nu M}p(\nu)-$$

$$-\frac{(\beta-\nu+1)m}{\nu M}\frac{\nu M}{(\beta-\nu+1)m}p(\nu)=$$

$$=\frac{(\beta-\nu)m}{\nu M}p(\nu);$$

$$p(\nu+j+1)=\frac{(\beta-\nu-j)m}{\nu M}p(\nu+j)=$$

$$=\frac{(\beta-\nu)!\,\rho^{j+1}}{(\beta-\nu-j-1)!\,\nu^{j+1}}p(\nu),\ 0\leqslant j<\beta-\nu.$$

Суммируя эти вероятности от $p(0)$ до $p(\beta$ и положив сумму равной единице, находим $p(0)$.

Пример. В [78] приведены два поучительных численных примера. В обоих примерах $m/M=0{,}1$. В первом примере $\beta=6$ и $\nu=1$; во втором примере $\beta=20$ и $\nu=3$. Результаты вычислений приведены соответственно в табл. 23.2 и 23.3.

Таблица 23.2

Вероятность простоя машин — шесть машин, один ремонтный мастер (перепечатано из книги Феллера [78])

Неработающие машины (n)	Простаивающие машины, не находящиеся в ремонте	$p(n)$
0	0	0,48
1	0	0,29
2	1	0,15
3	2	0,06
4	3	0,02
5	4	0,00
6	5	0,00

Хотя во втором случае абсолютное число машин значительно больше и относительное число машин на одного мастера несколько выше, однако вероятность того, что данная машина будет простаивать и не будет ремонтироваться, меньше. Действительно, в первом случае она равна 0,055, а во втором она равна 0,017.

Таблица 23.3

Вероятность простоя машин — двадцать машин, три ремонтных мастера (перепечатано из книги Феллера [78])

Неработающие машины (n)	Простаивающие машины, не находящиеся в ремонте	Бездействующие мастера	$p(n)$
0	0	3	0,14
1	0	2	0,27
2	0	1	0,26
3	0	0	0,16
4	1	0	0,09
5	2	0	0,05
6	3	0	0,023
7	4	0	0,011
8	5	0	0,005
9	6	0	0,002
10	7	0	0,001

С другой стороны, мастера используются более эффективно: они заняты 0,515 времени в первом случае и 0,596 времени во втором. Это показывает, что выгодно предусмотреть дополнительное число работников для сглаживания случайных колебаний.

23.6. Состояние теории массового обслуживания

Ясно, что случай многих каналов сложнее, чем случай одного канала, и что он может встретиться во многих интересных задачах проектирования систем. Эти задачи часто бывают разрешимы только при том условии, что можно принять какое-нибудь упрощающее допущение; обычно оно состоит в том, что промежутки между входами постоянны или распределены по закону Пуассона, а времена занятия постоянны или распределены экспоненциально. К счастью, в очень многих случаях эти допущения вполне обоснованны.

Например, прибытия автомобилей к заставе для сбора пошлины обычно распределены по закону Пуассона, потому что отдельные машины независимы одна от другой. Но если застава находится вблизи длинного искривленного участка холмистой дороги, то легковые машины обычно скопляются за грузовиком; это нарушает предположение о независимости, и входы уже не будут распределены по закону Пуассона. Машины на длинном прямом участке дороги, вероятно, распределены по закону Пуассона, но около светофора это уже не будет так.

Белловская телефонная система нашла [33], что телефонные вызовы распределены по закону Пуассона. Вычислительные бюро также нашли, что поступающие к ним задачи распределены по закону Пуассона. Многие другие входы оказались распределенными по закону Пуассона [129].

Экспоненциальное распределение времен занятия несколько труднее обосновать. Во многих случаях уравнения теории массового обслуживания, выведенные на основе этого допущения, оказались справедливыми, хотя сами времена занятия и не исследовались; это относится, в частности, к упомянутым выше вычислительным бюро.

Белловская телефонная система нашла, что продолжительность большей части телефонных вызовов чрезвычайно хорошо описывается экспоненциальным распределением. Может вызвать удивление большое число очень коротких вызовов, но, по-видимому, их следует отнести за счет неверно набранных номеров, или за счет отсутствия вызываемых абонентов, или за счет недостаточного количества соединительного оборудования на низшей ступени. Однако междугородные разговоры большей частью заканчиваются чуть-чуть раньше 3 *мин*, и экспоненциальное распределение здесь неприменимо.

Во многих случаях вероятность чрезвычайно короткого или чрезвычайно долгого времени занятия меньше предсказанной экспоненциальным законом, но тем не менее экспоненциальный закон еще является достаточно хорошим приближением. Например, Джонсон [85] сообщает по данным Отдела исследования операций Британского министерства гражданской авиации: «В практических случаях встреченные распределения времен обслуживания оказались двух основных типов. Распределение посадок самолетов можно рассматривать как распределение Пирсона III рода, а распределение сообщений по линиям связи можно считать экспоненциальным с отрицательным показателем и с довольно крутым обрывом с обоих концов».

Распределение Пирсона III рода представляет собой весьма общее выражение для группы распределений вероятностей, включающее и экспоненциальное распределение. Эрланг 40 лет назад пришел к заключению, что время обслуживания распределено согласно закону Пирсона III рода [80]:

$$p(t)=\frac{(kM)^k}{\Gamma(k)}e^{-kMt}t^{k-1}dt,\quad 0<t<\infty,$$

который переходит в экспоненциальный закон при $k=1$ и соответствует постоянному времени занятия при $k\to\infty$.

Должны проверяться в каждом случае и некоторые другие из принятых допущений. Одно из них состоит в том, что каждый клиент сохраняет свое место в очереди. Если это не соблюдается, формулу для распределения времен ожидания нужно изменить. Кроме того, было сделано допущение, что частота появления входов и время занятия не зависят от длины очереди (за исключением особо отмеченных случаев множественных каналов и множественных входных источников). В некоторых случаях эти допущения приводят к нелепости.

Например, если бы все эти допущения принять для двухконечной очереди такси и пассажиров, то никогда не получится устойчивого решения: либо средняя частота прихода пассажиров больше частоты прихода такси и в этом случае очередь пассажиров растет безгранично; либо имеет место обратное положение и очередь такси растет безгранично; либо они в точности равны и в этом случае обе очереди могут расти безгранично. Но в действительности люди отходят прочь от слишком длинной очереди, так что частота входов по существу зависит от длины очереди. Точно так же, когда каналом является оператор-человек, он будет стремиться работать быстрее,

если видит большую очередь, так что время занятия также зависит от длины очереди. Однако при автоматической аппаратуре эти допущения, вероятно, более обоснованны.

Если даже нельзя проверить все эти допущения, имеется много случаев, когда решение, основанное на этих допущениях, дает картину, не очень далекую от действительной ситуации в очереди. Например, Малькольм [128] утверждает, что пуассоново распределение входов «достаточно точно аппроксимирует ситуацию» в случае, когда реальный вход (поломки машин) распределяется случайно (предположительно нормально) со средним временем между поломками 9 *час* и стандартным отклонением 1 *час* для каждой из 10 машин, а среднее время ремонта (занятия) одной машины равно 1 *час*.

Однако все же есть много случаев, которые не поддаются аналитическому исследованию. Например, Точер [104], описывая работу аэропорта, говорит:

«Аэропорт, с современной точки зрения, представляет собой систему взаимосвязанных очередей. Как только самолет успешно пройдет один узел, он становится клиентом для другой очереди, причем другие клиенты этого последнего узла прибывают сюда по различным другим путям. Таким образом, входы отдельных пунктов образования очередей составляются из выходов или частей выходов группы соседних пунктов. В Лондонском аэропорту, который использовался для этого исследования, имелось около 25 узлов, обслуживающих самолеты примерно на 30 различных направлениях. Через один и тот же узел могло пройти до четырех различных транспортных потоков. Это связано с четырехкратной сверткой.

Следующее затруднение состоит в корреляции между событиями, происходящими в разное время..., так как вероятности задержки данного самолета в последовательных узлах на его пути, вообще говоря, не независимы.

Чтобы подойти к задаче в общем плане, нужно было бы рассматривать систему как целое и вывести вероятность того, что любая данная группа очередей будет существовать в случайный момент. Она будет зависеть от времени продвижения самолета между узлами, и можно показать, что задача приведет к системе интегро-дифференциальных уравнений очень высокого порядка».

Моделирование. Для исследования таких задач превосходным орудием является быстродействующая электронная цифровая вычислительная машина. Хорошим методом может служить метод «Монте Карло», рассмотренный в § 10.3; две модели, приведенные в § 10.3 в качестве примеров, были модели очередей.

Для выполнения такого моделирования группа ячеек в запоминающем устройстве машины отбирается в качестве очереди для каждого элемента системы. Арифметическое устройство осуществляет логику очередей и каналов, исследуя их все последовательно,

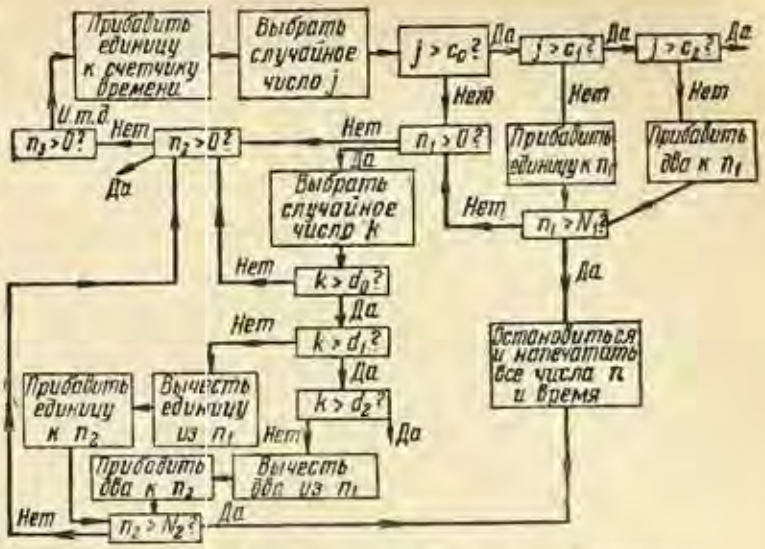


Рис. 23.3. Моделирование последовательных очередей и каналов.

а наступление случайных событий определяется случайными числами. Эта программа повторяется много раз, для каждого небольшого приращения реального времени задачи. На рис. 23.3 изображена типичная блок-схема.

На этом рисунке появление входа определяется случайным числом j. Если это j меньше некоторого числа c_0, то в течение этого интервала вход отсутствует; если оно лежит между c_0 и c_1, то в течение этого интервала появляется один вход; если оно лежит между c_1 и c_2, в течение этого интервала появляется два входа; и т. д. Значения c вычисляются заранее в соответствии с желаемым распределением. Аналогично, значения k определяют выходы первого канала (которые являются входами ко второй очереди и второму каналу), если первый канал занят.

Если длина первой очереди n_1 превосходит некоторое заданное число N_1, то положение выходит из нашей власти (например, переполнилось буферное хранилище) и процесс решения временно останавливается. Если входы распределены по закону Пуассона, а времена занятия распределены экспоненциально, то единственные сведения, хранимые в машине,—это длина очереди (значения n) и общее время задачи; если, например, времена занятия распределены нормально, то мы должны также запомнить время, которое находится в канале клиент, занимающий его в данный момент, так как от этого зависит вероятность появления выхода. В схеме нет параллельных каналов, однако легко распространить ее и на этот случай.

Составив подобные программы для каждого элемента, можно использовать их для построения модели большой системы. Затем на

основе модели можно исследовать интересующие нас величины и выполнить расчеты. При сложных задачах даже большие цифровые машины могут оказаться не в состоянии дать практические решения. В этом случае часто бывает необходимо принять упрощающие допущения. Если можно принять для некоторых величин пуассоновы и экспоненциальные распределения, то целые разделы задачи могут быть решены аналитически, и всю совокупность элементов можно заменить одной программой выборки случайных чисел.

Одно из затруднений при моделировании задачи с очередями состоит в том, что нас часто интересуют маловероятные события, а законы вероятности таковы, что для того чтобы получить сколько-нибудь надежные значения малых вероятностей, нужно провести очень большое число испытаний. Например, предположим, что некое событие имеет вероятность 0,001; мы хотим определить эту вероятность с высокой степенью уверенности в том, что мы не ошибемся более чем на $\pm 10\%$.

Высокая степень уверенности часто истолковывается в том смысле, что точки 3σ должны лежать в заданных границах (доверительная вероятность 99,7%). Число 0,001 равно k/n при биномиальном распределении; математическое ожидание этого распределения равно p, а дисперсия равна pq/n. Мы требуем, чтобы конечный результат лежал в пределах $d=0{,}0001$ от правильного ответа (0,001), и мы требуем, чтобы это отклонение равнялось 3σ. Отсюда

$$\sigma=\sqrt{\frac{pq}{n}}=\frac{d}{3},$$

$$n=\frac{9pq}{d^2}=\frac{9\times 0{,}001\times 1}{(0{,}0001)^2}\approx 10^6,$$

т. е. нам нужно провести миллион испытаний. Цифровые машины быстродействующие, но и они недостаточно быстры, чтобы выполнить миллион испытаний в приемлемое время, если каждое испытание продолжается достаточно долго.

Это показывает, что там, где возможно, желательно получить аналитические решения, если даже они применимы только к небольшим разделам задачи. Тем не менее метод моделирования применяется все больше и больше [16, 83, 84].

ЛИТЕРАТУРА

Для всякого, кто хочет вникнуть в теорию массового обслуживания, обязателен Кендалл [80]. Феллер [78] дал хороший разбор линий ожидания, а Фрай [33] — хороший разбор случаев, когда линии ожидания не образуются. Статьи по теории массового обслуживания начинают появляться довольно часто в технической литературе.

ЗАДАЧИ

23.1. Сравните среднюю длину очереди и среднее время ожидания для одиночного канала при экспоненциальном распределении времен занятия и при временах занятия, распределенных как хи-квадрат с одной степенью свободы. В обоих случаях нужно принять, что входы распределены по закону Пуассона с одним и тем же средним и что $p=0{,}5$. Для вычисления распределения χ^2 с одной степенью свободы подставить $\chi^2=T$ и $\nu=1$ в (7.27б).

23.2. Автомобили, прибывающие к пункту сбора пошлины на шоссе, распределены по закону Пуассона. Обслуживание автомобиля на каждой заставе занимает в среднем 5 *сек*, причем это время распределено экспоненциально. Инженеры-планировщики предсказали, что плотность движения не превысит 600 машин в час.

а) Определить необходимое число застав, чтобы вероятность образования линии ожидания была меньше 0,2.

б) Если построено столько каналов, как требуется в а), и предсказания оказались неправильными, а именно, интенсивность движения достигла 1500 машин в час, то какова вероятность образования линии ожидания?

ГЛАВА 24

СОСТЯЗАТЕЛЬНЫЕ АСПЕКТЫ. ТЕОРИЯ ИГР

Теория игр есть раздел математики. Подобно многим другим разделам математики, она включает большой класс задач, которые могут быть решены, и другой большой класс задач, которые не могут быть решены, по крайней мере сейчас. И подобно многим другим разделам математики, она аналогична некоторому классу задач реального мира. К несчастью, эта аналогия не всегда полная, и иногда нужно приложить определенные усилия, прежде чем мы сможем перевести реальную задачу в математические термины, с которыми мы умеем обращаться. Но это положение знакомо инженеру: прежде чем мы сможем применить такой простой раздел математики, как кинематика, мы должны сделать много нереалистических допущений и идеализаций, таких, как точечные массы и т. п.

Теория игр применима к таким практическим задачам, где наличествует столкновение

интересов и где участники до некоторой степени управляют исходом; все такие ситуации составляют предмет изучения теории игр. Противоположными интересами могут быть интересы игроков в игре, конкурентов в торговле, хозяина и служащих (столкновение интересов не обязательно должно быть полным) или противников в войне. В последнем случае приложения теории оказались ближе всего к теоретической модели, хотя теория была первоначально разработана для других конфликтов [38].

Может быть, лучше было бы назвать это учение *теорией конфликта;* многие авторы предпочитали длинное название *«теория стратегических игр»,* чтобы подчеркнуть, что ее предмет нетривиален. Но точно так же, как элементарные правила теории вероятностей легче всего вывести из задач, связанных с бросанием костей и вынниманием шаров из урн, элементарные правила теории игр легче всего вывести из игр. Всякая ситуация, содержащая конфликтные стороны, как например: закон против преступника, государство против налогоплательщика или конкурирующие промышленные корпорации, — имеет основу для применения теории игр, но теория игр на современной стадии развития редко применялась с пользой к другим системам, кроме военных. Зато в военных системах имеется много ситуаций, которые можно прояснить при помощи теоретико-игрового подхода: подводная лодка, старающаяся остаться необнаруженной, против патрульного самолета; бомбардировщик против перехватчика; выбор между простым и сложным криптографическим кодом; и весь вопрос о мерах, контрмерах и контрконтрмерах.

Когда в системе наличествуют состязательные аспекты, как во всех военных системах, система является особенно сложной. Однако состязательные аспекты можно выделить и изучить отдельно. Это предполагает, что интевые и нагрузочные аспекты системы уже хорошо понимаются, хотя, конечно, выводы теории игр могут несколько видоизменить их.

24.1. Определения

Класс задач теории игр, которые могут быть удовлетворительно решены, состоит в основном из так называемых *игр двух лиц с нулевой суммой.* Тем не менее это очень широкий класс, ибо «лицо» определяется как совокупность интересов. Так, почти всякую военную ситуацию можно рассматривать как игру двух лиц, в которых каждая из двух стран представляет собой «лицо». Термин «с нулевой суммой» относится к чистому платежу в игре (всякая конфликтная ситуация, изучаемая теорией игр, называется *игрой*): в игре должен быть какой-нибудь платеж, т. е. один или оба участника должны выиграть или проиграть в результате исхода какую-нибудь измеримую ценность, и если одно лицо выигрывает столько же, сколько проигрывает другое, игра называется *игрой с нулевой суммой.*

Большинство ситуаций, представляющих интерес, можно рассматривать как игры с нулевой суммой. Например, если город разрушен водородной бомбой, никто не получит реальной компенсации за причиненное разрушение; но мы можем с пользой исследовать тактику бомбоносителя (по отношению к противовоздушной обороне) теоретико-игровыми методами, предполагая, что разрушение города дает одной стране выгоду, точно эквивалентную убытку другой страны.

Игра с нулевой суммой не обязательно должна быть *справедливой,* т. е. ожидаемый доход от игры может быть положительным для одного игрока и отрицательным для другого.

Полная информация. В некоторых играх, как шахматы или шашки, каждому игроку полностью известна прошлая и настоящая ситуация; в других играх, как бридж и покер, имеются некоторые аспекты (карты противника), не известные в тот момент, когда игрок должен принять решение. Как мы увидим, это весьма существенное различие. Эти аспекты не нужно смешивать со случайными ходами, как жеребьевка в покере или результат бросания бомбы (когда распределение ошибок известно, но заранее не известно, будет ли попадание). Если настоящая ситуация и результаты прошлых ходов известны, то игра называется игрой с *полной информацией,* если даже исход будущих случайных ходов неизвестен. В действительности, как мы увидим, оптимальные стратегии часто намеренно вводят случайные исходы.

Партия, ход и стратегия. Мы применяли выражения *партия, ход и стратегия,* до сих пор не определив их. *Ход* есть момент партии, в котором у одного из игроков имеется ряд альтернатив; а его *выбор* есть та альтернатива, которую он выбирает (могут быть вырожденные случаи, когда выбор вынужден). *Партия* есть полная совокупность ходов обоих (или всех) игроков, после которой следует соответствующий платеж, зависящий от исхода (в вырожденных случаях платеж может быть нулем). *Правила* определяют, какие ходы закончены и какой будет производиться платеж по окончании партии. *Стратегия* есть полный

план ходов во всей партии, с учетом всякой возможной ситуации, вызванной случайностью или ходами противника.

Заметим, что некоторые из этих определений отличаются от используемых в обычной речи. Так, слово «игра» обычно употребляется в двух разных смыслах, которые мы здесь назвали «игра» и «партия»; а под словом «стратегия» обычно подразумевают искусный план, не обязательно подробный или учитывающий все альтернативы, тогда как в теории игр он может не быть искусным, но должен быть полным.

Цель теории игр — найти *оптимальную стратегию* (определяемую как стратегия, дающая игроку наибольшее математическое ожидание платежа) для одного или нескольких игроков и определить *цену игры* (т. е. ожидаемое значение платежа, если игроки применяют оптимальные стратегии).

Приводя какую-либо задачу по теории игр к математической форме, мы сперва перечисляем все возможные стратегии для каждого игрока и располагаем их в виде матрицы, как показано в табл. 24.1. Для удобства мы присваиваем одному игроку название «синие», и каждая его стратегия соответствует одной строке матрицы; другому игроку мы даем название «красные», и каждая его стратегия соответствует столбцу. В каждой точке матрицы мы будем указывать платеж в виде суммы, которую красные должны уплатить синим (если выигрывают красные, это число будет отрицательным).

Таблица 24.1

Платежная матрица для сравнения монет

Синие	Красные	
	Г	*Р*
Г	+1	−1
Р	−1	+1

Поскольку по определению стратегия является полной и мы перечислили все возможные стратегии, эта матрица описывает все возможные исходы. Если игра имеет ненулевую сумму, то должны быть две такие матрицы: одна, изображающая платеж синим, и другая, изображающая платеж красным; для игр с нулевой суммой одна матрица будет дополнением другой, и мы пишем лишь одну, которая указывает платеж синим.

Если игра имеет больше одного хода, то каждая стратегия должна учитывать все возможные перестановки ходов противником. Рассмотрим, например, игру, в которой синие могут выбрать *A* или *B*, после чего красные могут выбрать *C* или *D*; тогда игра заканчивается и производится платеж согласно правилам. Синие, очевидно, имеют две стратегии, а красные — четыре, хотя при одном ходе они имеют лишь два выбора. Четыре стратегии красных таковы:

C, если синие выбирают *A*, и *C*, если синие выбирают *B*;

C, если синие выбирают *A*, и *D*, если синие выбирают *B*;

D, если синие выбирают *A*, и *C*, если синие выбирают *B*;

D, если синие выбирают *A*, и *D*, если синие выбирают *B*.

В сложной игре вроде шахмат запись даже одной нетривиальной стратегии представляла бы колоссальную задачу, тем более запись всех стратегий; однако в принципе это можно сделать, и на практике часто оказывается возможным некоторое приближение к перечислению всех стратегий.

24.2. Минимакс, максимин и миниимаксный принцип

Табл. 24.1 изображает платежную матрицу для игры в сравнение монет; каждому игроку разрешается выбрать заранее, будет ли он ставить «на герб» или «на решку». Конечно, оба игрока показывают (или объявляют) свои выборы одновременно. Если синие выбирают то же, что и красные, они выигрывают монеты; если они выбирают другое, то проигрывают их. Это очень простая игра, в которой каждая партия состоит из одного хода для каждого игрока и у каждого игрока — лишь две стратегии. Но это не игра с полной информацией (так как игроки должны ходить одновременно и ни один не знает, что делает другой). По этой и другим причинам она не имеет простого решения в виде *чистой стратегии*, и, прежде чем вернуться к ней, мы рассмотрим некоторые более простые игры.

Таблица 24.2

Матрица 2×2 с седловой точкой

Синие	Красные		
	1	2	Минимумы строк
1	4	[2]	2*
2	2	1	1
Максимумы столбцов	4	2*	

Рассмотрим игру, платежная матрица которой изображена в табл. 24.2. Если угодно,

стратегию 1 можно рассматривать как выбор герба, а стратегию 2 — как выбор решки, и эту игру можно рассматривать как игру, в которой красные платят синим 4, 1 и 2 соответственно тому, совпадают ли их выборы на гербах, на решках или не совпадают. Однако матрица записана в весьма общем виде и может относиться ко многим другим играм. Игра несправедлива, но в действительной жизни состязательные ситуации не обязательно справедливы.

Будем считать теперь, что красные должны играть первыми, а синие будут играть вторыми, имея полную информацию. Такая игра называется *мажорантной игрой* для синих и *минорантной игрой* для красных. Красные могут выбрать любой из двух столбцов, но знают, что, какой бы столбец они ни выбрали, синие выберут строку, соответствующую максимальному платежу в этом столбце. Максимумы столбцов перечислены под матрицей.

Красные выберут столбец, отмеченный звездочкой, дающий наименьший из этих максимумов столбцов, или *минимакс*. Очевидно, оптимальная стратегия красных в их минорантной игре — свести к минимуму свой максимальный проигрыш. (Мы увидим, что критерий для игры во многие более сложные игры и, возможно, во все игры состоит в следующем: каждый игрок должен свести к минимуму свой максимальный проигрыш. Этот критерий называется *минимаксным принципом* и лежит в основе методов решения в теории игр; см. § 24.6). Цена мажорантной игры синих в этом случае равна 2.

Рассмотрим теперь минорантную игру синих. Они могут выбрать любую из двух строк, но знают, что красные выберут минимум строки, указанный справа от матрицы; поэтому они выбирают стратегию 1, которая дает им *максимин*.

Седловая точка. В действительной игре, где синие и красные должны играть одновременно, синие могут играть по меньшей мере столь же хорошо, как в своей минорантной игре с ценой 2, и не лучше, чем в своей мажорантной игре, которая также имеет цену 2.

Таким образом, синие имеют «чистую» оптимальную стратегию; они всегда применяют стратегию «синие-1». Они могут выиграть не меньше 2, а если красные сделают ошибку, синие могут выиграть 4. Точно так же, красные имеют чистую оптимальную стратегию; они всегда применяют стратегию «красные-2». Они могут проиграть не больше 2, а если синие сделают ошибку, красные могут проиграть лишь 1. Когда максимин равен минимаксу, говорят, что игра имеет *седловую точку;*

Таблица 24.3

Матрица 4×4 с седловой точкой

Синие	Красные				
	1	2	3	4	Минимумы строк
1	+4	−2	−3	0	−3
2	+1	−1	0	+3	−1
3	+2	[0]	+1	+3	0*
4	+3	−2	−4	0	−4
Максимумы столбцов	+4	0*	+1	+3	

в таблице 24.2 седловая точка обведена. Таблица 24.3 изображает более сложную игру, которая также имеет седловую точку; минимакс и максимин оба равны нулю, так что красные должны применять стратегию «красные-2», синие должны применять стратегию «синие-3» и платеж будет тогда всегда равен нулю.

Таблица 24.4

Платежная матрица для сравнения монет

Синие	Красные		
	1	2	Минимумы строк
1	1	−1	−1
2	−1	1	−1
Максимумы столбцов	1	1	

Понятие седловой точки имеет важное значение в теории игр. Когда платежная матрица имеет седловую точку, задача легко решается: оба игрока всегда должны применять стратегию, включающую седловую точку, и цена игры равна седловой точке. Очевидна аналогия с седловой точкой стереометрии — точкой, являющейся максимумом в одной плоскости и минимумом в другой, но эти понятия математически не тождественны. Игра может иметь несколько седловых точек с одинаковыми значениями (см. табл. 24.12).

Смешанные стратегии. К сожалению, не всякая платежная матрица имеет седловую точку; в самом деле, большая матрица, составленная из случайных чисел, почти наверное не будет ее иметь. Например, матрица табл. 24.1 не имеет седловой точки. Эта матрица воспроизведена в более общем виде в табл. 24.4, где указаны минимумы строк и

Таблица 24.5

Значения $E_{\text{син}}$

p	q / $2q-1$ / $2p-1$	1	0,9	0,8	0,7	0,6	0,5	0,4	0,3	0,2	0,1	0	Минимумы строк
		1	0,8	0,6	0,4	0,2	0	−0,2	−0,4	−0,6	−0,8	−0,8	
1	1	1	0,8	0,6	0,4	0,2	0	−0,2	−0,4	−0,6	−0,8	−1,0	−1,0
0,9	0,8	0,8	0,64	0,48	0,32	0,16	—	−0,16	−0,32	−0,48	−0,64	−0,8	−0,8
0,8	0,6	0,6	0,48	0,36	0,24	0,12	0	−0,12	−0,24	−0,36	−0,48	−0,6	−0,6
0,7	0,4	0,4	0,32	0,24	0,16	0,08	0	−0,08	−0,16	−0,24	−0,32	−0,4	−0,4
0,6	0,2	0,2	0,16	0,12	0,08	0,04	0	−0,04	−0,08	−0,12	−0,16	−0,2	−0,2
0,5	0	0	0	0	0	0	0	0	0	0	0	0	0
0,4	−0,2	−0,2	−0,16	−0,12	−0,08	−0,04	0	0,04	0,08	0,12	0,16	0,2	−0,2
0,3	−0,4	−0,4	−0,32	−0,24	−0,16	−0,08	0	0,08	0,16	0,24	0,32	0,4	−0,4
0,2	−0,6	−0,6	−0,48	−0,36	−0,24	−0,12	0	0,12	0,24	0,36	0,48	0,6	−0,6
0,1	−0,8	−0,8	−0,64	−0,48	−0,32	−0,16	0	0,16	0,32	0,48	0,64	0,8	−0,8
0	−1	−1	−0,8	−0,6	−0,4	−0,2	0	0,2	0,4	0,6	0,8	1,0	−1,0
Максимумы столбцов		1	0,8	0,6	0,4	0,2	0	0,2	0,4	0,6	0,8	1,0	

максимумы столбцов. Цена мажорантной игры синих равна +1, а цена их минорантной игры равна —1; цена самой игры должна лежать между этими значениями, и ввиду симметрии она должна быть равна нулю. Более существенно то, что ни одна из сторон не может применять чистую стратегию; например, синие должны применять в некоторых случаях стратегию 1, а в других случаях — стратегию 2; красные также должны применять *смешанную стратегию.*

Оказывается, существует оптимальный способ играть в эту игру, основанный на минимаксном принципе; иначе говоря, каждый игрок может найти стратегию, которая даст минимум его максимальных проигрышей (или даст максимум его минимальных выигрышей, что одно и то же). Далее, минимакс для одного игрока равен максимину для другого игрока (и равен цене игры), так что игра имеет устойчивое решение. Можно доказать, что можно высказать такое же положение о всякой игре двух лиц с нулевой суммой; это действительно является основной теоремой теории игр.

Допустим в качестве иллюстрации, что синие решают применять стратегию 1 с вероятностью p и стратегию 2 с вероятностью $1-p$, а красные соответственно применяют свои стратегии 1 и 2 с вероятностями q и $1-q$. Тогда вероятность того, что выиграют синие, равна $pq+(1-p)(1-q)$, а вероятность того, что они проиграют, равна $p(1-q)+q(1-p)$. Умножая первое выражение на +1, а второе на —1, получаем математическое ожидание выигрыша синих

$$E_{\text{син}} = pq+(1-p)(1-q)-p(1-q)-q(1-p) = (2p-1)(2q-1). \qquad (24.1)$$

Из этого выражения можно найти оптимальные стратегии; они, очевидно, состоят из $p=1/2$ и $q=1/2$, и каждая из них дает ожидаемый платеж 0. Ситуация поясняется табл. 24.5, в которой приведены ожидаемые исходы для различных значений p и q. Если каждый из игроков применяет свою оптимальную смешанную стратегию (случайный выбор герба и решки того и другого с вероятностью 0,5), то не имеет значения, что делает другой игрок. Это всегда справедливо для матриц 2×2 без седловых точек; когда в больших матрицах один игрок применяет свою оптимальную смешанную стратегию, он всегда получит по меньшей мере цену игры и в некоторых случаях может получить больше, если другой игрок не применяет своей оптимальной смешанной стратегии.

Следует отметить, что при обычном сравнении монет каждый игрок осуществляет случайный выбор гербов и решек (с вероятностью 0,5) путем бросания монет; они поступают так не ради справедливости, но для собствен-

ной защиты. Нужно также заметить (и это является общим положением), что противник не может использовать знание стратегии другого игрока, пока он не знает, какой выбор будет сделан на данном ходе. Это и следовало ожидать, так как нужно полагать, что противник достаточно умен, чтобы самому решить теоретическую задачу игры и определить таким образом оптимальную стратегию другого игрока.

Более крупные игры. Всякую игру с конечным числом стратегий можно привести к форме прямоугольной матрицы, как в таблицах этой главы; матрица не обязательно должна быть квадратной, так как у одного игрока может быть больше стратегий, чем у другого. Для таких игр, как бридж, шахматы или даже шашки, было бы безнадежно пытаться перечислить все стратегии, так как каждая стратегия включает каждый ход, который будет сделан в игре, и учитывает все возможные комбинации ходов противника и природы (т. е. случайные ходы, как в карточных играх).

Одним из более простых примеров является «трехпальцевая морра»*. В этой игре двух лиц оба игрока показывают одновременно один, два или три пальца и в то же время объявляют числа один, два или три. Таким образом, у каждого игрока имеется девять стратегий, указанных в табл. 24.6 (например, стратегия 12 означает: показать один палец и объявить «два»); имеется 81 возможный исход. Платеж определяется следующим правилом: если один и только один игрок объявляет число, равное числу пальцев, показанных его противником, то ему уплачивается сумма, равная общему числу показанных пальцев; в других случаях платеж равен нулю.

Из платежной матрицы (табл. 24.6) можно видеть, что максимин равен —3, минимакс равен +3 и седловой точки нет. Цена игры, очевидно, равна нулю. Оптимальная стратегия — применять стратегии 13, 22 и 31 с вероятностями соответственно 5/12, 4/12 и 3/12. Легко проверить, что если синие применяют эту оптимальную смешанную стратегию, а красные — чистые стратегии 12, 13, 22, 31 или 32 или любое их сочетание, то ожидаемый платеж синим равен нулю, а если красные применяют стратегии 11, 21, 23 или 33 или какую-нибудь смешанную стратегию, включающую одну или несколько из этих стратегий с ненулевой вероятностью, то ожидаемый доход синих будет положителен.

* Под таким названием она издавна известна в Италии, хотя, по-видимому, эта игра практикуется и в других странах. — *Прим. ред.*

Таблица 24.6

Платежная матрица трехпальцевой морры

Синие	Красные									
	11	12	13	21	22	23	31	32	33	Минимумы строк
11	0	2	2	—3	0	0	—4	0	0	—4
12	—2	0	0	0	3	3	—4	0	0	—4
13	—2	0	0	—3	0	0	0	4	4	—3*
21	3	0	3	0	—4	0	0	—5	0	—5
22	0	—3	0	4	0	4	0	—5	0	—5
23	0	—3	0	0	—4	0	5	0	5	—4
31	4	4	0	0	0	—5	0	0	—6	—6
32	0	0	—4	5	5	0	0	0	—6	—6
33	0	0	—4	0	0	—5	6	6	0	—5
Максимумы столбцов	4	4	3*	5	5	4	6	6	5	

Вильямс [62], у которого мы взяли описание этой игры и ее решение, говорит: «Нам кажется, что стóит научить этой игре ваших друзей, так как решение легко запомнить, но его трудно отыскать по интуиции». Но он указывает также, что метод, примененный им для отыскания решения, он выбрал «по причинам, имеющим отношение к теории чисел, и имея некоторые понятия об ответе». Как показано ниже, имеются конструктивные методы решения таких игр, но они в лучшем случае утомительны, и если размер матрицы не очень мал, для них нужна вычислительная машина.

24.3. Методы решения конечных игр двух лиц с нулевой суммой

Всякая игра с конечным числом ходов и конечным числом выборов на каждом ходе имеет конечное число чистых стратегий. Мы можем описанным выше методом привести любую игру двух лиц с нулевой суммой, состоящую из конечного числа чистых стратегий, к следующей *нормальной* форме [63]: «Первый игрок выбирает число из первых m натуральных чисел, а второй игрок, не зная, какой выбор сделал первый игрок, выбирает число из n первых натуральных чисел. Затем два числа сравниваются, и один из игроков платит другому сумму, зависящую от произведенных выборов и указываемую правилами игры» *.

Первый шаг решения игры состоит в том, что ее приводят к нормальной форме и затем ищут в матрице седловую точку; это самое важное, так как некоторые методы решения приведут к ошибочным решениям, если игра не имеет седловой точки. Если матрица имеет седловую точку, то у каждого игрока есть ука-

* Эквивалентную нормальную форму игр n лиц с ненулевой суммой см. в [38]. — *Прим. авт.*

Рис. 24.1. Графическое решение матрицы игры.

занная оптимальная чистая стратегия, цена игры равна седловой точке и игра решается. Если матрица имеет более одной седловой точки, то один или оба игрока имеют более одной оптимальной стратегии и могут применять одну из этих чистых стратегий или любое их сочетание.

Можно доказать, что все игры с полной информацией имеют седловые точки; к таким играм относятся, например, шахматы, но мы не в состоянии перечислить все стратегии и построить матрицу для шахмат и поэтому не знаем оптимальной стратегии (или стратегий).

Матрицы 2×2. Если матрица не имеет седловой точки, то решение игры может оказаться трудным. Рассмотрим матрицу из табл. 24.7. Минимакс и максимин не равны, и поэтому игра не имеет седловой точки. Поэтому синие должны будут применять смешанную стратегию. Допустим, что они применяют стратегию «синие-1» с вероятностью p. Тогда их выигрыш в том случае, если красные применяют стратегию «красные-1», определяется одной прямой на рис. 24.1, а выигрыш в том случае, если красные применяют стратегию «красные-2», определяется другой прямой.

Минимальный выигрыш синих для любого значения p определяется жирными прямыми. Если синие выбирают p в пересечении этих двух прямых, то они получат максимум своего минимального выигрыша при обеих чистых стратегиях красных и при любой смешанной стратегии красных. Следовательно, эта точка определяет оптимальную смешанную стратегию синих ($p=0{,}6$) и также указывает цену игры ($v=1{,}6$). Аналогичная пара прямых определяет оптимальную смешанную стратегию красных ($q=0{,}2$) и дает ту же цену игры.

Аналитический метод. Вместо этого графического способа можно применить простой аналитический метод. Стратегия «синие-1» должна применяться с вероятностью (в обозначениях табл. 24.10)

$$|a_{21}-a_{22}|/(|a_{11}-a_{12}|+|a_{21}-a_{22}|).$$

Стратегии красных можно определить по тому же правилу; например, стратегия «красные-2» применяется с вероятностью

$$|a_{11}-a_{21}|/(|a_{11}-a_{21}|+|a_{12}-a_{22}|).$$

Заметим, что числа в первой строке (или столбце) определяют частоту выбора второй строки (или столбца) и наоборот. Этот метод не годится, если матрица имеет седловую точку.

Таблица 24.7

Матрица без седловой точки

Синие	Красные		
	1	2	Минимумы строк
1	0	2	0
2	4	1	1*
Максимумы столбцов	4	2*	

Матрицы $n\times m$. Если матрица имеет порядок 2 на 3, а не 2 на 2, то рис. 24.1 нужно было бы заменить аналогичным двумерным графиком с тремя прямыми, а соответствующим графиком для стратегий красных будет трехмерный график с двумя плоскостями (как на рис. 25.3). Для большой матрицы график будет представлять собой гипермногогранник (гиперполиэдр) в n-мерном пространстве. Такие графики, конечно, нельзя использовать для отыскания решения, но они полезны для понимания рассмотренных ниже свойств решения.

Доминирование. Иногда большие матрицы можно привести к малым. Так, в игре, изображенной в табл. 24.8, синим не следует никогда применять стратегию «синие-3», потому что над этой стратегией *доминирует* стратегия «синие-2»; иначе говоря, синие всегда получат лучший исход, применяя стратегию «синие-2», чем если бы они применяли стратегию «синие-3», независимо от действий красных. На основании тех же рассуждений красным не следует никогда применять стратегию «красные-3», так как они всегда могут получить по меньшей мере столь же хороший исход, применяя *подчиненную* (доминируемую) стратегию «красные-2». Следовательно, стратегии «синие-3» и «красные-3» можно исключить из матрицы, и она сведется тогда к матрице 2×2 в табл. 24.7, которую мы уже решили.

Можно также исключить из матрицы строки (или столбцы), если над ними доминирует соответствующая смесь других строк (или они доминируют над смесью других столбцов). Так, в игре, изображенной в табл. 24.9, синие никогда не будут приме-

Таблица 24.8

Платежная матрица с подчиненными и доминирующими стратегиями

Синие	Красные			
	1	2	3	Минимумы строк
1	0	2	2	0
2	4	1	2	1*
3	3	0	0	0
Максимумы столбцов	4	2*	2*	

нять стратегию «синие-3», потому что они всегда могут получить столь же хороший исход, применяя соответствующую смесь (половину и половину) стратегий «синие-1» и «синие-2». Следовательно, стратегию «синие-3» можно исключить из матрицы, которая сведется тогда к табл. 24.7.

Таблица 24.9

Платежная матрица с подчиненной стратегией

Синие	Красные	
	1	2
1	0	2
2	4	1
3	1,5	1,5

Решение при помощи системы неравенств. Рассмотрим обобщенную матрицу из табл. 24.10. Основная теорема теории игр говорит нам, что у синих имеется оптимальная стратегия, а именно множество вероятностей $p_1, \ldots, p_n$. Если синие применяют стратегию этой смеси, а красные применяют свою оптимальную стратегию $(q_1, \ldots, q_m)$, то математическое ожидание выигрыша синих будет равно v; далее, если красные применяют какую-нибудь другую стратегию и, в частности, любую

Таблица 24.10

Обобщенная платежная матрица

Синие	Красные					
	1	2		j		m
1	a_{11}	a_{12}		a_{1j}		a_{1m}
2	a_{21}	a_{22}		a_{2j}		a_{2m}
i	a_{i1}	a_{i2}		a_{ij}		a_{im}
n	a_{n1}	a_{n2}		a_{nj}		a_{nm}

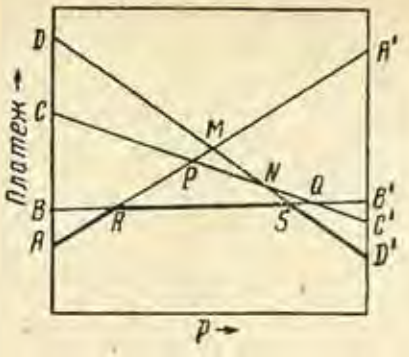


Рис. 24.2. Графическое решение матрицы игры 2×4.

чистую стратегию, то математическое ожидание выигрыша синих будет равно по меньшей мере v.

Обратно, красные, применяя свою оптимальную стратегию, получают математическое ожидание $-v$, если синие применяют свою оптимальную стратегию, и по меньшей мере $-v$, если синие применяют любую другую стратегию.

Это приводит к следующим неравенствам: n неравенствам вида

$$\sum_{j=1}^{m} q_j a_{ij} \leqslant v, \tag{24.2}$$

m неравенствам вида

$$\sum_{i=1}^{m} p_i a_{ij} \geqslant v, \tag{24.3}$$

n двойным неравенствам вида

$$0 \leqslant p_i \leqslant 1 \tag{24.4}$$

и m двойным неравенствам вида

$$0 \leqslant p_j \leqslant 1. \tag{24.5}$$

Имеются также два уравнения

$$\sum_{i=1}^{n} p_i = 1, \quad \sum_{j=1}^{m} q_j = 1. \tag{24.6}$$

Можно решить эту систему с $m+n+1$ неизвестными (p_i, q_j и v), но можно применить более удобные методы отыскания решения.

Общий метод решения. На рис. 24.2 показан доход синих в игре 2×4, если синие применяют одну из своих двух стратегий с вероятностью p, а красные — любую из своих четырех стратегий (AA', BB', CC' и DD'). Как и раньше, минимальный выигрыш синих изображается ломаной жирной линией ($ARSD'$),

а решение определяется точкой максимума на этой линии, а именно S. Если бы была большая матрица, скажем $n \times m$, то график состоял бы из m гиперплоскостей в n-мерном пространстве. Но можно показать, что минимальный выигрыш синих по-прежнему будет изображаться выпуклой поверхностью и решение (или решения, см. § 24.6) будет определяться наивысшей точкой (или точками) на этой поверхности. Ставится задача — найти эту точку аналитически.

Один метод таков: исключить всюду, где возможно, подчиненные стратегии (как стратегию CC' на рис. 24.2). Это приводит к уменьшению вершин, подлежащих исследованию; на рис. 24.2 исключается пять вершин и остается девять.

Следующий метод — исследовать систематически каждую вершину на графике (или оставшиеся после устранения подчиненных стратегий). Это можно сделать аналитически путем систематического исследования всех квадратных подматриц матрицы $n \times m$; в матрице 2×4 имеются 14 квадратных подматриц (шесть 2×2 и восемь 1×1), соответствующих 14 вершинам рис. 24.2. Так, приняв $m > n$, мы исследуем сперва все матрицы $n \times n$, которые можно составить, исключив $m - n$ столбцов; затем все матрицы $(n-1) \times (n-1)$, которые можно составить исключив одну строку и $m - n + 1$ столбцов; и т. д.

При проведении этого исследования неравенства (24.2) и (24.3) заменяются равенствами. Тогда мы получаем достаточное число уравнений для того, чтобы отыскать неизвестные, не используя (24.4) и (24.5); эти ограничения используются для того, чтобы определить, действительно ли решение этой системы уравнений является решением квадратной «подыгры». * Если это так, то проверяют, является ли оно решением первоначальной игры. Решение найдено, если: а) синие получают хотя бы некоторый доход (цену игры), применяя свою оптимальную смешанную стратегию против любой чистой стратегии красных, и получают в точности эту цену против оптимальной смешанной стратегии красных, и б) красные получают хотя бы эту цену со знаком минус, применяя свою оптимальную смешанную стратегию против любой чистой стратегии синих.

Этот метод является систематическим, конструктивным и наверняка приводит к успеху, но, разумеется, утомителен и при большой матрице может быть неосуществимым даже для большой цифровой вычислительной машины, так как число вершин быстро увеличивается при увеличении размера матрицы. Существуют другие способы, позволяющие исследовать меньшее число вершин, хотя необходимы некоторые вычисления для выбора надлежащих вершин, которые нужно исследовать.

Например, так называемый *метод двойного описания* дает возможность исследовать лишь те вершины, которые находятся на выпуклой поверхности (A, R, S, D' на рис. 24.2). В так называемом *симплекс-методе* число вершин еще меньше. Сначала берут вершину, лежащую на выпуклой поверхности, и затем переходят последовательно к другим вершинам на поверхности, причем каждая последующая находится на такой же или большей высоте, чем предыдущая. Кроме того, известно, когда достигнуто решение игры.

Метод приближения. Существует, кроме того, конструктивный метод отыскания приближенных решений матриц игры. Этот метод настолько прост, что его можно выполнить вручную при матрицах небольшого размера и можно запрограммировать для вычислительной машины при матрицах почти любого размера, если только машина имеет достаточную емкость памяти для хранения всей матрицы. Метод основан на вычислениях, эквивалентных последовательности партий, в которой каждый игрок: а) всегда действует в предположении, что противник должен играть в будущем совершенно так же, как в прошедшем, и б) проводит свою очередную партию так, чтобы получить при этом предположении максимальный платеж. Первая партия, очевидно, произвольна. Допустим, что применяется стратегия «синие-1». Возможны также другие случаи, в которых у игрока будет несколько выборов, и для определенности мы предположим, что он выбирает стратегию с наименьшим номером.

Вычисления проводятся следующим образом (табл. 24.11). Отмечаем звездочкой стратегию «синие-1», выбранную по нашему произвольному правилу в первой партии. Поскольку она соответствует первой строке, переписываем первую строку внизу под матрицей и отмечаем звездочкой наименьшее число (1). Поскольку оно в первом столбце, переписываем первый столбец справа от матрицы и ставим звездочку у наибольшего числа (4). Поскольку оно во второй строке, складываем вторую строку со строкой под матрицей и отмечаем звездочкой наименьшее число (2). Поскольку оно во втором столбце, складываем второй столбец со столбцом справа от матрицы и отмечаем звездочкой наибольшее

* Подыгра — игра, составляющая часть другой игры. — *Прим. ред.*

Таблица 24.11

Метод отыскания приближенного решения для матрицы без седловой точки

	1	2	3	*	1	3	6*	8*	9	12*	13	15*	17	20*	21*	22	24	27*	30*	31*	32	34*	36*	37
	4	0	1		4*	4	5	5	9	10	14*	14	14	15	19	23*	23	24	25	29	33*	33	33	37
	2	3	0		2	5*	5	8	10*	10	12	15	18*	18	20	22	25*	25	25	27	29	32	35	37
1	1*	2	3		4	5/2	6/3	8/4	10/5	12/6	14/7	15/8	18/9	20/10	21/11	23/12	25/13	27/14	30/15	31/16	33/17	34/18	36/19	37/20
2/2	5	2*	4																					
4/3	7	5	4*																					
7/4	8	7*	7																					
9/5	9*	9	10																					
10/6	11	12	10*																					
12/7	12*	14	13																					
14/8	16	14*	14																					
16/9	17	16*	17																					
17/10	19	19	17*																					
20/11	20*	21	20																					
21/12	21*	23	23																					
23/13	25	23*	24																					
24/14	27	26	24*																					
27/15	28	28	27*																					
29/16	29*	30	30																					
30/17	30*	32	33																					
32/18	34	32*	34																					
34/19	35	34*	37																					
36/20	36*	36	40																					

число (5). Поскольку оно в третьей строке, складываем третью строку с самой нижней строкой и отмечаем звездочкой наименьшее число (4). Затем складываем третий столбец с самым правым столбцом и отмечаем звездочкой наибольшее число (6). Затем складываем первую строку с самой нижней строкой, и поскольку здесь имеются два наименьших числа, применяем наше произвольное правило и отмечаем звездочкой число во втором столбце.

Этот процесс продолжается как угодно долго, в зависимости от желаемой степени приближения. Мы остановимся («имея некоторые понятия об ответе») после 20 шагов. Теперь мы определяем вероятности, считая число звездочек в каждой строке и каждом столбце. В первой, второй и третьей строках соответственно 12, 4 и 4 звездочек. Поэтому стратегии «синие-1», «синие-2» и «синие-3» нужно применять с вероятностями соответственно 12/20, 4/20 и 4/20 (точные оптимальные стратегии суть соответственно 11/20, 4/20 и 5/20). Вероятности для стратегий «красные-1», «красные-2» и «красные-3» равны соответственно 8/20, 7/20 и 5/20 (что совпадает с точными оптимальными стратегиями).

Следует отметить, что после 20 шагов редко достигается такое хорошее приближение. Кроме того, последовательность флюктуирует (на 21-м шаге стратегии синих и красных удалились бы от правильного результата), но она сходится, если у каждого игрока одна оптимальная стратегия.

Слева от чисел, написанных под матрицей, и снизу от чисел, написанных справа от матрицы, записаны отмеченные звездочкой числа, деленные на номер шага. Можно показать, что каждое из этих чисел указывает соответственно нижнюю и верхнюю грани цены игры. Наибольшая нижняя грань равна $20/11 = 1{,}818$, а наименьшая верхняя грань равна $37/20 = 1{,}85$. Точная цена игры равна 1,85 (опять-таки такое хорошее приближение при столь малом числе шагов необычно).

24.4. Бесконечные игры

На рис. 24.3 изображена состязательная ситуация [11] между патрульным самолетом и подводной лодкой. Подводная лодка должна пройти по указанному проливу из одного водного пространства в другое. Она может оставаться под водой бо́льшую часть времени, но пролив такой длинный, что когда-нибудь она должна выйти на поверхность для зарядки аккумуляторов. Ее стратегия—это та точка пролива, в которой она хочет всплыть на поверхность. Патрульный самолет совершает заградительный облет перпендикулярно к оси пролива и может обнаружить подводную лодку, если она покажется на поверхности в пределах 5 миль от самолета; во всех других случаях подводная лодка ускользнет необнаруженной. Стратегия самолета состоит в выборе положения заградительной линии.

Рис. 24.3. Игра на континууме.

Таким образом, каждому игроку нужно выбрать точку на оси x, и, следовательно, у каждого игрока имеется бесконечное число (в действительности несчетное * множество) стратегий.

Платеж самолету можно считать равным $+1$ при обнаружении лодки и -1 при необнаружении. Предполагается, что форма пролива описывается простой функцией. Подводной лодке следовало бы выбрать широкие участки пролива, где у патрульного самолета меньше вероятности ее обнаружить; с другой стороны, самолет может уделить поискам здесь больше времени и т. д. Иными словами, перед нами типичная задача теории игр, и ее решение отнюдь не очевидно.

Эту игру можно привести к нормальной форме (§ 24.3), заменив слова «первые m натуральных чисел» словами «интервал от x_1 до x_2» и соответственно для второго игрока. Чистая стратегия будет состоять в выборе одного значения x, а смешанная стратегия будет представлять собой плотность вероятностей для x. Игра может иметь седловую точку, и в этом случае у обоих игроков будут оптимальные чистые стратегии. В общем случае седловой точки может не быть, и нам нужно найти функцию, дающую максимальный платеж. Задача решается с помощью вариационного исчисления; подробности читатель может найти в учебниках по теории игр. Здесь доста-

* Счетным называется множество, элементы которого можно перенумеровать натуральными числами от 1, 2, 3 и т. д. до бесконечности; рассматриваемое множество стратегий несчетно потому, что множество всех точек в любом непрерывном геометрическом многообразии (линии, поверхности и т. д.) гораздо обширнее множества натуральных чисел и не может быть перенумеровано последним. — *Прим. ред.*

точно сказать, что эта игра и многие другие игры двух лиц с нулевой суммой и с бесконечным числом стратегий могут быть решены.

24.5. Игры с ненулевой суммой и игры n лиц

В игре двух лиц с ненулевой суммой у каждого игрока имеется стратегия, посредством которой он добивается максимального выигрыша в игре против несотрудничающего противника, и имеется также пара стратегий, посредством которых два сотрудничающих игрока максимизируют общую сумму своих выигрышей. Если эти стратегии совпадают, никакой задачи не возникает. Если они не совпадают, то это значит, что нужно будет разделить какие-то доходы, если игроки решают сотрудничать, и задача состоит в том, как разделить эти доходы.

Предположим, например, что если каждый игрок стремится обеспечить себе максимум против любой возможной стратегии своего противника, то минимакс синих равен —2 и минимакс красных равен +2, а если они сотрудничают, то красные выиграют +3 и синие выиграют +3. Одно решение таково: действовать совместно и синим уплатить красным *побочный платеж* +2, так как каждая из сторон выиграет тогда на 3 больше, чем если бы они не сотрудничали. Но синие могут сказать, что они уплатят только побочный платеж +1, ибо игра и так несправедлива и красные все равно получат больше. Кроме того, синие могут угрожать отказаться от сотрудничества, если красные не согласятся принять побочный платеж +1.

Сложность математических задач, связанных с такой игрой, находит свое отражение в трудности решения таких задач в реальной жизни. Рассмотрим переговоры о заработной плате между рабочими и администрацией; несотрудничество означает забастовку, сотрудничество — подписание договора, а побочный платеж — согласованную шкалу заработной платы. Очевидно, обе стороны получают выгоды от подписания договора, но каждый из игроков думает, что он добьется большего угрозой забастовки, чтобы в конце концов подписать более выгодный договор *. Чтобы добиться этого, каждая сторона должна не только угрожать, но и быть готовой выполнить свою угрозу; это значит, что игроки не уверены в том, что забастовки не будет, и поэтому каждый из них действительно принял смешанную (случайную) стратегию. Один из игроков или оба могут даже наметить короткую забастовку — бесконечная игра, в которой переменной является длительность забастовки.

В действительной жизни может оказаться возможным примирить разногласия, настолько усложнив ситуацию (например, дополнительными пособиями) **, что каждая сторона будет считать себя получающей немного больше; но математик не имеет такого утешения в своих попытках решить игру с ненулевой суммой. До сего времени не было получено общего удовлетворительного решения игр с ненулевой суммой.

Положение с играми n лиц (где $n>2$) примерно такое же; в самом деле, игру двух лиц с ненулевой суммой можно рассматривать как игру трех лиц с нулевой суммой, в которой третий игрок не имеет выбора, а лишь получает разность, на которую сумма отличается от нуля. В играх n лиц с нулевой суммой трудность заключается в *коалициях*. Если коалиции не разрешены, то каждый игрок может считать себя игроком в игре двух лиц, в которой все другие игроки против него, и может вычислить оптимальную стратегию. Но обычно два или несколько игроков могут образовать коалицию, которая позволяет им добиться лучших результатов, чем каждый из них мог бы получить по отдельности, и задача состоит в том, как разделить выигрыши.

Рассмотрим, например, игру трех лиц, состоящую в сравнении монет: «непарный» игрок берет все три монеты. Если коалиции не разрешены, игра допускает легкое решение: каждый игрок ставит случайно на герб или решку, назначая вероятность 0,5 для той и другой чистой стратегии. Но если игроки A и B образуют коалицию, они могут условиться всегда делать разные выборы, так что C никогда не выиграет. Затем A и B могут разделить свои выигрыши поровну. С другой стороны, A и B могут условиться дать C выиграть в небольшой доле конов, что приводит к бесконечному множеству решений. Существуют также два других множества решений,

---

* Здесь проблема, конечно, не столь проста, как достижение взаимовыгодной сделки; ситуация связана с классовой борьбой. — *Прим. ред.*

** Дополнительными пособиями (fringe benefits) в США называются различные компенсации и страховые пособия, выплачиваемые сверх установленной повременной или сдельной заработной платы, как-то: надбавки за сверхурочную работу, оплата отпусков, пенсии по старости, пособия по болезни, групповое страхование жизни и т. д. Состав и размеры таких дополнительных пособий определяются борьбой рабочих с предпринимателями, причем нередко для получения пособий рабочие должны сами делать регулярные взносы в соответствующие фонды. — *Прим. ред.*

когда A и B или B и C образуют аналогичные коалиции.

Согласно критерию решения, разработанному фон Нейманом и Моргенштерном [38], этим исчерпываются все решения. Однако и это множество интуитивно недостаточно, ибо если A и B делят все выигрыши поровну, то C может предложить A коалицию, в которой A получит 60%, а C получит 40%. Очевидно, это «решение» будет даже еще менее устойчиво, чем предыдущие. По этим и другим причинам в настоящее время не существует общепринятых решений игр n лиц, когда $n>2$.

24.6. Состояние теории игр

Теория игр имеет еще ряд сильных и слабых сторон, не отмеченных в этом кратком разборе.

Множественность решений. Игра может иметь более одного решения. Если говорить применительно к графику на рис. 24.2, то на выпуклой поверхности могут быть несколько вершин, которые все являются наивысшими. Если имеются две наивысшие вершины, соединенные прямолинейным отрезком, то каждая из этих вершин является *основным решением* и существует бесконечное число оптимальных стратегий, отмеченных разными точками на этом прямолинейном отрезке и изображающих все возможные смеси основных оптимальных стратегий. Если имеется больше двух наивысших вершин, то все они являются основными решениями и всякая точка на определяемой ими плоскости или гиперплоскости есть решение. Методы, описанные в § 24.3, всегда позволяют найти оптимальную стратегию, но может оказаться, что некоторые из этих методов не определяют всех оптимальных стратегий.

В трехпальцевой морре две стратегии (12 и 32) в данном нами решении выбираются с вероятностью 0, но их можно применять безнаказанно против оптимальной смешанной стратегии противника. Обычно с *неактивными* стратегиями так не бывает, но в рассматриваемой игре существует другое основное решение, в котором эти две стратегии применяются с ненулевой вероятностью.

Минимаксный принцип. Рассмотрим матрицу из табл. 24.12. Здесь имеются две седловые точки и, следовательно, две оптимальные чистые стратегии; синие могут применять стратегии 1 или 2, а красные будут всегда применять стратегию 1. Хотя у синих две оптимальные стратегии, одна из них лучше другой, так как стратегия «синие-2» доминирует над стратегией «синие-1», т. е. могут быть

Таблица 24.12

Игра с подчиненной оптимальной стратегией

Синие	Красные			
	1	2	3	Минимумы строк
1	[1]	2	3	1*
2	[1]	3	4	1*
Максимумы столбцов	1*	3	4	

оптимальные стратегии и *наилучшие оптимальные стратегии*.

Эта кажущаяся нелепость вызвана основным допущением теории игр, что критерием стратегии является минимаксный принцип, согласно которому каждый игрок должен добиваться минимума своего максимального убытка. В игре против искусного противника это обстоятельство не имеет значения, так как синие всегда выигрывают 1, какую бы стратегию они ни применяли; но в игре против неискусного противника они могут выиграть больше, если применят стратегию «синие-2».

Это допущение приводит к затруднениям и в практических ситуациях. В сложной игре предположение о том, что противник искусен в теоретико-игровом смысле, может оказаться неразумным либо потому, что он не может решить матрицу, либо—это более вероятно—потому, что у него другая информация, или у него другая оценка платежей, или он играет совсем в другую игру. Например, в сражении противник может неверно оценить нашу силу (или мы — его силу), или он может считать, что овладение неким холмом стоит дивизии, тогда как по нашему мнению оно стоит лишь батальона, или он может быть более заинтересован в том, чтобы получить повышение, нежели в том, чтобы выиграть войну. Иначе говоря, нельзя вполне точно перевести практическую ситуацию в математическую.

При этих условиях, конечно, нежелательно игнорировать возможность значительно большего выигрыша, нежели минимаксный, из-за теоретической возможности получить выигрыш, несколько меньший, чем минимаксный. Другими словами, теория игр не учитывает в достаточной мере рассчитанного риска, разумного азарта, смелых действий.

Есть и другие критерии, кроме минимаксного, служащие для того, чтобы обеспечить максимум ожидаемого дохода. Например, если некто имеет собственность (включая движимое имущество) общей стоимостью в 20 000 долларов и ему предложат поставить

эту собственность в игре в «орлянку» против 50 000 долларов, то он, очевидно, максимизирует ожидаемую стоимость своего состояния, приняв предложение (если стоимость денег линейна). С другой стороны, он может выбрать критерий «минимального сожаления» и отвергнуть предложение. См. [141].

Неточные оценки платежей. Если даже мы примем минимаксный принцип и его невыгодные стороны, все-таки выразить в числах платежную матрицу всегда бывает затруднительно. Если мы сравниваем потерю холма с потерей батальона, то как сделать квадратные футы земли соизмеримыми с человеческими жизнями? Тем не менее проектировщик системы должен решать задачи такого рода, как было указано в § 9.4.

Кроме того, сравнительно большие ошибки в платежах могут иметь очень малое влияние на стратегии. Например, если каждый платеж матрицы умножить на постоянное число, стратегии остаются без изменения (но цена игры умножается на постоянное число); точно так же стратегии не меняются при прибавлении постоянного числа к каждому платежу. Если даже не все платежи меняются одинаково, результат может быть ничтожным, особенно если относительные величины платежей не меняются. Поэтому теоретико-игровое решение можно дать для чисто качественной матрицы.

В табл. 24.13 исходы оцениваются по порядку только как отличный, хороший, удовлетворительный, посредственный и плохой (для синих); хотя ни одну стратегию нельзя вычеркнуть как подчиненную, игра имеет седловую точку и, следовательно, допускает оптимальные чистые стратегии для обоих игроков.

Если такая качественная матрица не имеет седловой точки, то во многих случаях мы можем найти качественное решение, немного видоизменив описанные выше методы. Конечно, в таком решении не могут быть указаны точные значения вероятностей выбора всех чистых стратегий. Но оно может указать, что некоторые стратегии не следует применять, а некоторые другие нужно применять чаще остальных. Например, если мы установили, что каждый игрок должен применять лишь две стратегии, которые приводят к такой подматрице, как в табл. 24.14, то мы будем знать, что стратегия «синие-2» должна применяться значительно чаще, чем стратегия «синие-1»; вероятности красных не столь определенны, но мы знаем по крайней мере, что они различаются меньше, чем вероятности синих.

Таблица 24.13

Качественная платежная матрица с седловой точкой

Синие	Красные					
	1	2	3	4	5	Минимумы строк
1	хор	уд	пос	хор	пос	пос
2	уд	хор	уд	отл	хор	уд*
3	пл	уд	пос	пл	отл	пл
4	отл	пл	пос	хор	уд	пл
5	отл	хор	уд	пос	пос	пос
Максимумы столбцов	отл	хор	уд*	отл	отл	

Таблица 24.14

Качественная платежная матрица без седловой точки

Синие	Красные	
	1	2
1	хор	пл
2	пос	уд

Многократно повторяемые игры. Мы молчаливо допустили, что игра будет иметь много партий; в частности, когда мы говорим о смешанной стратегии, то мы подразумеваем, что игра будет проводиться достаточно часто, чтобы закон больших чисел гарантировал нам в конце концов ожидаемое значение. Однако теория, изложенная на этих страницах, применима к игре, которая проводится лишь однажды (действительно, некоторые авторитеты утверждают, что она применима только к таким играм, потому что действительная партия приносит новую информацию, которую можно использовать для изменения стратегий). Тогда смешанная стратегия означает просто то, что мы не знаем заранее, какой именно выбор мы сделаем, и противник также не может этого знать. Тем не менее законы теории вероятностей позволяют вычислить ожидаемое значение платежа после партии, точно так же, как мы можем вычислить ожидаемое значение одиночного наблюдения по распределению вероятностей, и законы теории игр позволяют нам найти максимум этого ожидаемого значения на основании минимаксного принципа.

С другой стороны, когда мы рассчитываем на много партий, для нас может быть выгоднее изучить выборы противника и проверить, применяет ли он свою оптимальную стратегию; если нет, то нам может оказаться выгоднее изменить свою стратегию. В частности, если противник — природа, то она

обычно не будет применять оптимальной стратегии (т. е. не будет действовать коварно), и нам выгодно изучить ее стратегию.

Заключение. Мы можем резюмировать наше изложение теории игр цитатой из [62]:

«Теория игр, несмотря на ее ограничения, имеет в настоящее время приложения. Однако основная заслуга теории игр в том, что она дала ориентацию людям, которые сталкиваются с крайне запутанными проблемами. И хотя теория игр не дает строгого решения этих проблем по крайней мере в настоящее время и, вероятно, в течение неопределенного срока в будущем, тем не менее она указывает основу и направление усилий, предназначенных для их разрешения. Понятие стратегий, различие между игроками, роль случайных событий, матричное представление платежей, понятие о чистых и смешанных стратегиях и т. д. дают полезную ориентацию людям, которым приходится иметь дело со сложными конфликтными ситуациями».

ЛИТЕРАТУРА

Первоначальный труд по теории игр, с которым должен быть знаком всякий, серьезно изучающий этот предмет, принадлежит фон Нейману и Моргенштерну [38]; однако он довольно труден и содержит много материала по абстрактной экономической теории, не представляющего общего интереса для инженера-системника. Вильямс [62] написал чрезвычайно занимательную и приятную книгу, которая дает много знаний по теории игр, не выходя из рамок математики для средней школы; однако при этом ограничении она не может дать глубокого проникновения в предмет. Наиболее современны и наилучшим образом охватывают предмет Маккинси [63] и Блекуэлл и Гиршик [91].

ЗАДАЧИ

24.1*. Беглый арестант, приговоренный к смертной казни, был ранен при побеге и, прежде чем убежать через границу, должен найти медицинскую помощь. Он может либо пойти к местному врачу — и в этом случае шансы его выздоровления от пулевой раны составляют 50%, либо к местному ветеринару — и в этом случае шансы его выздоровления составляют 25%. Полиция располагает только одной автомашиной и будет пытаться захватить его у того или другого. Полицейские могут наверняка попасть к доктору вовремя, но шансы перехвата беглеца у ветеринара составляют лишь 50%. Какие стратегии должны выбрать арестант и полиция и каковы шансы арестанта на спасение (т. е. шансы выжить и убежать)?

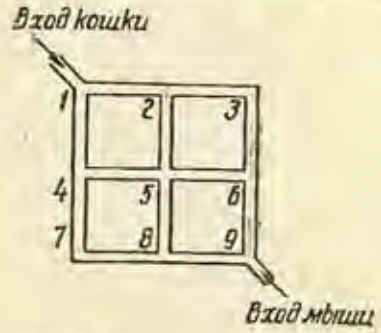
24.2*. Два брата согласились разыграть между собой на аукционе некоторое совместное имущество путем подачи запечатанных предложений на целое число сотен долларов. Давший бо́льшую цену должен заплатить эту сумму своему брату и забрать имущество. Если цены одинаковы, имущество будет отдано тому или другому в зависимости от исхода бросания монеты, без побочных платежей. Оба оценивают имущество в 800 долларов. Один брат имеет 800 долларов, а другой — лишь 500 долларов. Каковы оптимальные стратегии и цена игры?

24.3*. Решите следующую матрицу игры приближенным методом, изложенным в § 24.3.

Указание: процедуру следует провести по меньшей мере в 20 шагов и затем вычислить остаточную ошибку.

2	3	1	4
1	2	5	4
2	3	1	4
4	2	2	2

24.4*. Кошка и мышь входят одновременно в лабиринт, изображенный на прилагаемом рисунке. Обе идут с одинаковой скоростью к трем последовательным перекресткам; они не должны останавливаться и поворачивать обратно. Если они приходят к одному и тому же перекрестку одновременно, кошка съедает мышь. Если они приходят к третьему перекрестку не встретившись, мышь спасена. Каковы их оптимальные стратегии и какова для мыши вероятность остаться живой?



* Переработка из [62].

ГЛАВА 25

РУКОВОДЯЩИЕ ИДЕИ ПРИ ПРОЕКТИРОВАНИИ СИСТЕМ. ЛИНЕЙНОЕ ПРОГРАММИРОВАНИЕ, ГРУППОВАЯ ДИНАМИКА И КИБЕРНЕТИКА

Как было указано в гл. 3, имеются некоторые дисциплины, которые хотя и не служат непосредственными орудиями решения системотехнических задач, тем не менее дают общие положения и принципы, способные направлять мышление по системным линиям. К ним относятся три науки: линейное программирование, групповая динамика и кибернетика. Первая имеет свой источник в экономике, а вторая — в психологии, тогда как третья

является плодом союза физиологии и техники. В каждой из них встречаются задачи, аналогичные по своей природе задачам, встречающимся в системах. Считая, что эти передовые области науки обещают дать положительные результаты в системотехнике в будущем и что, даже если эти ожидания не исполнятся, их идеи действуют как катализатор на проектирование систем в настоящее время, мы проведем в этой главе краткий разбор каждой из этих наук.

25.1. Линейное программирование

Линейное программирование представляет собой новую математическую методику, разработанную лишь за последнее десятилетие*. Оно уже было применено к ряду практических задач, но, как и в теории игр, часто оказывается необходимым придать задаче довольно абстрактную форму и приписать численные значения величинам, точное измерение которых затруднительно. В отличие от теории игр, не существует аспекта систем, который был бы специфическим для применения линейного программирования, или во всяком случае этот аспект трудно определить. Наконец, как мы увидим, всякую задачу линейного программирования можно превратить в игру двух лиц с нулевой суммой (и обратно). По этим причинам мы посвятим линейному программированию сравнительно немного времени. Кроме того, наши примеры будут сравнительно тривиальны, так как более реалистические примеры были бы слишком длинны и сложны.

Даже класс задач, которые можно решать при помощи линейного программирования, трудно определить иначе, как на примере. Вообще говоря, существует много различных вещей (материалы, труд и т. д.), которые распределяются разными способами. На них накладываются некоторые ограничения; например, некоторые или все материалы могут иметься в ограниченном количестве, или допустимы только до определенных предельных количеств, или некоторые из них можно делить лишь поштучно. При этих ограничениях нужно получить максимум или минимум некоторого общего измерителя (как стоимость или прибыль). Мы поясним это на нескольких примерах.

Рассмотрим элементарную задачу по математическому анализу. Из всех прямоугольников, площадь которых равна или больше A, мы хотим найти прямоугольник, имеющий наименьший периметр. Задача поясняется на рис. 25.1. Прямоугольник вполне определен, если мы задали его длину x и высоту y. Всякая точка, лежащая на гиперболе $xy=A$ или над ней, изображает прямоугольник, удовлетворяющий нашему условию. Пунктирные прямые имеют уравнения $2x+2y=d$, где d — константа. Мы хотим найти наименьшее значение d, совместимое с ограничением $xy \geqslant A$. Легко видеть, что касательная к гиперболе дает решение и что это решение можно найти обычными аналитическими методами. Однако если мы заменим гиперболу ломаной линией, как на рис. 25.2, то искомая точка должна по-прежнему лежать на границе, но ее нельзя найти обычными аналитическими методами. Методом решения таких задач служит линейное программирование.

Рис. 25.1. Задача определения минимума в математическом анализе.

Задача минимизации. Допустим, мы хотим при помощи зенитных управляемых реактивных снарядов (ЗУРС) защитить город от воздушного нападения. Мы можем применять маловысотные снаряды, высотные снаряды или и те и другие. Высотные снаряды, стоящие дороже, более эффективны на больших высотах, но менее эффективны на малых высотах, как показано в табл. 25.1. Мы, конечно, не знаем, будет ли противник атаковать нас на больших, средних или малых высотах, но полагаем, что максимальное число атакующих единиц будет таким, как указано в таблице. Мы хотим, чтобы ожидаемое значение числа уничтоженных единиц было не меньше указанного максимального числа, и желаем добиться этого минимальной ценой.

Из приведенных данных мы получаем пять неравенств. Обозначим через x число

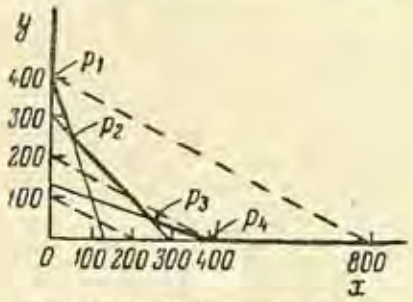


Рис. 25.2. Возможные решения и постоянные стоимости.

* В СССР методы линейного программирования разрабатывались еще в 1939 г. ленинградским математиком Л. В. Канторовичем [Д. 15]. — *Прим. ред.*

Таблица 25.1

Вероятность поражения

Тип ЗУРС	Высота атаки			Стоимость на снаряд, в тысячах долларов
	малая	средняя	большая	
Маловысотный	0,75	0,56	0,25	25
Высотный	0,25	0,56	0,75	50
Число атакующих единиц	100	150	100	—

маловысотных снарядов, а через y — число высотных снарядов. Тогда при атаке на малых высотах

$$0{,}75x + 0{,}25y \geqslant 100,$$

или, после умножения на 4 для устранения дробных величин,

$$3x + y \geqslant 400. \qquad (25.1а)$$

Два аналогичных неравенства для атаки на средних и больших высотах:

$$x + y \geqslant 300, \qquad (25.1б)$$

$$x + 3y \geqslant 400. \qquad (25.1в)$$

Кроме того, так как мы не можем хранить отрицательное число снарядов, то

$$x \geqslant 0,\ y \geqslant 0. \qquad (25.2)$$

Эти неравенства представлены графически на рис. 25.2. Координатные оси изображают наши границы (25.2), и точка решения должна лежать на них или над ними. Три другие сплошные прямые изображают границы областей (25.1), и точки решения также должны лежать на них или над ними. Соответствующие участки пяти прямых показаны жирными линиями; эта жирная ломаная определяет область *осуществимых* решений. В некоторых задачах линейного программирования область осуществимых решений ограничена (рис. 25.3), но в данном случае она не ограничена.

Функция стоимости для этой задачи имеет вид

$$C = 25x + 50y, \qquad (25.3)$$

и нам нужно найти ее минимум. Каждая из пунктирных прямых рис. 25.2 изображает линию постоянной стоимости, и мы хотим найти самую низшую. Представляется интуитивно очевидным, что эта низшая пунктирная прямая пересечет жирную линию в вершине, и это действительно так. Вообще, решение задачи линейного программирования всегда находится в вершине границы осуществимых решений, за исключением того случая, когда функция стоимости совпадает с граничной прямой (или с плоскостью, или с гиперплоскостью — в более сложных задачах), так что решения изображаются несколькими вершинами, а также всеми точками на прямой или плоскости, для которой уравнения совпадают, причем все решения имеют одинаковые стоимости.

В этой задаче мы вычисляем стоимость в каждой вершине; результаты приведены в табл. 25.2. Как можно видеть из рисунка, решение изображается вершиной p_3 и определяется пересечением выражений (25.1б) и (25.1в), рассматриваемых как равенства; решением будет $x=250$, $y=50$. Ожидаемое число уничтоженных единиц равно требуемому числу уничтоженных единиц при атаке на больших или средних высотах и больше требуемого числа уничтоженных единиц при атаке на малых высотах.

Таблица 25.2

Вершины на рис. 25.2

	p_1	p_2	p_3	p_4
x	0	50	250	400
y	400	250	50	0
Стоимость	20 000	13 750	8 750	10 000

Двойственность. Решение говорит, что нам следует применять пять шестых дешевых (маловысотных) и одну шестую дорогих (высотных) снарядов. Если бы это была задача из теории игр, то мы были бы вынуждены применять все снаряды одного или другого типа и решение состояло бы в том, чтобы применять маловысотные снаряды с вероятностью $^5/_6$ и высотные снаряды с вероятностью $^1/_6$ (причем ожидаемое число уничтоженных единиц было бы то же самое). Рис. 25.2 аналогичен рис. 24.1, который определял оптимальную стратегию; там было сказано, что если бы матрица была 2×3, то рисунку 24.1 у другого игрока соответствовала бы трехмерная фигура. Матрица из табл. 25.1 имеет порядок 2×3, и ей соответствует описанная ниже двойственная задача, которая приводит к трехмерному графику на рис. 25.3.

Задача максимизации. Задача состоит в том, чтобы распределить некоторое количество имеющегося материала и труда на изготовление определенных видов вооружения:

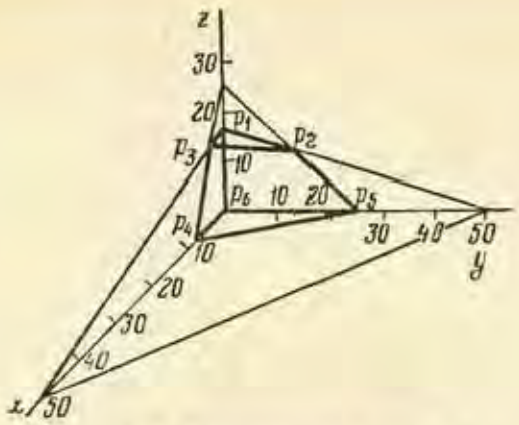


Рис. 25.3. Возможные решения.

зенитных орудий, самолетов-истребителей и ЗУРС. Нужно добиться максимального ожидаемого числа уничтоженных вражеских бомбардировщиков. Обозначив число тысяч пу-

Таблица 25.3

Требуемые материалы

	Единиц материалов на 1 000	Единиц труда на 1 000	Ожидаемое число уничтоженных единиц на 1 000
Зенитные пушки . .	3	1	400
Истребители	1	1	300
ЗУРС	1	3	400
Наличие	25	50	—

шек буквой x, число тысяч истребителей буквой y и число тысяч управляемых реактивных снарядов буквой z, находим из табл. 25.3 следующие неравенства:

$$\left.\begin{aligned} 3x+y+z &\leqslant 25 \\ x+y+3z &\leqslant 50 \end{aligned}\right\}. \qquad (25.4)$$

Ожидаемое число уничтоженных бомбардировщиков равно

$$K = 400x + 300y + 400z. \qquad (25.5)$$

Нужно определить максимум этого выражения. Поскольку число изготовленных единиц вооружения может быть только положительным, справедливы также неравенства

$$x \geqslant 0,\ y \geqslant 0,\ z \geqslant 0. \qquad (25.6)$$

В данном случае каждое неравенство изображает область в трехмерном пространстве, ограниченную плоскостью, а не область в двумерном пространстве, ограниченную прямой, как в предыдущем случае. Если мы будем считать их равенствами, то (25.6) изображает координатные плоскости, а (25.4) — две другие плоскости, как показано на рис. 25.3. В этом случае область осуществимых решений ограничена, так как она должна лежать на плоскостях (25.6) или выше, т. е. в первом октанте, и на плоскостях (25.4) или ниже. Она очерчена на рис. 25.3 жирными линиями.

Уравнение (25.5) для разных значений K описывает семейство параллельных плоскостей, и нам требуется найти максимальное значение K, соответствующее плоскости, у которой имеется по крайней мере одна общая точка с областью осуществимых решений. Как и раньше, точка решения должна быть в вершине многогранника, поэтому мы вычисляем и сводим в таблицу значения для каждой вершины (табл. 25.4) и замечаем, что решение изображается точкой p_2, а максимальное ожидаемое число уничтоженных бомбардировщиков равно 8 750.

Этот результат численно равен решению предыдущей задачи, потому что матрица в табл. 25.3 является двойником матрицы в табл. 25.1. Всякая задача линейного программирования имеет двойственный аналог, причем одна задача является задачей максимизации, а другая — задачей минимизации; они соответствуют задачам синих и красных в теории игр.

Таблица 25.4

Вершины на рис. 25.3

	p_1	p_2	p_3	p_4	p_5	p_6
x	0	0	25/8	25/3	0	0
y	0	25/2	0	0	25	0
z	50/3	25/2	125/8	0	0	0
K	6 667	8 750	7 500	3 333	7 500	0

Основная теорема. Подобно обобщенной матрице теории игр (табл. 25.10), мы можем составить обобщенную матрицу линейного программирования (табл. 25.5). Для задачи минимизации x_i обозначает число производимых единиц i-го изделия; a_{ij} — эффективность i-го изделия в j-й категории; b_j — общая требуемая эффективность в j-й категории; c_i — стоимость единицы i-го изделия; u_j не участвуют в задаче. Для задачи максимизации u_j обозначает число производимых единиц j-го изделия; a_{ij} — число единиц i-го предмета снабжения, используемых при производстве единицы j-го изделия; b_j — эффективность единицы j-го изделия; c_i — число имеющихся единиц i-го предмета снабжения; x_i не входят в задачу.

Таблица 25.5

Обобщенная матрица для линейного программирования

	u_1	u_2		u_j		u_m	
x_1	a_{11}	a_{12}		a_{1j}		a_{1m}	c_1
x_2	a_{21}	a_{22}		a_{2j}		a_{2m}	c_2
x_i	a_{i1}	a_{i2}		a_{ij}		a_{im}	c_i
x_n	a_{n1}	a_{n2}		a_{nj}		a_{nm}	c_n
	b_1	b_2		b_j		b_m	

Итак, в первой задаче величины, обозначенные нами буквами x и y, становятся теперь величинами x_1 и x_2; максимальные числа атакующих единиц становятся величинами b_1, b_2 и b_3; величины, названные нами «стоимостями», становятся величинами c_1 и c_2. Во второй задаче величины, обозначенные нами буквами x, y и z, становятся величинами u_1, u_2 и u_3; ожидаемые числа уничтоженных единиц противника становятся величинами b_1, b_2 и b_3; а наличные единицы соответствуют c_1 и c_2.

Общая задача минимизации состоит, таким образом, в том, чтобы получить минимум функции

$$C=\sum_{i=1}^{n} c_i x_i \qquad (25.7)$$

при выполнении m неравенств вида

$$\sum_{i=1}^{n} a_{ij} x_i \geqslant b_j \qquad (25.8)$$

и n неравенств вида

$$x_i \geqslant 0. \qquad (25.9)$$

Общая задача максимизации состоит в том, чтобы получить максимум функции

$$K=\sum_{j=1}^{m} b_j u_j \qquad (25.10)$$

при выполнении n неравенств вида

$$\sum_{j=1}^{m} a_{ij} u_j \leqslant c_i$$

и m неравенств вида

$$u_j \geqslant 0.$$

Аналогия с (24.2)—(24.5) очевидна.

Основная теорема линейного программирования гласит, что решение (если оно существует — см. ниже) любой задачи максимизации тождественно с решением ее двойственной задачи минимизации и наоборот; таким образом, минимум величины C в (25.7) равен максимуму величины K в (25.10). Мы не будем этого доказывать, но можно показать, что эта задача линейного программирования эквивалентна игре двух лиц с нулевой суммой и что C равно цене игры для синих, а K равно цене игры для красных. Тогда основная теорема линейного программирования вытекает из основной теоремы теории игр.

Эквивалентность с теорией игр. Примем, что цена игры v для синих положительна. Это не всегда имеет место, но всегда можно сделать так, что это будет справедливо. В самом деле, мы не изменим решения игры (т. е. оптимальных стратегий), прибавив постоянное число к каждому элементу матрицы; а прибавив достаточно большое число, мы можем сделать каждый элемент положительным, так что цена игры должна быть положительна. Таким образом, мы полагаем, что матрица игры была приведена к такому виду, что всякое $a_{ij}>0$.

Когда синие нашли свою оптимальную смешанную стратегию $(p_1, \ldots, p_n)$, то, если красные применяют какую-нибудь чистую стратегию, скажем j-ю стратегию, мы знаем, что

$$\sum_{i=1}^{n} p_i a_{ij} \geqslant v. \qquad (25.11)$$

Введем новую переменную, определенную равенством

$$x_i=\frac{p_i}{v} \text{ или } p_i=vx_i, \qquad (25.12)$$

и подставим ее в (25.11):

$$\sum_{i=1}^{n} vx_i a_{ij} \geqslant v. \qquad (25.13)$$

Из (25.12), зная, что v положительно, а p_i неотрицательны, мы получаем n неравенств вида

$$x_i \geqslant 0. \qquad (25.14)$$

Из (25.13), деля на постоянное число v, мы получаем m неравенств вида

$$\sum_{i=1}^{n} x_i a_{ij} \geqslant 1. \qquad (25.15)$$

Но цель игры синих — получить максимальное v, что равносильно достижению минимума величины $1/v$. Из (25.12) и того обстоятельства, что сумма p_i равна 1, вытекает

$$\frac{1}{v} = \sum_{i=1}^{n} x_i. \qquad (25.16)$$

Итак, задача состоит в минимизации величины (25.16) при выполнении условий (25.14) и (25.15). Это задача линейного программирования в своей стандартной форме.

Двойственная задача (максимизации) получается таким же путем из стратегии красных в теоретико-игровой задаче. Можно также произвести обратное преобразование (разрешимой задачи линейного программирования в задачу теории игр), но оно сложнее.

Задачи, не имеющие решения. В то время как задача теории игр, имеющая форму табл. 24.10, всегда допускает решение, существуют два (и только два) случая, когда задача линейного программирования не имеет решения.

Первый случай — когда имеются ограничения при максимизации и минимизации и они не пересекаются, т. е. нет осуществимых решений. Вот, например, типичная задача линейного программирования: найти самое дешевое питание из нескольких видов продуктов при определенной стоимости единицы каждого продукта и определенном содержании питательных веществ в каждом продукте, причем в итоге от пищи требуется некоторое минимальное количество каждого питательного вещества. Если других условий нет, то эта задача всегда имеет решение. Но в пище могут быть некоторые яды, и может существовать дополнительное ограничение, что пища не должна содержать больше определенного количества яда. Если во всех продуктах или в тех, которые содержат необходимые ингредиенты, достаточно большое количество яда, то задача может не иметь решения.

Другой случай — когда область осуществимых решений не ограничена. Например, в задаче минимизации один из предметов может иметь отрицательную стоимость (мы получаем премию за каждый такой купленный нами предмет). Если не установлено ограничение на количество покупаемых предметов данного вида, то минимальная стоимость будет достигнута при покупке бесконечного количества этих предметов. Ситуациям, соответствующим такому положению, можно придать математическую формулировку, и они приводят к задачам, не имеющим решения, но, конечно, они не имеют физического смысла.

Практическая задача. Нижеследующая задача линейного программирования, сформулированная и частично решенная М. Э. Сальвесоном [34] из «Дженерал Электрик Компани», была поставлена для производственных линий этой фирмы, и решение ее привело к экономии в сотни тысяч долларов.

На конвейерной сборочной линии выполняется N операций, которые по соглашению между фирмой и профсоюзом и на основании хронометража требуют для своего выполнения время $t_1, \ldots, t_N$. Эти операции нужно распределить между отдельными людьми, находящимися на рабочих местах; у каждого рабочего места движущаяся деталь будет находиться в течение времени $T = d/v$, где v — постоянная скорость конвейера, а d — пространство, которое может без затруднения охватить один человек (время, затрачиваемое им на обратное передвижение, не должно быть столь велико, чтобы рабочее место начало смещаться), т. е. длина рабочего места. Для обеспечения требуемого объема производства может применяться несколько линий (в этой формулировке задачи мы не рассматриваем случай, когда имеется несколько рабочих на одном рабочем месте, и не рассматриваем разветвлений или стыков в линии; все это значительно усложнило бы задачу). Кроме того, некоторые операции должны предшествовать другим или следовать за ними, тогда как некоторые другие группы операций взаимозаменимы; например, гайки должны завинчиваться после установки соответствующего болта, но из двух болтов любой может устанавливаться первым. Эти отношения называются *отношениями очередности.*

Задача состоит в том, чтобы выбрать сочетания операций на рабочих местах так, чтобы:

1) удовлетворялись отношения очередности;

2) удовлетворялось следующее неравенство:

$$\sum_{i=1}^{n_j} t_{ij} \leqslant \frac{d}{v},$$

где t_{ij} — время, необходимое для i-й операции на j-м рабочем месте, и n_j — число операций, выполняемых на j-м месте;

3) сумма простоев, т. е.

$$\sum_j \left(\frac{d}{v} - \sum_{i=0}^{n} t_{ij} \right),$$

была минимальна.

Нелинейное программирование. Если функция стоимости (25.7) или граничная функция (25.8) нелинейна по x_i (или u_j в двойственной задаче), то перед нами задача нелинейного программирования. Например, предположим, что человек с большими средствами имеет возможность поставить последнюю ставку на бегах. Предположим, далее, что он имеет точную информацию об априорной вероятности выигрыша каждой лошади. Бегут пять лошадей.

Пусть p_i — вероятность того, что выиграет i-я лошадь;

$$p_1 + p_2 + p_3 + p_4 + p_5 = 1.$$

Пусть s_i — сумма, уже поставленная на i-ю лошадь;

x_i — сумма, которую наш игрок ставит на i-ю лошадь;

q — доля, которую берет с каждой ставки тотализатор.

Далее, положим

$$s_0 = \sum_{i=1}^{5} s_i$$

$$x_0 = \sum_{i=1}^{5} x_i.$$

Ожидаемые выигрыши игрока равны

$$f(x_1, \ldots, x_5) = \sum_{i=1}^{5} \frac{p_i (1-q)(s_0 + x_0)}{(s_i + x_i)} x_i - x_0.$$

Он стремится получить максимум функции $f(x_1, \ldots, x_5)$ при условии $x_i \geqslant 0$.

Эту задачу решил Айзекс [65]. Вообще говоря, решение предусматривает распределение ставок на несколько лошадей. Для мелкого игрока наилучшая стратегия такова: ставить всю ставку на лошадь, для которой $p_i s_0 / s_i$ максимально, если вообще имеет смысл играть; но это не будет наилучшей стратегией, если ставка столь велика, что она существенно повлияет на относительные платежи.

25.2. Групповая динамика

Групповая динамика возникла из работ Курта Левита [68], который стремился найти математические (топологические) представления для отношений групп индивидуумов между собой и их взаимных влияний. Эти представления были применены ко всем видам отношений между людьми, таким, как эмоциональные воздействия и господство. Мы здесь не занимаемся этим широким аспектом предмета, называемого *групповой динамикой* *. Наше внимание будет сосредоточено на опытах, которые были выполнены недавно с целью изучения потока информации в простых сетях, состоящих из людей и линий связи.

Основные идеи. Это изучение было начато психологами, в частности Бавеласом [67], который основывался на идеях Левина. Бавелас рассматривал сети связи, подобные изображенным на рис. 25.4. Их можно считать схемами иерархической организации. В первой каждый человек сообщается только со своим непосредственным начальником и своими непосредственными подчиненными; во второй каждый сообщается с этими чинами и, кроме того, с другими непосредственными подчиненными своего непосредственного начальника; в третьей каждый сообщается со всеми этими чинами и со всеми другими на его уровне.

Мы интересуемся эффективностью связей, но, в отличие от задач, разрабатываемых в теории информации, природа каналов связи нас не касается. Мы считаем, что каждая линия, соединяющая два узла, одинакова со всеми другими и представляет единицу расстояния. Бавелас определил *расстояние* d_{ij} между любыми двумя узлами i и j как наименьшее число звеньев, по которым можно пройти от одного узла к другому. Для любого данного узла i имеется набор таких расстояний до других узлов, и наибольшее из этих расстояний обозначается символом D_i. *Центральная область* сети есть узел (или узлы) с наименьшим D_i. На рис. 25.4 узел a

* Групповая динамика — направление в американской социологии и психологии, пытающееся изучать топологическими средствами поведение индивидуумов в группах и взаимоотношения между членами групп. Представителями этого направления разработан ряд математических методов для анализа структуры сложных сетей, например сетей связи или сетей зависимости. Большое внимание уделялось проблеме относительной эффективности группы в зависимости от ее структуры. По общим социальным вопросам представители американской групповой динамики, понятно, стоят на буржуазных позициях. — *Прим. ред.*

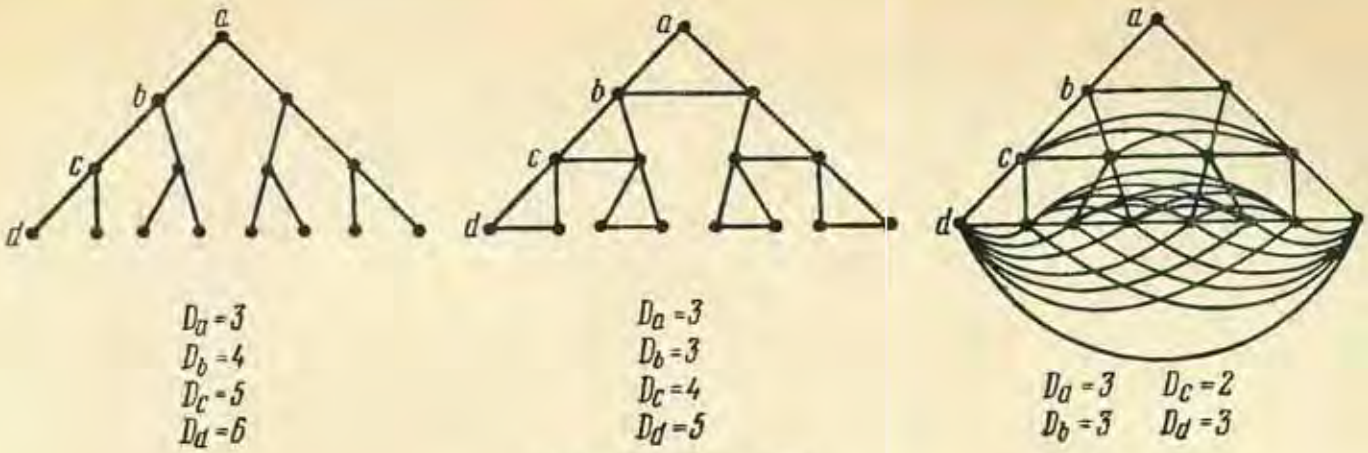


Рис. 25.4. Сети связи.

есть центральная область в первой сети, узлы a и b образуют центральную область во второй сети, узел c образует центральную область в третьей сети.

Понятие центральности важно, как мы увидим дальше, но это определение несовершенно. Если мы имеем весьма густую сеть (как на рис. 25.4, но такую, что каждая ветвь идет к многим подчиненным, а не только к двум) и добавим цепь из нескольких звеньев к одному ее концу, то центральная область, определенная указанным выше образом, переместится в эту цепь, что интуитивно неприемлемо. Позднейшие исследователи видоизменили это понятие и развили идею *показателя центральности* C_i для каждой позиции i. Этот показатель определяется как

$$C_i = \frac{\sum_i \sum_j d_{ij}}{\sum_j d_{ij}}.$$

На рис. 25.5 изображена цепная сеть, а в табл. 25.6 приведены вычисленные показатели центральности для этой сети.

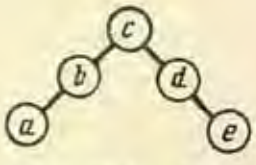


Рис. 25.5. Пятизвенная цепь.

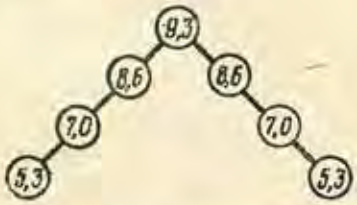


Рис. 25.6. Показатели центральности для семизвенной цепи.

В этом случае, если мы добавим позиции на конце, как на рис. 25.6, позиция с максимальной центральностью не меняется, но численное значение показателя изменится; на рисунке указаны показатели центральности. Во избежание этого вводятся две новые меры: относительная периферийность

$$p_i = C_i - C_{\text{мин}}$$

и полная периферийность

$$P = \sum_i p_i.$$

На рис. 25.7 показана относительная и полная периферийность для пятичленной и семичленной сетей предыдущих диаграмм. Это определение, по-видимому, хорошо совпадает с нашим интуитивным представлением. Большее значение имеет то обстоятельство, что оно, по-видимому, выражает существенные свойства сетей связи, как это видно из описанных ниже экспериментов.

Таблица 25.6

d_{ij} и C_i для рис. 25.5

i / j	a	b	c	d	e
a	0	1	2	3	4
b	1	0	1	2	3
c	2	1	0	1	2
d	3	2	1	0	1
e	4	3	2	1	0
$\sum_j$	10	7	6	7	10
$\sum_i \sum_j$	40	40	40	40	40
C_i	4,0	5,7	6,7	5,7	4,0

Подготовка эксперимента. В проводившихся до сих пор экспериментах по групповой

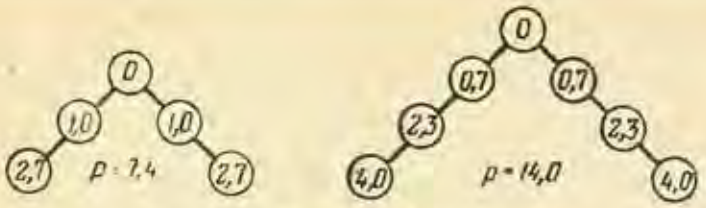


Рис. 25.7. Периферийность цепей.

динамике применяются человеческие сети таких форм, как на рис. 25.8. Группе дается некоторое задание. Например [45, 68]: каждому из пяти субъектов в группе могут быть даны по пяти различных символов; всего имеется шесть символов, так что у каждого субъекта отсутствует один из шести символов и лишь один из шести символов является общим для всех пяти участников. Задача состоит в отыскании общего символа, и эксперимент заканчивается, когда ответ найден каждым членом группы.

Организация группы не раскрывается индивидууму, и устная связь не допускается, а письменные сообщения могут посылаться только по линиям, указанным на рис. 25.8 (в обоих направлениях). Характер разрешенных сообщений более или менее ограничен, в зависимости от эксперимента. Можно ввести такие усложняющие факторы, как шум в канале (в виде искажений сообщений). Критериями поведения служат скорость решения, измеряемая временем или числом сообщений, и частота ошибок (ошибка будет тогда, когда кто-нибудь заявляет, что он нашел решение, а ответ оказывается неверным). По окончании эксперимента каждого участника спрашивают, что он думает о некоторых сторонах эксперимента, и эти субъективные оценки являются также критериями выбора между различными сетями.

Результаты. Оказалось, что почти все результаты эксперимента можно расположить по порядку соответственно возрастанию полной периферийности сети. Например, кратчайшее время решения для «колеса» было меньше кратчайшего времени для Y-образной сети, а оно было меньше, чем для цепи, а последнее было меньше, чем для окружности.

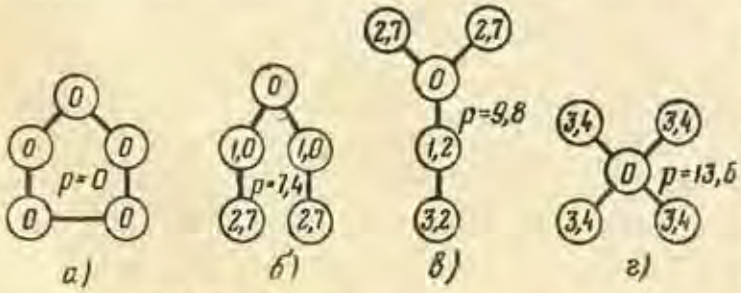


Рис. 25.8. Периферийность пятизвенных сетей.

В колесе участники скоро обнаружили организационную структуру, но еще до этого сложилась эффективная организация, в которой центральная позиция взяла на себя всю власть. В окружности участники ни разу не раскрыли организационной структуры и не создали единой организации, а посылали сообщения по обоим направлениям, пока кто-нибудь не получал ответа.

При опросе после эксперимента участников спрашивали, был ли у них руководитель; фактически все участники колеса сказали, что был, а в окружности ни один участник не сказал этого. Участники окружности сказали, что, по их мнению, они получили бы лучшие результаты, если бы у них была лучшая организация. Однако им нравилась их деятельность, тогда как участникам колеса она не нравилась (за исключением того, кто был на центральной позиции).

Возможно, самый существенный вывод состоял в том, что для легких задач лучше всего было колесо, но для трудных задач окружность иногда подходила больше. По-видимому, в организациях с большой полной периферийностью (колесо) все было эффективно организовано для того, чтобы центральный член выполнял свою работу; если он не был пригоден к этой работе, вся организация была неэффективной. В организациях с меньшей периферийностью (окружность) было больше ошибочных действий, но каждый мог получить представление о том, как будет действовать группа. Например, в данной задаче кто-нибудь из участников часто делал открытие, что целесообразнее сообщать недостающий символ, чем пять имеющихся символов. Когда это предложение выдвигалось в окружности, оно всегда принималось, и это приводило к улучшению; но однажды, когда его выдвинули в колесе, лицо, находившееся в центре, наложило на него запрет и предложение не прошло.

Конечно, сфера таких экспериментов ограниченна, и поэтому мы должны быть осторожны при обобщении их результатов. В описанном исследовании не учитывались моральные факторы, общественные ценности, влияние продолжительного знакомства членов бригады и принятая иерархия. Так, в реальной сети центральное лицо, вероятно, будет самым способным, что, вообще говоря, не имело места в описанных выше экспериментах. Тем не менее мы, вероятно, вправе сделать следующее предварительное заключение: человеку, который будет «хозяином», надо дать хорошие связи со всеми частями системы и для решения легких задач желательна

такая система, как колесо, тогда как для более трудных задач, может быть, лучше такая система, как окружность.

Заметим, что бригаду проектировщиков системы отнюдь не следует организовывать подобно системе. Система решает легкие задачи (т. е. организация системы такова, что задача автоматически разбивается на простые задания), тогда как бригада проектирования системы решает трудные задачи; поэтому в первой высокая степень авторитарности может оказаться желательной, а во второй — нет.

При проведении экспериментов по групповой динамике попутно освещались некоторые интересные стороны человеческой природы. В одной сети ни одному субъекту не разрешалось посылать сообщения прямо к тем лицам, от которых он получал сообщения. В одном эксперименте с этой сетью участникам была дана неразрешимая задача и им дали работать в течение часа, не требуя от них ответа. Затем каждого из них опросили по отдельности, почему, по его мнению, группа потерпела неудачу. Почти все ответили: «Если бы этот идиот слева от меня передавал мне надлежащие сообщения, мы бы нашли ответ моментально».

25.3. Кибернетика

Слово *«кибернетика»* было создано в 1947 г. Норбертом Винером, и в следующем году он выпустил книгу на эту тему [53]. Это слово происходит от греческого слова, означающего «кормчий», в знак того, что человек, управляющий кораблем, действует подобно сервомеханизму; в то же время оно родственно слову governor — «регулятор» * (Винер говорит [53]: «первой значительной работой по механизмам с обратной связью была статья о регуляторах, опубликованная Кларком Максвеллом в 1868 г.»).

Область. Винер нигде не определяет четко термин «кибернетика». Однако он указывает, что в электротехнике существует разделение на области, называемые в Америке *энергетической техникой* и *техникой связи*, а в Германии — *техникой сильных токов* и *техникой слабых токов*. Он приравнивает первую изучению консервативных систем (грубо говоря, таких систем, в которых сохраняются существенные параметры, например масса, энергия и момент), а последнюю — изучению информационных систем, в которых законы сохранения имеют второстепенное значение. Он указывает на два фундаментальных принципа: все системы, живые и механические, суть информационные системы, и все системы, живые и механические, суть системы обратной связи (т. е. следящие системы, или системы автоматического регулирования). Он заключает отсюда, что методы изучения обеих этих групп систем должны быть основаны на теории информации и теории следящих систем (теории автоматического регулирования).

Он связывает эти идеи с темой, которая кажется очень далекой, — со старым спором между механистами и виталистами. Механисты доказывали, что жизнь можно объяснить на основе механических законов; виталисты доказывали, что жизнь, и особенно человеческая душа, содержит нечто сверх этого. Винер говорит [53], что «весь спор между механистами и виталистами можно отложить в архив плохо сформулированных вопросов»; но он указывает, что в действительности спор был разрешен совершенно неожиданным образом, что нечто добавочное было найдено в жизни и даже в неодушевленном физическом мире. Это нечто добавочное Винер отождествляет с неопределенностью — не только с неопределенностью Гейзенберга, но и с более существенной неопределенностью статистики, которую он опять-таки связывает со статистической природой теории информации и входов управляющих систем.

Хотя сказанное не следует понимать как определение или описание кибернетики, но все же это дает некоторое представление о ее сфере. Мы не можем надеяться извлечь много непосредственных практических выводов из краткого разбора такой обширной темы. В самом широком смысле кибернетика включает все проектирование систем плюс понимание систем, как живых, так и механических. Мы остановимся здесь на очень маленьком разделе предмета — на аналогиях между челове-

---

* Слово «кибернетика» (κυβερνητική) образовано от греч. κυβερνήτης — «кормчий, рулевой» и, стало быть, должно означать «искусство кормчего» или — в широком смысле — «искусство управления». Винер считал слово «кибернетика» своим неологизмом, но впоследствии выяснилось, что оно уже использовалось в разных более или менее широких смыслах некоторыми авторами (Платон, Ампер).

Английское слово governor, обозначающее в частности «регулятор», происходит от латинского слова gubernator — «кормчий, правитель», этимологически родственного греческому κυβερνήτης.

После появления в 1947 г. книги Н. Винера «Кибернетика» вопрос о сущности и значении кибернетики стал предметом ожесточенных споров, в ходе которых ряд авторов выступил против кибернетики, а ряд других — в ее защиту, определяя ее в то же время нередко по-своему, не так, как у Винера. В последние годы кибернетика получила положительное признание у большинства советских ученых, хотя многие вопросы, и в частности о точном содержании кибернетики, продолжают оставаться дискуссионными; см., например, [Д. 6]. — *Прим. ред.*

ком и машиной — и надеемся, что они позволят лучше понимать те и другие, и в частности дадут некоторые указания на более эффективное проектирование больших систем.

Такие аналогии могут принести некоторую пользу. При проектировании первых автоматических телефонных систем была потеряна одна важная способность человека-телефонистки. Телефонистка может осуществить соединение при помощи сравнительно недорогого оборудования (соединительных шнуров и штепсельных гнезд) и затем использовать свои руки и мозг, являющиеся наиболее дорогими частями установки, для выполнения других обязанностей, пока поддерживается соединение. В первых автоматических системах все коммутационное оборудование было занято в течение всей продолжительности соединения. Только с введением координатной системы появилась машина, аналогичная в этом отношении человеку. В координатной системе № 5 (§ 2.2) дорогие части коммутационного оборудования, и в особенности маркер, используются лишь очень короткое время — лишь столько, сколько нужно, чтобы установить соединительный путь и возбудить реле в точке соединения, а затем их можно использовать для выполнения других задач, хотя соединение продолжается.

Нейроны и клапаны. Аналогию между человеком и *автоматом* (как называет Винер искусственную, сделанную человеком систему) можно провести для целого ряда основных функций, включая связь, управление, хранение (память) и организацию. Наиболее важным средством связи в организме служит нервная клетка, или *нейрон.* Хотя нейрон, подобно другим клеткам, имеет микроскопический диаметр, его длина может достигать нескольких футов. Когда он переносит сообщение, говорят, что он *возбужден*: по нейрону с одного конца до другого распространяется некоторая электрохимическая реакция.

В соответствии с важным физиологическим законом *«все или ничего»,* нейрон, если он возбужден, переносит максимальное сообщение, на которое он способен; если стимул не достаточно силен, чтобы заставить нейрон переносить это максимальное сообщение, то нейрон вообще не реагирует. В течение короткого времени после возбуждения нейрон совершенно заторможен (рефрактерен); после этого он опять способен переносить свое стандартное сообщение со стандартной задержкой. Возле конца нейрона находится ряд соединительных точек, называемых *си́напсами,* к которым поступают стимулы. Имеются доказательства, что некоторые синапсы имеют больший «вес», чем другие, и что пороговый уровень (т. е. число и расположение синаптических стимулов, необходимых для возбуждения нейрона) подвержен медленным изменениям.

Таким образом, нейрон в сочетании со своими многочисленными синапсами может рассматриваться как клапан со сложным сочетанием клапанирующих свойств И и ИЛИ. Существуют также доказательства, что некоторые синапсы являются запретительными, т. е. клапанами НЕ. Входами некоторых нейронов служат чувствительные концевые органы, но входами всех других нейронов, по-видимому, служат выходы других нейронов, через синапсы.

Память и обучение. Природа человеческой памяти неизвестна, но мы можем представить себе несколько возможных методов ее синтеза из нейронного механизма. Существенную роль в действии памяти в организме могут играть два элемента: циркуляционное хранение и модификация порогов.

Циркуляционное хранение должно быть подобно акустической линии задержки (§ 15.3 и рис. 15.10); оно может состоять из петли нейронов, возбуждающих друг друга, так что сообщение, однажды возникнув в петле, будет продолжать циркулировать неопределенно долго. Хотя мозг слишком сложен, чтобы можно было проследить все его нейронные соединения, такие петли, бесспорно, могут существовать. Модификация порогов означает, что либо меняется полный стимул, необходимый для возбуждения нейрона, либо некоторые синапсы имеют различные веса при возбуждении нейрона. Такой механизм может служить для объяснения *условного рефлекса.*

Рефлексный механизм, в организме или машине, представляет собой такой механизм, в котором действие происходит в ответ на внешний стимул, независимо от высших управляющих центров. В автомате это имеет вид реакции следящих устройств без связи с логическим управлением и обычно представляет собой действие нижней ступени без ожидания решения высшей ступени. В организме рефлексное действие есть действие, управляемое низшим нервным центром, например спинным мозгом или стволом мозга, независимо от головной (сознающей) части мозга.

Типичным рефлексным действием является выделение слюны при виде пищи. Павлов в своих знаменитых опытах с условным рефлексом изучал это рефлексное действие на собаках. Он раздражал собак видом и запахом пищи и одновременно звонил в звонок. При этом у собак выделялась слюна. После того как это было проделано несколько раз,

слюна у собак выделялась всякий раз, когда звонил звонок, даже при отсутствии пищи. Итак, рефлексное действие (выделение слюны под действием раздражителя-звонка) было обусловлено приобретенным собаками опытом.

Идея модификации порогов, по-видимому, объясняет довольно хорошо такие формы обучения. При каждом повторении последовательности стимул — реакция нейронный путь становится все более прохоженным и потому все легче проходимым. Возможно, что все обучение происходит таким способом — образованием освоенного пути через лабиринт нейронов и синапсов посредством непрерывного изменения отдельных синаптических порогов. Возможно даже, что такие пороги могут меняться только в одном направлении и поэтому обучение может происходить только до некоторого предела. Это явление, может быть, играет некоторую роль (хотя, бесспорно, не главную) в развитии старческого одряхления. С другой стороны, возможно, что пороги в неиспользуемых путях медленно возвращаются к нормальному уровню и что именно в этом состоит явление забывания.

Системы, которые учатся. Хотя точный механизм обучения не вполне объяснен, рассмотрение процесса обучения весьма поучительно. Имеющиеся сейчас автоматы могут учиться только очень примитивным способом. Рассмотрим телефонную систему. Если пункт A может быть соединен с пунктом C только посредством коммутационного оборудования в некотором промежуточном пункте B и если разговоры между A и C происходят достаточно часто, то было бы желательно облегчить образование соединения между A и C.

Такие ситуации, конечно, возникают, и они решаются разными методами. Если кто-либо подозревает, что возникла такая ситуация, то проводится статистическое исследование и оценка стоимости и, если это оправдано, устанавливаются соответствующие соединительные линии, или же если два абонента разговаривают между собой достаточно часто, то им становится экономически целесообразным арендовать особую частную линию.

Но если бы телефонная система была аналогична нервной системе человека, то всякий раз, когда устанавливалось бы соединение, делалось бы легче устанавливать соответствующий конкретный путь в следующий раз, так что часто используемые пути стало бы, по-видимому, легко восстанавливать. Разумеется, у нас сейчас нет искусственных систем большого масштаба, действующих по такому принципу, но, пожалуй, было бы поучительно рассмотреть возможные применения таких систем.

Мозг против вычислительной машины. Если бы человек воспринимал 100 битов информации в секунду в течение 100 лет и ничего из нее не забывал, он накопил бы около 3×10^{11} битов информации. Мы не знаем точно, какова емкость памяти у человека, но это число, по-видимому, составляет верхнюю границу. С другой стороны, нет определенных доказательств, чтобы кому-либо не хватило емкости памяти, кроме как в психопатических состояниях.

Время доступа для этой памяти велико по сравнению с тем, которое требуется в современных вычислительных машинах, порядка 0,1 *сек*, но эта память полностью является внутренней памятью и чрезвычайно эффективна по своей доступности. Мы можем, например, вспомнить нужные биты информации на основании весьма смутных намеков; чтобы запрограммировать вычислительную машину примерно с такими же способностями, нам нужно было бы иметь миллионы или миллиарды различных классов категорий при выборе.

Человеческий мозг работает при значительно меньшей мощности, чем автоматическая цифровая вычислительная машина, но ни машина, ни мозг не имеют высокого теплового коэффициента полезного действия (хотя это не имеет значения; как говорит Винер [53]: «Статистическая механика Гиббса, возможно, является достаточно адекватной моделью того, что происходит в живом теле; картина, которую нам подсказывает аналогия с обычным тепловым двигателем, конечно, не адекватна»).

Возможно, более замечательным, чем малое рассеяние энергии в вычислительном устройстве человека, является малый объем этого устройства. Сто тысяч битов быстродействующей памяти в наиболее компактной форме, какая только известна в настоящее время, заполнили бы целую комнату. Многие специалисты считают, что мы могли бы, по крайней мере в теории, построить вычислительную машину для выполнения какой-нибудь одной определенной функции человеческого мозга, и попытки описать, как будет построена машина, выполняющая ту или иную данную функцию (например, играющая в шахматы, или обучающаяся игре в шахматы, или тому подобное), уже привели к более глубокому пониманию сущности логического проектирования. Но нужно добавить ограничение «по крайней мере в теории», потому что такие машины всегда оказываются такими большими (и такими дорогими), что они становятся

неосуществимы. Очевидно, из изучения человеческого мозга можно многому научиться в вопросах конструирования компактных и эффективных компонентов вычислительных устройств.

При дальнейшем сравнении мозга с электронной вычислительной машиной мы замечаем, что вычислительная машина гораздо надежнее. С другой стороны, недостаточная надежность мозга компенсируется наличием многократных (избыточных) путей, по крайней мере для важных сообщений. Иногда у людей, пострадавших от несчастных случаев, при которых у них была разрушена значительная часть мозга, не было заметно никакой потери памяти или изменения личности или способностей. Винер говорит, что, сравнивая мозг с цифровой вычислительной машиной, нужно сравнивать жизнь мозга не со всем сроком службы машины, а с одним циклом ее работы, т. е. от считывания одной задачи до напечатания решения, ибо описание машины (или мозга, рассматриваемого как машина) неполно без указания содержащейся в ней информации. Память машины можно стирать в промежутке между двумя задачами, но полное стирание человеческой памяти возможно только со смертью.

Винер дает интересный разбор навязчивой тревоги, сопровождающей неврозы страха. Он предполагает, что в этом состоянии циркулирующие воспоминания, наполненные тревогой, «вовлекают все большее число нейронов в свою систему, пока не захватят значительную часть нейронной сети. В этом случае, возможно, у больного просто нет места, нет достаточного числа нейронов для выполнения нормальных процессов мышления».

Если бы сходное нарушение произошло в электрической вычислительной машине (и Винер утверждает, что оно встречается), то мы очистили бы машину от всех данных и пустили бы ее снова, или же встряхнули бы машину, или подали бы на нее ненормально большой электрический импульс, или при необходимости отсоединили бы часть аппаратуры, вызывающую затруднения, и попробовали решить задачу с оставшейся частью.

Для каждого из этих видов терапии имеется нечто подобное в лечении психопатических состояний. Конечно, полное очищение невозможно, но сон ближе всего к полному очищению, и сон действительно является превосходным способом лечения тревоги (хотя «сильная тревога не дает заснуть по-настоящему»).

Во-вторых, мы можем применять шок: фармацевтический (метразол, инсулин) или электрический, — чтобы разорвать недавние (циркулирующие) воспоминания, в дополнение к каким-то плохо понимаемым влияниям на постоянную (синаптическую пороговую) память.

Наконец, можно удалить переднюю долю коры головного мозга; такая «префронтальная лоботомия, по-видимому, действует на навязчивое состояние не тем, что она помогает больному разрешить мучащие его вопросы, а тем, что она повреждает или уничтожает способность к продолжительной тревоге ... Лоботомия, по-видимому, ограничивает все виды циркулирующих воспоминаний». Наконец, Винер говорит, что эффективность психоанализа в подобных случаях можно объяснить такими же соображениями.

Обратная связь. Винер анализирует также другую группу болезней — различные атаксии, разного рода нарушения способности координировать волевые движения. Человек не может совершать координированные движения без обратной связи. Если мы посмотрим на чашку, стоящую на столе, и затем попытаемся с закрытыми глазами взять ее за ручку, то, вероятно, промахнемся на два или три дюйма; при нормальном процессе протягивания руки за вещью существует непрерывная зрительная обратная связь для положения руки, с надлежащей корректирующей цепью, непрерывно замыкающей цепь наведения.

При ходьбе кинестетическая обратная связь позволяет нам правильно ставить ноги. Но при одной болезни, которая называется *локомоторной атаксией,* часть спинного мозга больного разрушена сифилисом и кинестетические ощущения уже не передаются от ног к мозгу. Больной все время следит за своими ногами при ходьбе, чтобы видеть, где они находятся; он выбрасывает ногу вперед и неуверенно опускает ее на землю, как бы скользя по полу. При другом виде болезни, известном под названием *церебеллярной атаксии,* мозжечок (часть мозга) больного поврежден вследствие опухоли или несчастного случая и тонкий механизм обратной связи изменяется в обратную сторону. Когда больной протягивает руку за чашкой с открытыми глазами, он промахивается, а когда пытается поправиться, его рука может впасть в сильные колебания.

Инженеру известна важность обратной связи управляющих сигналов в следящих системах, положительной обратной связи в регенеративных усилителях, отрицательной обратной связи в системах звуковоспроизведения и т. п. Но главное, что мы должны усвоить из этой аналогии, — это важность информацион-

ной обратной связи. Например, современные самолеты имеют световой сигнал или семафор, указывающий, когда колеса опущены и зацеплены; возможно ведь, что летчик приведет в действие надлежащие органы управления, а механизм не будет реагировать как нужно. И однако необходимость в такой обратной связи стала очевидной лишь после того, как произошли несчастные случаи из-за ее отсутствия.

Различные сигналы, которые слышны в телефонной трубке: готовности станции, занятости, вызова, вводятся как чисто информационная обратная связь. В частности, сигнал вызова есть чистая подделка: мы слышим не звонок, звонящий в аппарате вызываемого абонента, но наш аппарат присоединяется к генератору звука, напоминающего телефонный звонок. Эта информационная обратная связь оказалась необходимой, потому, что без нее абоненту надоедает ждать и он может повесить трубку до того, как получит ответ.

Можно принять за принцип проектирования систем, что всякий важный переданный сигнал, всякая важная отданная команда, всякое важное совершаемое действие должны иметь информационную обратную связь — определенное уведомление, что сигнал был принят и понят, что команда была исполнена или что действие было совершено.

Общие представления. Человек может распознать нечто общее между малым и большим квадратом; между черным квадратом на белом фоне, белым квадратом на черном фоне и зеленым квадратом на оранжевом фоне; между заполненным квадратом и квадратом, который обведен по контуру или просто намечен пунктиром; между квадратом с вертикальными и горизонтальными сторонами и квадратом, поставленным на один угол; и даже между квадратом, видимым в направлении, перпендикулярном к его плоскости (так что он выглядит как квадрат), и квадратом, видимым под углом (так что он выглядит как ромб). Все они имеют общее свойство квадратности, и человеческий глаз и мозг распознают это *общее представление,* или *гештальт**, почти мгновенно (т. е. за время около 0,1 *сек*).

Нам будет ясно, насколько замечательно это свойство, если мы попытаемся представить, как бы мы программировали вычислительную машину, чтобы исследовать картину (например, при помощи обегающего луча, как в телевидении) и определить, содержит ли она какие-нибудь квадратные фигуры.

Эту способность распознавать общие представления проектировщик автоматов хотел бы повторить в своих системах. Например, мы хотели бы построить системы, которые могли бы распознавать приближение какого-либо важного события (например, такого, как перегрузка). Мы хотели бы построить машину, которая реагировала бы на человеческий голос, независимо от высоты, тембра, силы, акцента и т. п., и просто выбирала бы «важные» звуки, определяющие произносимые слова. Мы хотели бы построить машину, которая различала бы радиолокационные сигналы, вызванные целями, от сигналов, вызванных помехами, и делала бы это по крайней мере так же хорошо, как тренированный оператор, смотрящий на индикатор кругового обзора. Мы хотели бы построить машину, которая просматривала бы напечатанную страницу и производила бы последовательность сигналов, зависящую только от напечатанных там букв, независимо от их размера или шрифта.

Распознавание образов. Последняя из названных задач в настоящее время интенсивно изучается многими группами как метод быстродействующего ввода информации в вычислительную машину. Однако первоначально ее изучение было предпринято с целью помочь слепым читать. Эту задачу изучали несколько групп, в частности У. Мак-Каллох и У. Питтс. Основная цель при этом — производить при сканировании букв 26 различных звуков, соответствующих 26 буквам алфавита (звуки не обязательно должны быть обычными названиями букв: слепые могут научиться истолковывать звуки, как они истолковывают при чтении знаки письма Брайля). Трудность состоит в том, чтобы построить машину, которая могла бы, например, соревноваться со зрительным механизмом человека в распознавании «А-образности» таких разных символов, как a, *a*, **a**, ***a***, A, *A*, **A**, ***A***, **A** и ***A***.

Мак-Каллох и Питтс поставили себе более простую цель: они ограничились печатными заглавными буквами и одним или двумя простыми шрифтами. Они предположили, кроме того, что каждая строка печатного текста строго горизонтальна и центрирована. Тогда они сканируют строку горизонтально тремя фотоэлементами: одним у верхней части букв, производящим высокий тон; одним у нижней части букв, производящим низкий тон; и одним промежуточным по положению и по тону. На рис. 25.9 показаны результирующие звуки, производимые некоторыми буквами. Хотя

* Гештальт (нем. Gestalt) — целостная форма, целостный образ; термин так называемой «гештальтпсихологии». — *Прим. ред.*

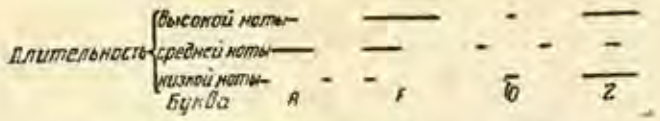


Рис. 25.9. Звуковые формы букв в приборе Мак-Каллоха—Питтса.

различение некоторых букв (например, D и O) может быть несколько затруднительным, в общем такие сочетания можно выучить и распознавать. Применив пять тонов вместо трех, можно достигнуть лучшего различения.

В качестве первого шага к тому, чтобы создать на этой основе практический прибор для чтения разнообразного печатного материала, Мак-Каллох и Питтс стремились отрегулировать его так, чтобы он читал шрифт разных размеров. Получившийся прибор изображен схематически на рис. 25.10. Группы переключателей *A*, *B*, *C* и т. д. срабатывают последовательно, пока не будет найдена надлежащая группа; эта группа затем удерживается замкнутой, пока фотоэлементы сканируют букву или строку. При этом даже наименьший шрифт, какой можно прочитать (т. е. шрифт, максимальная высота которого как раз простирается на три средних фотоэлемента), возбуждает по меньшей мере средний генератор и два крайних генератора; для большего шрифта, кроме этих генераторов, используются и некоторые промежуточные.

Интересно отметить в связи с этим один эпизод, о котором рассказывает Винер [53]. Когда схему на рис. 25.10 показали д-ру Гергардту фон Бонину, известному нейрофизиологу, он спросил: «Это схема четвертого слоя зрительной области коры головного мозга?» Имеются указания, что этот слой выполняет такое же назначение — сканировать образы разного размера и распознавать общие представления. Действительно, существует ярко выраженный ритм мозговых волн, так называемый *альфа-ритм*, связанный, согласно убедительным данным, со зрительным процессом и имеющий частоту около 10 *гц*; как утверждает Винер, этот ритм, по-видимому, связан с процессом сканирования, подобным применяемому в описанном приборе.

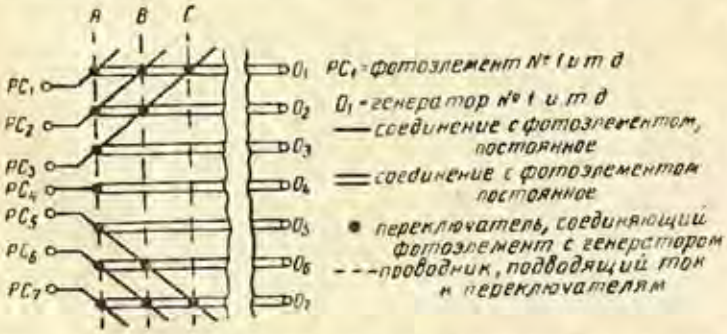


Рис. 25.10. Схема прибора Мак-Каллоха—Питтса для размера шрифта.

Харрингтон, Селфридж, Диннин и другие в Массачусетском технологическом институте [69] изучали другой метод распознавания образов. Образ подвергается развертке, подобной телевизионной развертке, при которой каждый элементарный участок отмечается как черный или белый. Разрешение соответствует 90 строкам (т. е. развертываемый участок разбивается на 8100 квадратов, «белых» или «черных»), и полученные сигналы вводятся в запоминающее устройство цифровой вычислительной машины, причем черному соответствует 1, а белому 0.

Затем результаты подвергаются операции *усреднения* для устранения шума и сглаживания случайных колебаний. Эта операция выполняется путем исследования всех квадратов, соседних с каждым квадратом (соседним квадратом считается квадрат, отстоящий от данного не больше чем на два квадрата); всего имеется 24 таких соседних квадрата, и если из них черных больше определенного числа (скажем, 10), то данный квадрат объявляется черным независимо от того, был ли он признан сначала черным или нет; если из них меньше 10 черных, то квадрат называется белым независимо от того, был ли он признан сначала белым или нет. Эта операция приводит к исключению рассеянных точек на изображении и заполнению пропусков в основном образе.

Рис. 25.11. Точки симметрии.

Результаты операции усреднения подвергаются операции *окантовки*, при которой выделяются элементы разрывности, скажем, узлы и конечные точки букв. Операция окантовки может превратить единицы в нули, но не может изменять нули; 1 превращается в 0, когда ситуация вокруг данного квадрата обнаруживает высокую степень симметрии (как для точек *s* на рис. 25.11, но не для точек *u*). Вычислительная машина с этой целью сканирует квадраты вокруг данного квадрата *p*. Найдя черный квадрат, она осведомляется, является ли диагонально противоположный квадрат (т. е. расположенный по другую сторону квадрата *p*) белым; если да, то это является элементом асимметрии. Машина считает общую степень асимметрии вокруг *p*, приписывая наибольший вес квадратам, самым близким к *p*, и если эта асимметрия не превысит

некоторый заданный уровень, символ квадрата p меняется на 0 (белый).

Результатом окантовки буквы А на рис. 25.11 будет выделение пяти точек u и устранение остальной части фигуры; этот результат будет иметь место независимо от того, было ли А первоначально напечатано курсивом, прямым или жирным шрифтом.

Следующий шаг — сосчитать точки и отличить эту точечную фигуру от фигуры, образованной какой-нибудь другой буквой (например, буквой К), состоящей также из пяти точек, — представляется выполнимой задачей для цифровой вычислительной машины. Для нас интересно следующее замечание автора [69а]: «В нейрофизиологии имеются доказательства, что нервные системы многих животных производят усреднение и окантовку зрительных образов».

Кибернетика и проектирование систем. Термин «кибернетика» истолковывался поразному разными авторами. В самом узком смысле она равнозначна теории следящих систем (теории автоматического регулирования). В широком смысле она охватывает по существу все естествознание. Мы разобрали здесь одну сторону — что нам дает для более глубокого понимания проектирования больших систем изучение механической организации животных и людей.

Но у этой кибернетической медали есть и другая сторона: мы можем также много узнать о людях путем изучения автоматов. Тривиальный пример: возможно, мы меньше бы понимали действие четвертого слоя зрительной коры головного мозга, если бы не было попыток построить читающую машину. Не столь тривиальным примером является надежда на то, что мы сможем достигнуть некоторых успехов в борьбе против психических болезней благодаря более основательному пониманию психических процессов, которое мы приобретем, изучая построенные человеком вычислительные машины.

Человеческий организм и физический мир являются очень сложно действующими системами. Неудивительно поэтому, что имеются поучительные аналогии между ними и сложными системами, которые в последнее время начал строить человек. Неудивительно и то, что, пытаясь строить такие системы, человечество узнало кое-что об их свойствах, чего бы оно, возможно, и не узнало при созерцании систем, построенных природой. В гл. 1 было указано, что уже вследствие количественных различий в масштабе и в сложности современные большие системы стали качественно отличными от меньших систем, которые изучались раньше; по той же причине они стали качественно более близкими к двум системам, всегда представлявшим главные объекты научного исследования: к самому человеку и к физическому миру, составляющему его окружение.

Итак, в той мере, в какой существует наука о проектировании систем, эта наука важна не только для инженера, но также и для представителей самых точных и самых антропоцентрических наук. И по мере того как системотехник будет строить все более сложные системы, он будет помогать человечеству не только управлять окружающей средой, но и понимать ее.

ЛИТЕРАТУРА

Хорошим введением в линейное программирование является книга Чарнеса, Купера и Гендерсона [64]. Более глубокие знания дает книга под редакцией Купманса [92]. Мы не знаем подходящей книги по групповой динамике. «Кибернетика» Винера [53], хотя и написана не для читателя, которого мы имеем в виду, является одной из самых значительных книг нашего поколения, и мы от всего сердца рекомендуем ее каждому инженеру; читатель без математической подготовки может пропустить гл. 2—4.

ГЛАВА 26

МОДЕЛИРОВАНИЕ

Мы коснулись моделирования в нескольких местах этой книги и описали более или менее подробно ряд примеров моделирования. При рассмотрении математических моделей (§ 10.3) мы представили две модели уличного движения для иллюстрации метода «Монте-Карло». Во введении в теорию вычислительных машин (§ 14.1) мы указали на важность моделирования в реальном времени и привели некоторые краткие примеры применений моделирования для испытания систем и тренировки операторов. В § 18.2 и 18.3 мы описали довольно подробно модель массы, подвешенной к пружине, а также менее по-

дробно — модель перехвата по кривой погони и более сложную модель прицела для воздушной стрельбы, в которой одним из элементов был вычислитель курса погони. В § 23.6 была представлена модель очереди.

В этой главе мы не будем заниматься применениями моделирования к частным типам задач и приводить дальнейшие примеры моделей. Вместо этого мы рассмотрим место моделирования в проектировании систем, выполнение программы моделирования и необходимые этапы, входящие в нее.

Чтобы ограничить наш разбор, мы определим здесь моделирование как изучение системы путем пробных исследований ее математического представления, проводимых при помощи большой вычислительной машины. Часто бывает трудно отличить моделирование от вычисления, но такие тонкие различия имеют главным образом семантический интерес; если ситуация такова, что мы не уверены, называть ли ее моделированием или вычислением, то нижеследующие рассуждения, вероятно, можно применить в обоих случаях.

26.1. Моделирование, анализ и испытание

Предполагается, что математическая модель системы сформулирована и имеет вид группы отдельных выражений, каждое из которых изображает «ящик», который может быть чем угодно — от элемента до подсистемы. Выходы одного такого выражения становятся входами следующего, с надлежащими коэффициентами и, конечно, с петлями обратной связи. Как было указано в связи с рис. 18.10—18.12, иногда можно провести аналоговое моделирование даже без формулировки соответствующих уравнений, но в этих случаях уравнения фигурируют неявно.

Моделирование как метод приобретения информации о работе системы следует сравнить с математическим анализом, с одной стороны, и с действительным экспериментальным испытанием, с другой. Анализ, конечно, гораздо дешевле моделирования, которое, в свою очередь, гораздо дешевле испытания; для большой системы различия в стоимости огромные. Эти выводы справедливы независимо от того, выражены ли стоимости в деньгах или во времени.

В то время как моделирование может стоить от нескольких тысяч до нескольких сот тысяч долларов и занимать от нескольких недель до многих месяцев, создание и испытание полной системы может стоить много миллионов долларов и занимать многие годы работы. При таких обстоятельствах метод проектирования систем путем проб и ошибок (попыток и пересмотров) просто недопустим; проектировщик еще до начала изготовления опытного образца должен быть в достаточной степени уверен, что система в основном построена правильно.

Анализ — мощное и решающее орудие там, где он применим, т. е. обычно для окружения, позволяющего с достаточной точностью применять линеаризацию. Но с ростом сложностей в виде нелинейности и даже разрывов, при шумных (или статистических) входах, при длинном пути через много ящиков и при участии людей в системе аналитик должен сдаться.

С другой стороны, испытание системы является наиболее реалистическим просто потому, что исследуется сама система. Но испытанием нельзя охватить широкий диапазон значений существенных переменных — отчасти из-за затрат и отчасти из-за того, что могут отсутствовать определенные условия [например, тропические и/или арктические метеорологические условия].

Таким образом, во всех этих отношениях—по стоимости, времени, реализму, диапазону переменных — моделирование занимает промежуточное положение между анализом и действительным экспериментальным испытанием. Значение этих двух этапов проектирования не следует преуменьшать, ибо они суть необходимые орудия при проектировании систем. Но моделирование — чрезвычайно важное дополнительное орудие. Можно уверенно сказать, что в настоящее время ни одна большая система не будет создаваться без какого-либо моделирования и в хорошо выполненном проекте системы моделирование будет применяться постоянно. Как указано в § 14.1, возможность моделировать отдельные блоки блок-схемы более или менее независимо от других означает, что моделирование можно сочетать разными способами с анализом или с испытанием. Так, мы можем испытывать один компонент и моделировать остальную систему или моделировать некоторые случаи и с помощью анализа экстраполировать (или скорее интерполировать) другие случаи.

26.2. Этапы моделирования

Мы приводим ниже контрольный перечень этапов, которые надо осуществить при выполнении какой-либо программы моделирования. Некоторые из них рассмотрены в других местах книги, в частности в ч. 4. Содержание остальной части этой главы составляют дальнейшие замечания о некоторых этапах моде-

лирования. Это следующие этапы: выбор вычислительной машины, программирование, предварительные (аналитические) решения, выбор исследуемых случаев, кодирование или путевое картирование, настройка машины и отладка программы, обработка данных, анализ результатов и составление отчета.

Выбор вычислительной машины. Этот выбор, естественно, будет зависеть от имеющихся машин. Если математическая модель состоит в основном из дифференциальных уравнений, обычно следует предпочесть аналоговую вычислительную машину; если модель состоит в основном из арифметических операций, или имеет логическую природу (игра «ним»), или состоит из разностных уравнений, часто более пригодной окажется цифровая вычислительная машина. Если моделирование должно быть проведено в реальном времени и, в особенности, если в моделировании действительно участвуют люди, аналоговая установка будет более привлекательна. При создании специализированной модели следует уделить внимание необычным установкам, использующим сочетания аналоговых и цифровых методов.

Необходимо рассмотреть емкость, точность и максимальное содержание частот. У аналоговых машин емкость будет определяться размером машины; у большинства цифровых машин, если емкость внутренней памяти оказывается недостаточной, дополнительные данные можно передавать во внешнюю память за счет некоторого увеличения времени. Содержание частот просто определяет время (и, следовательно, затраты), если только моделирование не должно проводиться в реальном времени. Требования к точности нужно исследовать уже на первых шагах моделирования. В аналоговых моделях они определяют, будет ли машина удовлетворять этим требованиям; в цифровых моделях точность можно увеличить методами двойной разрядности, но это дорого, и любое такое требование нужно знать заранее.

При моделировании, как и во всех задачах вычислительной машины, наблюдается тенденция уделять слишком много внимания центральным вычислительным устройствам за счет менее интересных входно-выходных устройств. В частности, там, где имеется большой объем данных и сравнительно небольшой объем вычислений над каждым входом, требования к входно-выходным устройствам будут иметь существенное влияние на выбор типа машины.

Программирование. Поскольку мы предполагаем, что математическая модель сформулирована, построить блок-схему вычислений довольно просто. Хотя это является технической задачей, которую должен решать специалист по вычислительным машинам, проектировщик системы должен принять участие в этом этапе, хотя он, возможно, и не будет следить за последующими подробностями кодирования или путевого картирования. Когда блок-схема закончена, она обычно оказывается слишком сложной, и в ней необходимы некоторые упрощения.

В общем случае трудно сказать, в чем должны состоять эти упрощения. Это могут быть линеаризации или подстановки среднего значения распределения вместо всего распределения. Часто там, где вход в группу блоков распределен вероятностным образом, форма выходов аналитически известна и всю группу можно заменить прямо распределением вероятностей выхода. В большинстве моделей подобные упрощения будут легко видны.

Обычно задача может быть решена разными методами. Так, мы можем генерировать на аналоговой машине какую-нибудь функцию, набрав на машине дифференциальное уравнение, решением которого она является, вместо того, чтобы применять специальный генератор функций; или мы можем представлять все переменные их логарифмами, если в задачу входит мало сложений и много умножений и извлечений корней. На цифровой машине при итерационных процессах можно применять несложный процесс (линейную интерполяцию) с малыми интервалами или более сложный процесс (криволинейную интерполяцию) с большими интервалами. Подобные вопросы обычно лучше всего передавать специалисту по вычислительным машинам.

Выход машины должен быть возможно ближе к окончательной форме, что приводит к уменьшению необходимых ручных вычислений. Однако такой выбор выхода не будет оправдан, если он требует чрезмерного увеличения аппаратуры.

Предварительные решения. По многим причинам, прежде чем проводить моделирование, необходимо знать приблизительно, какие будут ответы, по крайней мере знать их форму и по возможности их приблизительный диапазон. Это позволит избежать ошибок при кодировании, выбрать надлежащим образом значения параметров, выбрать надлежащим образом переменные, которые нужно записывать, и осуществить эффективную обработку данных.

Если моделирование сравнительно просто (в этом случае вычислительная машина при-

меняется потому, что она может быстро дать решение для очень большого числа случаев), то небольшое число решений можно вычислить вручную. При более сложном моделировании аналитические решения нужно находить либо путем грубых приближений, либо путем рассмотрения вырожденных случаев. Вырожденные случаи (например, когда некоторые параметры равны нулю) часто бывает очень легко решить, и, кроме того, их можно сравнить с физической интуицией, чтобы проверить модель. Грубые приближения могут выразиться просто в пренебрежении шумом и сохранении при вычислениях только одной значащей цифры; кроме того, можно вводить линеаризации, применять средние значения распределений и обрывать ряды.

Когда моделирование проводится на аналоговой вычислительной машине, почти всегда важно произвести полную проверку одного случая на цифровой машине. Это полезно не только для первоначальной настройки и отладки задачи, но и для последующих проверок работы машины. Через установленные промежутки времени или при появлении в машине неисправностей можно ввести в машину контрольные параметры и затем измерить контрольной аппаратурой напряжения во всех точках машины, пока не будет найден неисправный компонент или пока мы не убедимся в том, что все компоненты работают надлежащим образом.

Выбор случаев. Это важный этап, и для его разумного проведения необходим прежде всего практический опыт. Хотя число полных испытаний и конкретные значения параметров должны быть установлены заранее, но вообще нужно предусмотреть большую гибкость, так как первые результаты часто будут указывать на то, что нужно подчеркнуть другие пункты программы. По той же причине несколько первых испытаний нужно провести на широком диапазоне значений параметров, включая по меньшей мере значения, соответствующие минимуму, максимуму и предполагаемому оптимуму.

Часто наблюдается тенденция требовать продолжения опытов, пока не будет оставлена далеко позади точка, после которой дальнейшие опыты будут вносить лишь незначительные поправки. Гибкость не нужно доводить до такой степени, чтобы моделирование нельзя было довести до разумного завершения.

Разумное проектирование моделирования часто может сильно уменьшить число необходимых опытов [70]. Например, часто бывает возможно провести бо́льшую часть опытов при различных значениях параметров в отсутствие шума и затем провести несколько серий опытов с шумом при надлежащих наборах значений параметров. Затем можно учесть влияние шума, комбинируя аналитически эти результаты с результатами испытаний без шума. Если для получения статистически надежного результата при наличии шума требуется 20 полных испытаний, то этим способом можно уменьшить общее число необходимых полных испытаний почти в 20 раз.

Обработка и анализ данных. Фактически необходимо, чтобы во время проведения опытов по моделированию выполнялась в какой-либо форме и обработка данных. Если обработка данных состоит в основном в табуляции результатов, выданных машиной, то никаких затруднений не возникает, но если необходим сложный анализ, то обработку данных нужно делать в каком-нибудь упрощенном виде после каждого полного испытания. Если этого не делают, то после завершения анализа (возможно, через несколько месяцев) обнаруживается, что многие испытания были излишними, тогда как нужно еще проделать несколько решающих испытаний. Если моделирование проводилось на цифровой машине, может оказаться не слишком трудным снова пропустить ленты, чтобы получить добавочные серии результатов, но на аналоговой машине, после того как «вынуты штепсели», настройка дополнительного испытания может оказаться чрезвычайно утомительным делом.

Моделирование людей. Там, где это возможно, желательно имитировать человека-оператора какой-нибудь простой аналитической функцией, например сдвигом во времени. Это устраняет необходимость моделирования в реальном времени и позволяет использовать распределение, типичное для всей совокупности. Однако когда реакция человека зависит от различных переменных, его нужно включить в модель. Здесь нужно посоветоваться с инженером-психологом (гл. 30), чтобы удостовериться, что ситуация реальна и получаются надлежащие выборки реакций.

ЛИТЕРАТУРА

Большинство идей этой главы были опубликованы несколько лет назад [71]. По-видимому, хорошей книги по моделированию нет, но было несколько интересных комплектов трудов конференций по этому вопросу, например [72—75].

ГЛАВА 27

СОСТАВНЫЕ ЧАСТИ СИСТЕМЫ

До сих пор мы в основном занимались описанием системных задач и подходом к их решению, обращая мало внимания на «металл», т. е. на воплощение этих решений в оборудование. Как указано в § 3.3, оборудование подобного рода целесообразно разбить на шесть классов:

1) входные устройства, принимающие информацию или материал, с которыми будет работать система;

2) устройства связи, создающие информационные звенья для соединения различных частей системы;

3) устройства логического управления, направляющие поток информации, обрабатывающие информацию для извлечения новой информации и, если нужно, управляющие потоком материала в системе;

4) устройства рефлексивного управления, управляющие повторяющимися, кратковременными ответами;

5) устройства подачи, перемещающие материал по системе;

6) выходные устройства, куда входят как исполнительные механизмы, выполняющие конечные действия над материалом, так и индикаторы, доставляющие информацию людям, контролирующим систему.

Мы уже говорили, что хотя мы рассматриваем исключительно системы оборудования, главное значение имеют информационные аспекты этих систем. Поэтому устройства подачи, работающие исключительно с материалами, не рассматриваются нами нигде, за исключением этой главы. Входное и выходное оборудование обсуждалось в гл. 19 и с другой точки зрения будет обсуждаться в гл. 30. Однако мы рассматриваем исключительно информационные входы и выходы. Аппаратура логического управления была темой гл. 15 и 17. Аппаратура связи и аппаратура рефлексивного управления составляют соответственно содержание гл. 28 и 29.

Таким образом, настоящая глава, в которой по очереди кратко рассматриваются все эти классы оборудования, служит введением к следующим трем главам, в которых рассматриваются научные орудия, применимые к отдельным классам. Но у нее есть и более важная цель. Способ постановки задачи и поиски и нахождение решений зависят в значительной степени от того, насколько человек, стремящийся решить задачу, знаком с «состоянием техники».

Возьмем проектирование фондовой биржи, рассмотренное в § 9.3. В 1900 г. эта задача формулировалась во многом совсем иначе, чем в настоящее время, вследствие различий в методах хранения, передачи, обработки и индикации данных. Проектировщик системы, прежде чем начинать формулировку какого-либо решения, должен ознакомиться с общими типами имеющегося оборудования и с возможным диапазоном их рабочих характеристик. В этой главе упоминаются некоторые новые методы в каждом классе оборудования.

Нужно сразу отметить, что никакая классификация и описание оборудования, как бы обширны они ни были, не охватят всех компонентов, из которых системотехник будет строить свою систему. Новой чертой современного проектирования систем является регулярный порядок, с которым изобретаются необходимые устройства в результате формулировки потребностей системы. В известном смысле всякая попытка описать оборудование, которое можно использовать при построении систем, обречена на неудачу в самом начале, если при этом стремятся к полной каталогизации имеющихся устройств. Далее, при любой такой каталогизации приходится сталкиваться с высокой степенью специализации оборудования, проектируемого для конкретных систем. Рано или поздно почти любое устройство может понадобиться для какой-нибудь системы.

Поэтому в настоящей главе мы ограничимся рассмотрением устройств, поддающихся анализу на основе какой-либо единой теории, но даже и эта категория будет охвачена не полностью. Кроме того, мы постараемся привести примеры решений, принимаемых при выборе определенных классов оборудования. Рассматривать специализированное оборудование мы не будем.

27.1. Входная аппаратура

Оборудование требуется и для приема материальных входов системы, но мы обратим свое внимание на информационные входные устройства. Такие устройства удобно разбить на два класса, в зависимости от характера задачи. В некоторых случаях информация находится под рукой и трудность заключается в том, чтобы стандартизировать ее на возможно ранней стадии (как было рассмотрено в § 21.1 и для канадской почтовой системы

в § 2.4). В других случаях, в дополнение к проблеме стандартизации, объекты, относительно которых хотят получить информацию, находятся на расстоянии и основная трудность заключается в получении информации.

Стандартизация входов. Иногда можно изобрести специализированные автоматические устройства для стандартизации входов, как например: датчики давления для счета автомобилей, пневматические приборы, применяемые в автоматической почте, и механические щупы, применяемые в системах автоматического контроля в банках. Однако во многих случаях для перевода данных на язык системы используют человека; мы интересуемся здесь аппаратурой, которая придается этому оператору. Первое требование к любой такой аппаратуре, как было указано в § 25.3, состоит в том, чтобы имелась обратная связь к оператору от информации, которую он вводит в систему.

Устройства для ввода стандартизированной информации могут быть аналоговыми или цифровыми. К аналоговым устройствам относятся: пантограф (который может двигаться влево — вправо и вперед — назад, хотя только в одной плоскости), ручка управления (один конец которой может двигаться влево—вправо и вперед—назад, а другой закреплен в шарнирном соединении) и штурвал или колесо (которые могут вращаться по часовой стрелке или против часовой стрелки). Во многих самолетах ручка управления и штурвал объединены, так что прибор имеет три степени свободы; при желании можно предусмотреть даже четыре степени свободы (влево — вправо, вперед — назад, вверх — вниз и вращательное движение). Для того чтобы определить, насколько хорошо оператор сможет справиться с таким прибором, нужно посоветоваться с инженером-психологом.

Из цифровых устройств наиболее распространены диски (как номеронабирательный диск в автоматической телефонии) и клавиатуры. Клавиатуры могут быть специализированными и общей формы.

В специализированных клавиатурах панель имеет столько столбцов клавиш, сколько имеется категорий информации, и достаточное число позиций в каждом столбце, чтобы категорию можно было представить надлежащим символом. Так, для какой-нибудь категории может быть предусмотрено три позиции, обозначенные соответственно как «большая», «средняя» и «малая». Конечно, значение этих терминов должно быть установлено заранее. В некоторых категориях может требоваться ввод числа в диапазоне от 10 до 3500 с точностью до ближайшего десятка; в этом случае категория будет обозначаться тремя столбцами клавиш, а код будет вводиться в виде трех десятичных цифр. Каждый столбец (или группа столбцов) обычно отличается от остальных цветом, формой и/или размером.

Рис. 27.1. Клавиатура общего назначения.

Клавиатуры общего назначения, как, например, изображенная на рис. 27.1, имеют 10 клавиш для цифр от 1 до 9, исполнительную клавишу (*E*) и часто другие клавиши, для таких целей, как стирание («гашение») и подведение итогов. При нажатии любой клавиши в клавиатуру вводится число соответствующего порядка и появляется определенная последовательность цифр в окошке, на которое смотрит оператор. Как только вся информация набрана, оператор проверяет число в окошке и, если он удовлетворен, нажимает исполнительную клавишу. При этом информация посылается по своему пути в систему.

Такие клавиатуры [2] переводят информацию на язык, понятный системе. Этим языком могут служить импульсы, отверстия в бумажной ленте, отверстия в картах, пятна магнитных чернил или другие приспособления для автоматической обработки информации. При использовании подобных приспособлений от человека требуется, чтобы он был способен помнить и быстро применять код для перехода от языка внешнего мира к языку системы, ибо если применяются инструкции или книги кодов, ввод данных в систему чрезмерно замедляется. Эта способность операторов должна быть проверена совместно с инженером-психологом.

Устройства телевосприятия. Ситуации, в которых необходимо получить информацию об отдаленных объектах (называемых *целями* или *мишенями*), возникают в военных системах и в транспортных системах: аэронавигационных, морских и автомобильных. Восприятие на расстоянии связано с восприятием какой-либо энергии, излученной или отра-

женной от цели; хотя можно использовать весьма разнообразные виды энергии (инфракрасное излучение, механические колебания, магнитные или гравитационные поля и т. п.), больше всего применяются воспринимающие системы, основанные на использовании электромагнитной энергии либо в диапазоне видимых волн (свет), либо в диапазоне волн от 1 *см* до 1 *м* (радиолокация) или вблизи от него. Мы ограничимся рассмотрением радиолокации.

Трудно найти определение радиолокации, которое охватывало бы все современные значения этого слова. Обычно радиолокация понимается как измерение расстояния с помощью отраженных радиосигналов (т. е. как посылка сигнала, измерение времени до возвращения эхо от цели и определение по этим данным расстояния до цели). Однако этот термин охватывает также некоторые системы (маяки), которые не создают эхо в обычном смысле, и другие системы (непрерывного излучения), с помощью которых нельзя определять расстояние до цели *.

Классификация радиолокационных систем. Радиолокационные системы удобно делить на системы непрерывного излучения и импульсные системы. В первых передатчик излучает непрерывную последовательность волн, тогда как во вторых он накапливает энергию и периодически излучает ее короткими вспышками, или импульсами, состоящими из нескольких сот или нескольких тысяч волн. В обоих случаях, если излученная энергия попадает на цель и часть ее отражается, эта отраженная энергия может быть обнаружена приемником.

Нормально приемник и передатчик расположены в одном месте, и часто они являются дуплексными, т. е. используют одну и ту же антенну; но возможны также и *разнесенные* радиолокаторы, в которых приемник и передатчик отстоят друг от друга на большом расстоянии. В импульсных радиолокаторах можно измерять промежуток времени между излучением импульса и приходом переднего края отраженного сигнала и тем самым определять дальность до цели. Радиолокатор непрерывного излучения не может этого делать. Его преимущество — сравнительная простота, так как в нем не требуется высоких напряжений и накопления энергии и нет сложных схем формирования импульсов. Радиолокаторы обоих типов могут измерять направление на цель и могут быть снабжены СДЦ (селекцией движущихся целей). СДЦ основана на измерении допплерова сдвига, вызванного отражением от целей, имеющих отличную от нуля радиальную составляющую скорости **.

В обычном радиолокаторе происходит очень сильное ослабление энергии трояко рода: квадратическое ослабление при распространении энергии до цели, ослабление, обусловленное малым радиолокационным сечением цели (т. е. тем, что лишь небольшая доля энергии, падающей на цель, отражается к приемнику), и квадратическое ослабление на обратном пути. Поэтому типичная радиолокационная станция должна излучать импульсы порядка 1 *Мвт* и обнаруживать отраженные импульсы в 1 *пвт* (10^{-12} *вт*) или даже значительно меньше. Такие мощные передатчики и чувствительные приемники трудно и дорого строить и применять.

Если цель сотрудничает с радиолокационной станцией, как в навигационных системах коммерческих воздушных линий, то один или два из этих трех видов ослаблений можно устранить. Простейший способ — поместить на цели уголковый отражатель, отражающий значительную часть падающей энергии обратно к передатчику. Другой способ — поставить на цели маяк, или ответчик, и иметь передатчик для излучения кодированной серии импульсов, на которую маяк отвечает передачей серии импульсов, кодированной иначе. Это позволяет не только получить всю нормально доступную информацию от радиолокационной станции с гораздо меньшей мощностью и/или чувствительностью, но и опознавать цели (либо в военном смысле опознавания «свой — чужой», либо — при наличии достаточно полных кодов — в общем смысле индивидуального опознавания).

Телевоспринимающие системы можно подразделить также на активные, полуактивные и пассивные. В активных устройствах передатчик излучает энергию, а приемник воспри-

---

* Системы непрерывного излучения, использующие эффект Допплера, действительно не могут определять расстояние до цели, но системы непрерывного излучения, построенные по принципу фазовой или частотной модуляции, определяют расстояние до цели. См. другое примечание на этой странице. — *Прим. ред.*

** При методе непрерывного излучения обнаружение отраженного сигнала возможно только в том случае, когда частота отраженного сигнала отличается от частоты излученного сигнала. Этого можно достичь либо путем использования эффекта Допплера, либо путем применения фазовой или частотной модуляции. В первом случае можно определить радиальную составляющую цели и направление на цель, но не расстояние до нее. В случае же фазовой или частотной модуляции излучаемых колебаний в результате сложения излученного и отраженного сигналов возникают биения, частота которых зависит не только от радиальной составляющей скорости цели, но и от расстояния до цели. Поэтому такие системы непрерывного излучения способны определять дальность. — *Прим. ред.*

нимает отраженные сигналы от цели; в полуактивных устройствах приемник воспринимает энергию, отраженную от цели и излученную перед этим каким-нибудь другим устройством, входящим в систему; в пассивных устройствах передатчика нет и приемник воспринимает энергию, излучаемую самой целью.

Обычные радиолокационные и звуколокационные системы являются активными. Некоторые радиолокационные системы наведения реактивных снарядов являются полуактивными: бортовой радиолокатор снаряда управляет своим снарядом по отражениям от цели, которая *облучается*, или *сопровождается*, наземной радиолокационной станцией. Акустические и инфракрасные системы обнаружения самолетов являются пассивными. Обычное зрительное обнаружение самолетов нужно считать пассивным, потому что солнце — первичный источник воспринимаемой нами отраженной энергии — не является частью системы.

Функция телевосприятия в основном состоит из трех подчиненных функций: обнаружения, сопровождения и определения дополнительных данных об индивидуальности и/или свойствах цели. В простейших системах (например, в так называемых *ограждающих радиолокаторах* *) единственной функцией системы является обнаружение; основным выходом такой системы является констатация, что в интересующей нас зоне находится по меньшей мере одна цель или нет ни одной цели. Сопровождение бывает различного вида: от определения одной или двух координат положения до определения всех трех координат положения и еще трех координат скорости (для упреждения). В дополнительные данные может входить опознавание «свой — чужой» (т. е. проверка безопасности) и индивидуальное опознавание или другие весьма разнообразные виды информации, как, например, размер цели.

Редко бывает, чтобы всю необходимую информацию можно было получить при помощи одной радиолокационной станции; в действительности, как отмечалось в § 2.1 по поводу наземного управления посадкой, только одно сопровождение может потребовать нескольких установок. Когда информацию от одной цели должны получать несколько установок, в системе необходимо предусмотреть две дополнительные функции, а именно: захват (локация цели добавочным воспринимающим устройством) и сопоставление (проверка того, что две установки следят за одной и той же целью). Эти функции часто вызывают больше трудностей при проектировании системы, чем основные функции.

Соотношения между параметрами радиолокатора. Вообще говоря, проектировщик системы задает рабочие характеристики радиолокационной станции (дальность, точность, разрешающую способность, скорость передачи информации, допустимое ухудшение при неблагоприятных метеорологических условиях и т. д.), а специалист по радиолокации задает характеристики аппаратуры (длину волны, пиковую мощность, среднюю мощность, частоту повторения импульсов, размеры антенны, метод обзора и т. д.). Однако эти две группы характеристик так тесно связаны и так сильно влияют на другие части системы, что проектировщик системы часто задает некоторые характеристики аппаратуры, в частности длину волны.

Так как усилители радиолокационной станции могут усиливать сигнал до какой угодно величины, можно обнаружить столь слабый сигнал, что он еле отличим от шума. Мощность шума в первой ступени радиолокационной станции равна

$$N = FkTB, \tag{27.1}$$

где k — постоянная Больцмана,
T — абсолютная температура.
B — ширина полосы.

Величина F называется коэффициентом шума (шум-фактором). Этот коэффициент равен 1 (0 *дб*) для теоретически идеального приема и несколько больше для действительных приемников, в зависимости от состояния техники. При низких радиолокационных частотах (УКВ) можно получить коэффициент шума 3 или 4 *дб*, тогда как на сантиметровых волнах можно получить самое лучшее 10 *дб*.

Для мощности сигнала в гл. 10 была выведена формула

$$S = \frac{P_t G_t \sigma A_r}{(4\pi)^2 R^4}, \tag{27.2}$$

где P_t — излучаемая мощность,
G_t — усиление передающей антенны по отношению к изотропному излучателю,
σ — радиолокационное сечение цели,

** «Ограждающими» радиолокаторами (fence radars) в США называют радиолокаторы непрерывного излучения, использующие эффект Доплера и потому способные определять только скорость цели и направление на нее, но не дальность до цели. Такие радиолокаторы применяются в качестве «часовых», обнаруживающих движение целей в сложном окружении, когда импульсные радиолокаторы не могут выделить сигнала движущейся цели на фоне сигналов местных предметов. — *Прим. ред.*

$A_э$ — эффективная площадь приемной антенны,

R — расстояние от радиолокатора до цели.

Если для передачи и приема применяется одна и та же антенна, как обычно бывает, то формулу часто видоизменяют, подставляя

$$G = \frac{4\pi A}{\lambda^2}. \qquad (27.3)$$

Тогда отношение сигнал/шум равно

$$\frac{S}{N} = \frac{P_t \sigma A^2}{4\pi\lambda^2 R^4 FkT\Delta} = \frac{P_t \sigma G^2 \lambda^2}{(4\pi)^3 R^4 FkT\Delta}. \qquad (27.4)$$

Уравнение (27.4) показывает, что увеличение длины волны может привести к увеличению дальности радиолокации или к ее уменьшению в зависимости от того, сохраняются ли постоянными размеры антенны (в этом случае усиление увеличивается при уменьшении длины волны) или постоянное усиление (в этом случае размер антенны уменьшается при уменьшении длины волны). Размер антенны может быть ограничивающим фактором, в частности в самолетных установках, но усиление также должно быть не слишком большим (потому что при слишком узком луче локация целей может быть затруднительной, схемы сопровождения могут быть нестабильны, может ухудшиться работа СДЦ и могут появиться другие трудности) и не слишком малым (потому что при слишком широком луче ухудшается точность и разрешение по азимуту и/или углу места).

Кроме того, длина волны, как было отмечено выше, ограниченно влияет на коэффициент шума и может также оказывать большое влияние на радиолокационное сечение цели. Так, у целей, имеющих большие плоские поверхности и/или отражающие углы, сечение пропорционально $1/\lambda^2$; у больших целей с острыми носами или сильно обтекаемых при наблюдении со стороны носа сечение пропорционально λ^2; а у очень малых целей (в том числе у дождевых капель) сечение пропорционально $1/\lambda^4$.

Величины S, N и P_t в формуле (27.4) обозначают мощность; P_t может означать либо среднюю мощность непрерывного или импульсного излучения, либо пиковую мощность импульсных радиолокаторов. Можно показать при некоторых разумных допущениях, что средняя мощность в основном определяет дальность обнаружения (радиолокационной станции дальнего обнаружения), а пиковая мощность — дальность сопровождения (радиолокационной станции сопровождения). Иначе говоря, увеличение пиковой мощности при сохранении постоянной средней мощности приводит к увеличению дальности сопровождения, но мало влияет на дальность обнаружения. Это объясняется тем, что радиолокационная станция сопровождения приблизительно знает, где находится цель, и может исключить шум, поступающий от других участков пространства; радиолокационная станция дальнего обнаружения не может этого сделать.

Пиковая и средняя мощности импульсного радиолокатора связаны соотношением

$$\frac{P_a}{P_p} = \tau f, \qquad (27.5)$$

где P_a — средняя мощность,

P_p — пиковая мощность,

τ — длительность импульса,

f — частота повторения импульсов.

Частоту повторения импульсов обычно делают как можно большей, насколько это совместимо с требуемой максимальной дальностью радиолокации: нельзя излучать следующие импульсы, пока предыдущий не успеет пройти максимальное расстояние, на котором требуется обнаруживать цель, и вернуться назад, при скорости распространения туда и обратно 150 *м/мксек*. Длина импульса определяет разрешение по дальности; две цели, находящиеся на одном и том же азимуте и угле места и отстоящие по дальности меньше чем на длину импульса, нельзя разрешить. Это соображение определяет верхний предел длины импульса; нижний предел ставится шириной полосы, которую нельзя делать меньше приблизительно одной-двух обратных величин длительностей импульса.

Пиковая мощность является одним из основных факторов, определяющих трудность построения электрических цепей; поскольку пиковая мощность составляет киловатты в небольших радиолокационных станциях и мегаватты в больших, ясно, что при больших мощностях практическое выполнение аппаратуры связано с затруднениями.

Существует ряд других соображений, но вышеуказанные играют главную роль при выборе параметров радиолокатора. Ввиду их взаимосвязи и ввиду того, что они отражаются на многих других сторонах системы, инженер-системотехник не может их не учитывать. Поэтому его «функциональное» задание на радиолокационную станцию почти всегда будет включать приблизительную длину волны и приближенный расчет всех параметров, входящих в (27.4).

27.2. Аппаратура связи

Проектировщик системы в значительной мере определяет линии связи, соединяющие отдельные части системы, поскольку это касается конечных точек этих линий, т. е. он указывает, какие точки нужно соединить и какие не нужно. Саму аппаратуру выбирает проектировщик компонентов, но опять-таки «функциональное» задание проектировщика системы будет ориентировано на существующие возможности и обычно будет указывать, какой класс устройств нужно использовать. В частности, такие выборы, как между аналоговой или цифровой связью и между радиорелейной или проводной линией, не будут предоставлены проектировщику компонентов.

Локальными критериями эффективности оборудования связи служат запаздывание, количество передаваемых сообщений, частота ошибок и, конечно, стоимость. Все они зависят друг от друга. Скорость, пропускную способность и частоту ошибок можно изменить в лучшую сторону, увеличив ширину полосы или мощность или введя каналы, менее подверженные помехам, но все это связано с добавочными денежными расходами. Эти соотношения объясняются в гл. 28, посвященной теории информации.

Так как эти соотношения не всегда очевидны, может случиться, что из-за нереалистического задания будет разработана чрезмерно дорогая система связи. Связь должна быть в равновесии с остальной системой. Например, это не было соблюдено в одной системе, в которой телетайпы использовались для связи между быстродействующими цифровыми вычислительными машинами; если допустимы задержки, свойственные телетайпам (порядка минут), то на быстродействующие машины было напрасно потрачено много денег. С другой стороны, было бы столь же плохо применять радиорелейную линию между счетно-перфорационными машинами.

Подобная несогласованность была обычной в ранние дни вычислительных машин, когда, например, была построена вычислительная машина с параллельной логикой и мегагерцевой частотой повторения импульсов, так что вычисления можно было производить в течение микросекунд, но основным запоминающим устройством служил магнитный барабан, так что после каждого вычисления машина должна была целые миллисекунды ожидать доступа в память. Хотя несогласованность такого рода уже редко встречается, необходимость согласования скорости связи со скоростью других частей системы, по-видимому, еще не вполне осознана.

Аппаратуру связи можно классифицировать различными способами. Сообщение может иметь форму речи, изображения или кода. Средством передачи может служить почтальон, проволока, радио и прочие средства (в прочие входят такие средства, как семафор, инфракрасные лучи и почтовый голубь). Сама передача может быть непрерывной, дискретной или промежуточной (например, перемешанное или дискретизированное сообщение или кодирование типа импульсно-временной модуляции). Наконец, связь можно разделить на такие виды, как телеграф, телевидение, радио и т. д.

По телеграфу и телетайпу посылают дискретные кодированные сообщения, причем обычно подразумевается проволочная связь, хотя широко распространен также радиотелеграф и радиотелетайп. Телефонная связь является непрерывной и речевой, проволочной или беспроволочной. В телевидении и фототелеграфии передаются изображения, и обычно они считаются непрерывными, хотя в них есть дискретные элементы. Телевизионные передачи обычно происходят без проволоки, но могут происходить и по проволоке (по коаксиальному кабелю). Фототелеграфные изображения легче передавать по проводам, однако нелегко определить во всех случаях разницу между фототелеграфией и телевидением.

Почтовая связь состоит нормально в передаче изображений, хотя в том случае, когда сообщения содержат написанные слова, изображения представляют собой просто удобный способ кодирования сообщения. Нормально их передают почтальоны, но во время II мировой войны в системе «V-Mail» * письма фотографировались и передавались на микропленке; возможно сочетание такой системы с фототелеграфией.

Мы рассмотрим здесь ряд методов связи, приобретающих все возрастающее значение в течение последних нескольких лет,— так называемые методы импульсной модуляции. Большинство современных автоматических схем передачи данных основано на следующих трех методах: *импульсно-кодовой модуляции* (ИКМ), *импульсно-широтной модуляции* (ИШМ) и *импульсно-временной модуляции* (ИВМ), называемой также *импульсно-фазовой модуляцией* (ИФМ). Первый метод

---

* «V-Mail» (Victory Mail — «почта победы») применялась в США во время II мировой войны для пересылки корреспонденций за море. — *Прим. ред.*

цифровой, два другие по существу аналоговые, хотя в них и используются импульсы.

Допустим, что мы хотим одновременно передать 24 разговора по такому каналу, как коаксиальный кабель. Каждый разговор поступает из телефонной линии, в которой он ограничивается полосой от 300 до 3300 *гц*. Классический метод — модулировать по амплитуде каждый разговор на особой несущей, разнеся эти несущие друг от друга, скажем, на 4 *кгц*. Для этого потребуется общая ширина полосы примерно 100 *кгц*. В приемнике отдельные несущие фильтруются и демодулируются и таким образом воспроизводятся сообщения, хотя и искаженные до некоторой степени шумом в канале.

Если доступна ширина полосы только 100 *кгц* и шум мал, эта система может быть весьма приемлемой. Но предположим, что в канале большой шум, что доступна более широкая полоса (скажем, 2 *Мгц*) и мы хотим использовать эту увеличенную ширину полосы для повышения отношения сигнал/шум. Теория информации (гл. 28) говорит нам, что мы можем это сделать, но при этом нужно применить какое-нибудь кодирование.

В системе ИВМ это кодирование выражается в том, что от каждого сообщения берутся «дискреты» (образцы) с частотой, превосходящей наивысшую имеющуюся частоту несколько больше чем в два раза *, скажем 8000 раз в секунду. Тогда в каждые 125 *мксек* мы получаем 24 дискреты (по одной дискрете от каждого сообщения) и синхронизирующий импульс; все они отстоят друг от друга на 5 *мксек*. Синхронизирующий импульс проходит через ряд линий задержки и благодаря этому появляется через интервалы в 5 *мксек*; каждый раз он запускает наклонное напряжение, и когда это напряжение достигает уровня дискретизированного напряжения, посылается сигнальный импульс. Таким образом, момент (или положение) сигнального импульса изображает по аналогии величину дискреты.

В приемнике выполняется обратная операция: синхронизирующий импульс, соответственно задержанный, запускает наклонное напряжение, которое останавливается по приходе сигнального импульса, воспроизведя тем самым передаваемое напряжение. Различные напряжения, образующие каждое сообщение, соответственно коммутируются в одно целое, и теперь можно восстановить посланное сообщение.

В действительности, конечно, измеряется момент, в который сигнальный импульс при своем возрастании достигает известной величины (скажем, половины своей полной амплитуды). Этот отсчет искажается шумом. Показателем качества системы, определяющим ее верность, является отношение максимального значения периода повторения импульса (в данном случае 5 *мксек*) о времени возрастания импульса (которое можно уменьшить приблизительно до значения, обратного ширине полосы, в данном случае до 1/2 *мксек*). Увеличение ширины полосы можно использовать для повышения отношения сигнал/шум путем уменьшения времени возрастания импульса.

В системе ИВМ мы измеряем промежуток времени между появлением одного импульса (синхронизирующего импульса) и появлением другого импульса (сигнального импульса). Система ИШМ по существу такая же, но с той разницей, что мы модулируем ширину импульса и измеряем промежуток времени между появлением и исчезновением импульса.

В системе ИКМ мы отбираем дискреты, как и раньше, и выдаем по одной дискрете каждые 5 *мксек*. Каждая дискрета пропускается через 10-битовый аналого-цифровой преобразователь и тем самым квантуется на один из 1024 уровней амплитуды. При бите 1 посылается импульс, при бите 0 импульса нет. Это требует импульса каждые 1/2 *мксек*, что как раз и возможно при ширине полосы 2 *Мгц* (конечно, импульсы будут далеко не прямоугольные). Частота ошибок (отсчет нуля вместо единицы и наоборот) зависит от отношения сигнал/шум в канале. На практике это отношение обычно делают достаточно большим (скажем, 20 *дб*), чтобы частоту ошибок для всех практических целей можно было считать равной нулю. Тогда единственным искажением сообщения является искажение, обусловленное *шумом квантования*, которое возникает в аналого-цифровом преобразователе. Если доступная ширина полосы в два раза больше, то мы можем ввести в преобразователь добавочный разряд и тем самым увеличить отношение сигнал/шум.

Мы видим, что величину отношения сигнал/шум, ширину полосы и скорость передачи информации можно улучшить одну за счет другой. В теории информации выводятся количественные соотношения между этими величинами.

27.3. Аппаратура логического управления

В гл. 15 и 17 уже был дан довольно подробный разбор существующего оборудования

* В § 28.3 показано, что это достаточно для сохранения всей информации в сообщении. — *Прим. авт*

для логического управления, так что мы больше не будем касаться этой темы.

27.4. Аппаратура рефлексивного управления

Большинство соображений относительно выбора оборудования рефлексивного управления изложено в гл. 29, в рамках теории автоматического регулирования (теории следящих систем). Там сказано, что основными частями системы автоматического регулирования (следящей системы) являются: прибор, обнаруживающий ошибку; устройство для усиления этой ошибки; источник энергии и какой-нибудь исполнительный орган на входе и выходе. Хотя проектирование должно проводиться как одно целое, мы тем не менее приведем некоторые соображения относительно обнаружителей (детекторов) ошибки и исполнительных органов.

Выход и вход, которые совместно возбуждают обнаружитель ошибки, могут быть электрическими, механическими, гидравлическими или другими величинами, но в большинстве управляющих систем эти величины приводятся к электрическим величинам, таким, как напряжение или напряженность (или направление) поля. Физическими носителями этих электрических величин являются потенциометры, трансформаторы и сельсины. Для измерения химических, тепловых и ядерных величин применяются специальные методы, но все такие величины переводятся для сравнения опять-таки в электрические величины. Трудности при выборе обнаружителей ошибки связаны со стоимостью, с точностью и с обнаружением фазы (т. е. полярности ошибки).

Требования к качеству работы исполнительных органов делятся на два основных вида: требования к мощности и требования к точности. Что касается мощности, то исполнительный орган следящей системы должен развивать необходимую стационарную мощность, потребляемую нагрузкой во время проектируемого рабочего цикла, и пиковую мощность, требуемую во время возможных переходных состояний. Требуемая точность достигается тогда, когда исполнительный орган действует в заданном диапазоне скоростей и может сообщить запроектированные ускорения нагрузке и самому себе. Ясно, что мощность и точность связаны между собой.

В системах автоматического регулирования применяются двигатели постоянного и переменного тока. Двигатели постоянного тока при той же отдаваемой мощности обычно имеют меньший вес и больший пусковой и опрокидывающий момент, чем двигатели переменного тока. Однако двигатели переменного тока широко применяются в системах автоматического регулирования, потому что они просты, надежны и экономичны, обеспечивают быструю реакцию и не вызывают затруднений с коммутацией. Нагрузку можно также приводить в движение двигателем, вращающимся с постоянной скоростью, причем крутящий момент прикладывается к нагрузке через две муфты и зубчатую передачу. Это приложение крутящего момента к нагрузке достигается дифференциальным возбуждением двух муфт.

Электрические приводы имеют сравнительно небольшое отношение крутящего момента к моменту инерции и поэтому медленно ускоряются, вследствие чего снижается скорость реакции. Напротив, гидравлические, пневматические приводы и приводы с электромагнитными муфтами имеют большую величину отношения крутящего момента к моменту инерции и соответственно большую скорость реакции. Гидравлические приводы, кроме того, обладают дополнительными преимуществами: большей простотой и меньшими габаритами для данной отдаваемой мощности. Привод с электромагнитной муфтой несколько сложнее и требует большой резервной мощности.

Пневматические приводы имеют малый размер, но по сравнению с другими системами склонны к сильным колебаниям при постоянной скорости двигателя. Гидравлические двигатели, хотя они имеют более высокий к. п. д., чем пневматические, малопригодны в условиях, когда существует опасность пожара или взрыва.

27.5. Устройства подачи материала, включая транспортные средства

Устройства этого класса, предназначенные для больших систем, носят чрезвычайно специализированный характер по сравнению со всеми другими. В число устройств подачи входят пневматические, механические, химические и другие конструкции, тщательно приспособленные к данной цели и разработанные с большим искусством. Имеет смысл рассмотреть общие соображения относительно этих устройств лишь для авиационных систем.

Во всякой самолетной системе форма системы в значительной степени определяется типом самолета. Весовые ограничения всегда играли важную роль в технических заданиях на бортовые элементы, но в последнее время

существенное значение получил также объем и форма. Современные самолеты вследствие возрастающей мощности двигателей несут все больший и больший груз, так что места внутри самолета начинает нехватать; в то же время увеличение скорости и связанная с этим обтекаемость привели к тому, что фюзеляж самолетов становится все уже.

Поэтому, например, проектировщик радиолокатора, устанавливаемого на сверхзвуковом управляемом реактивном снаряде, сталкивается с рядом новых проблем. Все элементы, и, в частности, антенны, должны размещаться внутри пространства, меньшего в поперечнике, чем максимальный диаметр снаряда; вес и потребляемая мощность должны быть чрезвычайно малы, надежность — очень высока, время разогрева — мало; нельзя требовать ручных регулировок; работа аппаратуры должна быть одинаково удовлетворительна и на земле, и в холодной и разреженной атмосфере больших высот. Какие типы факторов приходится учитывать, хорошо видно из рассмотрения авиационных силовых (тяговых) установок.

Авиационные силовые установки. В летательных аппаратах применяются силовые установки четырех основных типов: обычная силовая установка (поршневой двигатель с воздушным винтом), турбореактивный двигатель, прямоточный реактивный двигатель и ракетный двигатель. Существуют также сочетания этих основных типов, как, например, турбовинтовой двигатель — сочетание воздушного винта и турбореактивного двигателя. Из четырех названных типов каждый последующий соответственно является менее экономичным и более мощным.

Винто-поршневая силовая установка дает максимальную энергию на единицу веса топлива; на практике, грубо говоря, наиболее экономичные винтомоторные группы дают в форме тяги энергию около 2 *л. с.-час* на фунт бензина (около 4,4 *л. с.-час* на 1 *кГ* бензина). Однако винто-поршневая установка не может отдать эту энергию очень быстро, если только она не будет очень большой. Так, большие двигатели, обычно применяемые на наших больших пассажирских самолетах, номинальной мощностью около 2000 *л. с.* и весом свыше тонны, при экономичной работе отдают самое большее около 1000 *л. с.*; при работе на полной мощности, которая возможна лишь в течение короткого времени, к. п. д. сильно уменьшается. Кроме того, двигатель имеет диаметр в несколько футов и длину во много футов, а винт описывает окружность диаметром 10—15 футов.

Возвратно-поступательный (поршневой) двигатель может иметь термический к. п. д. порядка 30% при к. п. д. воздушного винта около 90%. Мощность, отдаваемая такой комбинацией, не зависит в известных пределах от скорости самолета. Воздушный винт не рассчитан на высокий к. п. д. при очень малых скоростях [скажем менее 50 миль в час (≈80 *км/час*)]; при скоростях примерно свыше 400 миль в час (≈640 *км/час*) концы воздушных винтов (скорость которых равна векторной сумме скоростей поступательного и вращательного движения) движутся быстрее звука, что приводит к значительному снижению к. п. д.

На другом конце ряда находится ракетный двигатель. Ракетные двигатели обычно характеризуются тягой, а не мощностью, поскольку тяга, развиваемая таким двигателем, по существу не зависит от скорости, тогда как мощность прямо пропорциональна скорости. Ракетный двигатель, развивающий тягу в 1000 фунтов (≈450 *кГ*), может иметь диаметр меньше одного фута, длину немного больше фута и вес около 100 фунтов (≈45 *кГ*); такой двигатель будет развивать 1000 *л. с.* при скорости 375 миль в час (≈600 *км/час*). Однако он потребляет громадное количество топлива.

Хорошее ракетное топливо развивает удельный импульс в 200 *сек* (или *фунт · сек/фунт*, или *кГ · сек/кГ*); иначе говоря, 1 фунт топлива может развить тягу в 200 фунтов в течение 1 *сек*, или в 1000 фунтов в течение 1/5 *сек* (а 1 *кГ* топлива — в 200 *кГ* в течение 1 *сек*, или в 1000 *кГ* в течение 1/5 *сек*). Следовательно, упомянутый двигатель с тягой 1000 фунтов при своем теоретическом оптимуме будет сжигать самое меньшее 5 фунтов (≈2,25 *кГ*) топлива в 1 *сек*. Можно найти ракетные топлива с несколько большим удельным импульсом, но они содержат такие вещества, как жидкий водород (температура которого должна поддерживаться ниже 20° К), жидкий озон (который становится сильно взрывчатым веществом при весьма различных и не вполне выясненных условиях) и жидкий фтор (который очень ядовит и, по-видимому, является наиболее коррозийным из известных химических веществ).

Промежуточным типом между винто-поршневой силовой установкой и ракетным двигателем являются воздушно-реактивные двигатели: турбореактивный и прямоточный. Как и ракетный двигатель, они могут развивать некоторую максимальную тягу, и поэтому их максимальная мощность зависит от

скорости самолета; но, подобно двигателю поршневого типа, они должны «вдыхать» воздух, и отдаваемая ими мощность ограничена максимальным поступлением воздуха. Напротив, ракетному двигателю воздух не нужен; в его топливо входит жидкий или твердый окисляющий агент, который выполняет ту же функцию, что и воздух в воздушно-бензиновой смеси.

В турбореактивном двигателе воздух всасывается, сжимается (попутно нагреваясь) и затем смешивается с жидким или распыленным горючим (топливом), обычно каким-нибудь углеводородом вроде нефтепродуктов. Кислород воздуха соединяется химически с горючим (после зажигания) и выделяет энергию в виде тепла. Горячие газы проходят через турбину, отдавая часть своей энергии на вращение турбины, доставляющей механическую энергию компрессору. Выхлопные газы турбины направляются затем назад в виде струи. Будучи значительно более нагреты и более разрежены, чем воздух, который был всосан первоначально, они движутся значительно быстрее и обладают значительным количеством движения (импульсом); по закону противодействия этот импульс прикладывается к самолету и толкает его вперед.

Прямоточный реактивный двигатель в теории еще проще. Он состоит из полой трубки, которая перемещается поступательно в воздухе; в нее впрыскивается горючее и сжигается. Продукты сгорания, как и в турбореактивном двигателе, создают реактивную силу, которая толкает трубку вперед. Прямоточный двигатель работает всего экономичнее при скоростях 1000—2000 миль в час (≈1600—3200 *км/час*) и совсем не будет работать при скоростях, значительно меньших, чем скорость звука (равная приблизительно 700 милям в час, или 1224 *км/час*). Поэтому самолет с прямоточным реактивным двигателем необходимо разогнать до нужной скорости с помощью какой-нибудь вспомогательной силовой установки. Внутри двигателя обычно должны быть диффузоры для замедления поступающего воздуха, стабилизатор пламени, воспламенитель и т. д., так что на практике прямоточный двигатель представляет собой довольно сложный и тонкий аппарат.

Выбор основного типа силовой установки для самолета часто производится проектировщиком системы. Он определяется, как и всегда, стоимостью и, кроме того, сравнительными выгодами большой скорости и большой дальности полета. Экономия топлива в большинстве случаев представляет интерес не в связи со стоимостью топлива в долларах, а в связи со стоимостью запасания его на самолете; иначе говоря, большое потребление топлива означает малую дальность полета. Стоимость зависит от скорости и дальности нелинейно; даже небольшое увеличение скорости и/или дальности может обходиться чрезвычайно дорого.

Если требуется скорость больше чем примерно $M=3$ *, то нужно применять ракетный двигатель, и за необходимую дальность полета придется расплачиваться повышением стоимости. В другом крайнем случае, когда нужно оставаться в воздухе в течение многих часов (со скоростями в несколько сот миль в час и без пополнения топлива), будет применена обычная комбинация возвратно-поступательного двигателя и винта. Если необходимо оставаться непрерывно в воздухе еще дольше, то вес топлива приобретает еще большее значение и можно с полным основанием применить более тяжелый и более экономичный двигатель. Если бы потребовалось оставаться в воздухе один-два дня, то, вероятно, спроектировали бы какой-нибудь подходящий дизель, а если бы потребовалось оставаться в воздухе неделю или две, то, вероятно, был бы спроектирован какой-нибудь подходящий атомный двигатель (его термический к. п. д. был бы не больше, но зато, очевидно, его к. п. д., выраженный в лошадиных силах-часах на фунт потребляемого топлива, был бы чрезвычайно высок).

27.6. Выходная аппаратура

Как упомянуто раньше, к выходному оборудованию относятся исполнительные органы (материальный выход), которые являются специализированными, и устройства индикации (информационный выход). Мы ограничимся некоторыми существующими методами индикации, имеющими первостепенное значение в связи с информационной стороной системы.

В системе большого масштаба некоторые

---

* Термин «$M=3$» означает «в три раза быстрее звука». Поскольку скорость звука заметно меняется с изменением температуры, это не очень точная мера. Тем не менее этот термин имеет важное значение для специалиста по аэродинамике, и его приняли инженеры-системотехники ввиду удобства перевода числа Маха (M) в футы в секунду путем умножения на 1 000. Таким образом, когда проектировщик системы говорит $M=3$, он должен был бы подразумевать, что скорость в три раза больше скорости звука, но практически он, возможно, подразумевает, что скорость равна 3 000 футов в секунду. — *Прим. авт.* [Число Маха (число M) есть отношение скорости полета самолета к скорости распространения звука в данный момент; 1 000 *фут/сек*, принимаемые, как говорят авторы, за практическое значение $1M$, соответствует скорости 300 *м/сек*. — *Прим. ред.*].

решения всегда должны принимать люди, которым для этого должна быть доступна соответствующая информация. До появления автоматических устройств обработки данных и эффективных быстродействующих средств связи задача обычно состояла в том, чтобы получить достаточно точную и своевременную информацию для принятия надлежащих решений. Но теперь эта проблема, вообще говоря, решена, и вместо этого мы сталкиваемся с другой проблемой: оператор тонет в потоке излишней информации.

Когда информации слишком много для того, чтобы ее усвоил один человек, можно предпринять четыре меры: исключить часть информации, которая несущественна или не имеет отношения к делу; уплотнить часть информации путем суммирования; классифицировать информацию так, чтобы ее можно было брать по мере необходимости по категориям; и разделить ответственность за принятые решения между двумя или тремя людьми так, чтобы ни одному из них не требовалось всей информации. В хорошо спроектированной индикаторной системе, вероятно, будут применены все эти методы.

Как описано в гл. 30, небольшое количество информации можно представить органам слуха и еще меньшее количество информации — другим органам чувств, но главным методом индикации является визуальный. Визуальные методы удобно классифицировать двояким образом: 1) на табличные и изобразительные; 2) на переменные и постоянные. Табличная индикация означает представление с помощью условных символов, например чисел, или же обычную запись в виде букв и слов, а изобразительная индикация включает разнообразнейшие другие методы, как графики, карты и т. п. Типичный табличный материал можно, конечно, представить изобразительными методами, но не наоборот. Постоянная индикация обычно означает запись на бумаге или на фотопленке, переменная индикация означает стираемую индикацию. Переменную индикацию иногда называют *кратковременной*, но этот термин лучше оставить для запоминающих устройств, в которых хранимая величина теряется, когда снимается напряжение. Так, магнитные сердечники, хотя они являются долговременными запоминающими устройствами, следовало бы считать переменными накопителями.

Постоянные изобразительные индикаторы. В эту группу входит много обычных видов документов, как карты, графики, фотографии и т. д. Часто бывает удобно применять карту постоянного типа в качестве основы и прозрачные накладки, проецируемые изображения и т. п. — для обозначения меняющихся положений. Для определения наилучшей формы представления информации можно посоветоваться с инженером-психологом.

Существует ряд новых методов быстрого получения постоянных изображений. Один из них — метод фотографирования «Land Polaroid», при котором меньше чем за минуту получается готовая фотография. Для тех случаев, когда можно применять проекционные способы, и, в частности, когда достаточны негативные снимки, существует аппаратура, позволяющая снять, проявить и закрепить фотографию за 5—10 *сек* и затем спроецировать ее на экран, пока она еще влажная. Например, в системе управления воздушным движением, где положения самолетов определяются радиолокатором и изображаются на индикаторе кругового обзора в виде белых точек на темном фоне, такие методы вполне пригодны (в этом случае самолеты изображаются как черные точки на белом фоне).

В последнее время разработана другая группа фотографических методов, известных под общим названием *ксерографии* или *сухой фотографии*. В некоторых из них для проявления и закрепления скрытого фотографического изображения, образованного фотографическим способом на чувствительной поверхности, вместо обычных водных растворов применяются химикалии, например аммиачный газ. В других, так называемых *электрофотографических* методах применяются фотоэлектрические явления или явления фотопроводимости, вызывающие образование электростатического скрытого изображения (например, на полупроводящей поверхности), которое затем проявляют при помощи наэлектризованного порошка и закрепляют химическим способом (например, путем кратковременного нагревания).

Для передачи изобразительной информации на отдаленные пункты и создания там постоянных записей можно применять фототелеграфию. Термин *фототелеграфия* относится к функции, а не к конкретному способу ее выполнения. Нормальный метод заключается в развертывании материала (так же, как в телевидении) и модулировании этими сигналами несущей. Ширина полосы нормально значительно меньше, чем в телевидении, ибо на передачу одного изображения затрачиваются секунды и даже минуты, тогда как для передачи телевизионного изображения нужно 0,03 *сек*. Развертка на передающем конце обычно производится фотоэлектрическим элементом. Запись на приемном конце

часто осуществляется ксерографическими методами; при этом электрически заряженное перо, касаясь предварительно обработанного листа бумаги, вызывает химические изменения в поверхностном слое. Но запись можно осуществить любыми стандартными аналоговыми записывающими устройствами, например пером (с применением обычных чернил), которое управляется электромагнитом. Эти методы иногда можно отнести и к табличным, и к изобразительным.

Постоянные табличные индикаторы. Типичными изобразительными приборами, широко применяемыми для табличной индикации, являются телепостроитель кривых и телеавтограф. Первый представляет собой аналоговый прибор для построения кривых, управляемый на расстоянии. Второй представляет собой механический карандаш, который пишет на движущейся ленте; его движением управляет сигнал, создаваемый в другом месте перемещением подобного же карандаша, который держит оператор. Мертвый ход и искажения могут вносить значительные ошибки, но при больших буквах написанное вполне можно читать.

Типичным табличным прибором является телетайп. Для системы телетайпа характерны: особый код для обычных букв и чисел и некоторых других символов; бумажная лента особого типа для хранения сообщений в этом коде; особый вид модуляции для передачи символов — так называемая частотная модуляция; широкая и сложная сеть связи по всей территории Соединенных Штатов. Но лежащая в основе этого идея электрической пишущей машинки, печатающей в ответ на соответствующие кодированные сигналы любое требуемое сообщение, была применена также в различных других устройствах.

Переменные изобразительные индикаторы. Применение электронно-лучевой трубки для представления пригодного для этого изобразительного материала на переменной, меняющейся основе обычно называется *телевидением* — независимо от того, связано ли это с передачей из отдаленного пункта или нет. На переднюю поверхность воспроизводящей трубки можно постоянно или на время наложить сетки и другие опорные отметки. Кроме того, в трубку можно поставить два или даже три электронных прожектора, так что можно накладывать друг на дуга различные изображения, приходящие из разных источников. Наконец, можно использовать методы запоминания (описанные в § 15.3) для сохранения изображения на экране трубки в течение секунд или даже минут после отключения источников сигналов.

Всякое постоянное изображение можно проецировать при помощи соответствующих оптических приборов и световых источников. Легче, конечно, проецировать прозрачное изображение, но можно проецировать для целей индикации и непрозрачное изображение или изображение на электронно-лучевой трубке.

Переменные табличные индикаторы. Как говорилось выше, для представления табличного материала можно применить различные изобразительные индикаторы и любую постоянную запись можно индицировать на время с помощью проекционной техники. Однако существует группа весьма интересных новых методов индикации специально для переменного табличного материала.

Наиболее известным устройством для объявления курсов фондовой биржи является телерегистровая доска. Здесь названия или сокращенные обозначения ценных бумаг полупостоянны (их можно менять подобно названиям на щите театрального репертуара), а две-четыре цифры курса можно быстро менять. Для каждой цифры сделано маленькое окошко, через которое видна одна из граней многосторонней (например, 10-сторонней) призмы, устанавливаемая против окошка электромагнитным устройством. Установка любого числа производится при небольшой затрате мощности и с большой скоростью (меньше чем за 1 *сек*), причем возможно дистанционное управление. Недостатки этого способа — шум и необходимость достаточного пространства для надлежащего размещения призм (например, цифры нельзя размещать рядом в горизонтальном и вертикальном направлении), хотя размеры прибора чрезвычайно малы. Телерегистры обычно применяются только для индикации чисел, хотя их с таким же успехом можно применять для других символов, лишь бы эти символы не были слишком многих различных типов.

Существует ряд способов выбора данного символа при помощи переключающей матрицы с последующим освещением только этого символа, тогда как остальные остаются неосвещенными. Там, где достаточно места или времени, можно переместить в нужное положение соответствующую лампочку или щиток (трафарет), но мы предполагаем, что размеры символа желательно сделать такими, чтобы использовать все имеющиеся пространство, и что символ должен освещаться лишь небольшую долю секунды. В одном из методов для этого применяются семь неоновых трубок, расположенных как показано на

рис. 27.2. Выбирая соответствующие трубки, можно составить любую из 10 цифр и некоторые буквы. Получающиеся символы можно распознать, хотя и не очень легко; распознаваемость можно улучшить, увеличив число трубок и усложнив переключающую матрицу.

Вместо этого можно применить декатрон — одну неоновую трубку с 10 различными катодами, из которых можно выбирать

Рис. 27.2. Комбинация из семи неоновых ламп для образования всех цифр.

один. Катоды монтируются по окружности вокруг расположенного в центре анода, и перед каждым из них можно поместить отдельную маску (например, в форме одной из 10 цифр); при выполнении соответствующего переключения горит лишь выбранная цифра, а остальные не видны.

В индитроне фигурные катоды монтируются друг против друга. В другом подобном устройстве применяется столбик тонких пластмассовых пластин; нужная пластинка освещается с ребра и светится, становясь тем самым видимой, а другие прозрачны и не видны.

Электронно-лучевые трубки применяются для представления стандартных символов с помощью различных методов. Один из методов заключается в том, чтобы приближенно представить форму требуемого символа аналитической кривой, которая может быть создана аналоговыми вычислительными приборами. При приложении к отклоняющим пластинам соответствующих напряжений, зависящих от времени, нужный символ появляется на экране трубки или даже на любой части экрана, в зависимости от фиксирующих напряжений подсветки на отклоняющих пластинах. Созданный таким образом символ можно при желании проецировать.

В характроне применяется другой способ. Электронный луч сначала отклоняется одной группой пластин и проходит через трафарет, на котором при изготовлении трубки были нанесены постоянные символы. Затем луч фокусируется на другую группу пластин, направляющих его в заданную часть трубки. Символ можно записать таким способом на любой части экрана трубки за долю миллисекунды; благодаря послесвечению можно одновременно поместить в соответствующих местах достаточно большое число символов и выдать на экране любое сообщение.

Другое подобное устройство — типотрон; здесь применяется второй прожектор и техника запоминания для сохранения символов на экране трубки в течение нужного времени. Еще одно устройство — типсетрон, где трафарет помещен снаружи трубки (так что его можно изменять уже после изготовления трубки — это большое преимущество). Свет пропускается через трафарет, электронный луч создается с помощью фотоэлектронной, а не термоэлектронной эмиссии.

27.7. Проектирование аппаратуры

Процесс проектирования систем родствен процессу проектирования оборудования с тем отличием, что система очень велика и требует согласованных усилий многих людей. Проектировщик оборудования проходит примерно через ту же последовательность этапов, что и проектировщик систем; когда составлено функциональное задание и определены типы входов и выходов, проектировщик выбирает некоторое множество допустимых устройств, производит оценку, чтобы сузить область допустимых решений, предсказывает рабочие характеристики нескольких выбранных устройств и на основании этих предсказаний делает выбор тех устройств, которые нужно изготовить.

Всегда бывает так, что некоторых нужных компонентов не оказывается, но их вполне можно разработать за ограниченный отрезок времени. Изобретение приборов по заданию весьма обычно в проектировании современных систем. Однако есть существенное различие между разработкой нового прибора, находящегося на существующем уровне техники, и разработкой прибора, выходящего за рамки существующей техники.

Если это не вызвано настоятельной необходимостью, проектирование системы не следует основывать на идее, которая еще исследуется и не была проверена. Могут быть случаи, когда не совсем ясно, будет ли требуемый прибор переходом на новую техническую ступень. Кроме того, возможно, что не была испытана работа оператора в условиях предлагаемой конструкции. Обе эти ситуации приводят к идее критического испытания. Необходимость в таких испытаниях способностей человека или работы аппаратуры должна предусматриваться уже на первых шагах проектирования системы. Их выбирают совместно проектировщики системы и проектировщик оборудования.

ЛИТЕРАТУРА

Хороший разбор радиолокации, с особым упором на электрические схемы, можно найти

у Термена [114]. Более обширный разбор, в котором обращено внимание на «системные» аспекты, можно найти у Райднора [105], в первом томе серии книг по радиолокации Лаборатории излучений Массачусетского технологического института. 28 книг этой серии все еще являются весьма ценными справочниками. Краткий разбор схем импульсной модуляции со ссылками на литературу можно найти у Термена [114]. Разбор аппаратуры следящих систем, помещенный в этой главе, взят из Талера и Брауна [115]; см. также литературу к гл. 29. Пока еще не существует хорошей книги о методах индикации, но по этому вопросу имеется обширная литература, и книги, вероятно, скоро появятся.

ГЛАВА 28

СВЯЗЬ. ТЕОРИЯ ИНФОРМАЦИИ

В гл. 27 мы разбирали некоторые системы связи и отметили, что эти системы бывают разных типов, в частности: дискретные системы, как, например, телеграфные; непрерывные системы, как, например, телефонные, и смешанные системы, как, например, телефонные с импульсно-кодовой модуляцией и импульсно-временной модуляцией, которые являются непрерывными, хотя в них и применяются импульсы. В этой главе мы рассмотрим мощное орудие, применимое ко всем таким системам — *теорию информации*, или *математическую теорию связи*. Мы будем следовать обычному порядку и рассмотрим сначала дискретные системы без шума, затем дискретные системы с шумом и, наконец, непрерывные системы с шумом; попутно остановимся на смешанных системах.

Теория информации как наука основана в значительной мере на двух замечательных работах, опубликованных Клодом Шенноном в 1948 г. [40]. Заслуга Шеннона, помимо блестящей логической и математической разработки, состояла в том, что он исследовал вопрос при некоторых упрощающих условиях, а именно, игнорировал семантическое содержание (т. е. смысл сообщений) и анализировал только классы сообщений, а не отдельные сообщения. С одной стороны, это позволило ему прийти к некоторым определенным математическим выводам; с другой стороны, полученную теорию нельзя применять к практическим задачам, пока они не описаны статистически. Теория Шеннона позволяет вывести явные формулы, связывающие такие основные параметры, как отношение сигнал/шум и ширину полосы. Но эта теория, вообще говоря, не дает конструктивных формул, позволяющих решать конкретные задачи, она устанавливает главным образом верхние границы рабочих характеристик.

Вопросы, на которые отвечает теория, касаются скорости, с которой можно передавать информацию при некоторых заданных условиях. Эти условия касаются: свойств источника сигналов, и в частности того обстоятельства, является ли сигнал дискретным или непрерывным; свойств канала, и в частности его способности к передаче информации; свойств шума (если он имеется), искажающего передачу; и критерия верности, по которому судят об удовлетворительности передачи.

Разбираемая ситуация изображена на рис. 28.1. Устройства, которые мы называем *кодирующим* и *декодирующим* устройствами, обычно называются *передатчиком* и *приемником*, хотя эти термины в технике связи имеют другие значения. Идея кодирования играет важную роль в теории информации (ее не нужно смешивать с шифрованием, означающим кодирование для засекречивания).

Например, в телеграфной системе источником информации является письменное сообщение в словесной форме. Кодирующим устройством здесь является та часть системы, которая переводит слова в точки и тире и затем превращает их в соответствующие электрические сигналы. Каналом является прово-

Рис. 28.1. Скелетная схема общей системы связи (по Шеннону [40]).

лока (в радиотелеграфии — полоса частот или, если угодно, эфир). Декодирующим устройством является электрический приемник, превращающий сигнал обратно в слова.

На рисунке источник шума соединен только с каналом, но в действительности он искажает также работу передатчика и приемника. Слово «сообщение» употребляется в теории информации в двух различных смыслах: в одном смысле — как на рис. 28.1, где сообщение отличается от сигнала, в другом — как ниже, где сообщение отличается от алфавита.

28.1. Дискретная система без шума

Центральным пунктом теории является количественное определение информации. Наше интуитивное представление о сущности информации могло бы привести нас к тому, чтобы определить ее по семантическому содержанию, т. е. по смыслу слов в сообщении. Но никому еще не удалось найти надлежащее определение такого рода. Затруднение состоит в том, что если мы пытаемся выразить количественно величину информации, которую адресат получает из данного сообщения, мы должны учитывать его подготовку — его способность понимать сообщение, то обстоятельство, что он может уже знать все сообщение или часть, и т. д.

Чтобы избежать этих трудностей, Шеннон ввел меру информации, основанную на вероятностных соображениях. Мы различали в нашем разборе теории вероятностей генеральную совокупность и выборку. Теперь мы рассматриваем источник информации как генеральную совокупность, которую мы называем *алфавитом*. Переменная генеральной совокупности может принимать различные значения; в дискретной системе существует конечное число n таких значений, каждое из которых называется *символом алфавита*. Наблюденное событие, т. е. появление данного символа алфавита, называется *символом сообщения*. *Сообщение* состоит из выборки символов сообщения.

Энтропия информации. Предположим, что имеется алфавит из n символов (т. е. источник, способный производить n различных символов алфавита), и предположим, далее, что создается символ и что вероятность того, что это i-й символ, равна p_i. Пусть $H(p_1, p_2, ..., p_n)$ есть мера информации, произведенной этим процессом.

Необходимо пояснить, что именно мы здесь измеряем. Алфавит как таковой не содержит информации, и мы не говорим о количестве информации в данном символе сообщения, который уже был создан. Мы рассматриваем количество информации, появляющейся в процессе создания одного символа сообщения, и оно оказывается средним количеством информации на символ сообщения.

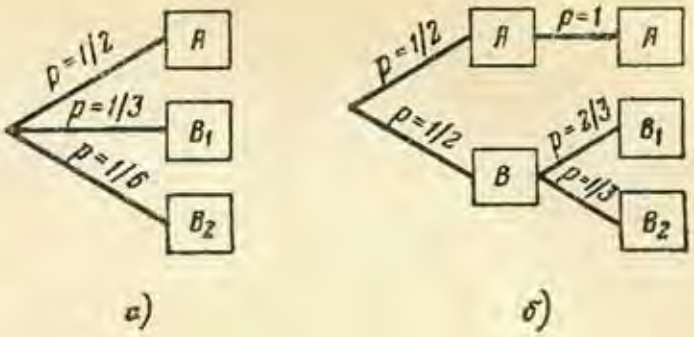


Рис. 28.2. Различные методы создания трех символов (по Шеннону [40]):
а) одноступенчатый процесс; б) двухступенчатый процесс.

Представляется разумным потребовать, чтобы величина H обладала следующими свойствами:

1) она должна быть непрерывной функцией от p_i;

2) если $p_1 = p_2 = ... = p_n = 1/n$, т. е. если все символы равновероятны, то H должна быть монотонной возрастающей функцией от n;

3) если каждое p_i имеет определенное значение, то H должна быть независима от способа, которым эти значения были достигнуты.

Первое правило говорит, что мы можем начертить график, как на рис. 28.3, на котором функция определена и непрерывна во всех точках. Второе правило говорит, что некоторый символ, скажем 1, дает нам больше информации, если мы знаем, что это десятичная цифра, чем если бы мы знали, что это двоичная цифра (или, более конкретно, символ дает нам больше информации, если он происходит из алфавита с 10 равновероят-

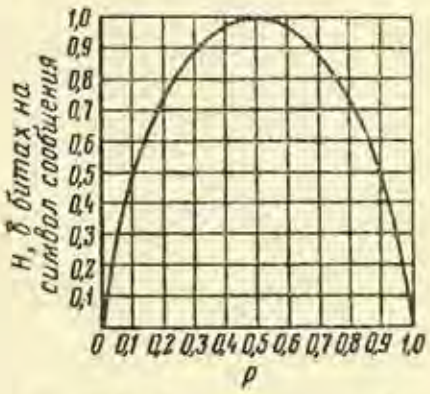


Рис. 28.3. Энтропия для алфавита из двух возможных символов с вероятностями p и $1-p$ (по Шеннону [40]).

ными символами, чем в том случае, когда он происходит из алфавита с двумя равновероятными символами). Третье правило говорит, что H одна и та же для двух процессов создания символов, изображенных на рис. 28.2.

В системе на рис. 28.2,*а* символы A, B_1 и B_2 создаются с вероятностями соответственно 1/2, 1/3 и 1/6. В системе на рис. 28.2,*б* символы A и B создаются с вероятностями соответственно 1/2 и 1/2, и если создается символ B, он обрабатывается таким образом, что вероятность его превращения в B_1 равна 2/3, а вероятность превращения в B_2 равна 1/3.

В работе [40] доказывается, что единственной функцией, удовлетворяющей этим критериям, является функция

$$H = -k\sum_{i=1}^{n} p_i \log p_i, \qquad (28.1)$$

где k — положительная константа. Поскольку p_i не могут быть больше единицы, все логарифмы отрицательны, и потому H всегда положительна.

Заметим, что

$$\lim_{p\to 0} p\log p = 0. \qquad (28.2)$$

Это можно доказать, представив выражение в виде $\frac{\log p}{p^{-1}}$ и продифференцировав числитель и знаменатель.

Выбор k в формуле (28.1) сводится к выбору основания логарифмов. Удобно выбрать основание 2; тогда равенство (28.1) переходит в следующее:

$$H = -\sum_{i=1}^{n} p_i \log_2 p_i, \qquad (28.3)$$

и H выражается в битах (двоичных единицах) на символ сообщения.

Это во многих отношениях аналогично прежнему употреблению слова «бит» (§ 14.3); так, если алфавит содержит два символа (скажем, 0 и 1), имеющие равные вероятности появления ($p_0 = p_1 = 1/2$), то в силу (28.3) количество информации на символ равно

$$H = -\left(\tfrac{1}{2}\log_2 \tfrac{1}{2} + \tfrac{1}{2}\log_2 \tfrac{1}{2}\right) =$$

$$= -\left[\tfrac{1}{2}\times(-1) + \tfrac{1}{2}\times(-1)\right] = 1,$$

или одному биту на символ сообщения. Для вычисления логарифмов по основанию 2 мы используем алгорифм (4.9).

H есть только одна из многих возможных мер информации. Шеннон назвал ее *энтропией информации*, по аналогии с термодинамической энтропией. Уравнение (28.1) тождественно с определением энтропии в статистической механике; в этом случае константа k становится постоянной Больцмана. Термодинамическая энтропия, как известно, есть мера дезорганизации, неопределенности или случайности. Подобно этому, H измеряет случайность появления символов в системе связи*.

На рис. 28.3 показано, как изменяется H для алфавита с двумя символами. Случайность, или неопределенность, становится наибольшей, когда $p = 0{,}5$. Вообще при n символах H является наибольшей, когда последовательные символы независимы и $p_1 = p_2 = \ldots = p_n = 1/n$; в этом случае

$$H_{\text{макс}} = -\sum_{1}^{n} \log_2 \frac{1}{n} = -\log_2 \frac{1}{n} =$$

$$= \log_2 n \text{ битов на символ сообщения.} \qquad (28.4)$$

Источники и *сообщения*. Простейшим из всех возможных источников является источник, выбирающий символ из алфавита с двумя символами, в котором оба символа имеют одинаковые вероятности появления, последовательные символы независимы и оба символа занимают одинаковое время. Мы исследуем по очереди все усложнения, получающиеся вследствие снятия этих ограничений, а именно: вследствие допущения большего числа символов, чем два; вследствие того, что последовательные символы уже не независимы; и вследствие того, что символы имеют различные длительности.

Мы замечаем, что источник, создающий только один возможный символ, не имеет смысла. Так, источник, который может создавать только один символ, скажем A, не дает никакой информации, когда он действительно создает A. Читатель может возразить, что момент, когда создан символ, также может иметь значение; но в этом случае источник

---

* В действительности можно с некоторым основанием сказать, что энтропия информации и термодинамическая энтропия не просто эквивалентны, но тождественны. Так, чтобы получить элементарную информацию о реальном мире, мы должны «возмутить» его по крайней мере настолько, чтобы получить, скажем, различимое перемещение стрелки. «Различимое» означает отличное от броунова движения, а такое перемещение стрелки, находящейся в равновесии с окружающей средой, означает изменение термодинамической энтропии системы на величину, равную постоянной Больцмана [120]. — *Прим. авт.*

в действительности способен производить два символа: A или пустое место.

Рассмотрим теперь наш простейший источник. При повторном появлении символов сообщения образуются *сообщения*. Существуют четыре возможных сообщения длиной в два символа, и поскольку все эти сообщения согласно нашим допущениям равновероятны, энтропия сообщения из двух символов ввиду (28.3) равна двум битам на сообщение, или одному биту на символ сообщения. Подобно этому существуют восемь возможных сообщений длиной в три символа или в общем случае 2^r сообщений длиной в r символов; энтропия таких сообщений равна r битам на сообщение, или одному биту на символ сообщения, как и в предыдущем случае. Итак, *скорость* создания информации равна одному биту на символ сообщения, независимо от времени, необходимого для создания одного символа, и одному биту на единицу времени, если каждый символ требует одной единицы времени.

В несколько более общем случае, когда существует источник n символов алфавита, которые все независимы и равновероятны, возможны n^r различных сообщений и вероятность любого из них равна $P_j = 1/n^r$.

Тогда энтропия сообщений длиной r равна

$$H = -\sum_{j=1}^{n^r} P_j \log_2 P_j = -n^r (P_j \log_2 P_j) =$$

$$= -n^r \frac{1}{n^r} \log_2 \frac{1}{n^r} = -\log_2 \frac{1}{n^r} =$$

$$= \log_2 n^r = r \log_2 n \text{ битов на сообщение, (28.5),}$$

а скорость создания информации равна $\log_2 n$ битов на символ сообщения, или, если каждый символ занимает единицу времени, $\log_2 n$ битов на единицу времени.

В более общем случае, когда каждый из n символов алфавита имеет вероятность появления p_i, не обязательно одинаковую для всех символов, но символы по-прежнему независимы, мы можем написать

$$H = -\sum_{j=1}^{n^r} P_j \log_2 P_j \text{ битов на сообщение,}$$

но мы уже не можем вычислить это выражение так легко. Каждое P_j выражается формулой

$$P_j = p_{i_1} p_{i_2} \cdots p_{i_r}$$

и не обязательно должно быть равно всем другим (здесь p_{i_1} есть соответствующее p_i для первого символа в j-м сообщении и т. д.).

Кроме того, мы можем написать

$$H = -\frac{1}{r} \sum_{j=1}^{n^r} P_j \log_2 P_j \text{ битов на символ сообщения,} \qquad (28.6)$$

и можно показать, что при независимости символов это выражение приводится к (28.3), как и в предыдущих простых случаях. При наличии зависимости это уже неверно. Мы рассмотрим ниже дополнительные усложнения, вносимые зависимостью (как в английском языке) и разными длительностями символов (как в азбуке Морзе), но прежде всего мы хотим ввести понятия канала и пропускной способности канала и затем пояснить эти понятия на примере.

Дискретный канал без шума. Дискретный канал без шума представляет собой до некоторой степени нереалистическую абстракцию. Представим себе, например, что существует канал, по которому можно посылать канальные символы двух видов, скажем импульс $+10$ *в* и импульс -10 *в*, и что эти символы можно различить на приемном конце. Кроме того, в канале не бывает ошибок, если только символы посылаются со скоростью не больше некоторой заданной величины, скажем 100 импульсов в секунду. Далее, в канале не может быть никаких других символов. В этом случае пропускная способность, очевидно, равна 100 битам в секунду (это мы могли бы сказать и без теории информации).

Если имеются три различных канальных символа, то каждый символ эквивалентен (по своей способности передавать информацию) одной цифре троичного числа; а в общем случае, при n различных символах, каждый символ эквивалентен одной цифре числа с основанием n. Число из r цифр с основанием n эквивалентно $r \log_2 n$ битам, так как оно равно двоичному числу из $r \log_2 n$ цифр; таким образом,

$$n^r = 2^{r \log_2 n},$$

как можно легко убедиться, взяв логарифмы обеих частей по основанию 2.

Следовательно, когда имеется n различных канальных символов, количество информации в битах на канальный символ равно $\log_2 n$ (а пропускная способность канала в битах в секунду равна произведению этого числа на число символов в секунду, которое может быть послано).

Но эта формула, которую мы также установили вне связи с теорией информации,

тождественна с выражением (28.4), представляющим максимальную скорость информации для источника, создающего n алфавитных символов. Итак, мы можем определить пропускную способность канала следующим образом: установить максимальную скорость (в символах в секунду), с которой некий источник может создавать символы и передавать их по каналу, что дает нам максимальную скорость (в двоичных единицах в секунду), с которой информация может проходить по каналу от источника; то же самое проделать для всех других возможных источников, соединенных с этим каналом, что даст нам набор максимальных скоростей; максимальная из этих максимальных скоростей называется *пропускной способностью канала.*

Это определение можно легко обобщить на каналы с шумом и на непрерывные каналы. Пока мы можем заметить, что его можно распространить на следующий случай: пусть канал может передавать алфавит из двух символов (скажем, импульсы -10 *в* и импульсы $+10$ *в*) со скоростью 100 символов в секунду и алфавит из трех символов (те же символы и, кроме того, 0 *в*) со скоростью 75 символов в секунду. Это практический пример (сравните рис. 15.28 и соответствующие рассуждения). В обоих случаях имеется определенная скорость информации в битах в секунду, и бо́льшая из этих скоростей информации называется *пропускной способностью канала.*

Эту скорость можно найти, рассматривая все возможные источники алфавитов из двух и из трех символов. Очевидно, те источники, которые создают символы, имеющие одинаковые вероятности появления, дадут бо́льшие скорости информации, чем источники, создающие символы с неодинаковыми вероятностями появления, и из двух оставшихся источников источник с тремя символами дает большую скорость и определяет пропускную способность канала.

Кодирование. Если дан источник и канал с фиксированной пропускной способностью, то часто сообщения можно кодировать так, чтобы полностью использовать пропускную способность канала.

Пример. Пусть имеется некая система управления воздушным движением [это может быть гражданская (§ 2.1) или военная система (§ 2.6)], и мы хотим послать с одного аэродрома на другой данные радиолокационного обзора. Чтобы уменьшить расходы, мы используем обычную телефонную линию с малой шириной полосы (это, как мы увидим, соответствует малой пропускной способности в битах в секунду) и импульсно-кодовую модуляцию. Мы решаем, что достаточно разделить радиолокационный экран на 256 частей и посылать только информацию о том, находится ли самолет в каждом из этих положений; эту информацию нужно посылать с максимально возможной частотой. Мы ожидаем в среднем 16 самолетов, так что вероятность того, что в одном из этих 256 положений будет самолет, равна $^1/_{16}$.

Если перевести эти данные на язык теории информации, то мы имеем алфавит из двух символов. Мы можем обозначить их как символ A (имеется цель) и символ B (цели нет). Вероятности образования символов в источнике равны $p_A = {}^1/_{16}$ и $p_B = {}^{15}/_{16}$. Нам нужно посылать сообщения длиной 256 символов сообщения. Передатчик должен превращать символы сообщения в канальные символы, состоящие в наличии или отсутствии импульса, которые мы обозначим как 1 и 0. Процесс превращения A и B в 1 и 0 называется *кодированием*, и, как мы увидим, у нас имеется большая свобода в выборе возможных типов кодирования.

Наиболее очевидный способ кодирования — закодировать A в 1 и B в 0. Так, если цель имеется в шестом положении и нет цели ни в одном из остальных первых девяти положений, то наше сообщение из 256 битов будет 000001000... Но возможны другие, более сложные способы кодирования. Один из способов — передавать только положения, в которых имеется цель. Поскольку существует $256 = 2^8$ положений, для указания любого положения требуется восемь битов; например, шестое положение выражается как 00000110. Поскольку мы ожидаем в среднем 16 самолетов и для каждого требуется восемь битов, для полного сообщения потребуется в среднем только 128 битов.

Этот код связан с двумя трудностями. Во-первых, мы можем с некоторой малой вероятностью встретиться с такой ситуацией на экране, когда в большинстве или во всех положениях имеются самолеты. В этом случае нам может потребоваться до $8 \times 256 = 2048$ битов, в то время как при простом способе, в котором в среднем требуется 256 битов, никогда не потребуется больше 256 битов. Этот недостаток типичен для «экономичных» кодов теории информации, о чем мы скажем еще раз ниже. Другая трудность состоит в том, что этот способ кодирования непригоден для сообщений произвольного размера. Так, если бы нам нужно было посылать сообщение длиной в 1 миллион символов из того же алфавита, то в этом коде потребовалось бы 20 битов для каждого появления маловероятного символа A, тогда как при простом коде потребовалось бы в среднем только 16. Это затруднение можно устранить при помощи сложных кодов. Ниже мы вернемся к этому примеру.

Избыточность. Как показывает формула (28.4), максимальная энтропия на символ сообщения в n-символьном алфавите равна $\log_2 n$, и она достигается тогда, когда все вероятности равны $1/n$ и последовательные символы независимы. Если снять одно из этих ограничений, источник будет иметь энтропию H, меньшую, чем $H_{\text{макс}}$, и получается некоторая избыточность, определяемая формулой

$$\text{избыточность} = 1 - \frac{H}{H_{\text{макс}}}. \qquad (28.7)$$

Отношение $H/H_{\text{макс}}$ называется *относительной энтропией* или *эффективностью.* Избыточность и относительная энтропия могут принимать значения от 0 до 1.

В приведенном выше примере последовательные символы, по предположению, незави-

симы, но их вероятности неодинаковы: они равны соответственно 1/16 и 15/16. Тогда энтропия определяется по формуле (28.3):

$$H = -(1/16 \log_2 1/16 + 15/16 \log_2 15/16) =$$
$$= 0{,}25 + 0{,}0873 = 0{,}337 \text{ бита на символ сообщения.}$$

Относительная энтропия равна 0,377, а избыточность равна 0,663. Как мы покажем ниже, можно найти значительно более экономичный код.

Очевидный пример избыточности, вызванной зависимостью последовательных символов, мы встречаем в английском языке. После символа q мы почти всегда вынуждены производить символ u; после последовательности символов thr мы почти всегда вынуждены производить один из символов a, e, i, o или u. Каждое такое правило уменьшает нашу свободу выбора и, следовательно, в какой-то степени увеличивает избыточность.

Так, если в приближении нулевого порядка мы представим английский язык кодом, в котором все 27 символов (26 букв и интервал) имеют одинаковые вероятности, то энтропия будет равна $\log_2 27 = 4{,}76$ бита на символ. В приближении первого порядка к языку мы учитываем безусловную вероятность каждого символа (например: 0,13 для наиболее распространенного символа e и 0,00077 для наименее распространенного символа z), и энтропия будет равна 4,2 бита на символ. В приближении второго порядка мы учитываем зависимость каждого символа от предшествующего, но не от других символов (как в § 4.4), и энтропия будет равна 3,6 бита на символ. Приближение восьмого порядка дает 2,4 бита на символ. Итак, можно сказать, что английский язык имеет избыточность около 50% *.

При кодировании английского языка в точки и тире в телеграфии приближение первого порядка выражается использованием точки для e и тире для t, тогда как редкие буквы q и z выражаются соответственно как тире—тире—точка—тире и тире—тире—точка—точка. Можно придумать более сложные коды, представляющие приближения более высокого порядка, но обычно считают, что это не имеет смысла. Исключением являются *шаблонные* сообщения **, дающие чрезвычайно большое уменьшение избыточности.

Степень уменьшения энтропии вследствие такой зависимости можно вычислить из (28.3). Если мы обозначим первый символ пары как x, а второй — как y, то для одиночных символов

$$\left.\begin{aligned} H(x) &= -\sum_i p(i) \log_2 p(i), \\ H(y) &= -\sum_j p(j) \log_2 p(j), \end{aligned}\right\} \quad (28.8)$$

а для пар

$$H(x, y) = -\sum_{i,j} p(i, j) \log_2 p(i, j). \quad (28.9)$$

Подставляя (4.15) в (28.8), получаем:

$$H(x) = -\sum_i \Big[\sum_j p(i, j) \log_2 p(i)\Big] =$$
$$= -\sum_{i,j} p(i, j) \log_2 p(i), \quad (28.10а)$$

$$H(y) = -\sum_{i,j} p(i, j) \log_2 p(j). \quad (28.10б)$$

Складывая и сравнивая сумму с (28.9), получаем

$$H(x) + H(y) =$$
$$= -\sum_{i,j} p(i, j) \log_2 [p(i)\, p(j)] \geqslant H(x, y), \quad (28.11)$$

причем знак равенства справедлив только тогда, когда x и y независимы.

Символы неодинаковой длительности. В выведенных ранее выражениях мы получали все результаты в битах на символ сообщения. Если для всех символов сообщения требуется одно и то же время, эти результаты нетрудно перевести в биты в секунду, что удобно для сравнения с пропускной способностью канала. Но в некоторых случаях символы занимают различное время. Так, в телеграфии используются символы: точка, тире, пробел между буквами, пробел между словами, — имеющие неодинаковые длительности. В этом случае скорость информации определяется как

$$\lim_{T \to \infty} \frac{\log_2 N(T)}{T},$$

где $N(T)$ — число различных сообщений длительностью T.

---

* В позднейшей работе [133] Шеннон указывает, что в английском языке большое значение имеет крупномасштабная структура и что примерно при 100 буквах энтропия уменьшается примерно до одного бита на букву. — *Прим. авт.*

** Имеются в виду стандартизированные поздравления и т. п., передаваемые по телеграфу просто своим номером плюс фамилии и даты. — *Прим. ред.*

Если все символы имеют одинаковую длительность, это выражение сводится к предыдущим. Таким образом, оно является более общим, и Шеннон использует его для определения пропускной способности канала. Если символы имеют разную длительность, то для оценки этого выражения следует применять исчисление конечных разностей. Для телеграфии этот случай подробно разбирается в [40].

Оптимальное кодирование. В процессе кодирования мы по существу стремимся создать такой набор символов для передачи по каналу, чтобы все они появлялись независимо и с равной частотой; если это выполнено, то выход кодирующего устройства будет подобен источнику с максимальной энтропией, и пропускную способность канала можно полностью использовать, если источник создает символы с надлежащей скоростью.

Пример. Рассмотрим источник, имеющий четыре символа A, B, C и D, с соответствующими вероятностями $1/2$, $1/4$, $1/8$ и $1/8$. Энтропия этого источника равна $7/4$ бита на символ; максимальная энтропия источника с четырьмя символами равна двум битам на символ, и, следовательно, относительная энтропия источника равна $7/8$, а избыточность равна $1/8$.

Допустим, что нам нужно кодировать эти символы для передачи по каналу в двоичной форме. Один очевидный код таков: 00 для A, 01 для B, 10 для C и 11 для D. При этом потребуются два бита на символ источника, а, как мы знаем, можно получить бо́льшую скорость. Другим кодом может быть такой: 0 для A, 10 для B, 110 для C и 111 для D. Но в длинной последовательности, состоящей из s символов, мы будем передавать однобитовый символ A $s/2$ раза, двухбитовый символ B $s/4$ раза и трехбитовый символ (C или D) $s/4$ раза. Тогда общее число битов при передаче s символов будет равно $s/2+2s/4+3s/4=7s/4$, или $7/4$ бита на символ, так что данный код является наиболее эффективным из всех возможных кодов. Нужно заметить, что $7s/4$ бита будут состоять из $7s/8$ нулей и $7s/8$ единиц.

Предположим теперь, что канал может передавать четыре символа вместо двух (т. е. A', B', C' и D' вместо 0 и 1). Очевидный «код» будет таков: A' для A, B' для B и т. д. Этот код также имеет избыточность 1/8, и поэтому мы можем найти код, который будет эффективнее в 8/7 раза. Например, мы можем закодировать буквы длинного сообщения в цифры 0 и 1 при помощи описанного выше кода и затем закодировать их обратно в буквы, подставив A' вместо 00, B' вместо 01, C' вместо 10 и D' вместо 11. Таким образом, мы закодировали символы A, B, C и D в символы A', B', C' и D', и в длинном сообщении мы будем применять лишь 7/8 от прежнего числа символов.

Указанные выше числа были выбраны так, чтобы было легко найти наиболее эффективное кодирование. Предположим теперь, что вероятности символов таковы: $p_A=0{,}9$, $p_B=0{,}08$, $p_C=p_D=0{,}01$. Энтропия источника равна теперь 0,56 бита на символ, и, очевидно, не существует никакого простого способа закодировать одну букву в какую-нибудь двоичную последовательность, который дал бы меньше одного бита на символ.

Однако если мы возьмем последовательность из двух букв, то последовательность AA будет встречаться наиболее часто, и для ее обозначения мы можем применить код 0; все другие последовательности из двух символов будут иметь коды из двух или больше битов, начинающихся с 1. Тогда в наиболее вероятной двухсимвольной последовательности (AA) мы использовали бы только один бит для двух символов, или 0,5 бита на символ; но нам нужно было бы использовать по меньшей мере 2 и 3 бита соответственно для AB и BA и еще больше битов для менее вероятных последовательностей, и наилучший возможный код имел бы в среднем 0,72 бита на символ.

Если мы могли бы закодировать сразу шесть символов, мы по-прежнему применили бы один бит (0) для наиболее вероятной последовательности ($AAAAAA$) и могли бы очень близко подойти к 0,56 бита на символ, но код был бы чрезвычайно сложен. Действительно, существует 4096 различных возможных последовательностей из шести символов, и коды для многих из них были бы длиной в тысячи битов. При появлении одной из маловероятных последовательностей кодирующее устройство должно было бы накапливать последовательные символы от источника (которые, по предположению, создаются с постоянной скоростью) и выдавать последовательности канальных символов длиной в тысячи битов. Для этого потребовалось бы очень большое буферное хранилище. Для декодирующего устройства также потребовалось бы достаточно большая буферная память для хранения самой длинной возможной последовательности канальных символов. Далее, если бы нам встретилась длинная последовательность символов C и D (не из-за какого-нибудь чрезвычайного события — такая возможность слишком отдаленная, чтобы ее учитывать, — но вследствие неправильной оценки вероятностей), мы почти переполнили бы любое буферное хранилище в кодирующем устройстве.

Конструктивные формулы для отыскания наилучшего кода известны лишь в немногих случаях, включая предыдущий пример. Значительная часть пионерской работы в области теории информации посвящена отысканию таких формул.

Основная теорема. Предположим теперь, что у нас имеется канал, способный передавать четыре различных символа (которые, например, можно интерпретировать как A, B, C и D) со скоростью 10 символов в секунду, но не быстрее. Если этот канал используется для передачи символов с вероятностями 1/2, 1/4, 1/8 и 1/8 (1,75 бита на символ), то он будет передавать максимум 17,5 бита в секунду. Но мы, несомненно, можем вообразить себе, что этот канал используется для передачи символов (например, упомянутых выше A', B', C' и D') с вероятностями 1/4, 1/4, 1/4 и 1/4; в этом случае он будет передавать 20 битов в секунду. Ввиду (28.4) это есть максимальная скорость, с которой этот канал создает информацию на приемном конце. Это есть также *пропускная способность канала* в битах в секунду.

Таков смысл основной теоремы для бесшумного канала, которая гласит [40]: «Пусть источник имеет энтропию H битов на символ, а канал обладает пропускной способностью C битов в секунду. Тогда можно закодировать

выход источника таким образом, чтобы передавать символы по каналу со средней скоростью $C/H-\varepsilon$ символов в секунду, где ε сколь угодно мало. Передавать со средней скоростью, большей чем C/H, невозможно».

Пример. Вернемся теперь к примеру алфавита из двух символов с вероятностями p_A и p_B, равными соответственно 1/16 и 15/16. Мы показали, что энтропия этого источника равна 0,337 бита на символ сообщения. Основная теорема говорит, что для длинных последовательностей всегда можно найти эффективный способ кодирования, устраняющий по существу всякую избыточность. Таким образом, для сообщений длиной в 256 символов нам нужно только $0{,}337\times256=86{,}6$ бита на сообщение или, возможно, чуть больше; следовательно, существует способ кодирования, требующий в среднем именно этого числа битов, и мы никогда не сможем найти более эффективного кода.

Для того чтобы найти улучшенный код, предположим, что мы передаем просто число символов B, появляющихся между символами A (число незанятых положений индикатора между занятыми положениями). Так, поскольку в нашем примере первый самолет появился в шестом положении, мы будем передавать двоичное число пять (101). Если подряд появляются два A, мы вводим символ 0 (двоичное обозначение нуля). Редко будет больше 31 символа B подряд и иногда будет меньше 16, и поэтому мы можем с некоторым запасом надежности сказать, что нам нужно в среднем пять битов для A, или в среднем 80 битов на сообщение. Но здесь что-то неверно, так как теорема утверждает, что не существует кода, требующего меньше 86,6 бита, и ошибка, очевидно, состоит в том, что не было предусмотрено разделение чисел.

Так, если первое A появилось в шестом положении, а следующее — в тридцатом, то числа 5 и 23 будут переданы как 101 и 10111; однако в действительности нужно будет передавать 10110111, а это нельзя отличить от 10110 (22) и 111 (7).

Следующее предложение — применить в качестве разделительной отметки между числами три цифры 0. Для этого систему двоичного счисления нужно изменить так, чтобы исключить числа (такие, как 8, 16 и 24), содержащие три нуля подряд; например, символ 1001, нормально изображающий 9, будет изображать 8; 23 будет изображаться как 11010. Последовательность, приведенная в примере, будет теперь изображаться как 10100011010000. Этот код потребует в среднем для 256 символов сообщения несколько меньше 128 битов и несколько больше 86,6 бита и имеет то преимущество, что его можно, не меняя, применить к более длинным сообщениям.

Этот код можно еще слегка усовершенствовать. Последовательность 10000001 не может появиться в этом коде. То обстоятельство, что какая-то последовательность невозможна, указывает на избыточность. Эту избыточность можно устранить и сделать код более эффективным, исключив какое-нибудь одно число, например 0. Это означает, что если появятся подряд два A, между двумя группами по три нуля ничего не будет передано и, таким образом, вместо последовательности семи нулей будет шесть нулей. Нуль в действительности наиболее вероятное число (§ 6.7), но он требует только одного бита; число 2 почти столь же вероятно и требует двух битов, поэтому можно бы опустить его (т. е. шесть нулей подряд означали бы, что имелась группа из двух B между символами A).

Мы не будем затруднять себя решением этих сложных вопросов, но просто укажем на следующие обстоятельства: а) теория информации *не дает* в этом случае конструктивного метода отыскания оптимального кода; б) она дает нам эталон (в данном случае 0,337 бита на символ), с которым мы можем сравнивать любой придуманный нами код и который мы никогда не сможем превзойти; в) Шеннон гарантирует нам, что последний из упомянутых методов, основанный на применении надлежащего числа нулей (не обязательно трех, как было предложено выше), обеспечивает оптимальное кодирование для последовательностей бесконечной длины, когда p приближается к нулю (вместо 1/16, как мы предположили); г) более сложный код, хотя в нем используется в среднем меньше битов, может использовать в какой-нибудь конкретной выборке больше битов, чем простой код; д) при этом мы до сих пор ничего не сказали о том, какие будут последствия, если будет сделана одна ошибка.

Предположим, теперь, что вероятность одного A равна 1/2 вместо 1/16. В этом случае энтропия равна $-1/2\log_2 1/2-1/2\log_2 1/2=1$, и наиболее эффективный возможный код потребует 256 битов для 256 символов сообщения. Итак, в этом случае нельзя найти лучшего способа кодирования, чем простой код.

28.2. Дискретная система с шумом

Действие шума вызывает ошибки, и шум будет описываться вероятностью появления ошибок всех возможных видов (а именно, всякого возможного изменения переданного символа в отличный от него принятый символ).

Пример. Рассмотрим источник, передающий по каналу 1000 символов в секунду, причем каждый символ есть 0 или 1, с равной вероятностью. Если все эти символы принимаются правильно, скорость передачи информации равна 1000 битам в секунду. Предположим теперь, что в среднем происходит 10 ошибок в секунду (т. е. 0 превращается в 1 и наоборот с вероятностью 0,01); какова скорость передачи информации? 990 битов в секунду? Конечно, меньше! Если бы было 500 ошибок в секунду, мы могли бы с таким же успехом бросать монету на приемном конце без наблюдения принимаемого сигнала, т. е. никакой информации не передается и скорость передачи информации равна нулю. Ниже приведено численное решение этой задачи.

Ненадежность. Для численного определения скорости передачи информации при наличии ошибок мы должны исследовать, как энтропия зависит от условных вероятностей. Пусть передан символ x, а принят символ y. Если принятый символ неизвестен, энтропия передатчика равна

$$H(x)=-\sum_i p(i)\log_2 p(i), \qquad (28.3)$$

но если известно, что был принят символ j, то остаточная неопределенность для переданного символа будет меньше и энтропия будет равна

$$H_j(x)=-\sum_i p_j(i)\log_2 p_j(i),$$

где $p_j(i)$ — условная вероятность (вероятность того, что послано i, если известно, что принято j), определяемая как

$$p_j(i)=\frac{p(i,j)}{p(j)}. \qquad (28.12)$$

Величину $H_j(x)$ можно назвать *условной энтропией*.

Теперь найдем математическое ожидание, умножая эту условную энтропию для принятого сигнала j на вероятность $p(j)$ появления символа j и суммируя по j:

$$H_y(x)=-\sum_j p(j)\sum_i p_j(i)\log_2 p_j(i)=$$

$$=-\sum_{i,j} p(i,j)\log_2 p_j(i) \text{ битов на символ.} \qquad (28.13)$$

Результирующая условная энтропия $H_y(x)$ называется *ненадежностью*; она служит мерой неопределенности, вводимой шумом в канале.

Подставляя (28.12) в (28.13), получаем

$$H_y(x)=-\sum_{i,j} p(i,j)[\log_2 p(i,j)-\log_2 p(j)]=$$

$$=-\sum_{i,j} p(i,j)\log_2 p(i,j)+$$

$$+\sum_{i,j} p(i,j)\log_2 p(j). \qquad (28.14)$$

Но первый член в правой части равен (28.9), а второй член равен (28.10). Следовательно,

$$H(x,y)=H(y)+H_y(x)=$$

$$=H(x)+H_x(y)\leqslant H(x)+H(y), \qquad (28.16)$$

причем второе равенство очевидно ввиду симметрии. Условная энтропия $H_x(y)$ может и не быть равна ненадежности.

Скорость передачи информации относительно какого-либо источника равна энтропии источника минус ненадежность, если обе величины измерять в битах в секунду:

$$R=H(x)-H_y(x). \qquad (28.16а)$$

Пропускная способность канала определяется по-прежнему как максимальная скорость, получаемая для данного канала при учете всех источников, которые могут применяться с этим каналом:

$$C=\max_{\text{источник}} R=\max_{\text{источник}}[H(x)-H_y(x)]. \qquad (28.16б)$$

Если канал бесшумный, то ненадежность равна нулю, $H(x)=H(y)=H(x,y)$ и этот случай сводится к предыдущему.

В табл. 28.1 приведены численные значения, найденные по формуле (28.13) для предыдущего примера. Ненадежность равна 81 биту в секунду, а действительная скорость передачи информации равна 1000—81=919 битам в секунду. Если вероятность ошибки равна 0,5, то ненадежность равна одному биту на символ и по каналу вообще не передается никакой информации. Если вероятность ошибки больше 0,5, то информация опять передается и в пределе (когда вероятность ошибки равна 1,0) ненадежность равна нулю. Это понятно: когда принимается нуль, мы знаем наверняка, что была передана 1, и можем безошибочно воспроизвести сообщение.

Таблица 28.1

i	j	$p_j(i)$	$\log_2 p_j(i)$	$p(i,j)$	$p(i,j)\log_2 p_j(i)$
0	0	0,99	—0,01447	0,495	—0,00721
0	1	0,01	—6,64	0,005	—0,03320
1	0	0,01	—6,64	0,005	—0,03320
1	1	0,99	—0,01447	0,495	—0,00721
$\sum_{i,j}$	. . .	. . .	. . .	1,000	—0,0808

Приводимый ниже пример дискретного канала с шумом показывает, каким путем нужно сопоставлять канал со всеми возможными источниками для определения пропускной способности канала.

Пример. Пусть имеется канал, способный передавать три различных символа, которые мы обозначим как A, B и C, со скоростью один символ в секунду. Свойства канала таковы, что если передается A, то принимается A с вероятностью 1, а если передается B или C, то вероятность их правильного приема равна p, а вероятность их превращения соответственно в C или B вследствие шума в канале равна $q=1-p$. «Все возможные источники» в этом случае означают источники, производящие эти три символа с различными вероятностями (случай источников, производящих два символа, представляет собой частный случай, когда одна из этих вероятностей равна 0). Обозначим вероятности создания A, B и C соответственно через P, Q_1 и Q_2.

Задача состоит в том, чтобы определить P, Q_1 и Q_2 как функции от p так, чтобы получить максимальную скорость передачи информации по каналу; найденная максимальная скорость будет равна пропускной способности канала. Вместе с тем, найденные значения P, Q_1 и Q_2 мы можем использовать при кодировании, ибо они изображают выходы оптимального кодирующего устройства для случая, когда реальный источник имеет другие вероятности.

Прежде всего замечаем, что, по соображениям симметрии, Q_1 и Q_2 при максимальной скорости должны быть равны; поэтому мы можем опустить индексы. Затем замечаем, что задача не тривиальна. Если $p=1$ (случай канала без шума), то решение будет $P=Q=1/3$, а пропускная способность канала равна $\log_2 3$; если $p=1/2$, то приемник может лишь знать, что если

он принимает B или C, то было послано B или C, и не существует никакого способа выбрать между ними. Таким образом, задача сводится по существу к выбору одного из двух символов, и источник должен выбрать $P=1/2$, что дает пропускную способность канала $\log_2 2$. Для значений p между 1/2 и 1 будут найдены промежуточные значения для P и для пропускной способности.

В общем случае энтропия источника равна

$$H(x) = -P\log_2 P - 2Q\log_2 Q;$$

ненадежность равна

$$H_y(x) = -2Q(p\log_2 p + q\log_2 q) = 2Q\alpha,$$

где

$$\alpha = -(p\log_2 p + q\log_2 q);$$

а скорость передачи равна

$$R = -P\log_2 P - 2Q\log_2 Q - 2Q\alpha. \qquad (28.17)$$

Мы хотим найти максимум этой скорости при условии

$$P + 2Q = 1. \qquad (28.18)$$

Как известно из вариационного исчисления, для этого нужно найти производную от функции

$$G = -P\log_2 p - 2Q\log_2 Q - 2Q\alpha + \lambda(P + 2Q)$$

и положить ее равной нулю (λ—множитель Лагранжа).

$$G = -\log_2 e(P\log_e P + 2Q\log_e Q) - 2Q\alpha + \lambda(P+2Q),$$

$$\frac{\partial G}{\partial P} = -\log_2 e(1 + \log_e P) + \lambda = 0, \qquad (28.19)$$

$$\frac{\partial G}{\partial Q} = -2\log_2 e(1 + \log_e Q) - 2\alpha + 2\lambda = 0. \qquad (28.20)$$

Разделив (28.20) на 2 и вычтя полученное выражение из (28.19), исключим λ:

$$-\log_2 e(\log_e P - \log_e Q) + \alpha = 0,$$

$$\log_2 P = \log_2 Q + \alpha = \log_2 Q + \log_2 2^\alpha$$

$$P = Q2^\alpha$$

Подставляя это значение P в (28.18), получаем:

$$Q2^\alpha + 2Q = 1,$$

$$Q = \frac{1}{2+2^\alpha}, \quad P = \frac{2^\alpha}{2+2^\alpha}.$$

Подставляя эти значения в (28.17), мы находим максимальную скорость и, следовательно, пропускную способность канала:

$$C = R_{\text{макс}} = \log_2 \frac{2+2^\alpha}{2^\alpha} =$$

$$= \log_2 \frac{2+2^{-p\log_2 p - q\log_2 q}}{2^{-p\log_2 p - q\log_2 q}}.$$

Для предельных случаев $p=1$ и $p=1/2$ эта формула приводится соответственно к $C=\log_2 3$ и $C=\log_2 2$, как и следовало ожидать.

Основная теорема. Основная теорема для дискретного канала с шумом гласит [40]: «Пусть дискретный канал обладает пропускной способностью C, а дискретный источник — энтропией в секунду H. Если $H \leqslant C$, то существует такая система кодирования, что выход источника может быть передан по каналу с произвольно малой частотой ошибок, или сколь угодно малой ненадежностью. Если $H > C$, то можно закодировать источник таким образом, чтобы ненадежность была меньше $H-C+\varepsilon$, где ε сколь угодно мало. Не существует способа кодирования, обеспечивающего ненадежность, меньшую чем $H-C$».

Эта теорема весьма замечательна и интуитивно почти неправдоподобна. В приведенном выше примере, когда нули и единицы передавались с частотой 1000 символов в секунду и ошибки происходили случайно с частотой 10 ошибок в секунду, очевидным методом кодирования было бы повторение каждого символа. Для этого в первом примере потребовалась бы избыточность 500 символов на 500 информационных символов плюс добавочная избыточность для проверки несовпавших символов, и тем не менее мы ожидаем в среднем одной ошибки каждые 10 *сек* плюс различные ошибки при самой проверке. Теорема говорит, что существует способ кодирования с избыточностью только 81 символ на 919 информационных символов, при котором ошибки будут происходить реже, чем 1 ошибка в 10 *сек*, по существу даже сколь угодно редко. Конечно, за это нужно заплатить усложнением кода, большой буферной памятью и неприменимостью кода к источникам и каналам, обладающим другими свойствами, чем описанные выше.

В приводимом ниже примере представлен способ кодирования, позволяющий посылать символы по шумному каналу с нулевой ошибкой. Для упрощения задачи мы взяли помехи несколько нереалистического типа.

Пример. Пусть имеется источник, создающий сообщения длиной в 7 двоичных цифр. Свойства канала таковы, что либо в сообщении вовсе нет ошибок, либо есть ошибка только в одной из семи цифр, причем все эти восемь возможностей равновероятны. Таким образом, возможны 2^7 различных принятых сообщений и каждому из них соответствуют восемь возможных переданных сообщений, причем вероятность каждого переданного сообщения равна 1/8. Ввиду (28.13) ненадежность равна

$$H_y(x) = -\sum_j p(j)\sum_i p_j(i)\log_2 p_j(i) =$$

$$= \sum_{j=1}^{2^7} 2^{-7}\sum_{i=1}^{8} 1/8\log_2 1/8 = 3 \text{ бита на сообщение.}$$

Энтропия источника равна семи битам на сообщение, и это есть максимальная энтропия, так как все сообщения независимы и равновероятны. Следовательно, скорость информации (и пропускная способность канала) равна четырем битам на сообщение. Следовательно, оптимальный код будет содержать четыре полезных символа и три избыточных символа на сообщение.

Ниже описан следующий код, который имеет такое соотношение символов и передает каждое сообщение без ошибки. Обозначим семь двоичных цифр через $X_1, \ldots, X_7$. Пусть X_3, X_5, X_6 и X_7 — символы сообщения; три избыточных символа будут выбираться по следующим правилам:

X_4 выбирается так, чтобы $\alpha = X_4 + X_5 + X_6 + X_7$ было четным;

X_2 выбирается так, чтобы $\beta = X_2 + X_3 + X_6 + X_7$ было четным;

X_1 выбирается так, чтобы $\gamma = X_1 + X_3 + X_5 + X_7$ было четным.

На приемном конце эти величины α, β и γ вычисляются снова и обозначаются как 0, если они четные, и как 1, если они нечетные. Тогда последовательность цифр $\alpha\beta\gamma$ указывает (в двоичной системе счисления) индекс неправильно переданного символа; если будет 000, то ошибки не было.

28.3. Непрерывная система

Как показано ниже, непрерывный сигнал всегда допускает квантование, и поэтому его можно математически представить эквивалентным дискретным сигналом. Тогда можно непосредственно применить основную теорему для дискретного случая, по существу без изменения (за исключением очевидных переделок в определениях энтропии и ненадежности). К сожалению, это не приносит пользы.

Во-первых, шум приходится описывать более сложным образом — как непрерывную функцию времени, амплитуда которой определяется вероятностной функцией. Во-вторых, определение пропускной способности канала уже не является очевидным: пропускную способность почти всякого канала можно увеличить, если увеличивается мощность сигнала, и это нужно учитывать. Мы выведем одну теорему Шеннона — теорему о пропускной способности канала в случае, когда средняя мощность сигнала ограничена и шум относится к одному определенному типу. Существуют другие теоремы для случая, когда сигнал характеризуется предельной пиковой (а не средней) мощностью и шум имеет другие свойства. И в-третьих, определение «ошибки» в принятом сигнале не является очевидным: мы можем интересоваться верностью воспроизведения сигнала и тогда пропускная способность канала будет зависеть от того, как мы определяем верность.

Мы дадим краткие замечания по некоторым из этих вопросов.

Теорема о дискретизации. Рассмотрим любой непрерывный сигнал вида $G(t)$, не имеющий составляющих с частотой выше W *гц*.

*Теорема о дискретизации** говорит, что такая функция полностью определяется своими ординатами в дискретных точках, отстоящих друг от друга на $1/2W$ *сек*, т. е. значениями функции при

$$t = 0, \pm\frac{1}{2}W, \pm\frac{2}{2}W, \ldots, \pm\frac{n}{2}W.$$

Для доказательства рассмотрим интеграл Фурье

$$G(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{i\omega t} F(\omega)\, d\omega.$$

Так как функция $G(t)$ не содержит частот свыше $2\pi W$ *рад/сек*, то мы можем сместить пределы интегрирования до $\pm 2\pi W$:

$$G(t) = \frac{1}{2\pi}\int_{-2\pi W}^{2\pi W} e^{i\omega t} F(\omega)\, d\omega.$$

Но точки $t = n/2W$ определяют коэффициенты Фурье для $F(\omega)$, где n принимает значения $0, \pm 1, \ldots, \pm\infty$. Например, для $n=3$ ордината равна

$$G\left(\frac{3}{2W}\right) = \frac{1}{2\pi}\int_{-2\pi W}^{2\pi W} e^{\frac{3i\omega}{2W}} F(\omega)\, d\omega.$$

Следовательно, эти наблюденные значения функции G полностью определяют $F(\omega)$, а так как $F(\omega)$ полностью определяет $G(t)$, то мы можем восстановить функцию $G(t)$ по значениям функции $G\left(\frac{n}{2W}\right)$.

В результате получаем следующую формулу [87]:

$$G(t) = \sum_{n=-\infty}^{\infty} G\left(\frac{n}{2W}\right)\frac{\sin(2\pi W t - n\pi)}{2\pi W t - n\pi}. \quad (28.21)$$

Итак, если мы хотим найти значение функции в некоторый момент t, не являющийся одной из точек выборки $t = \frac{n}{2W}$, мы можем вычислить ее точно, суммируя бесконечный ряд (28.21). В практических случаях, когда длительность сигнала конечна, формула (28.21) еще дает превосходное приближение, если длительность сигнала велика по сравне-

* Эта теорема была сформулирована русским радиоинженером В. А. Котельниковым (ныне академиком) в 1933 г. — *Прим. ред.*

нию с $1/2\,W$. Когда полоса ограничена шириной W *гц*, но не обязательно начинается от нуля, тогда необходимы и достаточны по-прежнему $2W$ выборочных значений (дискрет) в секунду, хотя в этом случае нужно применять другую формулу.

Квантование непрерывного канала. Если непрерывный сигнал можно заменить $2W$ дискретными сигналами в секунду, то мы можем заменить непрерывный источник дискретным источником, создающим $2W$ символов в секунду, и непосредственно применить основные теоремы для дискретного случая. Различные дискретные символы становятся сигналами различной амплитуды, и разность амплитуд, определяющая два *различных* сигнала, будет представлять собой наименьшую *различимую* разность для нашего непрерывного сигнала, которая, в свою очередь, будет зависеть от шума в канале.

Однако выбор этого уровня квантования не повлияет на наши математические выводы. Если мы возьмем очень мелкое квантование, то будет много уровней и, следовательно, большая энтропия источника, но вероятность ошибки (когда один уровень принимается за другой) увеличится и тем самым увеличится ненадежность, а потому скорость передачи информации по каналу останется той же самой. Например, если мы возьмем уровень квантования 0,1 *в* и пошлем в качестве «символа» 1,3 *в* и если принятые сигналы имеют нормальное распределение амплитуд с математическим ожиданием 1,3 *в* и стандартным отклонением 0,1 *в*, то вероятность принять символ 1,3 *в* равна 0,383; вероятность принять 1,2 *в* равна 0,242; вероятность принять 1,1 *в* равна 0,061; и т. д.

С такой же самой ситуацией мы уже встречались в задаче предыдущего параграфа о символах P, Q_1 и Q_2, и пропускную способность канала можно вычислить таким же способом, если у нас есть достаточно данных о распределении вероятностей амплитуд сигнала и шума и если известны автокорреляционные функции сигнала и шума. Автокорреляционная функция непрерывного сигнала соответствует избыточности дискретных сигналов, подобной описанной в § 28.1 при рассмотрении английского языка; уменьшение энтропии, вызванное этими условными вероятностями, определяется формулами (28.13)—(28.15). Автокорреляционная функция описывает, как изменяется коэффициент корреляции двух амплитуд сигнала, определяемый формулой (6.28), с изменением временного интервала между ними.

Например, в телевидении существует большая вероятность того, что данное световое пятно кадра будет неизменным в двух последовательных кадрах и что два соседних пятна на одной строке также будут одинаковы. Поэтому избыточность телевизионных изображений большая (Виснер [88] говорит, что она равна 80—95%), и теоретически можно получить большую экономию, кодируя их до передачи. Основные теоремы для дискретных систем могут указать нам, какое улучшение достижимо посредством кодирования, если мы найдем способы вычисления избыточности и пропускной способности канала.

Непрерывный источник. Дискретные источники описывались заданием вероятностей $p(i)$, связанных с каждым возможным символом; непрерывные источники описываются заданием плотности вероятностей возможных амплитуд. Таким образом, $p(x)$ есть вероятность того, что амплитуда [которую можно представить себе как $G(t)$ или $G\left(\frac{n}{2W}\right)$ в (28.21)] лежит между x и $x+dx$. Определение энтропии (28.3), очевидно, переходит в следующее:

$$H(x) = -\int_{-\infty}^{\infty} p(x)\log_2 p(x)\,dx, \qquad (28.22)$$

а вместо (28.9) и (28.13) у нас появляется

$$H(x, y) = -\int_{-\infty}^{\infty}\int p(x, y)\log_2 p(x, y)\,dxdy \qquad (28.23)$$

и

$$H_y(x) = -\int_{-\infty}^{\infty}\int p(x, y)\log_2 \frac{p(x, y)}{p(y)}\,dxdy. \qquad (28.24)$$

С помощью (6.21) можно вывести соотношения между этими величинами, аналогичные (28.11) и (28.15).

Имеется существенное различие между энтропией, определяемой формулой (28.22), и энтропией, определяемой формулой (28.3); последняя есть абсолютная и всегда положительная величина, а первая зависит от координатной системы и может быть даже отрицательной. Это подразумевалось выше, где мы указали, что энтропия источника зависит от того, как мы его квантовали; но в связи с этим мы показали эвристически, что скорость передачи информации и тем самым пропускная способность канала не зависят от уровней квантования.

Рассмотрим функцию с энтропией в се-

кунду $H(x)$ и произведем преобразование переменной $y=f(x)$. Ввиду (6.31)

$$q(y)=\frac{p(x)}{f'(x)}$$

и, конечно,

$$dy=f'(x)\,dx.$$

Следовательно,

$$H(y)=-\int_{-\infty}^{\infty} q(y)\log_2 q(y)\,dy=$$

$$=-\int_{-\infty}^{\infty}\frac{p(x)}{f'(x)}\log_2\frac{p(x)}{f'(x)}f'(x)\,dx=$$

$$=-\int_{-\infty}^{\infty}p(x)\log_2 p(x)\,dx+$$

$$+\int_{-\infty}^{\infty}p(x)\log_2 f'(x)\,dx=H(x)+$$

$$+E[\log_2 f'(x)]=H(x)-E(\log_2 J), \quad (28.25)$$

где $J=\frac{1}{f'(x)}$ — якобиан рассматриваемого преобразования [см. (6.32)]. Такое же уравнение справедливо, когда x и y суть многомерные случайные величины, так что

$$H(x)=-\int_{-\infty}^{\infty}\dots\int_{-\infty}^{\infty}p(x_1,\dots,x_n)\times$$

$$\times\log_2 p(x_1,\dots,x_n)\,dx_1\dots dx_n.$$

При вычислении скорости передачи информации или пропускной способности канала математическое ожидание логарифма якобиана равно нулю, так как эти вычисления включают вычитание одной энтропии из другой. Следовательно, полученные величины не зависят от координатной системы.

Максимальная энтропия. В связи с (28.4) было отмечено, что энтропия источника максимальна, если все символы равновероятны. Мы могли бы предположить, что максимальная энтропия от непрерывного источника получится при равномерном распределении $p(x)=\text{const}$, но это не так.

Рассмотрим равномерное распределение с плотностью вероятностей, равной 1 на интервале от 0 до 1 и нулю вне этого интервала. Тогда

$$H(x)=-\int_0^1 1\times\log 1\times dx=0,$$

и энтропия равна нулю. Любая функция распределения с меньшей случайностью (например, более узкое равномерное распределение с плотностью вероятности больше 1) будет иметь отрицательную энтропию. Для отыскания функции распределения* с максимальной энтропией мы должны найти максимум выражения (28.22) при соблюдении условий (6.2) и (6.5), которые устанавливают соответственно, что $p(x)$ есть плотность вероятностей, а дисперсия распределения имеет фиксированное значение σ^2. Допустим, что система координат выбрана так, что математическое ожидание распределения равно нулю.

Согласно вариационному исчислению, искомый максимум находится путем дифференцирования функции

$$G(x)=-p(x)\log_2 p(x)+\lambda x^2 p(x)+\mu p(x)$$

по $p(x)$, где λ и μ — множители Лагранжа. Это дает:

$$\frac{\partial G}{\partial p}=-\log_2 e\,[1+\log_e p(x)]+\lambda x^2+\mu=0,$$

$$\log_e p(x)=\frac{\lambda x^2}{\log_2 e}+\frac{\mu}{\log_2 e}-1=c_1x^2+c_2,$$

$$p(x)=e^{c_1x^2+c_2}=e^{c_1x^2}e^{c_2}+c_3e^{c_1x^2}.$$

Но мы показали (§ 6.4), что любая плотность вероятностей такого вида изображает нормальное распределение.

В свете центральной предельной теоремы этот вывод не является неожиданным. Он говорит, что нормальное распределение более «случайно», чем любое другое распределение с той же дисперсией. Для наших непосредственных целей это означает, что сигнал с нормально распределенными амплитудами обладает большей энтропией, чем любой другой непрерывный сигнал с той же средней мощностью переменного тока. Это вытекает из нашего допущения о нулевом математическом ожидании, которое равносильно тому, что мы отсчитываем напряжение (так как амплитуды суть напряжения) от уровня постоянного смещения. Тогда мощность переменного тока равна среднему квадрату отклонения напряжения, а это и есть дисперсия, которую мы считаем постоянной.

---

* Отдельно взятая функция, собственно говоря, не имеет энтропии, в том смысле, что не может быть никакой неопределенности относительно известной функции. В действительности мы определяем здесь плотность вероятностей, описывающую «ансамбль» (или множество) функций, который обладал бы свойством максимальной случайности. Математический анализ ансамблей функций занимает значительное место в работе Шеннона. — *Прим. авт.*

Белый шум. Сигнал, определяемый формулой (28.21), в которой значения $G\left(\frac{n}{2W}\right)$ распределены нормально и независимо, называется *белым шумом с ограниченной полосой частот.*

Он называется *белым* потому, что плотность вероятностей, определяющая его амплитуду, постоянна в полосе частот шума (Шеннон указывает, что это название неудачно, так как «белый свет» обычно означает спектр, равномерно распределенный по длине волны, а это не совпадает с равномерным распределением по частоте). *Шум** означает сигнал, амплитуда которого в любой момент непредсказуема, т. е. сигнал, автокорреляционная функция которого равна нулю для любого промежутка времени.

Однако сигнал с ограниченным спектром требует конечного времени для изменения амплитуды, вследствие чего здесь существует какая-то автокорреляция, по крайней мере за малые промежутки времени. Теорема о дискретизации по существу говорит, что автокорреляционная функция для белого шума с ограниченной полосой частот принимает нулевое значение через интервал времени $1/2W$. Таким образом, слова «*с ограниченной полосой частот*» указывают нам, в какой мере последовательные наблюдения из этого нормального распределения независимы одно от другого; чем больше ширина полосы, тем больше независимость, больше случайность и больше энтропия. *Мощность* такого белого шума с ограниченной полосой частот обычно обозначается как $N=\sigma^2$.

Энтропия нормального распределения с дисперсией σ^2, имеющего максимальную энтропию из всех распределений с такой же дисперсией, определяется формулой (28.22), в которой $p(x)$ выражается формулой (6.11):

$$p(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2\sigma^2}}$$

и

$$-\log_2 p(x)=-\log_2\frac{1}{\sqrt{2\pi}\sigma}-\log_2 e^{-\frac{x^2}{2\sigma^2}}=$$

$$=\frac{1}{2}\log_2 2\pi\sigma^2+(\log_2 e)\frac{x^2}{2\sigma^2}.$$

Следовательно, энтропия равна

$$H(x)=-\int_{-\infty}^{\infty}p(x)\log_2 p(x)\,dx=$$

$$=\frac{1}{2}\log_2 2\pi\sigma^2\int_{-\infty}^{\infty}p(x)\,dx+\frac{\log_2 e}{2\sigma^2}\int_{-\infty}^{\infty}x^2p(x)\,dx.$$

Но первый интеграл равен 1 ввиду (6.26), а второй интеграл равен σ^2 ввиду (6.5). Отсюда

$$H(x)=\frac{1}{2}\log_2 2\pi\sigma^2+\frac{1}{2}\log_2 e=$$

$$=\frac{1}{2}\log_2 2\pi e\sigma^2=\log_2\sqrt{2\pi eN}\,. \quad (28.26)$$

Равенство (28.26) определяет энтропию на дискрету, соответствующую энтропии на символ дискретного источника. Так как существует $2W$ независимых дискрет в секунду, то энтропия в секунду белого шума с ограниченной полосой частот равна

$$H=W\log_2 2\pi eN. \quad (28.27)$$

Это есть максимальная скорость, с которой создает информацию источник с шириной полосы W и средней мощностью N.

Пропускная способность канала как функция мощности шума, средней мощности сигнала и ширины полосы. Равенство (28.27) дает основу для вычисления пропускной способности канала по формуле (28.16), но мы сначала должны связать ненадежность $H_y(x)$ с энтропией шума, который, по предложению, создает помехи в канале. Обозначим ее через $H(n)$. Мы делаем естественное допущение, что этот шум не зависит от переданного сигнала x и что принятый сигнал y равен сумме

$$y=x+n. \quad (28.28)$$

Тогда ввиду (28.25)

$$H(x,y)=H(x,n)-E(\log J)=H(x,n), \quad (28.29)$$

так как якобиан от (28.28) равен единице. Подставляя (28.15) в обе части равенства (28.29), получаем

$$H(y)+H_y(x)=H(x)+H_x(n)=H(x)+H(n),$$

так как $H(n)=H_x(n)$ в силу предположения о независимости x и n.

Итак, (28.16) принимает вид

$$R=H(x)-H_y(x)=H(y)-H(n),$$

* Термин «шум» используется в технике связи в двух различных смыслах: в смысле ненужного сигнала и в описанном выше смысле — *Прим. авт.*

и пропускная способность канала будет определена, когда мы найдем максимум этого выражения. Поскольку $H(n)$ не зависит от наших действий, мы должны найти максимум $H(y)$ энтропии принятого сигнала.

По крайней мере один важный случай допускает простое математическое решение. Это тот случай, когда шум является белым шумом с ограниченной полосой (и с мощностью N), а передаваемые сигналы имеют максимальную среднюю мощность S. Принятые сигналы имеют максимальную энтропию, когда они распределены нормально (т. е. подобны белому шуму); так как принятый сигнал имеет мощность $S+N$, то ввиду (28.27) его энтропия равна

$$H(y)=W\log_2 2\pi e(S+N).$$

Энтропия шума равна

$$H(n)=W\log_2 2\pi eN,$$

а скорость передачи информации

$$R=H(y)-H(n)=W\log_2\frac{S+N}{N}.$$

Так как эта скорость максимальна, она равна пропускной способности канала:

$$C=W\log_2\left(1+\frac{S}{N}\right)\text{ битов в секунду.}\qquad(28.30)$$

Равенство (28.30) представляет собой, быть может, наиболее важную формулу в теории информации, поэтому целесообразно повторить допущения, необходимые для ее вывода: средняя мощность источника сигнала ограничена; шум является белым шумом с ограниченной полосой и не зависит от сигнала; канал имеет ограниченную ширину полосы W; кроме того, делаются определенные, более философские допущения относительно природы информации. Предположение о белом шуме часто оказывается верным; для тех случаев, когда оно не выполняется, выведены верхние и нижние границы пропускной способности канала. Предположение об ограниченной средней мощности часто не выполняется; для случая, когда ограничена пиковая мощность, полного решения нет, но известны некоторые пределы, особенно для частных случаев (например, при очень больших или очень малых отношениях сигнал/шум).

Основная теорема утверждает, что можно найти способ кодирования, при котором любой источник будет создавать информацию и посылать ее по непрерывному каналу со скоростью, равной пропускной способности канала, и с произвольно малыми ошибками. Свойства кода должны быть таковы, чтобы переданный сигнал, сложенный с шумом канала, производил принятый сигнал, обладающий свойствами белого шума. В § 7.4 мы показали, что если сложить два нормальных распределения с нулевым средним, то их сумма будет представлять собой нормальное распределение с нулевым математическим ожиданием и дисперсией, равной сумме отдельных дисперсий. Следовательно, переданный сигнал должен также иметь свойства белого шума с ограниченной полосой частот, и для достижения наибольшей эффективности кодирующее устройство должно создавать такие сигналы.

Пример [40]. Мы предполагаем, что сообщение дискретно или, если оно непрерывно, что оно было квантовано, как описано выше. Предполагается, что это, уже дискретное, сообщение было закодировано в оптимальную двоичную форму методами, рассмотренными в § 28.1. Пусть теперь генератор шума производит, скажем, восемь образцов (выборок) белого шума с шириной полосы W, каждый длительностью $3/C$ *сек*, где C — пропускная способность канала в битах в секунду. Эти восемь образцов нумеруются от 0 (в двоичном обозначении 000) до 7 (в двоичном обозначении 111) и запоминаются как на передающем, так и на приемном конце канала.

Когда сообщение в двоичном коде приходит к кодирующему устройству, мы берем сразу три двоичных разряда, выбираем соответствующий из восьми образцов шума и передаем его.

На приемном конце искаженный сигнал сравнивается с восемью образцами и образец с наименьшим среднеквадратическим отклонением признается правильным; затем три цифры восстанавливаются по известному номеру этого образца шума.

Если отношение сигнал/шум мало, то пропускная способность канала будет мала, а длительность $3/C$ — велика, вследствие чего возрастает возможность опознавания сигнала. Конечно, при этом способе будут случайные ошибки; если частота ошибок больше допустимой, мы используем 16 образцов шума длительностью в $4/W$ и кодируем сразу по четыре двоичные цифры. Если продолжать эту процедуру, то вероятность того, что будет выбран не тот образец и, следовательно, совершена ошибка, может быть сделана сколь угодно мала, как установлено основной теоремой.

«Это рассуждение, однако, — говорит Шеннон, — обходит действительное положение вещей. Практически при непрерывном источнике может интересовать не точная передача, а передача с определенным допуском. Вопрос заключается в том, можно ли приписать непрерывному источнику конечную скорость в том случае, когда требуется только определенная верность воспроизведения, измеренная подходящим способом».

Шеннон определяет критерий верности очень широко с помощью *оценивающей функции* v; определение является настолько общим, что оно может включать такие разные измерители, как например: обычное среднеквадратическое отклонение, т. е. среднее значение от $[x(t)-y(t)]^2$, где $x(t)$ и $y(t)$ — переданный и принятый сигналы соответственно; среднеквадратический критерий с частотным взвешиванием; критерий абсолютного отклонения; «разборчивость речи»; наконец, критерий ошибки в дискретном случае. По отношению к этому весьма общему критерию Шеннон определяет скорость создания информации.

Такое определение обратно определению пропускной способности канала (28.16); по существу оно состоит в том, что рассматриваемый источник берется в сочетании со всеми возможными каналами и скорость создания информации определяется как минимальная пропускная способность, при которой можно передать сигналы этого источника с требуемой верностью.

Затем Шеннон доказывает основную теорему для этого случая: «Если источник при данной оценке v_1 имеет скорость создания информации R_1, то можно закодировать выход источника и передавать его по каналу пропускной способности C при верности воспроизведения, как угодно близкой к v_1, если только $R_1 \leqslant C$. Это невозможно, если $R_1 > C$».

Раньше трудность состояла в вычислении пропускной способности канала, здесь же трудность заключается в вычислении скорости создания информации источником по отношению к критерию верности. В одном частном случае задача решена: источник производит белый шум с ограниченной полосой частот мощностью Q, критерием верности служит среднеквадратическое отклонение и допустимый средний квадрат ошибки между переданным и принятым сообщением равен N. В этом случае скорость информации равна

$$R = W \log_2 \frac{Q}{N}. \qquad (28.31)$$

Для других случаев найдены пределы.

28.4. Теория информации и техника связи

Теория информации в том виде, как она изложена в этой главе, возникла почти полностью из статей Шеннона 1948 года. Конечно, многие основы были заложены другими. Например, логарифмическая мера информации была предложена еще в 20-е годы, были разработаны методы кодирования, подобные рассмотренным в § 28.1, для повышения эффективности (экономичности) передачи, а Винер показал, что теория связи есть статистическая задача, и разработал некоторые статистические орудия, необходимые для ее анализа. Кроме того, была известна теорема о дискретизации и было признано значение ширины полосы и отношения сигнал/шум, как параметров, определяющих работу систем связи. Наконец, был разработан ряд весьма сложных систем связи, в том числе частотная модуляция и различные системы импульсной модуляции, и при помощи сложного математического анализа были выяснены их свойства.

Тем не менее именно Шеннон собрал все это вместе. Он определил энтропию информации и пропускную способность канала, сформулировал и доказал основные теоремы и вывел такие формулы, как (28.30). Одним из направлений исследований в теории информации является отыскание более фундаментальных формул такого рода.

Некоторые авторы пытались расширить теорию, устраняя часть ограничений, наложенных Шенноном.

Например, Шеннон утверждает, что для источника с двумя символами энтропия на символ равна

$$H = -p_1 \log_2 p_1 - p_2 \log_2 p_2.$$

Это — средняя величина (точнее, математическое ожидание логарифма, взятое со знаком минус), которая применима математически точно только к последовательностям бесконечной длины. Шеннон не определил энтропию или дополнительную информацию, которая создается при возникновении символа 1. Другие авторы, например Голдман [90] и Вудворд [89], предположили в порядке экстраполяции, что информация, создаваемая при возникновении символа 1, равна $H = -\log_2 p_1$. Такая формулировка может быть шагом к разработке теории информации, принимающей во внимание семантический аспект.

Мы описывали здесь применение теории информации к связи, и в частности к задаче кодирования при связи. Этим вопросам Шеннон уделял главное внимание, и к ним теория информации была применена с наибольшим успехом, но, по-видимому, область ее применения гораздо шире. Много усилий было потрачено, например, на применение ее к *теории шумов* и к задаче расчета фильтров, восстанавливающих сигнал заданного типа, искаженный шумом с заданными свойствами. Она была применена также к задачам *предсказания*, т. е. экстраполяции функции времени, наблюдавшейся в течение известного периода,

причем все наблюдения более или менее искажены шумом.

Как отмечалось в начале главы, теория информации указывает верхние границы и в большинстве случаев позволяет вычислить, как близко мы подходим к ним, но она, вообще говоря, не дает конструктивных методов повышения эффективности связи или предполагает явно непрактичные методы. Однако сейчас наблюдаются значительные успехи как в разработке конструктивных математических методов (которые мы отличаем от теорем о существовании), так и в разработке аппаратуры для кодирования. За повышение эффективности связи, конечно, нужно платить увеличением аппаратуры, в частности для буферной памяти, и запаздыванием. Но выгоды получаются большие; как указывает Виснер [88], если бы можно было найти способ сократить в два раза ширину полосы, необходимую для телевизионных передач, то это дало бы весьма большое сокращение денежных расходов на трансконтинентальные релейные линии.

ЛИТЕРАТУРА

Всегда интересно и обычно весьма полезно обратиться к оригинальным статьям по любому вопросу. Особенно хороши статьи Шеннона [40], которые обязательны для любого изучающего теорию информации. Из позднейших работ по этому вопросу Вудворд [89] содержит краткое, ясное изложение одного из приложений теории информации. Голдман [90] разрабатывает некоторые частные стороны, как кодирование и шум. Джексон [86] в общем дает обзор состояния теории информации к концу 1952 г.

ЗАДАЧИ *

28.1. Система радиозонда измеряет температуру и давление. Результаты измерений показывают два прибора: термометр имеет 100 отметок шкалы (различных значений) и его отсчет может измениться до любого допустимого значения за 0,05 *сек*; барометр имеет 10 отметок шкалы и его отсчеты могут измениться до любого допустимого значения за 0,01 *сек*. Какая требуется пропускная способность кода в битах в секунду, чтобы передать всю возможную информацию, которая может появиться на приборах?

28.2 *. Какова энтропия источника в задаче 28.1, если показания барометра будут появляться не с одинаковой частотой, но со следующими вероятностями?

Отметка шкалы	0	1	2	3	4	5	6	7	8	9
Вероятность	0,05	0,05	0,05	0,05	0,1	0,2	0,3	0,1	0,05	0,05

28.3. Если в задаче 28.1 шум белый и во время приема сигналов отношение сигнал/шум по средней мощности равно 10, то какая потребуется ширина полосы?

28.4. Думаете ли вы, что на практике будет предоставлена такая большая ширина полосы? Дайте свои соображения.

28.5*. Алфавит состоит из восьми согласных и восьми гласных. Предположим, что все буквы алфавита равновероятны и нет зависимости между символами. Согласные всегда воспринимаются правильно, а гласные воспринимаются правильно лишь в половину времени и при ошибке истолковываются как другие гласные, с одинаковой вероятностью для каждой гласной. Если в секунду передается одна буква, то какова средняя скорость передачи информации?

28.6*. По каналу могут передаваться два символа: 0 и 1, со скоростью 1 символ в секунду. Источник может производить три символа: A, B и C, с равными вероятностями и с любой скоростью до двух символов в секунду.

а) Какова максимальная пропускная способность этого источника в сочетании с любым каналом?

б) Какова пропускная способность канала?

в) Если символы источника кодируются в канальные символы по схеме $A \to 100$, $B \to 0,1$, $C \to 10$, то какова максимальная скорость передачи?

г) Если B кодируется в 01, а B — в 10, как и раньше, но A кодируется в 00 в течение одной трети всего времени и в 11 в течение двух третей времени, то какова максимальная скорость передачи?

* Взято с некоторыми изменениями из [90]. — *Прим. авт.*

ГЛАВА 29

РЕФЛЕКСИВНОЕ УПРАВЛЕНИЕ. ТЕОРИЯ АВТОМАТИЧЕСКОГО РЕГУЛИРОВАНИЯ

Сердцем всякой автоматической системы является управляющее устройство того или иного типа. В больших системах, составляющих предмет этой книги, оно принимает форму логического управления, рассмотренного в гл. 22, и обычно включает большую автоматическую вычислительную машину. В малой системе оно обычно принимает форму системы автоматического регулирования, или, иначе говоря, «следящей» системы. Как мы указали, нет резкой границы между большими и малыми системами. Кроме того, большинство больших систем включает много вспомогательных малых цепей управления, поэтому теория автоматического регулирования (теория следящих систем) является важным орудием проектировщика систем.

Применения систем автоматического регу-

лирования. Системы автоматического регулирования, или *следящие системы*, называемые также *сервомеханизмами* *, применяются в больших системах всюду, где требуется рефлексивное управление — управление из какого-нибудь более или менее периферийного пункта, без доступа к центральному логическому управлению. Такое управление характеризуется частым повторением выходов одного и того же типа и сравнительно малым временем реакции.

Примеры управления этого типа в больших системах: управление положением радиолокационной антенны в системе управления воздушным движением; управление положением, скоростью и ускорением реактивного снаряда в системе управляемых реактивных снарядов; контроль над температурой, давлением и скоростью воздуха и над положением модели в аэродинамической трубе; управление положением резца фрезерного станка на автоматическом заводе. Кроме того, существуют некоторые другие задачи управления (такие, как установление положения самолета в аэронавигационной системе), в которых длительность реакций сравнительно велика, но которые тем не менее удобно анализировать при помощи теории следящих систем.

Основную задачу системы автоматического регулирования всегда можно свести к следующему: при данной команде, которая может быть функцией времени, нужно управлять некоторой физической величиной так, чтобы она была близка к значению командной функции. Этой «регулируемой» физической величиной может быть положение, скорость или ускорение в механической системе. В других системах это может быть температура, напряжение, поток нейтронов и т. д., однако такие величины можно считать аналогами положения, скорости или ускорения, и, если не оговорено противное, мы будем иметь в виду системы автоматического регулирования, управляющие перемещением и его производными по времени.

Управление по разомкнутой и замкнутой петле. Самым простым видом системы управления является управление по *разомкнутой петле* (*разомкнутой цепи*). Например, представим себе автомобиль с несколькими отмеченными положениями акселератора; если нажать акселератор до определенного положения, автомобиль номинально должен идти со скоростью точно 50 миль в час. Если даже все время машина идет по ровной дороге, этот способ управления неудовлетворителен (например, потребуется очень много времени для достижения установленной скорости). Обычно цепь управления замыкается зрительной обратной связью (по наблюдению дороги или по спидометру); иначе говоря, водитель прикладывает значительную мощность, пока не достигнет желаемой скорости, и затем регулирует мощность так, чтобы поддерживать эту скорость.

Однако во многих случаях этот тип управления (с участием человека в цепи управления) недостаточен: либо потому, что нежелательно использовать человека подобным образом, либо потому, что такая система имеет чрезмерное запаздывание или недостаточную точность. Для улучшения системы управления ее часто механизируют следующим образом: 1) дается команда; 2) проверяют, насколько выходная величина согласуется с командой; 3) если между ними нет хорошего согласования, прилагают мощность для выправления этого расхождения. Если полученная таким образом система управления по *замкнутой петле* (*замкнутой цепи*) получает более или менее постоянную команду и лишь регулирует свой выход при изменении нагрузки, то она называется *системой автоматического регулирования в узком смысле* или *собственно регулятором* (который можно включать или нет в понятие «следящей системы»); если же это — воспроизводящее устройство, в котором команда меняется с течением времени, то перед нами *система автоматического регулирования в широком смысле* или *собственно следящая система* (следящее устройство, сервомеханизм) **.

Итак, система автоматического регулирования содержит устройство измерения выхода, соединенное обратной связью с прибором, измеряющим ошибку, усилитель для усиления сигнала ошибки и источник энергии для

---

* Авторы, как обычно в американской литературе, пользуются исключительно термином «сервомеханизмы», точным русским эквивалентом которого является термин «следящие системы». Термин «следящие системы» применяется довольно широко в русской технической литературе, используется в ней и термин «сервомеханизмы», особенно в более ранних работах. Однако в русской науке теория следящих систем обычно называется иначе — теорией автоматического регулирования, а следящие системы в общем случае обычно именуются системами автоматического регулирования. В соответствии с этим мы при переводе большей частью либо заменяем «сервомеханизмы» на «системы автоматического регулирования», либо говорим о «следящих системах», либо даем синонимы, что приводит иногда к небольшим изменениям текста. — *Прим. ред.*

** В русской оригинальной литературе под следящими системами большей частью подразумевается именно этот второй случай. — *Прим. ред.*

выполнения команды. В целом система — чувствительное к ошибке, воспроизводящее, усиливающее устройство между входом (командой, или заданием) и выходом (исполнением, или отработкой, или реакцией). Обычно все же существует некоторое расхождение по времени, или по величине, или по тому и другому между управляемой переменной и управляющим сигналом, потому что запасающие энергию физические элементы, применяемые в таких системах, имеют запаздывающую реакцию из-за трения и т. д. Далее, введение цепи обратной связи и усилителя приводит к тому, что команда может поступать в такие моменты по отношению к запаздывающей нагрузке, что эти факторы оказываются не в фазе; другими словами, корректирующий сигнал может увеличивать первоначальную команду, так что она становится все больше и больше. Таким образом, появилась новая трудность — возможность неустойчивости. Наконец, команда может содержать помеху (нежелательный сигнал), которая может подействовать нежелательным образом на управляющую систему; однако в настоящей главе это не рассматривается.

Критерии качества работы. При проектировании системы автоматического регулирования нужно, следовательно, удовлетворить три группы критериев качества: устойчивости, стационарного состояния и переходного процесса.

Вообще говоря, неустойчивых устройств не применяют. Стационарные условия характеризуются статической ошибкой — расхождением между входом и реакцией, остающимся после затухания всех переходных явлений. Если входная величина синусоидальная, статическая ошибка включает сдвиг фаз и разность амплитуд; если вход — ступенчатая функция по положению, статическая ошибка (если она вообще возникает) является ошибкой положения; если вход — ступенчатая функция по скорости (наклонная функция по положению), нужно учитывать статическую ошибку по положению, как и статическую ошибку по скорости, причем последняя может быть равна нулю, когда первая не равна нулю.

Переходная характеристика показывает, каким образом устройство достигает стационарного состояния. В частности, после подачи на вход ступенчатой функции по положению система автоматического регулирования либо будет медленно приближаться к стационарной величине, что нежелательно, либо будет приближаться к ней быстро, но при этом переходить за нее и колебаться около стационарного уровня, что также нежелательно. Критерии качества часто включают ограничения на пиковые отклонения при таком «перерегулировании», а также на время, необходимое для установления заданного значения с заданной точностью (часто равной 2% от заданного изменения).

Стандартные входы. Для оценки качества работы по этим критериям применяются определенные стандартные наборы входов. Одним из таких стандартных входов служит ступенчатая функция * по положению, при которой реакцию системы на возмущение в некотором смысле всего труднее определить. Другим стандартным входом служит ступенчатая функция по скорости.

Большое значение имеют также синусоидальные входные функции. Теорема Фурье говорит, что любая функция при довольно слабых ограничениях может быть разложена на множество синусоид, частоты которых суть целые кратные основной частоты. Далее, если элементы системы *линейны* (т. е. реакция на сумму входов равна сумме реакций на каждый вход в отдельности) и если известна реакция на все частоты, то можно предсказать реакцию на любой вход. При проектировании частотный спектр входа (относительные амплитуды синусоидальных функций всех частот) является удобным математическим описанием его характеристик.

Наконец, может представлять интерес случайно изменяющаяся функция, которая при проектировании может быть описана своей автокорреляционной функцией и распределением вероятностей своих амплитуд, связанными непосредственно с ее частотным спектром [98, 144].

Временна́я область и частотная область. Двум основным типам входов (ступенчатая функция и синусоида) соответствуют в целом два метода исследования действия входов: анализ временно́й области системы и анализ частотной области. Первый состоит в решении дифференциального уравнения системы и поэтому позволяет установить историю выходной величины во времени, или, что то же, историю расхождения, измеряемого между входом и выходом. По этой истории можно определить пиковые отклонения при перерегулировании, скорость реакции и — с некоторыми ограничениями — статические ошибки и частоту колебания в течение переходного периода.

Второй метод анализа предназначен для исследования реакции устройства на синусои-

* Иначе говоря, скачок. — *Прим. ред.*

ды при различных частотах. Результаты формулируются в терминах коэффициента усиления (отношение выходной амплитуды к входной амплитуде) и сдвига фаз (угол, на который выходная величина отстает от входной величины) как функций частоты.

Эти две области связаны между собой соотношениями, которые дают возможность по частотной характеристике указать величину отклонения при перерегулировании и скорость реакции, а по переходной характеристике — получить данные о резонансе системы на определенные входные частоты.

Анализ и синтез. Существуют два основных подхода к проектированию систем автоматического регулирования. При *анализе* проектировщику дают набор физических элементов, образующих замкнутую петлю, и просят предсказать поведение всей системы. При *синтезе* проектировщику дают набор необходимых рабочих характеристик и просят указать, какие можно выбрать физические элементы, чтобы реализовать требуемые рабочие характеристики. Желателен, конечно, второй подход, но настоящий синтетический подход в технике редко бывает возможен, и показателем разработанности любой теории является ее способность синтезировать решения поставленных задач. Нужно признать, что проектирование больших систем еще столь новое дело, что нашим единственным оружием является анализ.

Теория автоматического регулирования началась с анализа; в последние годы был разработан ряд методов, позволяющих до некоторой степени синтезировать системы. Например, при анализе, после того как определена реакция данной комбинации физических элементов, можно изменить эти элементы и включить новые так, чтобы модифицировать необходимые рабочие характеристики по временно́й и частотной характеристике, но обычно только по одной из них. В конце этой главы мы кратко разберем один из наиболее разработанных (т. е. синтетических) методов — метод корневого годографа, где эти модификации производятся таким образом, что при модификации временно́й зависимости приблизительно поддерживается требуемая частотная характеристика.

29.1. Временна́я область

Мы разберем сначала простую систему управления с разомкнутой петлей, затем — такую же систему с замкнутой петлей и, наконец, — несколько более сложную следящую систему и для нее исследуем довольно подробно дифференциальное уравнение.

Рис. 29.1. Система управления с разомкнутой петлей.

Система управления с разомкнутой петлей [95]. Рассмотрим механическое устройство (рис. 29.1), регулирующее нагрузку на вал. Вал имеет вязкое демпфирование f (вращающий момент на единицу угловой скорости) и коэффициент упругости k (вращающий момент на единицу угла). Нагрузка вызывает вращающий момент T, который мы считаем постоянным, за исключением возможного дискретного изменения в начальный момент. Выходом является угол θ_0 нагрузки. Номинальным входом является угол θ_i, который может быть функцией времени, но T также можно считать входом. Массой можно пренебречь.

Дифференциальное уравнение этой системы имеет вид

$$(\theta_i - \theta_0)\, k - f \frac{d\theta_0}{dt} - T = 0.$$

Переходная характеристика при ступенчатой входной функции по положению θ_i и нулевом моменте нагрузки равна

$$\frac{\theta_0}{\theta_i} = 1 - \exp\left(-\frac{k}{f} t\right) = \theta_0, \qquad (29.1)$$

где вход принят равным единице. *Постоянная времени* системы равна f/k *сек*. В этом случае статическая ошибка отсутствует.

Если подается ступенчатая входная функция по угловой скорости, скажем Ω *рад/сек*, то выходная скорость будет приближаться к входной скорости без статической ошибки, а выходное положение будет запаздывать и возникнет статическая ошибка положения, равная

$$\theta_i - \theta_0 = \varepsilon \to \varepsilon_{ст} = \frac{f}{k}\Omega + \frac{T}{k}.$$

Если входной угол поддерживается постоянным, скажем нулевым, и момент возрастает ступенями, то переходная характеристика системы равна

$$\theta_0 = -\frac{T}{k}\left[1 - \exp\left(-\frac{k}{f} t\right)\right] \qquad (29.2)$$

и дает, очевидно, ненулевую статическую ошибку.

Система управления с замкнутой петлей [95]. Если выход соединен обратной связью со входом и сравнивается с ним, а сигнал ошиб-

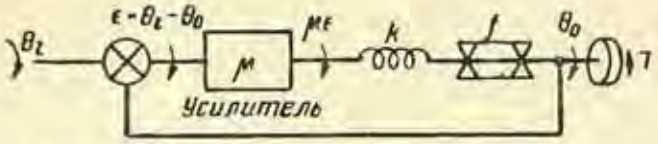


Рис. 29.2. Система управления с замкнутой петлей.

ки подается на усилитель с усилением μ, получается устройство, изображенное на рис. 29.2. Дифференциальное уравнение устройства теперь имеет вид

$$k[\mu(\theta_i-\theta_0)-\theta_0]-f\frac{d\theta_0}{dt}-T=0,$$

а уравнения соответствующих переходных характеристик имеют вид:

$$\frac{\theta_0}{\theta_i}=\frac{\mu}{1+\mu}\left\{1-\exp\left[-\frac{k(1+\mu)}{f}t\right]\right\} \quad (29.3)$$

для ступенчатого входа θ_i при $T=0$ и

$$\theta_0=\frac{-T}{k(1+\mu)}\left\{1-\exp\left[-\frac{k(1+\mu)}{f}t\right]\right\} \quad (29.4)$$

для $\theta_i=0$ и ступенчатого приращения момента.

В обоих случаях постоянная времени уменьшается теперь в $1+\mu$ раз, а скорость реакции пропорционально увеличивается. Далее, вал (соединение между входом и выходом) сделан по существу более жестким [вместо коэффициента k теперь коэффициент $(1+\mu)k$], поэтому статическая ошибка при ступенчатой входной функции по моменту уменьшается в $1+\mu$ раз.

К сожалению, статическая ошибка появляется при ступенчатой входной функции по θ_i. Ошибку можно уменьшить, увеличив усиление μ. Однако в действительности инерцией нельзя пренебрегать, и поэтому получается дифференциальное уравнение по меньшей мере 2-го порядка. В этом случае, как мы увидим, возможны колебания. Это значит, что если мы слишком увеличим усиление усилителя, чтобы увеличить скорость реакции, то мы чрезмерно увеличим колебания. Рассмотрим более подробно другую систему, чтобы исследовать влияние этих параметров.

Простая следящая система. На рис. 29.3 показана типичная следящая система с кулоновым трением (постоянным), вязким трением (пропорциональным скорости) и инерцией. Ошибка ε, равная разности между θ_i и θ_0, умножается на коэффициент усиления μ. Это произведение создает в магнитном поле двигателя вращающий момент, пропорциональный $\mu\varepsilon$, скажем момент $k\varepsilon$. Если при $t=0$ происходит внезапное изменение θ_i до величины H, то вызванные этим моменты связаны в любое мгновение соотношением

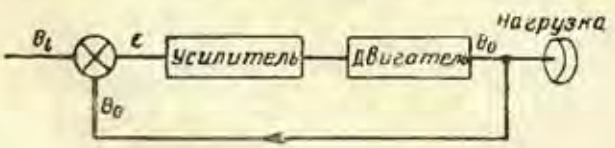


Рис. 29.3. Схема следящей системы.

$$k\varepsilon=J\frac{d^2\theta_0}{dt^2}+f\frac{d\theta_0}{dt}+T. \quad (29.5)$$

Так как выход не может следовать за входом мгновенно из-за элементов, запасающих энергию, то в любое время после $t=0$

$$\theta_0=\theta_i-\varepsilon=H-\varepsilon,$$

$$\frac{d\theta_0}{dt}=0-\frac{d\varepsilon}{dt},\quad \frac{d^2\theta_0}{dt^2}=0-\frac{d^2\varepsilon}{dt^2}.$$

Таким образом, (29.5) принимает вид

$$J\ddot{\varepsilon}+f\dot{\varepsilon}+k\varepsilon=T, \quad (29.6a)$$

где точками обозначены производные по времени.

Решение дифференциального уравнения. Это уравнение знакомо каждому инженеру; действительно, перед нами уравнение массы с пружиной, показанной на рис. 29.4. Очевидно, после внезапного изменения угла θ_i система массы с пружиной (как и рассматриваемая следящая система) при сильном затухании будет медленно подходить к своему положению равновесия, а при слабом затухании будет подходить к своему положению равновесия очень быстро, переходить через него и затем колебаться около этого положения с уменьшающейся амплитудой.

На рис. 29.5 показана ступенчатая входная функция по положению (для того и другого устройства) и три возможных типа переходных характеристик. Из этого рисунка видно, что в следящей системе может оказаться желательным слабое затухание (недодемпфирование), чтобы согласовать реакцию с командой за короткое время; с другой стороны, мо-

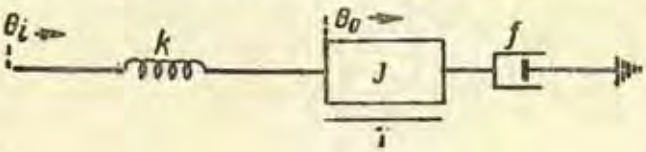


Рис. 29.4. Колеблющаяся масса с вязким трением f, кулоновым трением T, массой J и коэффициентом упругости k.

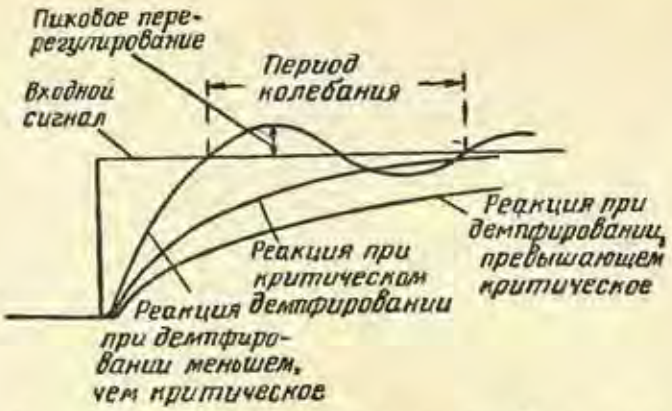


Рис. 29.5. Реакции простой следящей системы на ступенчатый вход по положению.

жет быть желательно критическое (или близкое к критическому) затухание, чтобы как можно скорее заглушить колебания. Выбор того или другого затухания зависит от обстоятельств, но очевидно, что затухание выше критического (передемпфирование) почти никогда не желательно.

Переходное решение. Для того чтобы решить уравнение (29.6), мы решаем сперва однородное уравнение, получающееся при предположении, что кулоновым трением можно пренебречь:

$$J\ddot{\varepsilon}+f\dot{\varepsilon}+k\varepsilon=0, \qquad (29.6б)$$

Решение уравнения (29.6б) имеет общий вид

$$\varepsilon=C_1e^{r_1t}+C_2e^{r_2t}, \qquad (29.7)$$

где C_1 и C_2 — произвольные константы, а r_1 и r_2 — корни характеристического уравнения

$$r=-\frac{f}{2J}\pm\sqrt{\frac{f^2}{4J^2}-\frac{k}{J}}. \qquad (29.8а)$$

Корень может быть действительным, нулевым или мнимым. Случай, когда он действительный, соответствует сильному затуханию и представляет для нас мало интереса. Поэтому мы примем, что корень мнимый, и перепишем (29.8а) в следующем виде:

$$r=-\frac{f}{2J}\pm j\sqrt{\frac{k}{J}-\frac{f^2}{4J^2}}=-\alpha\pm j\beta, \qquad (29.8б)$$

где* $j=\sqrt{-1}$, а α и β — положительные действительные числа.

---

* Мы используем в этой главе обозначение j, а в гл. 7 обозначение i, согласно обычаю, установившемуся соответственно среди инженеров-электриков и среди математиков. — *Прим. авт.*

Используя тождество для комплексных чисел

$$e^{j\omega t}=\cos\omega t+j\sin\omega t, \qquad (29.9)$$

мы можем переписать (29.7) в таком виде:

$$\varepsilon=e^{-\alpha t}(C_3\cos\beta t+C_4\sin\beta t). \qquad (29.10)$$

Мы видим, что это затухающее синусоидальное колебание. Поскольку ε — действительное число, все величины в (29.10) также должны быть действительными, если β — действительное [если β — мнимое, то будут действительными все величины в (29.7) и мы можем применять эту последнюю форму].

Интересно исследовать случай критического затухания. В этом случае корень в (29.8) равен нулю и

$$\frac{k}{J}=\frac{f_{кр}^2}{4J^2},\quad f_{кр}=\sqrt{4Jk}.$$

Удобно ввести безразмерное отношение

$$\xi=\frac{f}{f_C}=\frac{f}{2\sqrt{Jk}},$$

которое называется *декрементом затухания*. В интересующих нас случаях слабого затухания декремент затухания меньше единицы.

Интересно также сравнить фактическую частоту β с собственной частотой

$$\omega_n=\sqrt{\frac{k}{J}},$$

равной частоте колебания при отсутствии затухания. Фактическая частота равна

$$\beta=\sqrt{\omega_n^2-\frac{f^2}{4J^2}}=\omega_n\sqrt{1-\xi^2}.$$

Константа α определяется через эти параметры соотношением

$$\alpha=\frac{f}{2J}=\xi\omega_n.$$

Итак, решение дифференциального уравнения можно записать в виде

$$\varepsilon=e^{-\xi\omega_nt}[C_3\cos(\omega_n\sqrt{1-\xi^2})t+$$
$$+C_4\sin(\omega_n\sqrt{1-\xi^2})t].$$

Константы определяются из начальных условий в момент $t=0+$, т. е. в момент $t=0$, но после того, как была подана ступенчатая функция. В это время ошибка равна начальному возмущению H. Подставляя значения

в (29.10) при $t=0$ и $\varepsilon=H$, получаем $C_3=H$. Поскольку нагрузка имеет инерцию, ее ускорение не может быть мгновенным, так что $\dot{\varepsilon}=0$ в момент $t=0+$. Дифференцируя (29.10) и подставляя эти значения, получаем

$$C_4=H\alpha/\beta=H\xi/\sqrt{1-\xi^2}.$$

Синусоидальную и косинусоидальную функции можно соединить вместе, и мы получим

$$\frac{\varepsilon}{H}=\frac{e^{-\xi\omega_n t}}{\sqrt{1-\xi^2}}\sin(\omega_n\sqrt{1-\xi^2}t+\varphi), \quad (29.11)$$

где

$$\varphi=\operatorname{arctg}\sqrt{\frac{1-\xi^2}{\xi^2}}$$

и ошибка выражена по отношению к возмущению H.

Полное решение. Соотношение (29.11) есть решение однородного уравнения (29.6б). Решение уравнения (29.6а) состоит из переходной и стационарной части. Чтобы избежать нелинейностей, усложняющих решение, мы примем, что T имеет одно направление. Переходное решение — такое же, как и (29.11), с той разницей, что при вычислении констант C_3 и C_4 нужно подставить $(H-T/k)$ вместо H. Стационарное решение получается при подстановке в (29.6б) $\ddot{\varepsilon}=0$ и $\dot{\varepsilon}=0$, так как, по определению, ошибка не меняется по достижении стационарного состояния. Этим определяется константа

$$\varepsilon_{ст}=\frac{T}{k},$$

которую нужно прибавить к решению. Итак, полное решение имеет вид

$$\varepsilon=\frac{H-T/k}{\sqrt{1-\xi^2}}e^{-\xi\omega_n t}\sin(\omega_n\sqrt{1-\xi^2}t+\varphi)+\frac{T}{k}. \quad (29.12)$$

Следовательно, система устойчива и при ступенчатой входной функции по положению дает стационарную ошибку, зависящую только от кулонова трения, которую можно уменьшить, увеличивая усиление k. Переходный процесс представляет собой затухающее синусоидальное колебание с отставанием по фазе, зависящим от ξ, с частотой, зависящей от k, J и f (и приближающейся к собственной частоте ω_n, когда f приближается к нулю), и наложенной убывающей экспоненциальной огибающей, определяемой произведением $\xi\omega_n$.

Отсюда оцениваются изменения усиления

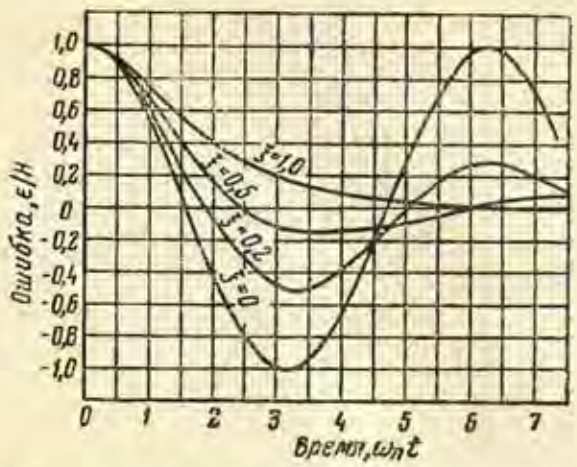


Рис. 29.6. Кривые безразмерной ошибки для следящей системы с рис. 29.3 (по Кейджу [96]).

усилителя, коэффициента вязкого трения и момента инерции двигателя и нагрузки. На рис. 29.6 показано влияние этих параметров; ошибка измеряется в единицах возмущения, а время измеряется в единицах, обратных собственной частоте. При помощи таких кривых можно исследовать скорость реакции и пиковое отклонение при перерегулировании.

Вход по скорости. Если входное возмущение представляет собой ступенчатую функцию по скорости, а не по положению, то

$$\frac{d\theta_0}{dt}=\frac{d\theta_i}{dt}-\frac{d\varepsilon}{dt}=\Omega-\dot{\varepsilon}$$

и

$$\frac{d^2\theta_0}{dt^2}=\frac{d^2\theta_i}{dt^2}-\frac{d^2\varepsilon}{dt^2}=0-\ddot{\varepsilon}=-\ddot{\varepsilon}$$

и вместо (29.6) будет

$$J\ddot{\varepsilon}+f\dot{\varepsilon}+k\varepsilon=f\Omega+T. \quad (29.13)$$

Решение этого дифференциального уравнения показывает, что скорость будет колебаться около заданного значения с уменьшающейся амплитудой, подобно тому как показано на рис. 29.6. Мы можем определить статическую ошибку по положению, подставив в (29.13) $\ddot{\varepsilon}=0$ и $\dddot{\varepsilon}=0$, что дает

$$\varepsilon_{ст}=\frac{f_\Omega+T}{k}.$$

Итак, возможна заметная статическая ошибка даже в том случае, если скорость регулируется идеально и кулоновым трением можно пренебречь. Для устранения ошибки необходимо ввести в управляющую систему какое-нибудь устройство. Например, можно подавать обратно в усилитель интеграл по

времени от ошибки. Тогда ошибка должна свестись к нулю, так как в противном случае интеграл станет в конце концов бесконечным. Однако для быстрого уменьшения ошибки постоянная интегратора должна быть большой, и этот сигнал будет также усиливать колебания и может даже вызвать неустойчивое состояние. Поэтому нужно будет дать дополнительное демпфирование. Но увеличение демпфирования означает просто добавление постоянной величины к члену с $d\varepsilon/dt$ и может быть достигнуто путем подачи обратно на вход производной сигнала по времени вместо того, чтобы ставить реальный демпфер. Эти и другие сложные системы описываются дифференциальными уравнениями высокого порядка, и решать их было бы очень утомительно. Мы обратимся теперь к другому методу анализа систем автоматического регулирования.

29.2. Частотная область

Как было сказано раньше, существует тесная связь между временно́й и частотной областью. Рассмотрим функцию $f(t)$ из временно́й области, определяющую в явном виде значение функции f для любого момента t. Вообще говоря, существует единственное представление этой функции в частотной области, которое мы обозначим через $F(s)$ и которое определяет значение функции F для любой частоты s (вообще говоря, s — комплексное число). Существует простая математическая операция для преобразования $f(t)$ в $F(s)$, называемая *преобразованием Лапласа*.

Преобразование Лапласа можно применить к дифференциальному уравнению по t, и полученное уравнение по s после чисто алгебраических (и, следовательно, простых) манипуляций дает выражение, которое может быть преобразовано обратно в выражение по t, являющееся решением дифференциального уравнения. Поэтому мы сначала исследуем преобразование Лапласа, как полезное орудие решения дифференциальных уравнений, и попутно обнаружим некоторые интересные вещи относительно рассматриваемой следящей системы, которые позволят нам выполнить анализ частотной области.

Преобразование Лапласа. Лапласов образ (или изображение) функции $f(t)$ указанного типа определяется как

$$L[f(t)]=F(s)=\int_0^\infty f(t)\,e^{-st}dt, \quad (29.14)$$

где L обозначает преобразование Лапласа*, F есть функциональная форма алгебраического выражения, а s — комплексная переменная, равная

$$s=\sigma+j\omega. \quad (29.15)$$

Следующие равенства справедливы при $t\geqslant 0$:

$$L[af(t)]=aL[f(t)], \quad (29.16)$$

$$L(f_1+f_2)=L(f_1)+L(f_2), \quad (29.17)$$

$$L\left[\frac{df(t)}{dt}\right]=sL[f(t)]-f(0+), \quad (29.18)$$

$$L\left[\int f(t)\,dt\right]=\frac{L[f(t)]}{s}+\frac{1}{s}\int f(t)\,dt\Big|_{t=0+}, \quad (29.19)$$

$$L\text{ [ступенчатая функция } f(t)=1]=\frac{1}{s}, \quad (29.20)$$

$$L(e^{-\alpha t})=\frac{1}{s+\alpha}, \quad (29.21)$$

$$L(\sin\beta t)=\frac{\beta}{s^2+\beta^2}, \quad (29.22)$$

$$L(\cos\beta t)=\frac{s}{s^2+\beta^2}, \quad (29.23)$$

$$L(t)=\frac{1}{s^2}. \quad (29.24)$$

Преобразование Лапласа для дифференциального уравнения. Пусть нам нужно решить дифференциальное уравнение (29.6б), которое можно записать в виде

$$\ddot{\varepsilon}+\frac{f}{J}\dot{\varepsilon}+\frac{k}{J}\varepsilon=0,$$

с начальными условиями $\varepsilon=H$, $\dot{\varepsilon}=0$ при $t=0$.

Выполняя преобразование обеих частей по формулам (29.16) и (29.17), получаем

$$L(\ddot{\varepsilon})+\frac{f}{J}L(\dot{\varepsilon})+\frac{k}{J}L(\varepsilon)=0. \quad (29.25)$$

Ввиду (29.18)

$$L(\dot{\varepsilon})=sL(\varepsilon)-\varepsilon(0+).$$

Применяя опять (29.18), получаем

$$L(\ddot{\varepsilon})=sL(\dot{\varepsilon})-\dot{\varepsilon}(0+)=s[sL(\varepsilon)-\varepsilon(0+)]-\dot{\varepsilon}(0+)=s^2L(\varepsilon)-s\varepsilon(0+)-\dot{\varepsilon}(0+).$$

* Стало быть, лапласов образ (или изображение) функции есть другая функция, получаемая из первой преобразованием Лапласа. Обратно, первая функция есть лапласов прообраз (или оригинал) второй функции. — *Прим. ред.*

Начальные условия указывают, что $\varepsilon(0+)=H$ и $\dot{\varepsilon}(0+)=0$.

Таким образом, из (29.25) получаем

$$s^2L(\varepsilon)-sH+\frac{f}{J}sL(\varepsilon)-\frac{f}{J}H+\frac{k}{J}L(\varepsilon)=0.$$

Решая относительно $L(\varepsilon)$, находим

$$L(\varepsilon)=\frac{H(s+f/J)}{s^2+(f/J)s+k/J}. \qquad (29.26)$$

Этим заканчиваются манипуляции над алгебраическим выражением. Для окончательного решения дифференциального уравнения остается только преобразовать обратно функцию от s (т. е. найти соответствующую функцию от t). Знаменатель выражения (29.26) есть многочлен по s, который можно записать как произведение множителей, линейных относительно s и содержащих корни многочлена; таким образом, правую часть уравнения (29.26) можно разложить на дроби, знаменателями которых будут эти множители. Тогда для обратного преобразования нужно лишь использовать (29.21).

Разложение на дроби должно иметь вид

$$\frac{L(\varepsilon)}{H}=\frac{A}{s-r_1}+\frac{B}{s-r_2}, \qquad (29.27)$$

где корни r_1 и r_2 определяются точно так же, как и раньше, — формулой (29.8б). Деля выражение (29.26) на H и полагая его равным (29.27), получаем

$$\frac{L(\varepsilon)}{H}=\frac{s+2\alpha}{(s-r_1)(s-r_2)}=\frac{A}{(s-r_1)}+\frac{B}{(s-r_2)},$$

или

$$s+2\alpha=A(s-r_2)+B(s-r_1)=$$
$$=s(A+B)+A(\alpha+j\beta)+B(\alpha-j\beta). \qquad (29.28)$$

Поскольку (29.28) есть тождество по s, то коэффициенты при s в обеих частях и постоянные члены должны быть одинаковы. Это дает уравнения, которые можно решить относительно A и B. Подставляя полученные значения в (29.27), находим

$$\frac{L(\varepsilon)}{H}=\frac{(\alpha+j\beta)/2j\beta}{s-r_1}-\frac{(\alpha-j\beta)/2j\beta}{s-r_2}. \qquad (29.29)$$

Производим обратное преобразование обеих частей по формуле (29.21):

$$\frac{\varepsilon}{H}=\frac{1}{2j\beta}[(j\beta+\alpha)e^{r_1t}+(j\beta-\alpha)e^{r_2t}]=$$
$$=e^{-\alpha t}\frac{(j\beta+\alpha)e^{j\beta t}+(j\beta-\alpha)e^{-j\beta t}}{2j\beta}.$$

Теперь с помощью (29.9) можно получить (29.11).

Частотная интерпретация корней. Преобразование Лапласа позволяет найти единый подход к решению линейных уравнений высших порядков. Кроме того, выражение (29.26) от s имеет физическое истолкование.

Для линейных уравнений с постоянными коэффициентами знаменатель функции $(L\varepsilon)$ всегда является многочленом по s, и ее можно разложить на дроби. Корни в знаменателях этих дробей могут быть действительными или комплексными, кратными или одиночными. Они образуют в решении члены вида

$$(c_qt^q+c_{q-1}t^{q-1}+\ldots+c_0)e^{rt},$$

где r — корень, действительный или комплексный, а q — кратность корня.

Эти корни соответствуют разным типам колебаний, причем входная величина может возбудить колебание любого из этих типов. Если действительная часть корня r отрицательна и колебание соответствующего типа вызвано возмущением, то оно будет постепенно затухать вследствие гасящего действия действительной части, так как показательная функция уменьшается быстрее, чем возрастает любая степень переменной t. Поэтому система является устойчивой. Если действительная часть r равна нулю, а мнимая не равна нулю, то входная величина вызовет возрастающее колебание данного типа. Если корень r равен нулю и не является кратным, то он входит в решение как постоянное число, так что получается колебание постоянной амплитуды; если же корень кратный, то коэффициент с t постепенно возрастает.

На рис. 29.7 два корня уравнения (29.8) представлены на комплексной плоскости (действительная часть отложена параллельно горизонтальной оси, мнимая часть — параллельно вертикальной оси). Корень, лежащий в правой полуплоскости или на вертикальной (мни-

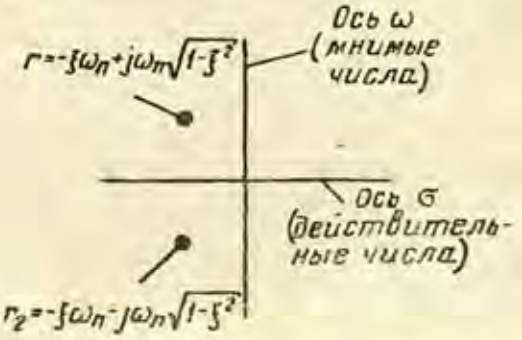


Рис. 29.7. График корней в плоскости s ($s=\sigma+j\omega$).

мой) оси, может вызвать неустойчивость, как и кратные корни, лежащие в начале координат. Корни, изображенные на рисунке, не могут привести к неустойчивости. Ниже мы рассмотрим график такого вида подробнее, но сначала обратимся к другой интересной интерпретации уравнения (29.26).

Передаточная функция. Уравнение (29.26) определяет $L(\varepsilon)$ для ступенчатой функции, у которой лапласов образ ввиду (29.20) и (29.16) равен H/s. Но ввиду (29.17)

$$L(\varepsilon)=L(\theta_i)-L(\theta_0).$$

Отсюда

$$\frac{F_0(s)}{F_i(s)}=\frac{L[\theta_0(t)]}{L[\theta_i(t)]}=\frac{L(\theta_i-L(\varepsilon)}{L(\theta_i)}=\frac{H/s-L(\varepsilon)}{H/s}=$$

$$=1-\frac{s}{H}L(\varepsilon)=1-\frac{s^2+(f/J)s}{s^2+(f/J)s+k/J}=$$

$$=\frac{k/J}{s^2+(f/J)s+k/J}=KG'(s), \quad (29.30)$$

где K — константа, называемая *усилением* или *коэффициентом чувствительности*, а $G'(s)$ — функция комплексной частоты $s=\sigma+j\omega$. Функция $KG'(s)$ называется *передаточной функцией системы*, она не зависит от входной величины и зависит лишь от параметров системы.

Если синусоидальная входная величина подается на следящую систему, описываемую линейным дифференциальным уравнением с постоянными коэффициентами, то выходная величина после затухания всех переходных колебаний и установления стационарного состояния будет также синусоидальной. Это синусоидальное колебание может отличаться от входной синусоиды по величине и фазе, но будет всегда иметь ту же частоту. Оказывается, что для этого стационарного выходного синусоидального колебания в (29.30) исчезают члены с σ. Поэтому передаточная функция системы может быть записана как функция от действительной частоты ω. Итак, вместо (29.30) получается

$$\frac{L[\theta_0(t)]}{L[\theta_i(t)]}=KG'(j\omega)=\frac{k/J}{(j\omega)^2+(f/J)j\omega+k/J}. \quad (29.31)$$

Таким образом, выходная величина следящей системы зависит только от входной частоты, и эта зависимость описывается передаточной функцией системы $KG'(j\omega)$. Анализ частотной области систем автоматического регулирования почти всегда производится с помощью удобного понятия передаточной функции. Например, если два элемента системы соединены последовательно, то общая передаточная функция равна произведению их передаточных функций, так как входной величиной второго элемента является выходная величина первого.

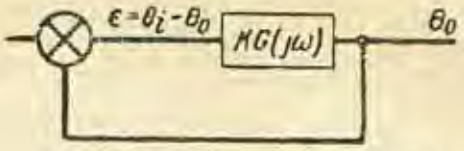


Рис. 29.8. Система с замкнутой петлей.

На рис. 29.8 представлен другой пример. Здесь устройство с передаточной функцией $KG(j\omega)$ служит элементом системы с замкнутой петлей, усиление которой (помимо усиления этого элемента) равно единице. Тогда

$$\frac{\theta_0(s)}{\varepsilon(s)}=KG(j\omega), \quad \varepsilon(s)=\theta_i(s)-\theta_0(s)$$

и, следовательно,

$$\frac{\theta_0(s)}{\theta_i(s)-\theta_0(s)}=KG(j\omega), \quad (29.32)$$

откуда

$$\frac{\theta_0(s)}{\theta_i(s)}=\frac{KG(j\omega)}{1+KG(j\omega)}=KG'(j\omega). \quad (29.33)$$

Рассматривая систему на рис. 29.8 как самостоятельную, $KG'(j\omega)$ называют *передаточной функцией системы* или *частотной характеристикой системы*, а функцию $KG(j\omega)$, которая выражает связь выходной величины с ошибкой, называют *передаточной функцией петли* *. Передаточная функция системы может являться частью общей передаточной функции более широкой системы, элементом которой является данная система.

Полярная диаграмма. Выражения, подобные (29.31) и (29.33), суть комплексные функции действительной переменной; иначе говоря, они содержат действительную и комплексную части, и поэтому их можно представить кривыми в прямоугольной системе координат, изображающей комплексную плоскость, как на рис. 29.7. Когда ω принимает значения от нуля до бесконечности, каждое значение передаточной функции системы или петли соответствует точке в комплексной области и получаются геометрические места точек, подобные кривым на рис. 29.9.

Ввиду тождества (29.9) комплексную передаточную функцию можно также записать

* В русской технической литературе передаточная функция системы в этом смысле называется обычно *передаточной функцией замкнутой системы*, а передаточная функция петли — *передаточной функцией разомкнутой системы*. — *Прим. ред.*

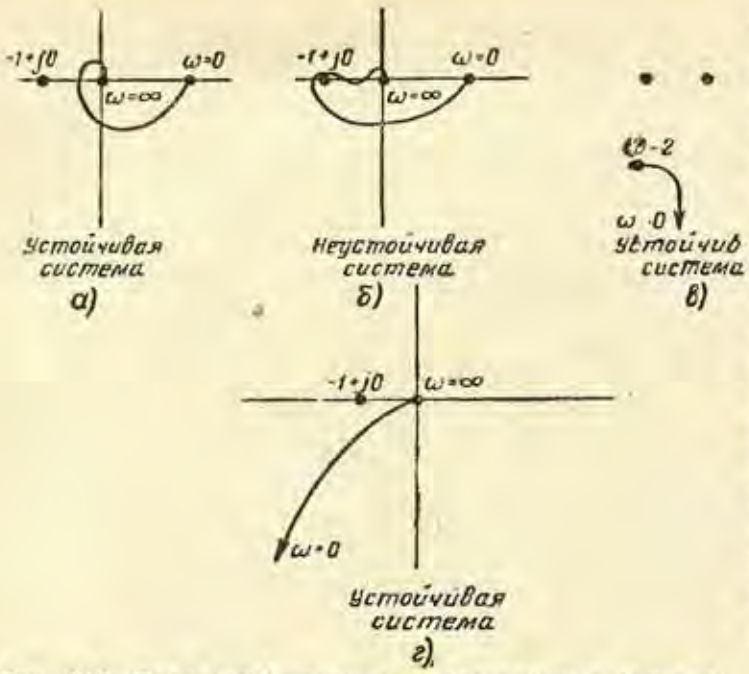


Рис. 29.9. Полярные диаграммы передаточных функций в плоскости $KG(j\omega)$ (по Талеру [94]).

как векторное выражение, характеризуемое длиной и углом. Если вход есть единичная синусоида, то длина этого вектора равна амплитуде выходной синусоиды, а угол равен фазовому отставанию выходной синусоиды. Следовательно, графики на рис. 29.9 можно истолковать как годограф этого вектора в полярных координатах (длину — как радиус-вектор, фазу — как полярный угол); такие годографы обычно называются *полярными диаграммами* *.

Эквивалентность временной и частотной области содержится в выражении для передаточной функции, т. е. в выражении для всех частот, а не для какой-либо одной частоты. Поэтому значение передаточной функции для одной частоты определяет амплитуду и сдвиг фаз выходной величины для этой частоты, а сама передаточная функция, относящаяся ко всем частотам, содержит в неявном виде переходную характеристику и всю другую информацию, получаемую из решения во временной области. Для раскрытия этой информации обычно чертят график передаточной функции петли и исследуют его форму. Наиболее распространенным типом графика является полярная диаграмма.

На каждой диаграмме, приведенной на рис. 29.9, отмечена точка $-1+j0$; это та точка комплексной плоскости, у которой действительная часть равна —1, а мнимая часть равна 0; это также та точка, для которой длина вектора равна 1, а фаза равна 180°. Если $KG=-1+j0$, то $KG'=\infty$. В силу важной теоремы, известной под названием *критерия Найквиста*, устойчивость системы всегда можно определить по положению этой точки относительно графика передаточной функции петли **.

В упрощенной форме теорема гласит, что если проходить полярную диаграмму передаточной функции петли от $\omega=0$ до $\omega=\infty$, то система устойчива, когда точка $-1+j0$ находится слева (рис. 29.9,*а, в* и *г*), и неустойчива, когда точка находится справа (рис. 29.9,*б*). Далее, если система устойчива, но кривая проходит вблизи точки $-1+j0$, то система будет близка к неустойчивости, т. е. отклонения перерегулирования будут очень большие. Вследствие широкого применения критерия Найквиста полярная диаграмма часто называется *диаграммой Найквиста*.

Общий вид полярной диаграммы иногда можно выяснить и без подробного вычисления. Так, рис. 29.9,*г* есть полярная диаграмма передаточной функции петли, соответствующей передаточной функции системы (29.31). В то время как функция петли становится бесконечно большой при $\omega=0$, передаточная функция системы конечна при всех значениях ω. В этом случае очевидно, что кривая никогда не может пересечь действительную ось (так как мнимая часть не может быть равна нулю, за исключением случая, когда $\omega\to\infty$), и поэтому система будет устойчивой для всех значений параметров f, J и k. В других случаях, как показано на рисунке, кривая обычно пересекает отрицательную действительную ось один или несколько раз и систему можно сделать устойчивой или неустойчивой, изменяя параметры, определяющие точки пересечения.

Другие графики. Итак, в полярной диаграмме передаточной функции петли длина вектора, проведенного из начала координат в какую-нибудь точку кривой, равна абсолютной величине передаточной функции петли KG; аналогично, длина вектора, проведенного из точки $-1+j0$ в точку кривой, равна абсолютной величине выражения $1+KJ$, а отношение этих длин равно абсолютной величине передаточной функции системы (29.33). Далее, угол между этими векторами равен фазовому запаздыванию системы. Можно показать, что семейство кривых, для которых

$$\frac{KG}{1+KG}=M\,(\text{const}),$$

* В русской технической литературе их принято называть *амплитудно-фазовыми характеристиками*. — *Прим. пер.*

** Критерий Найквиста, сформулированный американским инженером Х. Найквистом для регенеративных радиоустройств в 1932 г., впервые был применен для исследования систем автоматического регулирования русским ученым А. В. Михайловым [Д. 30]. — *Прим. ред.*

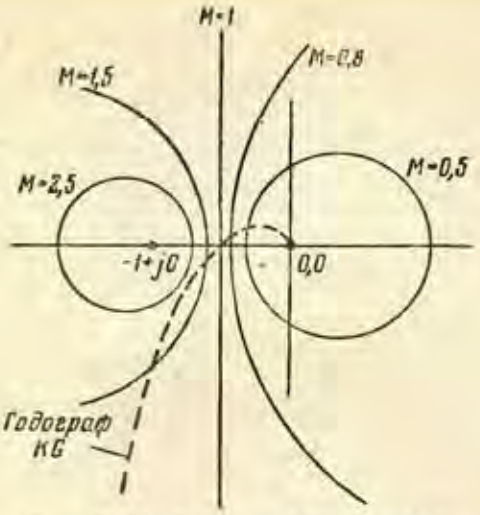


Рис. 29.10. Кривые $KG/(1+KG)=M$ (const) в полярной плоскости (по Талеру [94]).

состоит из окружностей на полярной диаграмме, как показано на рис. 29.10.

По значениям M в точках пересечения окружностей с годографом KG можно определить частотную характеристику. Максимум

Рис. 29.11. Действительная частотная характеристика, полученная по рис. 29.10 (по Талеру [94]).

получается при наибольшем встреченном значении M. Так, по указанной диаграмме можно найти *действительную частотную характеристику*, подобную изображенной на рис. 29.11. Конечно, действительная частотная характеристика представляет собой важный показатель выполнения проектных требований.

Рис. 29.12. График логарифмической амплитуды и фазового угла против частоты (по Талеру [94]).

Есть и другие общепринятые методы вычерчивания частотной характеристики для получения нужной информации. Один из наиболее обычных способов — это график зависимости логарифмической амплитуды и фазового угла от частоты, изображенный на рис. 29.12. Существуют методы [94] достаточно надежной аппроксимации этих кривых.

29.3. Другие вопросы теории авторегулирования

Этим заканчивается наше изложение простого (т. е. аналитического) подхода к простым следящим системам (т. е. системам, подчиняющимся линейным дифференциальным уравнениям с постоянными коэффициентами) при простых (т. е. непрерывных и свободных от шума) входных условиях. Теперь мы кратко разберем несколько более трудных вопросов.

Синтез. Основная цель синтеза—перевести заданные рабочие характеристики в физические данные. При этом исходят, как и при проектировании больших систем, из формулировки задачи: из заданных характеристик нагрузки и требуемых рабочих характеристик и во многих случаях из командного сигнала. Задача заключается в том, чтобы выбрать конкретные устройства, из которых будет состоять система управления, измерить или вычислить ее рабочие характеристики и затем изменить систему так, чтобы удовлетворить этим требованиям. В характеристиках могут быть заданы: условия стационарной нагрузки с допусками, скорость реакции с ее верхним пределом, максимальное отклонение при перерегулировании и, возможно, допустимое число отклонений перерегулирования.

Элементарный метод состоит в том, чтобы выбрать привод (скажем, электрический двигатель), способ измерения выхода (скажем, тахометр), устройство для регулирования мощности (скажем, электронный усилитель) и способ обнаружения ошибки (скажем, уравновешенную потенциометрическую схему). Главная цель этого предварительного решения — составить грубую оценку статической ошибки и определить желаемую скорость реакции. При этом учитывается перерегулирование, резонансная частота и качественно определяется стационарное состояние и скорость реакции.

В системе с одной замкнутой петлей можно менять усиление K передаточной функции. Это можно делать путем изменения усиления усилителя, что увеличивает всю диаграмму, как показано на рис. 29.13. Кроме того, можно включить компенсирующие че-

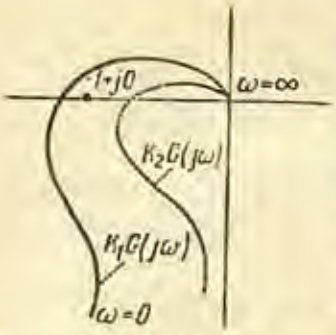


Рис. 29.13. Изменение усиления с целью сместить полярную диаграмму передаточной функции.

тырехполюсники или фильтры и тем самым деформировать диаграмму, как показано на рис. 29.14. Рисунок показывает результат применения фильтра верхних частот для уменьшения фазового сдвига в интересующем нас диапазоне частот. За фазовое опережение, вводимое в диапазоне нижних частот, расплачиваются усилением.

График можно, кроме того, изменить посредством цепей обратной связи, которые вводятся в добавление к последовательным элементам или вместо них. Выбор тех или других зависит от многих факторов, включая стоимость, физическую осуществимость, спектр шума и действие возмущений нагрузки.

Усовершенствованный метод синтеза при помощи корневого годографа. Без глубокого исследования можно было бы предположить, что по заданным параметрам сравнительно легко определить заранее частотную и переходную характеристики следящей системы. Недостаток рассмотренных методов в том, что в них не рассматривается сразу вся информация о системе. При попытке применить такой подход и определить влияние изменения одного параметра системы возникают столь большие трудности, что это оказывается неосуществимым. С другой стороны, вычисление по множеству точек требует очень много времени.

Были предприняты успешные попытки разработать методы синтеза, учитывающие сразу и частотную, и временну́ю область. Одним из этих методов является так называемый *метод корневого годографа.* В нем применяются передаточные функции, в которых переходные явления, обусловленные затуханием σ, не игнорируются; т. е. вместо мнимой переменной $j\omega$ в передаточной функции для разомкнутой и замкнутой петель берется комплексная переменная s. Таким образом, выражение (29.33), связывающее

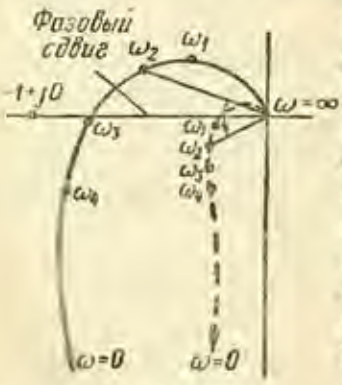


Рис. 29.14. Применение фильтра для деформации полярной диаграммы передаточной функции (по Талеру [94]).

Рис. 29.15. Полюсы передаточной функции петли.

передаточную функцию системы и передаточную функцию петли, заменяется выражением

$$\frac{\theta_o}{\theta_i}(s)=\frac{KG(s)}{1+KG(s)}=KG'(s). \quad (29.34)$$

Функция (29.34) есть комплексная функция комплексной переменной, т. е. для каждого значения s, состоящего из действительной и мнимой части, эта функция принимает значение, состоящее из действительной и мнимой части.

Нужно отметить, что $KG(s)$ также есть комплексное число и, кроме того, оно состоит из комплексного многочлена по s в числителе и аналогичного комплексного многочлена в знаменателе. Если мы обозначим числитель через $q(s)$, а знаменатель — через $p(s)$, то вместо (29.34) получается

$$KG'(s)=\frac{q(s)}{p(s)+q(s)}.$$

Это выражение становится бесконечным, когда знаменатель равен нулю, т. е. когда

$$KG(s)+1=0.$$

Если мы возьмем, например, выражение

$$KG(s)=\frac{K}{s(s+1)(s+2)}, \quad (29.35)$$

то мы можем определить нули знаменателя в плоскости s, как показано на рис. 29.15.

Очевидно, всякому значению s соответствует значение передаточной функции системы, но нас в данный момент интересуют только те значения s, при которых передаточная функция системы становится бесконечно большой. Это будет зависеть не только от K, но и от $G(s)$; иначе говоря, $KG'(s)$ будет бесконечно большим, когда $KG(s)$ равно —1. Если K очень велико, то для того, чтобы произведение было равно —1, $G(s)$ должно быть близко к нулю; следовательно, при больших значениях K корни знаменателя передаточной функции системы должны находиться в нулях функции $G(s)$. Беря K как пара-

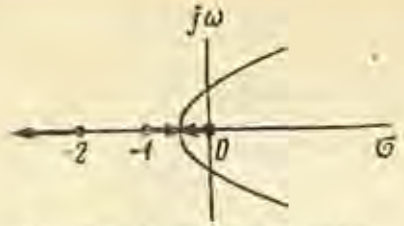


Рис. 29.16. Годограф полюсов передаточных функций системы в плоскости s.

метр, мы можем теперь начертить в плоскости s геометрические места всех значений s, при которых передаточная функция системы становится бесконечно большой. Мы получаем диаграмму, подобную изображенной на рис. 29.16.

Напомним, что знаменатель передаточной функции системы есть также характеристическое уравнение для дифференциального уравнения, описывающего систему. Кривая на рис. 29.16 имеет параметр K, представляющий усиление системы. Известно, что система неустойчива, если корни лежат в правой полуплоскости. Далее, соотношения между действительными и мнимыми частями комплексных корней определяют, будут ли эти корни затухать. Таким образом, мы определяем устойчивость и переходный процесс по тому, в какой области плоскости находятся корни; с другой стороны, частотную характеристику можно определить путем подстановки корней в (29.35).

Наконец, если в годографе отсутствуют корни, обеспечивающие требуемое поведение системы, то нужно изменить самый годограф путем введения дополнительных полюсов (точек, в которых функция обращается в бесконечность) и нулей (точек, в которых функция обращается в нуль). В [97] описаны соотношения между формой годографа и положением этих полюсов и нулей.

Нелинейные системы автоматического регулирования. Мы определили линейную систему автоматического регулирования (линейную следящую систему) как систему, подчиняющуюся принципу суперпозиции: сумма реакций на отдельные входные величины равна реакции на сумму этих входных величин. Другое, равносильное определение гласит: линейная система автоматического регулирования есть система, описываемая линейным дифференциальным уравнением. Мы разбирали только уравнения с постоянными коэффициентами — единственные линейные уравнения, имеющие практическое значение для систем автоматического регулирования. Однако нелинейные явления часто играют значительную роль.

Во-первых, всякая система автоматического регулирования может быть сделана нелинейной просто путем увеличения амплитуды входной величины до тех пор, пока в каком-нибудь элементе системы не наступит насыщение. Такие явления насыщения часто играют важную роль, так как требование, чтобы система автоматического регулирования никогда не входила в нелинейный режим, как правило влечет за собой необходимость применять слишком дорогие компоненты.

Во-вторых, некоторым параметрам системы, таким, как начальное трение и мертвый ход, присущи нелинейные характеристики, которые трудно устранить при изготовлении и нужно учитывать при анализе.

В-третьих, желаемую рабочую характеристику иногда лучше всего осуществить намеренно вводя нелинейный элемент, при условии, что проектировщик способен анализировать получающуюся нелинейную систему. Крайний пример такой намеренной нелинейности — регулирование включением и выключением (релейное регулирование), как в термостате в домашней отопительной системе или в «контакторе» во многих сервомоторах, который доводит мотор до полного вращающего момента, когда ошибка превышает некоторую величину, но не вводит поправки, когда ошибка меньше этого значения.

Трудности анализа нелинейных следящих систем аналогичны тем, с которыми мы вообще сталкиваемся при попытке анализировать другие нелинейные системы. Временной анализ здесь неприменим, потому что в этом случае нет стандартных входных величин. Поскольку принцип суперпозиции не соблюдается, система может быть, например, устойчивой для входной величины одного типа и неустойчивой для входной величины другого типа. Но частотный анализ также неприменим, потому что нельзя дать точного определения передаточной функции; в частности, синусоидальная входная величина не обязательно вызовет синусоидальную выходную величину.

Для этой задачи имеются два основных аналитических подхода. Если нелинейные явления можно считать «медленными», т. е. их воздействия медленны по сравнению с постоянной времени системы, то можно провести обычный анализ путем исследования полюсов и нулей, медленно перемещающихся в плоскости s; если нелинейные явления «быстрые», то их часто можно приближенно представить медленными нелинейными явлениями с помощью методов *анализа описывающей функ-*

*ции**. Таким образом, можно применять типичные линейные методы.

Второй аналитический метод, который действительно применим при нелинейном дифференциальном уравнении, — это *анализ фазовой плоскости*, когда нелинейные дифференциальные уравнения 2-го порядка любой степени изучаются посредством диаграмм в фазовой плоскости, в которых одной координатной осью является выходной сигнал, а другой осью — его производная.

Хотя эти аналитические методы очень мощны, с их помощью можно исследовать надлежащим образом лишь небольшую часть практических нелинейных явлений. Однако в том случае, когда нет хорошего аналитического метода, всегда можно прибегнуть к моделированию. Это самое универсальное орудие исследования нелинейных следящих систем.

Дискретизированные входы и шумные входы. Во многих случаях сама команда поступает не в непрерывной форме; по той или иной причине, из команды время от времени могут отбираться дискреты, и на вход системы поступают лишь эти выборочные данные. Если эта дискретность появляется внутри системы автоматического регулирования (например, если сигнал ошибки подается обратно не непрерывно, а в виде дискрет), то получается нелинейная система; однако такие дискретизированные системы часто бывают линейными. Тогда основная задача проектирования состоит в том, чтобы заставить дискретизированную систему автоматического регулирования вести себя по возможности так же, как система автоматического регулирования, получающая непрерывный командный сигнал.

Это вызывает два вопроса: каково лучшее приближение к непрерывному командному сигналу и какое придумать управление, чтобы оно соответствовало непрерывному командному сигналу? На первый вопрос можно ответить на основе статистических соображений (подобных рассмотренным в теории информации, см. § 28.3), а на второй — на основе соображений, изложеных в этой главе. Иногда эти вопросы неразделимы. Если при проектировании выбирают прерывистость и хотят отбирать дискреты как можно реже, то задача состоит в определении наименьшего количества данных, обеспечивающих требуемые рабочие характеристики.

Если на командный сигнал накладывается шум, то справедливы рассуждения, аналогичные применяемым в теории информации. Сначала нужно описать шумовую и командную функции. При этом нужно пересмотреть критерии работы системы, так как система автоматического регулирования, реагирующая на всякую входную величину с малым запаздыванием или ошибкой, будет слишком сильно реагировать на шум. Поэтому должны быть предусмотрены фильтры для уменьшения шума без ухудшения сигнала.

Вообще говоря, высокочастотные составляющие шума сравнительно легко отфильтровать (по существу сама система автоматического регулирования действует как фильтр нижних частот: передаточная функция всегда стремится к нулю при бесконечной частоте). При проектировании фильтров для устранения низкочастотных составляющих шума наиболее обычный критерий — свести к минимуму среднее значение по времени для среднеквадратической ошибки.

Многопетлевые системы. Большинство встречающихся на практике систем управления и систем большого масштаба имеют много петель (контуров) обратной связи. Мощным средством исследования таких систем является анализ частотной области, так как для определения общей передаточной функции во многих случаях можно аналитически комбинировать передаточные функции отдельных петель. Примером к этому методу служит уравнение (29.33). К сожалению, во многих практических системах с множественными петлями обратной связи аналитические выражения слишком сложны для практического использования. В таких случаях лучше всего прибегнуть к моделированию.

ЛИТЕРАТУРА

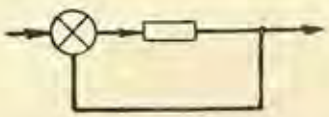
В отличие от других вопросов, разбираемых в этой книге (за исключением теории вероятностей и статистики), по теории автоматического регулирования и следящим системам, существует чрезвычайно обширная литература. Классические работы в этой области — Браун и Кэмпбелл [95] и Джеймс, Никольс и Филипс [113]. Эта глава была написана в основном по этим источникам и по Талеру [94], Кейджу [96] и Тракселу [97]. Талер — это хорошее, краткое введение в тему, стоящее вполне на современном уровне, особенно полезное для тех, кто не имеет хорошей подготовки по теории комплексных переменных. Траксел — прекрасный современный учебник повышенного типа. Вероятно, лучший разбор следящих систем, на которые воздей-

* Описывающая функция есть отношение амплитуды основной гармоники выхода к амплитуде входа. — *Прим. ред.*

ствуют случайные входные величины, дается Бэттином и Ланингом [114].

ЗАДАЧИ

29.1. Дана простая следящая система, изображенная на приложенном рисунке, у которой нет нагрузочного вращающего момента, а вращающий момент при-



вода пропорционален разности между входным и выходным положением; $\xi = 0{,}5$. Найти:

а) собственную частоту незатухающих колебаний;

б) частоту переходных колебаний;

в) константу жесткости k;

г) статическую ошибку при входной скорости 0,3 *рад/сек*;

д) нагрузочный момент, который вызвал бы такую же статическую ошибку, как входная скорость 0,3 *рад/сек*.

29.2. Найти лапласовы образы функций:

а) $\theta(t) = 6(1 - e^{0,2t})$;

б) $\theta(t) = 10 \sin \omega t$;

в) $\theta(t) = 10 \cos(t + 30°)$.

29.3. Валы механизмов часто имеют такую упругость, что их следует рассматривать как пружины. Пусть в системе на рис. 29.1 величина $k(\theta_i - \theta_o)$ есть вращающий момент, приложенный к валу махового колеса, где k — коэффициент упругости. Покажите, что передаточная функция, связывающая положение махового колеса с положением входного вала, равна

$$\frac{\theta_o}{\theta_i}(j\omega) = \frac{k_1}{(j\omega)^2 + (j\omega)2\xi\omega_n + \omega_n^2}$$

ГЛАВА 30

ВХОД-ВЫХОД. ТЕХНИЧЕСКАЯ ПСИХОЛОГИЯ

Техническая психология (называемая также *техникой человеческих факторов, прикладной экспериментальной психологией, прикладной психофизикой* и другими именами) представляет собой сравнительно новую отрасль технических наук. Конечно, какая-то грубая форма технической психологии применялась с давних времен; строитель амфитеатра в древней Греции, принимая решение о том, какой высоты делать сидения, должен был, по-видимому, учитывать различный рост людей и применял при этом методы, подобные тем, которые мы здесь рассматриваем. Но, как и во многих науках, разбираемых в этой книге, такие решения стали критическими и сложными только во время II мировой войны.

Главным предметом внимания в технической психологии является связь между машиной и человеком. По мере того как системы становятся более сложными, машины, которые мы ставим в эти системы, также становятся более сложными, но люди, работающие на этих машинах, не меняются. Мы строим великолепный самолет невиданной сложности; затем происходит авария с человеческими жертвами, и экспертная комиссия составляет заключение, что авария была вызвана «ошибкой летчика». Но во многих случаях, как мы увидим, летчик совершил ошибку из-за чрезмерных требований, которые были к нему предъявлены; задача, которую он должен был выполнить, была чрезмерно сложной без особой в этом необходимости.

И такие ситуации отнюдь не ограничиваются сложными задачами, подобными пилотированию самолета. Автоматические машины часто бывают неправильно спроектированы с точки зрения операторов, которые ими управляют; радиолокаторы бывают плохо спроектированы с точки зрения наблюдателей, которые должны извлекать информацию из изображений на экранах индикаторов. Потенциальные возможности любой машины будут плохо использованы, если пренебретать связью между человеком и машиной.

Например, если в самолете произойдет авария, летчик часто бывает вынужден увеличить мощность двигателя до полной величины. Для увеличения мощности двигателя он увеличивает давление во всасывающем патрубке дросселем, увеличивает скорость вращения винта регулятором шага и увеличивает процент бензина в воздушно-газовой смеси регулятором состава смеси. При этом его внимание обычно направлено на находящиеся перед ним приборы или на внешнее окружение самолета, так что ему приходится воздействовать на соответствующие органы управления (которые обычно находятся сбоку) на ощупь и по памяти.

Но три указанных органа управления расположены в некоторых самолетах очень близко друг к другу, имеют одинаковую форму и устроены так, что для увеличения мощности летчик должен вести два из них вперед, а третий назад. Кроме того, в некоторых военных самолетах дроссель находится слева, а регулятор шага винта — справа, тогда как в других самолетах подобного типа они расположены наоборот. Если бы при проектиро-

вании подобных органов управления обращали больше внимания на хорошо известные способности человека, то ошибки, происходящие при обращении с ними, можно было бы свести к минимуму.

Некоторые меры исправления в подобных ситуациях очевидны: органы управления должны быть разделены и они должны кодироваться своей формой. Другие меры не столь очевидны. Какие формы легче всего различить на ощупь и сколько различных форм можно различить друг от друга без затруднения? Если нужно двигать одновременно два органа, то что лучше: двигать их в одном направлении или перемещать каждый орган в его естественном направлении? (Например, орган, управляющий движением шасси или щитков, естественно перемещать вверх, чтобы поднять их, и вниз, чтобы опустить; движение в обратном направлении неестественно.)

Таким образом, специалист по технической психологии выполняет две важные обязанности в группе проектирования системы: он должен добиться того, чтобы такие вопросы были поставлены, и должен отвечать на более сложные вопросы на основании опыта, по литературным источникам или по данным эксперимента. Область технической психологии широка, и даже при элементарном разборе оказываются затронутыми многие связанные с ней дисциплины. За недостатком места мы рассмотрим здесь лишь немногие из ее наиболее интересных разделов. Бо́льшая часть этой главы основана на книге [76].

30.1. Зрение

Классические пять внешних чувств суть зрение, слух, вкус, осязание и обоняние. Из них вкус и обоняние имеют сравнительно малое значение для проектировщика систем; мы остановимся на трех других. Каждое из этих чувств является по существу довольно сложной функцией, которую целесообразно рассматривать с нескольких сторон. Выходные приборы, с которыми имеет дело инженер-психолог, связаны с этими чувствами, и, прежде чем он сможет заняться улучшением согласования между человеком и машиной, он должен сначала основательно исследовать природу процесса восприятия.

Глаз. Человеческий глаз есть оптическая система с большим раствором, состоящая из века, хрусталика, радужной оболочки и сетчатки. Веко закрывается на мгновение каждые несколько секунд для смазки глазного яблока. Хрусталик снабжен мышцами, которые изменяют его форму, чтобы фокусировать глаз на расстояния от нескольких дюймов до бесконечности. Радужная оболочка открывает зрачок до максимального диаметра около 7 *мм*, чтобы пропускать как можно больше света в темноте, и сокращает его примерно до 2 *мм*, чтобы предохранить сетчатку при сильном освещении. Сетчатка представляет собой светочувствительную поверхность на задней стенке глазного яблока, примерно в виде полушария; ее природе и свойствам мы уделим основное внимание.

Глаз можно перемещать, чтобы направлять луч зрения вверх, или вниз, или в стороны; и, действительно, он почти постоянно находится в движении. Нормальные движения глаза быстрые и резкие, от одного фиксированного положения до другого. Например, при чтении глаз пробегает каждую печатную строчку путем фиксирования по полудюжине последовательных точек; он находится в движении между этими точками примерно 10% общего времени и видит только тогда, когда фиксирован.

Все предыдущее относится только к одному глазу. Нормальное зрение, конечно, бинокулярно, т. е. в нем участвуют оба глаза. Это значительно увеличивает поле зрения, компенсирует некоторые дефекты сетчатки и дает ощущение глубины.

Центральная часть сетчатки называется *желтым пятном*. Однако сетчатка не симметрична относительного этого «центра». *Носовая область* сетчатки (которая получает свет, приходящий со стороны носа) простирается с уменьшающейся чувствительностью почти на 100° от желтого пятна; с другой стороны (*височная область*) сетчатка простирается меньше чем на 60° от желтого пятна. Примерно в 15° от желтого пятна с носовой стороны находится *слепое пятно* шириной около 7° и высотой около 5°, совершенно нечувствительное, потому что оно занято кровяными сосудами и зрительным нервом, переносящим зрительные сигналы к мозгу. У каждого человека есть эти слепые пятна, но большинство из нас живет всю жизнь даже не догадываясь об этом (слепое пятно было открыто только в 1668 г.); однако можно представить себе такую ситуацию, когда проектировщик, не посоветовавшись с инженером-психологом, поместит предупредительный световой сигнал против слепого пятна оператора.

Палочки и колбочки. Светочувствительные приборы сетчатки состоят из клеток двух типов, которые из-за своего вида под микроскопом называются *палочками* и *колбочками*. Колбочки связаны с дневным зрением, а палочки — с ночным. Колбочки сконцентрирова-

ны в области желтого пятна, и их концентрация постепенно уменьшается по мере удаления от желтого пятна. Палочки совершенно отсутствуют в желтом пятне, их концентрация возрастает до углового расстояния примерно в 20° от желтого пятна и затем снова уменьшается.

Наибольшая острота зрения при ярком дневном свете достигается тогда, когда изображение предмета попадает на желтое пятно. Анатомически однородный участок желтого пятна составляет около 1°; это означает, что если смотреть на человека, находящегося на расстоянии 1,5 *м*, и рассматривать сначала его глаз, а затем переносицу, то необходимо переместить глаз из одного фиксированного положения в другое. Однако область вне желтого пятна также обладает некоторой чувствительностью. В частности, внезапную вспышку света или отчетливое перемещение можно обнаружить даже тогда, когда оно происходит на краю поля зрения. (Точная природа способности глаза к восприятию движения еще не вполне выяснена; это связано не только с глазом, но и с мозгом).

С другой стороны, при слабом, сумеречном освещении — не более, примерно, чем при полном лунном свете — зрение осуществляется почти исключительно при помощи палочек. Так как в желтом пятне нет палочек, а возле него их лишь немного, то очень трудно видеть слабо освещенный предмет, смотря прямо на него. Для лучшей видимости (способности обнаружения) в полутьме нужно смотреть примерно на 20° в сторону от интересующего предмета, но наилучшая острота зрения (разрешающая способность) получается ближе к желтому пятну даже в полутьме. Распознавание, восприятие формы и т. п. могут быть связаны и с видимостью, и с остротой зрения.

Колбочки наиболее чувствительны к желтому и зеленому цвету, а палочки — к синему. Если два источника: один синий (4700 Å), а другой желтый (6000 Å) — излучают одинаковую световую энергию, то первый будет виден приблизительно в шесть раз лучше, когда оба слабые, и в шесть раз хуже, когда оба яркие. Из этого положения вытекают очевидные указания для выбора цветов, чтобы получить наилучшую видимость при различных условиях.

Далее, палочки совершенно нечувствительны к красному свету; поэтому можно ясно видеть при ярком свете (т. е. использовать колбочковое зрение), не нарушая адаптацию к темноте (т. е. не уменьшая чувствительности палочек), если носить красные защитные

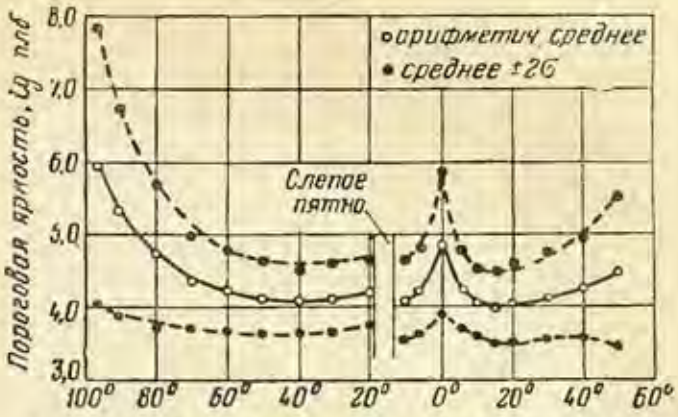


Рис. 30.1. Чувствительность различных частей глаза (по Слону [108]); 1 пиколамберт=10^{-12} *лб*.

очки. После экспозиции на ярком свету даже на короткое время палочки теряют чувствительность, и им нужно долгое время, чтобы снова стать чувствительными: около 10 *мин* для того, чтобы палочковое зрение сравнялось с колбочковым зрением, и еще 20 *мин* для того, чтобы оно достигло максимальной чувствительности (которая более чем в 1000 раз превосходит чувствительность колбочек).

Отклонения от среднего уровня. Нужно заметить, что все приведенные выше данные изображают средний уровень. Многие люди имеют значительные отклонения от этих средних величин. Например, на рис. 30.1 приведены данные о видимости посредством палочек. Пунктирные линии изображают точки 2σ и, таким образом, охватывают около 95% населения; значения, лежащие между этими двумя линиями, отличаются друг от друга в 10 раз для центрального зрительного поля и в гораздо большее число раз для периферии зрительного поля. Двадцать пять человек из каждой тысячи покажут еще худшие данные, чем на верхней кривой.

Инженер-психолог должен учитывать такие различия и проектировать аппаратуру с достаточным запасом надежности или же выбирать операторов на основании соответствующих испытаний; обычно он должен делать и то и другое.

Восприятие цвета. До сих пор мы рассматривали только одну сторону зрительного процесса, а именно глаз. Имеются еще три стороны: обнаруживаемый или анализируемый предмет, освещенность этого предмета и среда, сквозь которую он виден. Среду мы здесь не будем рассматривать; хотя она имеет важное значение для некоторых задач (например, для зрительного обнаружения кораблей и самолетов), она не представляет общего интереса при проектировании систем, где нас интересуют большей частью такие предметы, как

индикаторы и органы управления, находящиеся вблизи от зрителя.

Говоря о свойствах предмета или о его освещенности, мы должны располагать некоторыми определениями, относящимися к природе различных видов света. Источник света характеризуется яркостью (которая соответствует амплитуде световой волны), окраской (которая соответствует длине волны, если это спектрально чистый источник) и рядом других характеристик, большинство которых можно выразить суммарно таким словом, как «качество» (это соответствует примеси световых волн различных длин).

Классические спектральные цвета — это, конечно, красный, оранжевый, желтый, зеленый, синий и фиолетовый (в порядке уменьшающейся длины волны). Пределы видимого света обычно указываются как примерно 4000 Å в фиолетовой части и от 7000 до 8000 Å в красной части, но эти пределы не являются строго определенными; чувствительность глаза постепенно падает при приближении длины волны к этим пределам. За исключением света паров натрия с длиной волны около 5900 Å, по существу нет ни одного знакомого предмета, который давал бы хотя бы приблизительно спектрально чистый свет.

Кроме того, глаз есть интегрирующий прибор, который создает одно впечатление из смеси спектральных цветов и не может распознавать отдельные компоненты; так, весьма различные спектральные смеси вызывают одинаковые зрительные ощущения «белого» света. Мы располагаем несколькими прилагательными для описания некоторых спектральных смесей: цвет может быть бледным или глубоким (розовый против красного), чистым или грязным (красный против кирпично-красного), ярким, блестящим, металлическим и т. д. Эти термины полезны при сравнении двух цветов, но редко бывают достаточны для того, чтобы дать абсолютное описание цвета; для этой цели нужно иметь полную спектрофотометрическую запись света.

Указанные характеристики, как бы они ни казались сложными, достаточны только для описания источника света, который может быть либо первичным источником, либо каким-нибудь предметом, освещенным заданным первичным источником и отражающим свет от него. Если же мы попытаемся описать цвет предмета, который сам не является источником света, то ситуация станет еще сложнее, так как внешний вид предмета зависит от рода источника света и от способа смотреть на предмет. В частности, предметы часто кажутся различными по цвету, если на них смотрят при дневном свете и при свете лампочки накаливания, при излученном и при отраженном свете и, как сказано выше, при различной силе света. Наконец, смешение цветов производит очень сложные действия. Чтобы описать цвет предмета, лучше всего обратиться к одному из принятых наборов стандартных цветовых таблиц.

Количество света. При измерении количества света нужно различать *силу света* источника, *световой поток* из источника, *освещенность* предмета и *яркость* освещенного предмета или источника.

Сила света выражает количество световой энергии, приходящей от источника, независимо от его размера. Ее можно измерять в физических единицах (например, ваттах на стерадиан), но лучше применять психологические единицы, т. е. учитывать различную чувствительность глаза к световым волнам разной длины и считать, что два источника разного цвета имеют равную силу света, если они представляются на глаз одинаково яркими при определенном уровне светового потока. Для этой цели в качестве эталона силы света была взята *свеча* — сила света стандартной свечи, хранимой в Национальном бюро стандартов. Стандартная свеча производит полный световой поток в 4π люмен, или 1 люмен на стерадиан, при равномерном излучении во всех направлениях*.

Эталоном освещенности является *фут-свеча*, равная освещенности поверхности, находящейся на расстоянии одного фута от равномерного источника силой в одну свечу (или

---

* Световой поток есть количество энергии, переносимой в единицу времени сквозь какую-либо площадку и оцениваемой по производимому ею световому впечатлению; сила света равна световому потоку, приходящемуся на единицу телесного угла. Единицей светового потока служит *люмен* (*лм*, lm), единицей силы света — *свеча* (*св*, cd); 1 *св* = 1 *лм/стер*.

Свеча принимается за основную световую единицу; все другие световые единицы определяются в зависимости от свечи. В различное время эта единица определялась разными способами. Длительное время применялась так называемая международная свеча (*м. св.*), установленная Международной комиссией по освещению в 1909 г. Однако в 1948 г. IX Международная конференция по мерам и весам предложила новый (воспроизводимый) световой эталон и установила новое определение свечи (*новая свеча*, обозначаемая просто через *св*); 1 международная свеча = 1,005 новой свечи. С тех пор в СССР и других странах повсеместно употребляется новая свеча. Соответственно изменились все световые единицы; единицы, соотнесенные международной свече, именуются «прежними», единицы, соотнесенные новой свече, — «новыми». Таким образом, 1 прежний люмен = 1,005 нового люмена и т. д. — *Прим. ред.*

на расстоянии двух футов от источника силой в четыре свечи) *.

Яркость всегда является психологическим (а не физическим) свойством. Яркость источника определяется светимостью (которая равна световому потоку с единицы поверхности источника); яркость освещенного предмета определяется произведением освещенности предмета на его коэффициент отражения (в обоих случаях при равномерном световом потоке). Единицей яркости является *ламберт*; 1 *футо-ламберт*, или *кажущаяся футо-свеча*, равен приблизительно 1 (точнее, 1,076) миллиламберту **.

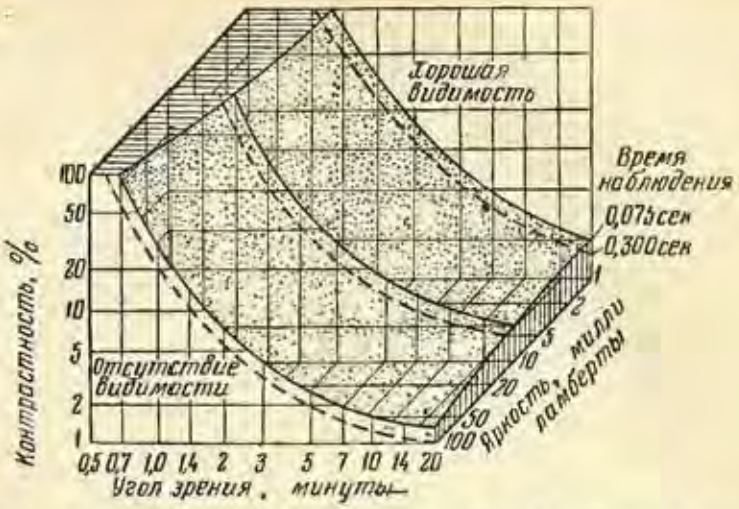


Рис. 30.2. Зависимость между остротой зрения, контрастом, яркостью фона и временем наблюдения (по Лукишу [109])

Факторы, определяющие видимость. Переходя к определению величин, характеризующих наблюдаемый предмет, мы встречаемся с еще большими трудностями. Видимость точечного источника света определить сравнительно просто; она зависит только от кажущейся силы света, цвета, длительности наблюдения и наблюдателя. Но видимость большинства интересующих нас предметов зависит также от их размера (или их яркости), от контраста между предметом и его фоном, от резкости границы между предметом и фоном, от формы предмета и от других факторов. На рис. 30.2 представлены зависимости между четырьмя из этих факторов. В дополнение к физиологическим факторам существуют также психологические факторы, как например: побуждения наблюдателя; знает ли наблюда-

---

* Освещенность равна полному световому потоку, падающему на единицу поверхности освещаемого тела. В СССР применялись единицы освещенности, основанные на метрической системе мер: *люкс* (*лк*, lx) и фот (*ф*, phot). Люкс — освещенность площадки в 1 $м^2$ равномерно распределенным световым потоком в 1 *лм* (т. е. 1 *лк*=1 *лм*/$м^2$); за рубежом люкс иногда называют также *метро-свечой*. Фот — освещенность площадки в 1 $см^2$ равномерно распределенным световым потоком в 1 *лм* (т. е. 1 *ф*=1 *лм*/$см^2$); 1 *лк*=10^{-4} *ф*=0,1 *мф*. ГОСТ 7932-56, введенный в 1956 г., оставил только люкс. Футо-свеча — единица освещенности, основанная на американской традиционной системе мер; 1 футо-свеча (f. c.)=1 *лм*/$фут^2$=10,764 *лк*, 1 *лк*=0,0929 футо-свечи. — *Прим. ред.*

** Светимость (или светность) обычно определяется как полный световой поток, испускаемый с единицы поверхности светящегося или освещенного тела. В последнем случае она равна произведению освещенности на коэффициент отражения тела. Светимость измеряется в тех же единицах, что и освещенность.

Яркость же в собственном смысле отлична от светимости: яркость поверхности светящегося (или освещенного) тела в данном направлении равна силе света, даваемой с единицы видимой поверхности тела в этом направлении. Поверхность, у которой яркость во всех направлениях одинакова, называется идеальной рассеивающей поверхностью.

На основе метрической системы мер построены единицы яркости двух родов: 1) *нит* и *стильб*; 2) *апостильб* и *ламберт*. В СССР чаще применялись единицы первого рода, в США — второго. Нит (*нт*, nt) — яркость равномерно светящейся поверхности, дающей в нормальном к ней направлении силу света в 1 *св* с площади в 1 $м^2$ (т. е. 1 *нт*=1 *св*/$м^2$). Стильб (*сб*, sb) — яркость равномерно светящейся поверхности, дающей в нормальном к ней направлении силу света в 1 *св* с площади в 1 $см^2$ (т. е. 1 *сб*=1 *св*/$см^2$); 1 *нт*=10^{-4} *сб*=0,1 *мсб*. ГОСТ 7932-56, введенный в 1956 г., оставил только нит.

Апостильб и ламберт определяются как яркости идеальной рассеивающей поверхности. Апостильб (*асб*, asb) — яркость идеальной рассеивающей поверхности, обладающей светимостью в 1 *лк*; эта единица носит также наименование «блондель». Ламберт (*лб*, L) — яркость идеальной рассеивающей поверхности, обладающей светимостью в 1 *ф*; 1 *асб*=10^{-4} *лб*=0,1 *млб*.

Для идеальной рассеивающей поверхности справедливо соотношение

$$B = \frac{R}{\pi\ стер},$$

где B — яркость в нитах, R — светимость в люксах. От сюда $1\ асб = \frac{1}{\pi}\ нт = \frac{1}{\pi}\ св/м^2$, $1\ лб = \frac{1}{\pi}\ сб = \frac{1}{\pi}\ св/см^2$, или $1\ асб \approx 0{,}3183\ нт$, $1\ лб \approx 0{,}3183\ сб = 3183\ нт$; $1\ нт = \pi\ асб \approx 3{,}142\ асб$, $1\ сб \approx 3{,}142\ лб$.

В системе световых единиц, построенной на основе американской традиционной системы мер, единицей яркости первого рода служит *свеча на кв. фут* (cd/ft^2), а единицей яркости второго рода — *футо-ламберт* (ft-L), называемый также *кажущейся* или *эквивалентной футо-свечой*; 1 *св*/$фут^2$=10,764 *нт*=1,0764 *мсб*; 1 *нт*=0,0929 *св*/$фут^2$; 1 футо-ламберт=0,3183 *св*/$фут^2$=10,764 *асб*=1,7064 *млб*; 1 *асб*=0,0929 футо-ламберта.

Все указанные соотношения между различными единицами освещенности, светимости и яркости не зависят от выбора основной световой единицы — свечи. Так как 1 международная свеча=1,005 новой свечи, то каждая прежняя единица освещенности, светимости или яркости равна 1,005 соответствующей новой единицы; это относится как к метрическим, так и к английским единицам. — *Прим. ред.*

тель куда и когда смотреть; знает ли наблюдатель, что он может увидеть, и т. п.

Практические задачи могут быть связаны не только с видимостью или остротой зрения. Не подлежит сомнению, что летчик может видеть свой альтиметр, если смотрит на него; точность, с которой он может прочесть его показание, находится под вопросом; сомнительно также, сможет ли он прочесть показание без грубой ошибки, одним быстрым взглядом; и еще сомнительнее, сможет ли он сделать это в спешке. Таковы некоторые из наиболее важных вопросов, на которые должен ответить инженер-психолог; ответы на них можно в некоторой степени предсказать на основе сведений о сущности зрительного процесса, однако обычно для этого требуются еще специальные испытания и эксперименты.

30.2. Шкалы

При проектировании связи между человеком и машиной визуальные индикаторы нужно рассматривать применительно к функции управления (или другому действию), для которой они предназначены. Для простоты мы отделили рассмотрение индикации и рассмотрение управления.

Шкала представляет особый вид индикации, при котором дается визуальное указание одного параметра. Мы не будем здесь разбирать такие существенные вопросы, как размещение шкал группами, нужно ли использовать один индикатор только для одного параметра и дает ли визуальная индикация наилучшую информацию, хотя все эти соображения очень важны для инженера-психолога. Мы рассмотрим лишь вопрос о читаемости шкал.

Назначение шкалы может быть более или менее сложным. В простейшей форме она просто указывает: «включено» или «выключено»; к этому типу относится красная лампочка, указывающая включение верхнего луча на фарах автомобиля. Индикатор следующего по сложности типа указывает одно из нескольких дискретных состояний (как индикатор переключения скоростей в автоматическом приводе) или дает качественное указание (как индикатор давления масла в современном автомобиле). Наиболее сложный индикатор дает количественный отсчет (как спидометр в автомобиле). Когда количественный отсчет должен быть точным в широком диапазоне, как на чувствительном альтиметре самолета, имеющем отметки через каждый 20-футовый интервал от уровня моря до 40 000 футов, спроектировать надлежащую шкалу труднее всего. (Часы со второй стрелкой могут давать еще более точный отсчет.)

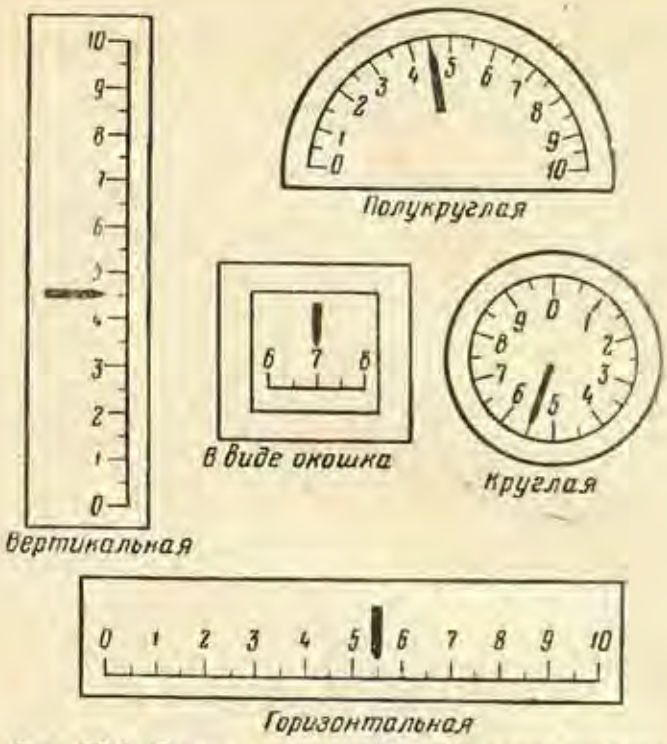


Рис. 30.3. Формы шкал, над которыми проводились испытания (по Слейту [110]).

Важное значение надлежащего проектирования шкал показывает опыт, описанный в [76], в котором были исследованы 270 катастроф и «почти катастроф» с самолетами, вызванных ошибками летчиков. Большинство этих ошибок состояло в неверном чтении шкалы, и бóльшую часть из них можно было связать тем или иным образом с неудовлетворительностью самой шкалы. Например, 40 ошибок были связаны с грубой ошибкой в чтении показаний такого сложного прибора, как чувствительный альтиметр, который имел две или три стрелки на одной шкале; 24 ошибки были связаны с тем, что были перепутаны два похожих по внешнему виду прибора; 25 ошибок были связаны с тем, что было прочтено показание бездействующего прибора, на котором не было указания о неисправности; 15 ошибок были связаны с отсчетом неправильной цифры с простой шкалы; 5 ошибок были связаны с тем, что летчик не заметил предупредительного светового сигнала (или спутал его с другим предупредительным сигналом), и т. д.

Как пример экспериментов, которые проводятся инженером-психологом над шкалами, укажем на следующий опыт. Пять шкал, изображенных на рис. 30.3, сравнивались в отношении их читаемости. Числа, интервалы и размер стрелки были одинаковы на всех шкалах. Шкалы выставлялись на очень короткое время, и испытуемого попросили прочитать их. Наилучшим типом шкалы оказалась шкала со смотровым окошком; следующей за ней, дав-

шей в 20 раз больше ошибок, была обычная круглая шкала; самой худшей была вертикальная, давшая примерно в три раза больше ошибок, чем круглая шкала.

По поводу этого эксперимента можно высказать много замечаний. Во-первых, из довольно простого испытания можно получить много полезной информации. Во-вторых, нужно быть весьма осторожным при экстраполяции результатов такого испытания без проведения экспериментов в рабочих условиях; то, что человек может заметить в лаборатории, когда шкала промелькнет перед ним в течение 0,12 *сек*, может очень отличаться от того, что он заметит в самолете при беглом взгляде на шкалу. В-третьих, здесь не учитывались производственные требования: хотя остальная часть шкалы с окошком и не видна, но для нее нужно предусмотреть место за щитом, а его может не оказаться. Кроме того, на перемещение цифр вместо стрелки затрачивается больше работы, и это может привести к уменьшению чувствительности прибора. В-четвертых, рассматривалась только читаемость и только при нормальном расстоянии для чтения и неподвижном индикаторе. Обычная шкала круглого типа может приближенно читаться на таком расстоянии, что цифр уже нельзя различить, а при шкале с окошком это невозможно. Кроме того, круглую шкалу можно прочитать, когда стрелка колеблется или движется быстро (как при «установке» шкалы), а при шкале с окошком ни то, ни другое невозможно.

Все эти замечания тем более применимы к шкалам, совершающим много оборотов; у этих шкал число оборотов обычно указывается вспомогательными стрелками (как в часах или в чувствительном альтиметре). Если шкала должна читаться только в стационарном состоянии, то шкала с окошком далеко превосходит все остальные типы и приводит к значительному уменьшению ошибок, но если ее нужно читать при движении или быстро устанавливать, то следует предпочесть обычную круглую шкалу.

При проектировании шкал обнаружился ряд особенностей весьма общего характера. Например, увеличение числа отметок на шкале на практике не приводит к повышению точности отсчета (когда индикаторы имели слишком тесно расположенные отметки, теоретическая точность часто превышала практически полученную точность), напротив, она повышает число грубых ошибок.

Для большинства целей наилучшей круглой шкалой с одним оборотом оказалась шкала диаметром около 3 дюймов * с десятью основными отметками (соответствующими единицам, десяткам, сотням и т. д.), из которых каждая обозначена цифрой, и десятью меньшими отметками (соответствующими 0,5; 1,5 и т. д.), более короткими и без цифровых обозначений. Должен быть пропуск между 0 и 10.

Ширина стрелки (или указателя) несущественна, но длина стрелки имеет значение: стрелка должна простираться от центра шкалы до отметок. Отсчеты должны возрастать по часовой стрелке (если нет основательных причин делать иначе), и между ними должны быть одинаковые интервалы (в большинстве радиоприемников это не соблюдается, потому что одно и то же изменение емкости вызывает различные изменения частоты на разных участках шкалы; но в этом случае точность отсчета шкалы не является критической). Нормальное положение стрелки, если у нее есть нормальное положение, должно быть одно и то же для каждой из расположенных рядом шкал; так, на указателях давления масла, температуры воды и амперметре на приборных щитах многих автомобилей при нормальных условиях стрелки стоят на середине шкалы и приблизительно по вертикали.

30.3. Слух

Звук, как и свет, можно описать тремя величинами; для звука этими величинами служат громкость (которая грубо соответствует амплитуде колебания), высота (которая в случае чистого тона приблизительно соответствует частоте) и тембр (который зависит от примеси звуковых волн различной длины). Хотя некоторые звуки, называемые *шумом*, представляют собой более или менее случайную смесь волн различной длины, обычный звук большей частью можно характеризовать определенной высотой, потому что он состоит из *основной частоты* и *различных гармоник*, или *обертонов*, частоты которых кратны основной частоте. Чистые звуки, как и чистые цвета, редко встречаются в природе, но они известны нам в виде «воя» гетеродина или «фона» в 60 или 120 *гц* в некоторых приемниках и проигрывателях.

Возможности уха. Ухо, подобно глазу, интегрирует ощущение, получаемое им в данное мгновение, но не так совершенно. Хотя относительные амплитуды обертонов в таком одиночном тоне, как нота рояля, различить нельзя, хороший музыкант может распознать восемь или девять нот, одновременно ударяемых на

* Около 75 *мм*. — *Прим. ред.*

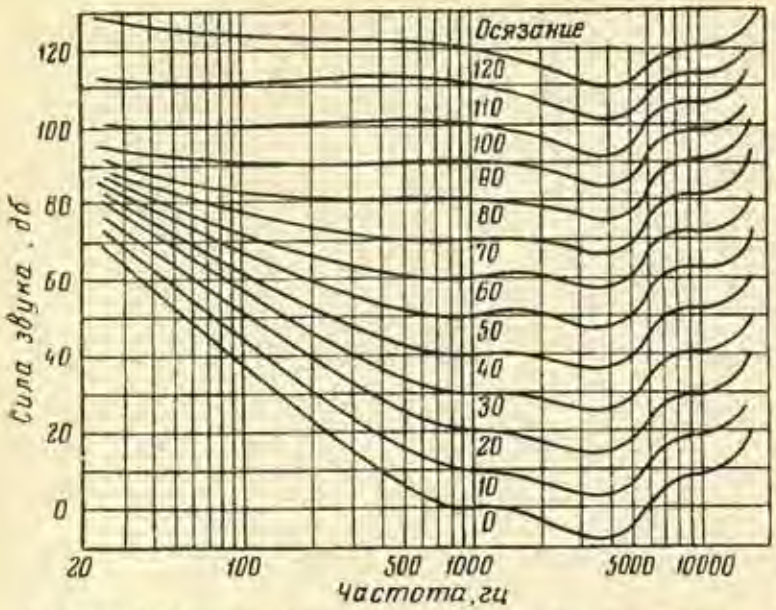


Рис. 30.4. Кривые равной громкости (по Флетчеру и Мансону [111]).

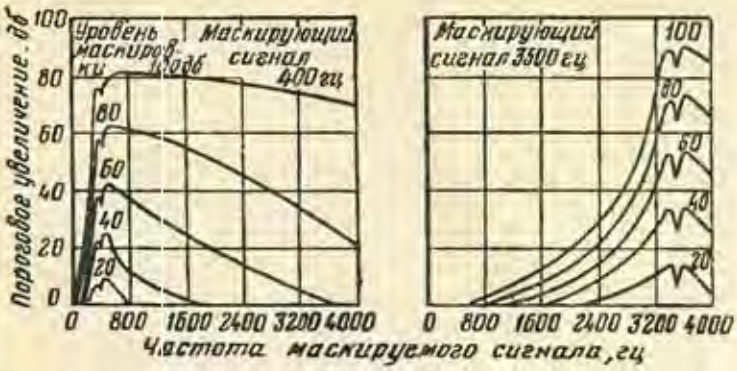


Рис. 30.5. Увеличение пороговой силы сигнала при маскировке чистыми тонами двух частот (по Флетчеру [112]).

рояле. Способность слышать звук в присутствии другого звука имеет большое значение для инженера-психолога и рассматривается ниже. Внутреннее ухо ответственно также за чувства равновесия и головокружения, но эти вопросы, интересные для инженера-психолога, мы не можем здесь рассматривать.

Высоту чистого тона или обычного сложного тона можно в достаточной мере характеризовать частотой основного колебания, хотя с изменением силы звука происходит небольшое изменение высоты тона (т. е. получаемого ощущения). Но силу звука нельзя охарактеризовать так просто, так как при изменении частоты происходит очень большое изменение громкости (т. е. получаемого ощущения). Это явление изображено на рис. 30.4; шкала ординат нанесена в *децибелах*, которые первоначально были введены для этой цели и затем приняты инженерами-электриками. Как и в электротехнике, шкала децибел является шкалой энергии; точно так же, как отношение напряжений 10 : 1 равно 20 *дб*, отношение звуковых давлений 10 : 1 равно 20 *дб*. Нуль децибел обычно определяют как 10^{-16} *вт/см*^{2} или * как 0,0002 *дин/см*^{2}.

Шкала параметров на рис. 30.4 также выражена в децибелах, но в децибелах громкости. Нуль децибел громкости определяется как нуль децибел силы звука для тона в 1000 *гц* и представляет собой самый слабый звук, какой можно услышать в абсолютной тишине. Ухо ощущает силу звуков в диапазоне, превышающем 130 *дб*, или, иными словами, звуков, различающихся по энергии в 10^{13} раз; звук громкостью выше максимальной вызывает болевое ощущение и не слышится, а осязается как давление.

Ухо также чрезвычайно чувствительно к изменениям силы и частоты звука. Хотя 1 *дб* номинально есть наименьшая разница в силе звука, которую можно различить (это и есть первоначальное определение децибела), степень различия, которое может быть обнаружено, сильно зависит от частоты и силы звука. При 20 *дб* и 4000 *гц* или 40 *дб* и 71 *гц* наименьшая различимая разность силы звука равна примерно 1 *дб*; но при большой силе звука (80—90 *дб*) и оптимальных частотах (4000 *гц*) могут быть обнаружены различия менее $^1/_3$ *дб*. При высоких частотах (1000—8000 *гц*) и большой силе звука (60 *дб*) могут быть обнаружены столь малые изменения частоты, как 0,2—0,3%; при более низких частотах (примерно ниже 500 *гц*) минимальное обнаружимое изменение частоты составляет около 2,5 *гц*, что при низких частотах составляет больший процент.

Маскировка звуков. На рис. 30.5 и 30.6 изображается способность слышать один звук в присутствии другого. Значения параметра обозначают силу маскирующих звуков. Следует отметить, что высокие частоты оказывают очень слабое маскирующее действие на низкие частоты, тогда как низкие частоты оказывают значительное маскирующее действие на высокие частоты, особенно при большой силе звука. Шум оказывает значительное маскирующее действие на все частоты и, кроме того, сглаживает кривые с рис. 30.4, т. е. приводит к тому, что все частоты слышимы более или менее одинаково.

Данные этого рода полезны для того, что-

* Сила (или интенсивность) звука есть количество энергии, переносимое звуковой волной в единицу времени через единичную площадку, расположенную перпендикулярно к направлению распространения волн. Силе звука 10^{-16} *вт/см*^{2}=10^{-9} *эрг/см*^{2}·*сек*, принимаемой за нулевой уровень на шкале децибел, соответствует амплитуда звукового давления 0,0002 *дин/см*^{2} = =0,2 *мбар*. — *Прим. ред.*

Рис. 30.6. Увеличение пороговой силы сигнала при маскировке шумом (по Хокинсу и Стивенсу [121]).

бы определить, может ли оператор обнаружить слуховой сигнал, например предупредительный звук, и может ли оператор понять звуки, кодированные по тону или по времени, в присутствии шума, помех или намеренных мер противодействия. Было проведено много специальных экспериментов по этим вопросам. Выводы из них можно подытожить в весьма грубом виде так: колебание одной частоты можно обнаружить при чрезвычайно большом уровне помех, но оператору очень трудно реагировать одновременно на слуховые сигналы двух или трех разных видов (хотя ему нетрудно различать большое число одновременных видимых сигналов, таких, как отсчеты шкал, предупредительные световые сигналы, стрелки для указания направления и т. п.).

Разборчивость речи. Речь есть сложный звук, состоящий из смеси основных частот, каждая из которых сопровождается своей группой обертонов. Процесс восприятия речи — весьма замечательный процесс, как мы уже отмечали при рассмотрении общих представлений в § 25.3. Одиночное слово, скажем «система», можно распознать как одно и то же слово, произносит ли его мужчина, женщина или ребенок, поют ли его, кричат или шепчут, говорят ли его быстро или медленно, громко или тихо и с любым из дюжины возможных акцентов. Однако по своему спектральному составу эти звуки имеют больше различия, чем сходства. Поскольку не вполне ясно, какие именно частотные составляющие являются решающими для понимания слова, задача понимания слова при наличии маскирующих звуков или искажений не имеет простого аналитического решения.

Бóльшая часть энергии речи сосредоточена в полосе от 100 до 1000 *гц*, и сравнительно мало энергии содержится в диапазоне выше 2000 *гц*. Тем не менее, если из речи удалить фильтром верхних частот всю энергию частот ниже 2000 *гц*, она все же будет довольно понятна, хотя то или иное слово может оказаться непонятным и звук речи будет искажен. Это объясняется двумя причинами: одна заключается в том, что можно распознавать обертоны низких частот, другая — в том, что звуки низких частот вообще связаны с гласными, а звуки высоких частот — с согласными, а согласные понимать важнее, чем гласные.

С другой стороны, если вырезать фильтром нижних частот все составляющие речи выше 2000 *гц*, речь останется примерно такой же понятной, как и при фильтре верхних частот. При фильтре верхних частот на 3000 *гц* или фильтре нижних частот на 1000 *гц* степень разборчивости речи будет меньше, но многие звуки все же можно будет распознать.

Ряд других искажений также представляет особый интерес. На рис. 30.7 показаны два вида амплитудных искажений, известных под названием *пикового* ограничения (срезания) и *центрального* ограничения. Небольшое центральное ограничение, например на 4 *дб*, делает речь почти непонятной, потому что оно уничтожает составляющие с малой энергией, которые содержат согласные, имеющие такое большое значение для разборчивости.

С другой стороны, срезание пиков на 20 *дб* или больше мало влияет на разборчивость, потому что оно не уничтожает согласных. В некоторых условиях ограничение пиков даже увеличивает разборчивость. Оно часто применяется в связных приемниках, потому что приводит к подавлению коротких, резких вспышек атмосферных помех и сравнительно мало влияет на речь. Иногда оно применяется в связных передатчиках, потому что позволяет передать гораздо бóльшую среднюю мощность при ограничении пиковой мощности и получить большее отношение сигнал/шум в приемнике. Но это срезание, конечно, не применяется в коммерческом радио, потому что вызываемые им искажения неприятны.

Интересен еще один вид искажения — искажение, производимое прерыванием речи много раз в секунду, так что она слышна в действительности лишь часть времени. Если частота прерывания большая, так что дискре-

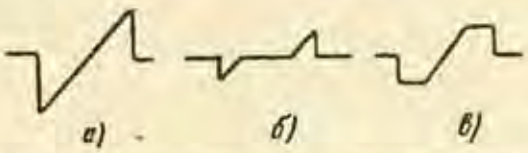


Рис. 30.7. Ограничение: *а* — входной сигнал, *б* — центральное ограничение, *в* — пиковое ограничение.

тизация (выборка) речи происходит по крайней мере восемь раз в секунду, разборчивость мало уменьшается в том случае, когда речь слышна 50% времени, и речь еще можно частично понять, когда она слышна 20% времени. Напротив, при более редкой дискретизации речь становится неразборчивой, потому что отдельные звуки речи длятся около $^1/_8$ *сек* и критические звуки совершенно теряются вследствие малой частоты дискретизации.

Маскировка речи другими звуками подобна маскировке чистых тонов: низкочастотные тоны производят большее маскирующее действие, чем высокочастотные, а шум производит большее действие, чем те и другие. Еще большее действие, чем шум, производит другой голос, а еще сильнее действие трех или четырех голосов.

Наиболее действенной мерой против маскировки речи шумом является увеличение отношения сигнал/шум. К сожалению, это не всегда возможно; например, регулируя усиление связного приемника, мы пропорционально увеличиваем сигнал и шум так, что отношение сигнал/шум остается без изменения. Тем не менее существует оптимальная регулировка усиления, так как существует оптимальная сила звука для разборчивости при любом отношении сигнал/шум. Когда отношение сигнал/шум велико (например, 15 *дб*), то и оптимальная сила звука велика (85 *дб*); когда отношение сигнал/шум мало (например, 5 *дб*), то и оптимальная сила звука меньше (65 *дб*). Конечно, оптимум силы звука при малом отношении сигнал/шум дает не столь хорошую разборчивость, как оптимум при большом отношении сигнал/шум.

Интересно практическое применение этого принципа в шумной обстановке, например на самолете или на заводе. В этих случаях можно получить большее отношение сигнал/шум, увеличивая усиление связного приемника до максимума, но и тогда сила звука будет больше оптимальной. В таких случаях часто можно значительно повысить разборчивость при помощи ушных затычек, благодаря которым отношение сигнал/шум остается без изменений, но сила звука уменьшается.

30.4. Осязание

Классическое *чувство осязания* в действительности состоит из нескольких различных чувств, включающих ощущения тепла и холода, различение предметов на ощупь и т. д. Мы применяем этот термин в широком смысле, включающем кинестетическое чувство. Для инженера-психолога особенно интересны *тактильное чувство* (что мы ощущаем кончиками пальцев) и *кинестетическое чувство* (которое позволяет нам знать положение наших конечностей, когда мы не можем их видеть). Мы рассмотрим один эксперимент, связанный с обоими этими чувствами.

Во многих случаях оператор пользуется органами управления не глядя на них; мы упомянули о подобной ситуации при управлении в самолете дросселем, регулятором шага винта и регулятором смеси. Был выполнен такой эксперимент: испытуемые надели цветные защитные очки, так что они могли видеть контрольные сигналы, но не могли видеть предметы, которые им нужно было достать. Этими предметами были расположенные различным образом «глазки», которых испытуемый должен был коснуться карандашом. Они находились все на одном расстоянии, но в различных положениях: впереди, сбоку и сзади, выше, ниже и на одном уровне с плечами испытуемого.

Оказалось, что точность значительно лучше, когда глазки находились впереди, самая худшая, когда они были сзади, и промежуточная, когда они были сбоку; средняя ошибка была от 2 до 2,5 дюйма при расположении впереди и от 3,5 до 4 дюймов при расположении сзади. Высота расположения имела меньшее значение. Очевидно, эти результаты можно использовать в случае, когда нужно поместить близко друг к другу сходные органы управления. Но эти числа являются средними, и поэтому органы управления нужно помещать на расстоянии, в три или четыре раза большем (например, на 8 дюймов впереди или на 15 дюймов сзади), если мы хотим быть уверены, что они не будут перепутаны.

Конечно, существуют и другие способы различать органы управления. Так, органы управления можно обозначить (кодировать) условными цветами, если ожидается, что оператор будет смотреть на них. Их можно обозначить различием в габаритах. При этом инженер-психолог может сказать нам, насколько ручки приборов должны отличаться по величине, чтобы различать их. В этих условиях относительные и абсолютные определения будут весьма отличаться. Если человеку нужно ощупать две ручки и сказать, какая из них больше (относительное определение), он ошибется только один раз из сотни, если диаметры отличаются на 50% при 5-дюймовой ручке или несколько больше чем на 20% при 1-дюймовой ручке. Но если человек должен ощупать лишь одну из двух ручек, с которыми он уже знаком, и сказать, какая это ручка из двух (абсолютное определение), то разница должна быть гораздо больше.

Наиболее очевидный способ кодирования ручек управления — это обозначать их по форме. При этом возникает вопрос, какие формы легче всего различать на ощупь и сколько разных форм можно различить, не путая их. Результаты экспериментов, проведенных с этой целью, оказались несколько удивительными. Некоторые пары, например призмы квадратного и прямоугольного сечения, ни разу не были перепутаны, тогда как в том же эксперименте призма прямоугольного сечения была перепутана с L-образным предметом 34 раза, а крест был перепутан с шестиконечной звездой 87 раз. Из 22 форм, испытанных в эксперименте, восемь ни разу не были перепутаны друг с другом.

30.5. Движения

Физиологи делят движения на две группы: баллистические и фиксированные. При фиксированных движениях одна группа мышц действует против другой группы, что приводит к напряженному, строго управляемому движению. При баллистических движениях одна группы мышц выполняет всю работу, а противодействующие мышцы полностью расслаблены. Баллистические движения применяются в атлетике, например при бросании или ударении мяча, а фиксированные движения применяются при точных регулировках, например при слежении за целью.

Представляется, что баллистические движения гораздо эффективнее, чем фиксированные, и значительная часть упражнений в атлетике имеет целью устранить фиксированные движения. Представляется также, что большинство других движений становятся более эффективными, когда они делаются более баллистическими, — даже такие движения, как управление машиной. Например, опытный водитель поворачивает рулевое колесо в основном баллистическими движениями, без напряжения. Однако поскольку мы не знаем, как заставить людей расслаблять мышцы, неясно, как практически использовать эти сведения.

С точки зрения инженера-психолога, удобнее классифицировать движения другими способами: движения к корпусу и от корпуса, движения установки предметов вслепую и движения со зрительной обратной связью, действительные и «статические» движения, повторяющиеся и неповторяющиеся движения, вращательные и прямолинейные движения и т. д.

Статические движения имеют место при держании предмета в устойчивом положении без дрожания. Этому вопросу было посвящено много исследований. Некоторые выводы очевидны: дрожание (тремор) меньше, если человек преодолевает какое-либо трение, если его мышцы расслаблены и если он не устал. Другие выводы не столь очевидны: дрожание меньше, когда человек сидит, чем в том случае, когда он стоит, и меньше, когда рука покоится на твердой поверхности, чем в том случае, когда она свободна.

Относительно движений установки вслепую по направлению к корпусу и от корпуса в работе [76] говорится: «Реакции установки предметов точнее при движениях от корпуса, чем при движениях к корпусу. Экспериментаторы всегда приходили к выводу, что процентная ошибка или средняя относительная ошибка меньше в том случае, когда движения установки совершались в направлении от корпуса». Подобные утверждения, основанные на статистических исследованиях, нужно всегда принимать с осторожностью. Вывод применим к «средним относительным ошибкам», которые являются систематическими ошибками и могут быть устранены тренировкой (т. е. если человек видит, что он всегда переоценивает расстояние, он может ввести компенсацию для этой ошибки); но гораздо труднее осуществлять коррекцию случайных ошибок. Исследуя дисперсию ошибок в этих экспериментах, мы обнаруживаем, что она больше для малых движений к корпусу и для больших движений от корпуса.

При установке предметов с помощью зрительной обратной связи в конце концов всегда можно найти правильное положение, так что основной вопрос состоит в том, какое время затрачивается на движения. Здесь был получен неожиданный вывод. Ускорение, максимальная скорость и замедление увеличиваются при установке предмета на большом расстоянии, поэтому при увеличении расстояния время достижения правильного положения увеличивается очень медленно.

Например, в одном опыте испытуемого просили установить по сигналу стрелку на отметку; расстояние до отметки менялось примерно от 1 до 15 дюймов. Все необходимое время было разбито на три части: время реакции (перед тем как испытуемый начал движение), время на быстрое движение до приблизительно правильного положения и время на подгонку. Время реакции составляло около 0,25 *сек* независимо от расстояния (это типичное время реакции людей во всяком опыте, в котором люди сосредоточились и приготовились реагировать определенным образом на определенное раздражение; следовательно, это минимальное время для любой реакции человека — важная особенность,

которую нужно помнить при проектировании любой связи между человеком и машиной). Время окончательной подгонки при большинстве расстояний было несколько меньше 0,2 *сек* и было еще меньше при наименьших расстояниях, при которых быстрые движения были точнее. Время быстрого движения менялось от 0,20 *сек* при 1 дюйме до 0,56 *сек* при 15 дюймах.

Типичным повторяющимся движением является стучание. Некоторые могут стучать до 13 раз в секунду, другие — не больше чем примерно 8 раз в секунду. Эта разница иногда использовалась как мера проворства при ручной работе, хотя она, по-видимому, мало связана с ловкостью при других ручных операциях. Для данного индивидуума наилучшая скорость стучания получается при горизонтальном, а не при вертикальном положении и при движении кисти или всей руки, а не только пальцев. Эти выводы были применены при конструировании улучшенных телеграфных ключей.

Типичным вращательным движением является вращение коленчатой рукоятки. Максимальная скорость вращения составляет около 4,5 *об/сек* и достигается при диаметре около 2,5 дюймов (63,5 *мм*) в условиях, когда отсутствует сопротивление трения. Если существует значительное сопротивление трения, то скорость уменьшается примерно до 4 *об/сек* и оптимальный диаметр увеличивается до 3 или $3^1/_2$ дюймов (76,2 или 88,9 *мм*).

Для инженера-психолога наиболее важным видом движения является слежение, при котором проводятся непрерывные подгонки со сравнительно малой скоростью, чтобы не упустить непрерывно движущуюся цель. В слежении мы имеем человеческий аналог сервосистемы, и по аналогии здесь имеются ошибки тех же двух типов: передемпфирование (которое может привести к недостаточной скорости или чувствительности) и недодемпфирование (которое может привести к неустойчивости). Задача слежения является крайне сложной и была предметом весьма многообразного анализа, особенно при автоматическом слежении или ручном слежении с машинной помощью. Мы укажем здесь только два вывода: в то время как при статическом движении трение желательно, при слежении оно весьма нежелательно; напротив, инерция в следящем механизме весьма полезна.

30.6. Усталость

Способность работника выполнять свою работу в течение долгого времени без снижения эффективности имеет большое значение. Здесь приходится учитывать как умственную усталость и скуку, так и физическую усталость; последняя связана с появлением судорог вследствие недостатка места или неудобства сидения, с непосредственным физическим трудом и с отсутствием отдыха. Все эти факторы заслуживают изучения как по отдельности, так и в совокупности. Одна из целей рабочего испытания любой системы заключается как раз в совокупном исследовании таких вопросов; инженеру-психологу нужно будет исследовать их также по отдельности.

Мы рассмотрим один такой эксперимент и отметим ловушки, в которые можно попасть, если не проводить рабочих испытаний. В этом эксперименте испытуемые спали очень мало или совсем не спали в течение нескольких дней и периодически проверялись их способности: умственные (память, решение задач) и физические (скорость стучания, тонкие подгонки). Испытуемые не обнаружили никакого ухудшения в течение испытаний. Это как будто указывает на то, что работоспособность человека не страдает при недостаточном сне.

Такой вывод, однако, противоречит общеизвестным фактам и в действительности является неверным. Дело в том, что при испытаниях измерялись потенциальные способности, а мы хотели бы предсказывать поведение в рабочей обстановке. Под воздействием такого сильного побуждения, как при данном испытании, люди, лишенные сна, могут показать — по крайней мере на короткое время — значительные умственные и физические способности. Однако проявление этих способностей требует чрезвычайного усилия. Человек не в состоянии производить такое усилие в течение долгого времени, и, поскольку недостаток сна отрицательно влияет на его личность, готовность человека к таким усилиям также будет меньше, чем в обычных условиях. Поэтому показатели работы операторов, лишенных сна, сильно ухудшаются, хотя простые испытания умственных способностей не обнаруживают этого.

30.7. Планировка размещения

Планировка щита управления и планировка помещения, наполненного людьми и машинами, представляют собой родственные задачи, имеющие большое значение в конечных стадиях проектирования систем (когда изготовляется опытный образец). Эти задачи имеют много общего. Некоторые из наибольших успехов в технической психологии были достигнуты как раз в планировке размещения систем.

При проектировании щита управления

нужно иметь в виду три вещи: 1) сравнительную важность каждой шкалы или органа управления; 2) как часто он применяется, и 3) *частоту связей* для любой пары органов управления, т. е. частоту последовательного применения двух органов. Органы управления, применяемые при работе, должны находиться в пределах досягаемости руки, которая должна на них воздействовать, и наиболее важные и/или наиболее часто используемые должны быть наиболее доступны. Далее, органы управления, обычно используемые один за другим, должны быть расположены рядом.

Частоту связей можно определить простым подсчетом при типичных рабочих условиях, а важность можно определить из расспросов. Нужно составить взвешенное среднее какого-то вида для этих величин (например, каждую величину можно обозначить на относительной шкале как 1, 2 и 3 и затем сложить числа в каждой группе). Затем чертится диаграмма, в которой связи обозначаются своими взвешенными средними значениями, и эта диаграмма переделывается различным образом до тех пор, пока наиболее важные связи не будут наивозможно короткими.

Если это оправдывается точностью собранных данных, то для отыскания оптимального решения можно применить более сложные математические методы.

При планировке большого машинного зала применяются аналогичные методы. Если в помещении происходит транспортировка (подача) материалов, то машины, между которыми должен проходить материал, должны, очевидно, находиться рядом. Существуют также несколько стандартных принципов рациональной организации труда, которые следует применять, например максимальное использование бункеров с самотеком. Проектировщик систем чаще всего имеет дело с такими ситуациями, в которых движется информация, а не материалы. Если передача информации происходит по телефону, то размещение не имеет особого значения, но когда один человек должен общаться с другим человеком лично или смотреть на свой индикатор или когда человек должен управлять несколькими различными машинами, мы имеем дело с проблемой организации связей, подобной описанной выше. Существенные связи опять заносятся в таблицу, с соответствующими весами по частоте и важности, и чертится диаграмма с этими значениями связей. После этого сравнительно просто, даже в сложных системах с участием людей и машин, менять планировку размещения, пока не будут найдены приблизительно минимальные расстояния для всех важных связей.

30.8. Рациональная организация труда *

Хотя рациональная организация труда предшествовала технической психологии примерно на полстолетие и еще остается вполне законченной и полноправной специальностью, мы можем рассмотреть ее здесь, под рубрикой связи между человеком и машиной. Обычно инженер по рациональной организации труда появляется на сцене после пуска машины в работу и стремится найти наиболее эффективный метод связи между человеком и машиной, однако нет причин, препятствующих включению его уже в проектирование. Он проводит, во-первых, подробное изучение работы данного оператора непосредственно в процессе ее выполнения. В некоторых случаях это можно сделать с помощью хронометра, но часто требуется более детальный анализ, и тогда работа изучается с помощью замедленной съемки движений. Движения разбиваются на основные элементы, называемые *терблигами* (написанная обратно фамилия Гилбрета — инициатора такого исследования движений), например: «искать», «выбрать», «взять», «установить», «проверить», «собрать» и т. д. Каждый терблиг точно хронометрируется. Из простого рассмотрения такой записи можно получить много сведений об операции **.

Был также установлен ряд принципов. Наиболее важный из них гласит: «распределять работу по возможности между обеими руками и обеими ногами». Ввиду того что наши руки такие ловкие, мы склонны пренебрегать ногами, которые способны выполнять многие задачи с большой скоростью и точностью (так, хотя можно построить автомобиль, которым можно управлять одними руками, но, конечно, гораздо эффективнее управлять педалями тормоза и газа, переключателем ближнего света и педалью сцепления, если она имеется, ногами, а не руками). Часто бывает так, что введение в машину ножных педалей позволяет рукам выполнять остальные задачи быстрее и точнее.

---

* Словами «рациональная организация труда» мы передаем американский термин «time-and-motion study», «time-and-motion engineering» (буквально: «изучение времени и движений», «проектирование времени и движений»). Имеются в виду методы анализа и хронометража движений рабочего при различных операциях с целью устранения лишних усилий и установления наиболее рационального способа работы. — *Прим. ред.*

** Такой метод хронометража называется «изучением микродвижений». — *Прим. ред.*

Далее, обе руки должны быть заняты одновременно и, если возможно, должны двигаться симметрично. Наконец, различные ящики, органы управления и т. п. должны быть расположены в логической и правильной последовательности — так, чтобы те предметы, которые нужно доставать чаще всего, были наиболее доступны и чтобы последовательность движений протекала по простому плану и не нужно было блуждать туда-сюда по столу или щиту.

Многие из этих выводов кажутся очевидными, однако инженеры по организации труда путем применения именно таких простых принципов часто могли повышать эффективность во много раз, не делая работу более утомительной или более неприятной.

30.9. Обучение

Отбор и тренировка операторов имеет большое значение при проектировании систем и относится к ведению инженера-психолога. Если связи между машиной и человеком правильно спроектированы, трудности при отборе и тренировке сильно уменьшаются.

Обучение представляет собой сложный и до сих пор не вполне выясненный процесс. Однако известно, что обучение некоторым операциям протекает по определенным кривым, подобным кривой на рис. 30.8. Так, в типичных случаях обучение является быстрым вначале и затем асимптотически приближается к постоянному уровню. Нередко случается, что достигнутый таким образом уровень представляет собой не истинный максимум, а горизонтальный участок («плато»), как показано на рисунке. В этом случае после дальнейшего периода инструктирования происходит новый резкий подъем в обучении. Так происходит, например, при обучении телеграфистов, где учащийся сначала учится распознавать буквы. Затем идет ровный участок, когда учащийся проходит период *интеграции*, после которого он начинает распознавать гештальты (§ 25.3) целых слов и скорость его обучения снова заметно возрастает.

Рис. 30.8. Типичная кривая обучения с плато.

В некоторых случаях оказывалось возможным повысить эффективность учебного процесса, прекратив инструктирование на то время, когда учащийся находится на ровном участке. Ввиду того что формы таких кривых сильно меняются, их следует исследовать с помощью инженера-психолога во всякой системе, где работа оператора имеет решающее значение.

ЛИТЕРАТУРА

Инженеры склонны порицать инженера-психолога за недостаточное понимание в некоторых случаях практических инженерных задач. Тем не менее книга Чапаниса, Гарнера и Моргана [76] представляет собой чрезвычайно хороший учебник, который послужил основой большей части этой главы. Стивенс [139] является более подробной и строгой справочной книгой. Работы [77] и [93] представляют собой полезные справочники с данными по технической психологии.

ЧАСТЬ 6
ЭПИЛОГ

ГЛАВА 31

ЭКОНОМИКА, ИСПЫТАНИЕ И ОЦЕНКА, РУКОВОДСТВО

В этой заключительной главе разбираются вперемежку различные вопросы, требующие подробного рассмотрения, но еще недостаточно разработанные в разрезе системотехники, чтобы их можно было изложить регулярным образом. Стоимость является центральным пунктом при рассмотрении систем, но методы оценки стоимости еще не стандартизированы в форме, полезной для инженеров-системотехников, и в разбираемых здесь вопросах мы основывались во многом на собственном опыте.

Весь вопрос испытания большой системы качественно отличен от тех вопросов, которые ставились в прошлом для малых систем, и здесь необходимо сделать некоторые замечания, хотя бы в порядке первого подхода к теме. Наконец, поскольку проектирование систем большого масштаба является сравнительно новой областью, целесообразно сделать некоторые замечания о руководстве проектировочными работами. Мы говорим здесь не об управлении системой, чем мы занимались в предыдущих главах, а об управлении бригадой проектирования системы. Поскольку эти вопросы только начинают исследоваться в разрезе больших систем, настоящая глава представляет собой скорее эпилог, чем заключение.

31.1. Экономика

Проектировщик системы должен оценивать стоимости. Основной интерес представляет стоимость самой системы — стоимость ее приобретения (или производства) и стоимость ее эксплуатации. Однако представляет интерес также стоимость ее разработки. Расходы на разработку, которые всего труднее оценить, разбираются в конце этого параграфа. В обоих случаях общим знаменателем всех расходов служат, конечно, доллары, однако существует ряд других факторов, которые легче уяснить, если не переводить их прямо в доллары. К таким факторам относятся, например, время, рабочий персонал и вспомогательные средства; мы разбираем все эти величины по очереди.

Стоимость производства и эксплуатации. Стоимость производства (или приобретения) и эксплуатации системы потребует многократной калькуляции, и, как во всех задачах оценки, применяемые методы и соответствующая точность будут меняться в зависимости от стадии разработки и соответствующего уровня понимания системы. Грубо говоря, оценка, сделанная во время предварительного проектирования, будет удачной, если она правильна с точностью до коэффициента 2, тогда как оценка с точностью до нескольких процентов будет возможна лишь к тому времени, когда приступают к производственному образцу. Эти цифры (которые меняются в зависимости от того, насколько новая система отклоняется от существующих систем) относятся только к абсолютным оценкам стоимости; относительные оценки часто можно сделать со значительно большей точностью даже на первых стадиях проектирования. Поэтому на вопросы, указанные в § 3.3 как центральные при проектировании систем (аналоговые или цифровые системы и т. д.), можно ответить на основе сравнительного анализа стоимости.

С этой целью часто бывает желательно постулировать какое-нибудь разумное проектирование остальной части системы в той мере, в какой это необходимо для выбора изучаемых частей, если только можно установить (обычно это возможно), что исследуемые стоимости не зависят от принятых кон-

кретных допущений. Полученные таким образом сравнительные оценки стоимостей могут приниматься с доверием, и разности, даже если они составляют лишь несколько процентов, часто складываются в большую сумму долларов.

В системе большого масштаба нередко какая-нибудь одна операция выполняется столь много раз, что небольшое относительное сокращение стоимости операции дает большую разницу в абсолютной стоимости. Поэтому такие потенциальные выигрыши могут оправдать широкие исследования при разработке. Аналогичные замечания применимы и к выбору оборудования на позднейшей стадии проектирования (например, выбор типа реле в коммутационной системе), если данный компонент повторяется в системе много раз.

Необходимо различать первоначальные вложения и эксплуатационные расходы. Так как проектировщику систем часто придется сравнивать альтернативные системы (например, автоматическую и ручную систему), из которых одна имеет большую первоначальную стоимость, а другая — большую эксплуатационную стоимость, то требуется коэффициент перевода. Обычно сравнение производится на основе годовых расходов. Начальную стоимость можно перевести в годовой расход, предположив, что первоначальный капитал получен взаймы и за него уплачиваются проценты на вечные времена; так, если деньги получены из расчета 5% годовых, то 100 000 долларов первоначальных затрат равноценны 5000 долларов годовых расходов. Мы предполагаем при этом, что оборудование все время поддерживается в первоначальном состоянии, а необходимые расходы включены в стоимость обслуживания. Конечно, всякое оборудование изнашивается, но если части заменяются по отдельности или отдельные блоки заменяются по мере износа, то эти расходы можно вполне включить в стоимость обслуживания.

Более распространенный метод — выбрать опорный период и формулу списывания стоимости за указанное число лет либо в соответствии с действительным сроком службы (обесценение), либо в соответствии с фиктивным сроком службы, принятым для целей калькуляции (амортизация). В этих формулах принимается во внимание уменьшение стоимости оборудования под действием нескольких различных механизмов, включая эксплуатационный износ (число раз применения), износ от времени (коррозия) и моральное старение. По каждому из них подсчитывается процент от необесцененного (или неамортизированного) остатка стоимости.

Бухгалтеры-профессионалы разработали такие формулы весьма подробно. Однако их цели отличаются от целей проектировщика систем. Они занимаются определением абсолютной стоимости единицы продукта или единицы услуг, чтобы дать сведения для установления цены, и при этом они стремятся манипулировать цифрами так, чтобы свести к минимуму стоимость финансирования, налоги и т. п. Проектировщик же систем интересуется прежде всего определением относительных стоимостей, чтобы выбрать одну из альтернативных схем. Далее, бухгалтер часто интересуется текущими операциями (например, распределением расходов на несколько изделий, производимых на одном заводе, чтобы определить, какие изделия выгоднее), тогда как проектировщик систем почти всегда интересуется сравнением процедур на основании предсказания. По этим причинам классические методы калькуляции могут оказаться неприменимыми.

Существует, однако, более важное различие. В деловой жизни решение о наилучшей формуле калькуляции имеет большое значение, а при проектировании систем справедливо совершенно противоположное. Проектировщик систем должен удостовериться, что выбор между разумными формулами не изменит суждения в ту или другую сторону. Если суждение изменяется, то проектировщик должен поставить под вопрос реальность различий между двумя рассматриваемыми системами: если система A должна заменить систему B, то замена еще не вполне обоснована; если ни A, ни B еще не существуют, то мы не вправе основывать выбор между ними на сравнительной стоимости. Если в конце концов для выбора надо бросить монету, то исход будет не столь неожиданным, когда отдают себе отчет в основе, на которой принимается решение.

Например, при исследованиях по обработке деловых данных, проводившихся непосредственно после того, как вычислительные машины начали применяться всерьез, большинство доводов в пользу замены было основано на ожидаемом уменьшении расходов. Последующее исследование показало, что расходы оценить нелегко, но что во всяком случае относительное изменение стоимости обработки деловых данных в большой системе не является решающим фактором при таком выборе. Однако быстродействие автоматической системы обработки данных и вытекающая отсюда ее способность доставлять ранее отсут-

ствовавшую информацию, полезную при принятии решений, вынуждают принимать новую систему независимо от ее стоимости и в некоторых случаях несмотря на ее меньшую гибкость [122].

Могут быть случаи, когда проектировщик системы должен делать абсолютные оценки стоимости, в частности когда он исследует желательность введения нового продукта или нового вида услуг. Спрос есть функция цены, стоимость производства есть функция спроса, а цена есть функция стоимости производства. В своих экономических расчетах проектировщик систем должен учитывать этот круг не только при оценке расходов и оценке рынка или области применения системы, но также при оценке влияния ошибок в этих оценках. Для этого опять-таки требуется расширение методов за рамки обычных бухгалтерских расчетов.

Статистические методы оценки стоимостей. В прежнее время обращали недостаточное внимание на стохастические модели стоимости. Современный экономист при оценке стоимости производства или эксплуатации вводит методы, в которых используется аппарат статистики для получения информации как об оценке, так и о ее точности. До сих пор было собрано мало пригодных данных по этим вопросам, но тем не менее мы поясним методику на следующем примере.

При проектировании большой электронной системы в 1952 г. нужно было добиться минимальной стоимости, изменяя компоненты схем, но сохраняя рабочие характеристики системы постоянными. Это задача на отыскание минимума при дополнительном условии. Достаточной информации не было, но к задаче подошли путем отыскания функций стоимости, выраженных через компоненты схем. Ввиду того что подобные системы не строились (как обычно бывает), нельзя было вычислить сравнительную стоимость.

Чтобы получить данные о стоимостях, за основу выбрали *электронный компонент* (например, блок питания, усилитель, индикатор, генератор или модулятор). Выбор того, что должно входить в компонент, был до некоторой степени произволен, хотя при этом и соблюдалось некоторое соответствие размеров. Компоненты, в свою очередь, состояли из известного числа ламп, переключателей, сопротивлений, катушек, конденсаторов и т. д. Теперь задача состояла в предсказании стоимости компонента по минимуму проектных данных о его элементах. Испытав много методов, нашли (в то время), что за показатель стоимости компонентов можно взять число ламп при стоимости лампы около 30 долларов (т. е. приблизительно как для передающих и приемных ламп). Следовательно, стоимость y можно выразить как функцию от числа ламп x простым линейным уравнением

$$y = ax + b,$$

где a и b — постоянные, определяемые статистическими методами. Было принято во внимание влияние на стоимость запасов надежности в технических требованиях, области применения — в бортовой или наземной аппаратуре — и производственного опыта. Кроме того, было исследовано, что может дать при предсказании стоимости использование нескольких «частей» вместо одной, и установлено, что новых данных здесь не получается. В итоге исследований были получены таблицы для различных компонентов, подобные табл. 31.1. Последняя строка включена для того, чтобы показать, что для блоков питания лампы оказались лучшим показателем стоимости, чем трансформаторы.

В итоге этой работы можно было сделать очень грубые проекты двух альтернативных систем, оценить число ламп в разных компонентах и затем оценить общие стоимости двух систем. Результаты, конечно, были не очень точные, но они позволили сделать выбор между альтернативными системами (например, если стоимости различались в два раза) при сравнительно небольшой затрате проектной работы.

Рабочий персонал. Количество рабочего персонала можно, конечно, перевести в доллары годовых эксплуатационных расходов, но нужно всегда оценивать эту величину отдельно, и часто бывает целесообразно включать эти цифры в описание стоимости системы. Так, могут быть затруднения при наборе, обучении и размещении необходимого персонала, помимо его непосредственной стоимости. Если даже нужно перевести количество персонала в доллары, то число долларов на человека может меняться сильнее, чем другие расходы, и оно всегда гораздо больше, чем заработная плата.

Если система должна работать круглые сутки (как часто бывает при автоматических установках), то требуются по крайней мере четыре смены рабочего персонала. Далее, к заработной плате каждого человека нужно добавить расходы на дополнительные посо-

Таблица 31.1

Статистическая оценка стоимостей в предположении линейных зависимостей

Блок	Число элементов в выборке	x	y	y начальное	Наклон	Коэффициент корреляции между x и y
Усилитель	31	Лампы	Стоимость	31 доллар	33, 85 доллара на лампу	0,734
Блок питания	23	Лампы	Стоимость	420 долларов	27, 74 доллара на лампу	0,665
Блок питания	23	Трансформаторы	Стоимость	360 долларов	153, 44 доллара на трансформатор	0,4175

бия, накладные расходы (здание, надзор, обслуживание и т. д.) и стоимость вербовки и обучения, которые должны быть амортизированы за ожидаемый срок его работы. Так, если для машины требуется один оператор с заработной платой 5000 долларов в год, то годовая стоимость ее непрерывной эксплуатации составит, вероятно, от 35 000 до 70 000 долларов только на заработную плату и разные накладные расходы для четырех смен операторов.

При подсчете требуемого персонала нужно проводить различие между рабочим и обслуживающим (ремонтным) персоналом. В пользу автоматического оборудования иногда выставляют тот довод, что оно уменьшает количество требуемого персонала, тогда как в действительности оно нередко уменьшает лишь число рабочего персонала за счет равного увеличения обслуживающего персонала, необходимого для поддержания оборудования в рабочем состоянии. Кроме того, когда группа работников ручного труда заменяется машиной, операторы и ремонтники на машине обычно должны быть более квалифицированными и, следовательно,более высоко оплачиваемыми, и их труднее набрать и обучить, чем работников ручного труда.

Критические материалы. Под *критическим* (или *дефицитным)* материалом мы понимаем здесь любое сырье или любой компонент системы, которые трудно или невозможно достать или которые могут стать таковыми. Например, хотя в 1954 г. можно было предсказать с достаточной уверенностью, что транзисторы в конце концов заменят многие электронные лампы, в то время было бы нецелесообразно рассчитывать на них ввиду их недоступности. Точно так же требование о применении титана в 1953 г. привело бы к большим затруднениям из-за чрезвычайной трудности его обработки, если бы даже свойства титана определенно указывали на желательность его применения. Конечно, не вызывало сомнений, что благодаря исследованиям эта проблема в конце концов будет решена; но проектировщик систем не должен поддаваться тенденции к допущениям, что ожидаемое изобретение уже является производимым товаром.

С другой стороны, проектировщик систем не может позволить себе быть слишком консервативным; спроектировав систему без транзисторов или без титана, он мог бы обречь ее на быстрое моральное старение. Конечно, если он может придать системе гибкость, ему это и следует сделать, и тогда у него будут и волки сыты, и овцы целы. Но если он не в состоянии это сделать, то у него нет другого выхода, как собрать все необходимые сведения о будущих возможностях, вычислить стоимости ошибок I и II рода и затем принять решение, отдавая себе отчет в том, что это азартная игра, и сообщив руководству, что ситуация именно такова.

Классические примеры с критическими материалами даются военными системами и вызываемыми войной нехватками материалов. Так, во II мировой войне сталь была дефицитной, алюминий — более дефицитным, а олово еще более дефицитным. Однако II мировая война была выиграна в значительной мере оружием, выпущенным уже после начала войны, тогда как по крайней мере одна школа теоретиков утверждает, что III мировая война, если она произойдет, будет выиграна оружием, которое будет в наличии в начале войны. Если это верно, то проектировщик систем должен беспокоиться лишь о нехватках материалов в мирное время. Однако такие решения принимаются более высокой инстанцией, чем проектировщик системы, как было указано в § 9.2.

В другой форме вопрос о критических материалах возникает вследствие нелинейности стоимости любого материала. Это становится заметным, когда требуемое количество делается сравнимым с общим национальным потреблением. Такое положение возникло по отношению к титану в связи с развитием авиации в начале 50-х годов, и нужно было рассмотреть возможные дополнительные источники снабжения, чтобы определить, можно ли будет удовлетворить спрос, вызванный проектируемой системой, без чрезмерного повышения цен. Когда в системе большого масштаба требуется новый сырой материал или новый продукт, проектировщику следует убедиться в том, что продукт может быть получен в нужном количестве.

Вспомогательные средства. На проектировщика системы часто ложится обязанность предусмотреть помещение для системы. При этом либо может иметься в виду лишь размещение системы при испытаниях, либо размещение рабочего оборудования законченной системы. Так, проектировщика может интересовать размещение антенных мачт для радиорелейной линии, топография радиолокационной установки, размещение складов в системе материально-технического обеспечения и т. д.

Этот интерес вызывается тем, что составляемые проектировщиком технические условия (задания) должны всегда на практике допускать некоторую свободу и он должен

найти наилучший компромисс в отношении размещения, топографии и других факторов. На этой стадии ему выпадают на долю все трудности разбивки планов, приобретения участков, снабжения энергией, расходов на подготовку участка и т. п. Эти вопросы лучше всего решать в других отделах проектной организации; но когда ответственность за них лежит на группе проектирования системы, лучше всего придать ей для этой цели эксперта на требуемое время.

Минимизация стоимости. В заключение нашего разбора стоимости систем мы приведем одну выдержку [117] об уменьшении издержек производства, причем эти замечания в еще большей мере применимы к системам.

> Стало аксиомой промышленного производства, что стоимость производства должна быть как можно меньше. Возникает вопрос: какая именно стоимость? *При комплексном производстве может оказаться нецелесообразным и слишком дорогим уменьшать каждую стоимость по отдельности; нужно рассматривать составную стоимости как целое и уменьшать именно ее.*
>
> Например, изготовитель, выпускающий вещательные радиоприемники двух типов, может решить, что производство каждой модели должно стоить как можно меньше, независимо от другой модели. Затем он, возможно, установит одну линию компонентов с минимальной стоимостью для первой модели и более дорогую линию для обеспечения различных потребностей второй модели. Но в действительности может оказаться в конечном счете дешевле применять в более дешевой модели компонент лучшего качества, чем безусловно необходимо. Это может быть вызвано рядом причин: изготовителю теперь нужно делать одним компонентом меньше; он может производить один компонент в большем количестве и, следовательно, снизить окончательную стоимость одной единицы; становится проще учет; снижаются конторские и складские расходы; уменьшаются потери производственного времени, вызываемые переводом линии с одного компонента на другой.
>
> Взаимозаменимость частей и методы массового производства существовали в промышленности и раньше, но до последнего времени они применялись в основном для однотипных изделий. Самуэль Кольт ввел взаимозаменимость частей для одного типа револьвера; Генри Форд ввел массовое производство одного типа автомобиля. Но те же принципы применимы к общему производству завода или группы заводов, занятых изготовлением многих родственных изделий. Обычно требуются группы изделий с некоторыми общими основными характеристиками, и при наличии этих общих характеристик справедливы выводы о преимуществах взаимозаменимости и массового производства. *Приносить прибыль должен завод как целое, а не отдельный продукт* (курсив наш).

Стоимость разработки. Существует, конечно, резкое различие между стоимостью системы и стоимостью разработки. Фактически стоимость разработки очень мало связана со стоимостью окончательной системы, но зависит прежде всего от того, насколько новая система отклоняется от прежних систем. Нужно оценить стоимость и время разработки хотя бы для того, чтобы позволить проектировщику системы представить своим начальникам обоснованные предположения о том, чего они могут ожидать.

К счастью, стоимость разработки можно довольно точно предсказать, если выразить ее в требуемых человеко-годах инженерного времени*. Иначе говоря, если определить этот показатель, то обычно можно предсказать с большой точностью для довольно широкого класса систем все другие расходы, как-то: время техников, стоимость материалов, транспортные расходы и накладные расходы. К сожалению, еще не существует адекватной формулы для предсказания требуемых человеко-лет инженерного времени.

В настоящее время этот расчет делают по аналогии с другими сравнимыми разработками, и, как указано в § 3.2, обычно получаются низкие оценки, если только бригада проектирования системы не является очень опытной. Однако если придерживаться правила, что в проектирование системы включается только разработка, но не исследования, то можно получить точные оценки. Хотя основной единицей служит человеко-год, но это не значит, что, удвоив количество людей, можно сократить в два раза общее время. По-видимому, существует оптимальная величина проектной группы, и увеличение рабочей силы сверх этой величины мало влияет на уменьшение времени разработки.

В этой книге рассматриваются системы, состоящие из компонентов, для которых требуется различное время на составление технических условий, разработку, изготовление опытного образца и окончательное изготовление. Так, клавиатура, возможно, потребует сравнительно незначительных видоизменений, а специализированная центральная вычислительная машина вполне может потребовать одного года на составление технического задания, еще одного года на разработку и еще года на изготовление и ввод в эксплуатацию окончательного образца (но это уже большой шаг по сравнению с концом 40-х годов, когда один специалист по вычислительным машинам утверждал, что время, остающееся до завершения любой изготовляемой цифровой вычислительной машины, постоянно). Чтобы обеспечить равномерную

* В 1951 г. общие годовые расходы на исследования и разработки в Соединенных Штатах составляли в среднем 9 000 долларов на работника в этой области (включая вспомогательный персонал) и 22 000 долларов на инженера или научного работника. Первая цифра была одинакова для многих отраслей промышленности, а вторая изменялась от одной отрасли к другой [118]. — *Прим. авт.*

разработку системы, нужно оценить время, необходимое для каждой подсистемы, и не допускать чрезмерного неравновесия отдельных частей разработки.

31.2. Испытание и оценка

При проектировании автоматических устройств не слишком большой величины критической стадией проектирования является момент, когда устройство испытывается и оценивается: включают питание, отмечают, что устройство работает, и сравнивают способ, каким устройство выполняет свои функции, с другими методами выполнения того же самого. Мы намеренно поставили рядом слова «испытывается» и «оценивается»: это как будто представляется естественным. Но следующее предложение указывает, что одной операцией «включение питания» преследуются две разные цели: одна — установить, делает ли устройство то, что оно должно делать, и вторая — выяснить, будет ли то, что оно должно делать, самым подходящим способом делать это.

Неосуществимость оценки испытанием. Пока устройство сравнительно мало, не возникает никакой путаницы или затруднений из-за такого пренебрежения двойственным характером целей. Легче построить модель, чем оценить идею теоретически, и стоимость изготовления модели для достаточно малого устройства остается в пределах расходов, на которые можно согласиться, чтобы решить, является ли данное устройство правильным способом решения задачи. Можно построить модели двух способов решения задачи, и тогда испытание переходит в сравнительную оценку.

Но при увеличении размеров устройства (т. е. при больших системах) стоимость изготовления модели быстро возрастает, и в конце концов становится нецелесообразным изготовлять устройство и затем отказываться от него, если оценка показывает, что устройство неправильно. Эта стоимость измеряется в деньгах и времени: мы не можем позволить себе истратить зря ни миллион долларов, ни год работы. Кроме того, хотя стоимость оценки путем изготовления и испытания модели возросла вследствие увеличения наших систем, стоимость оценки путем анализа уменьшилась благодаря появлению новых методов (в особенности, моделирования и вычисления на быстродействующих машинах). Именно это положение дел и выявило потребность в использовании методов системотехники.

К сожалению, еще держится одно представление, унаследованное от малых устройств. Проектные организации до сих пор планируют и говорят так, как будто испытание для оценки больших систем можно провести полностью в натуральном масштабе. В действительности такие испытания слишком дороги и их слишком трудно координировать. Планирование и проведение испытаний часто по меньшей мере столь же трудны, как проектирование и эксплуатация самой системы.

Существует такое распространенное мнение: «Теоретически все хорошо, но мы построили систему на основе теории, а теперь ее нужно испытать на практике. Может быть, теория неверна. Оценим систему во время испытания, определив, как она фактически реагирует на входы в реальных условиях». К сожалению, без математической модели как основы это просто невозможно для любой системы большого масштаба. Иначе говоря, в оценку входит определение того, как система будет реагировать на все сочетания входов, с которыми она встретится в действительности.

Это можно сделать исследованием математической модели системы, которая была разработана во время проектирования, с помощью различных аналитических методов и методов моделирования, рассмотренных в предыдущих главах. Но если мы попытаемся сделать это путем измерения входов и выходов в реальных условиях без помощи математической модели, то результаты, бесспорно, будут бесполезны из-за слишком большой дисперсии, если только число экспериментов и степень контроля при экспериментах не будут очень велики — гораздо больше, чем это осуществимо практически. Правда, в случае сложной системы оценка путем исследования модели будет также трудной, но оценка путем испытания, без помощи модели, будет гораздо труднее.

В известном смысле оценка и испытание относятся друг к другу так же, как теория и эксперимент в научном исследовании. Оценка с помощью изложенных выше методов предсказывает результат изготовления определенных устройств и соединения их определенными способами с людьми. Экспериментом является испытание. Нет необходимости проверять все разделы теории — достаточно испытать лишь сомнительные пункты прогноза.

Итак, испытание все же требуется для подтверждения правильности модели, определения сомнительных значений параметров и отыскания «сбоев» — неизбежных и непредвиденных незначительных затруднений. Необходимо еще выяснить, работает ли изготовленное оборудование так, как предсказали проектировщики. На стадии проектирования

рабочие характеристики системы можно предсказать, основываясь на допущении, что все компоненты и люди будут работать так, как предполагалось. Остаются под сомнением допущения о рабочих характеристиках операторов и оборудования. Оба эти допущения можно в большинстве случаев проверить на небольших группах оборудования, не затрагивая всей совокупности оборудования системы. Проект не может быть закончен без наличия надежной информации о всех частях системы, зависящих от рабочих характеристик оператора или оборудования.

Испытательная аппаратура. Работа оператора и работа оборудования могут отличаться от заданных в проекте только в следующих основных направлениях: 1) работа может быть слишком медленной или 2) работа может приводить к ошибкам двух видов: непрерывным (слишком большое или слишком малое значение какой-нибудь непрерывной переменной) и дискретным (выбор неверной альтернативы). Эти два критерия: скорость и погрешность — дают всю требуемую информацию о том, как операторы и оборудование выполняют требования проектировщика системы, поэтому на них должно быть основано аппаратурное оформление испытаний. Вся другая информация будет вспомогательной.

Из сказанного ясно, что основная испытательная аппаратура должна быть связана с измерением времени (хронометры, осциллоскопы, аппаратура синхронизации и т. д.), записью непрерывных переменных и вычислениями над ними (самописцы Бруша, осциллографы, аналоговые вычислительные машины и т. д.) и записью выборов (бланки, перфокарты, быстродействующие счетчики, счетчики Видера — Рута и т. д.). Испытательная аппаратура должна работать по возможности автоматически, а в том случае, когда не требуется непрерывной работы, должно предусматриваться кнопочное управление.

Работу по проектированию испытательного оборудования нужно начинать заблаговременно и строго координировать ее с работой групп проектирования системы, проектирования оборудования и полевых испытаний. Испытательная аппаратура должна составлять часть испытательной установки (а не собираться наспех), должна быть весьма надежной и давать воспроизводимые результаты, приблизительно на один порядок более точные, чем аппаратура системы. Эта высокая точность необходима для того, чтобы получить вполне определенные выводы о точности аппаратуры системы; так, если аппаратура системы и испытательная аппаратура дают стандартное отклонение 1 мил*, то вряд ли возможно разделить погрешности испытания и погрешности системы при не слишком большом числе опытов.

В некоторых случаях аппаратура системы работает с наивысшей точностью, достижимой при существующем уровне техники. В таких случаях следует, если это возможно, применить в испытаниях другие способы измерения (например, точность радиолокатора можно проверить посредством оптических измерений). Если этого сделать нельзя, проектировщик системы должен признать тот факт, что, возможно, он никогда не узнает, как работает данный компонент.

Замечания о планировании испытаний. Опыт показывает, что если слишком полагаться на то, что отдельные единицы оборудования работают, как требуется (если, например, испытательные точки предусмотрены только в начале и конце длинных цепочек компонентов), то испытательную аппаратуру придется проектировать заново с большой затратой времени и денег. Испытательные точки должны быть предусмотрены во всех существенных местах соединений между разными видами оборудования.

Один важный принцип проектирования испытательных установок состоит в применении замкнутой петли, т. е. в сравнении входной команды и выходной реакции. Так, в одной системе, где цифровая вычислительная машина использовалась для управления реактивным снарядом, команда, выдаваемая машиной, записывалась и сравнивалась непосредственно с сигналом, принятым в снаряде после передачи. Благодаря такому замыканию петли вокруг линии передачи данных была обеспечена превосходная проверка работы этой линии и одновременно была достигнута лучшая проверка наведения снаряда, так как входные сигналы снаряда были известны достаточно точно.

Испытание оборудования следует проводить по возможности отдельно от испытания операторов. У оператора должен быть прибор, пусть даже не обязательно входящий в рабочее оборудование, но дающий пренебрежимо малую погрешность по сравнению с погрешностью как рабочего оборудования, так и самого оператора. Тогда можно легче разделить ошибки оборудования и ошибки оператора (хотя для этой цели могут оказаться дешевле методы дисперсионного анализа).

Конечно, бригада проектирования системы должна начать планирование испытаний обо-

---

* 1 мил (mil) = 0,001 дюйма = 25,4 *мк.* — *Прим. ред.*

рудования и испытаний операторов еще до аппаратурного оформления испытания. Эти подготовительные работы будут включать подбор операторов для оборудования, приемку законченного оборудования системы для испытания, составление сценария испытаний, обеспечение условий испытаний и подготовку средств обработки данных.

Оборудование, с которым могут работать только инженеры и лаборанты, совершенно непригодно для работы в системе. Нужно выбрать операторов самого низкого уровня, какой может встретиться на практике, и операторов среднего уровня. С другой стороны, нужно предусмотреть обучение операторов за некоторое время до испытаний, возможно еще в период изготовления оборудования. При этом целесообразно использовать имитаторы (тренажеры). Обучение должно быть достаточно продолжительным, чтобы кривая обучения (§ 30.9) достигла своего оконечного плато.

Частая ошибка при испытании больших систем — это приемка оборудования, надежность которого не была проверена до поставки. Планировщик испытаний предохранит себя от многих неприятностей, если напишет свои приемочные технические условия именно для завода, а не для испытательного полигона. Условия для проектируемого оборудования на заводе лучше, и испытания иногда проваливаются потому, что оборудование принимается после того, как оно проработает 1-2 раза под управлением инженера-разработчика и под наблюдением первоначального проектировщика. Каждое оборудование нужно проверить на время работы и на погрешность в группе разного оборудования и с разными людьми. К сожалению, инженер-испытатель иногда считает, что испытал прибор, если проверит, «включается» ли прибор, когда повернут выключатель.

Сценарий серии испытаний представляет собой именно сценарий. В нем должны быть указаны: время, сигналы, величины, произносимые слова и инструкции по ремонту, как профилактическому, так и текущему. Если проверяемая группа представляет собой существенное звено в цепи, то типичным временем нужно выбрать *наиболее загруженный час*, как в телефонии или на транспорте. На испытании системы нет места людям с розовыми очками. В большинстве организаций принимаются меры, чтобы испытательные функции выполнялись людьми с критическим складом ума.

Конечно, все предыдущие замечания об оценке и испытании совершенно неприменимы к плохо спроектированной системе, ибо хотя безусловно верно, что и плохо спроектированная система может оказаться при испытании способной делать то, что она должна была делать, но ее работа будет все же недостаточно хороша. Позволительно утверждать, что такой исход можно было бы заранее предсказать, так как после испытания оборудования мы не узнали ничего такого, чего мы не знали бы до испытания. Результатом такого провала является либо период модификации, иногда более продолжительный, чем первоначальное проектирование, либо потеря большого количества времени, труда и денег.

Надежность. Надежность системы является одним из факторов, которые можно действительно определить на стадии испытания, если даже она была несколько недостоверна в стадии проектирования. Надежность системы была определена [143] как «вероятность того, что система будет выполнять свое назначение при данных условиях в течение требуемого рабочего времени». Отсутствие надежности может быть вызвано либо ошибками, либо непригодностью; мы имеем здесь в виду вторую причину.

Надежность является важным критерием эффективности любой автоматической системы и, как указывалось в § 9.4, может быть заменена другими критериями эффективности. Существовала достойная сожаления тенденция, особенно при проектировании электронных систем, жертвовать надежностью ради рабочих характеристик и времени разработки. Это означает, что не допускаются коэффициенты запаса, что лампы, например, постоянно нагружаются выше их проверенной мощности. Это, в свою очередь, означает, что систему можно заставить работать в течение короткого времени под тщательным надзором ее проектировщиков, но что она бесполезна как рабочая система. Такую ситему не следует допускать до стадии испытаний.

Однако даже когда система запущена в производство и повседневную эксплуатацию, все равно будут возникать какие-то вопросы, связанные с обеспечением исправности аппаратуры. Этим вопросам проектировщик системы должен уделять внимание на самых ранних этапах проектирования. Там, где возможно, в системе должны быть запроектированы рабочие контрольные приборы (например, встроенные измерительные приборы и контрольные точки). Эта рабочая контрольная аппаратура может быть полезной на стадии испытаний, но при испытаниях системы проверить эту рабочую контрольную аппаратуру было бы гораздо важнее.

Проектировщик должен, кроме того, пре-

дусмотреть легкий доступ к оборудованию, вставные блоки (кассеты) и сменные стойки и шасси, встроенные процедуры проверки в предельных режимах, надлежащую защиту от окружения (фильтрация, кондиционирование воздуха и т. д.) и все остальное, что необходимо для обеспечения надежности его системы в реальных эксплуатационных условиях. Эти вещи гораздо легче предусмотреть на стадии проектирования, чем вводить их при испытаниях, и если на них не обратить должного внимания, они вызовут увеличение времени испытаний и большие расходы.

31.3. Руководство проектированием

Системотехника есть совокупность орудий, этапов и частей, рассмотренных в этой книге, плюс взгляд, который связывает все это воедино. Этот взгляд особенно нужен при руководстве проектированием систем, ибо он отразится на всей организации проектирования. Наибольшая трудность, с которой может встретится проектировщик, — это не сама задача, но руководство, у которого нет сочувствия или ясного понимания.

Первым условием такого понимания и правильного взгляда является сознание того, что прежнее отставание разработки и производства от теоретического исследования уменьшилось теперь на несколько лет. Это сокращение сроков привело к новым требованиям в отношении уровня знаний производственников и к большой терпимости со стороны исследователей. В действительности происходит обмен людьми между этими тремя областями деятельности, и существующие затруднения, вероятно, преходящие. Но раздражение производственника осторожной процедурой исследователя-теоретика «в башне из слоновой кости» и, наоборот, отвращение ученого к «торгашескому» подходу производственника, идущего на снижение качества ради выигрыша во времени, деньгах и простоте, не могут привести к добру. Оба должны пойти на компромисс.

Вторым условием понимания является момент, подчеркнутый в § 1.4, — бригадный метод работы. Не только инженеры работают бригадами при проектировании больших систем, но и ученые теперь обычно работают бригадно. Это ставит такие требования к обмену информацией и уровню группового понимания, как никогда раньше, и значительная часть усилий проектировщика должна быть затрачена на самообразование — как по данной работе, так и вообще. В еще большей мере это относится к инженеру-системотехнику, интересы которого охватывают бо́льший диапазон, чем у других инженеров.

Исследование в соотношении с разработкой. Как указано в § 27.7, нужно различать проектирование системы и исследовательскую работу. Если элементы предлагаемой системы не были полностью изучены при исследовании, то системная бригада, вероятно, будет барахтаться в трясине возможных идей. Шизофрения календарного плана, с одной стороны, и неизученная исследовательская идея, с другой стороны, не позволят проекту иметь определенное лицо. Всякая работа, будь то исследование или проектирование системы, полезна, но если это — исследование, пусть оно так и называется, и тогда шансы на успех будут больше. Келли [119] говорит даже, что работа над проектированием системы должна прекратиться, если обнаружится, что не закончено исследование над существенным компонентом.

Это замечание допускает одно исключение. Иногда полезно упражнение в проектировании систем, чтобы определить, как будет построена система, если еще не проверенные идеи окажутся успешными. В этом случае проектирование также должно иметь соответствующее название.

Дублирование работ. Правительственные комитеты, администраторы и руководители проектов часто пытаются выявлять и устранять «дублирование» в исследованиях, тогда как считается вполне обычным и нормальным начинать конкурирующие разработки и выбирать «лучшую» из двух или трех разработок, когда они достаточно продвинулись. Такой взгляд господствует, в частности, в военной области.

По нашему мнению, это противоречит логике. Не найдется и двух исследователей, работающих над одной и той же задачей, которые подошли бы к ней одинаково; вероятность успеха, во всяком случае, мала (Кеттеринг однажды сказал, что исследования обходятся столь дорого потому, что из сотни идей оправдывает себя лишь одна), и увеличить эту вероятность можно лишь повторными независимыми попытками. Кроме того, такие попытки будут дешевле, чем повторные попытки разработок.

С другой стороны, исход разработки можно довольно точно предсказать по прежним разработкам проводящей ее организации. Если разработка основана на доказанных идеях, можно легко достичь вероятности успеха до 0,95 и дублирование (если оно независимо, что обычно не имеет места) лишь немного увеличивает и без того высокую ве-

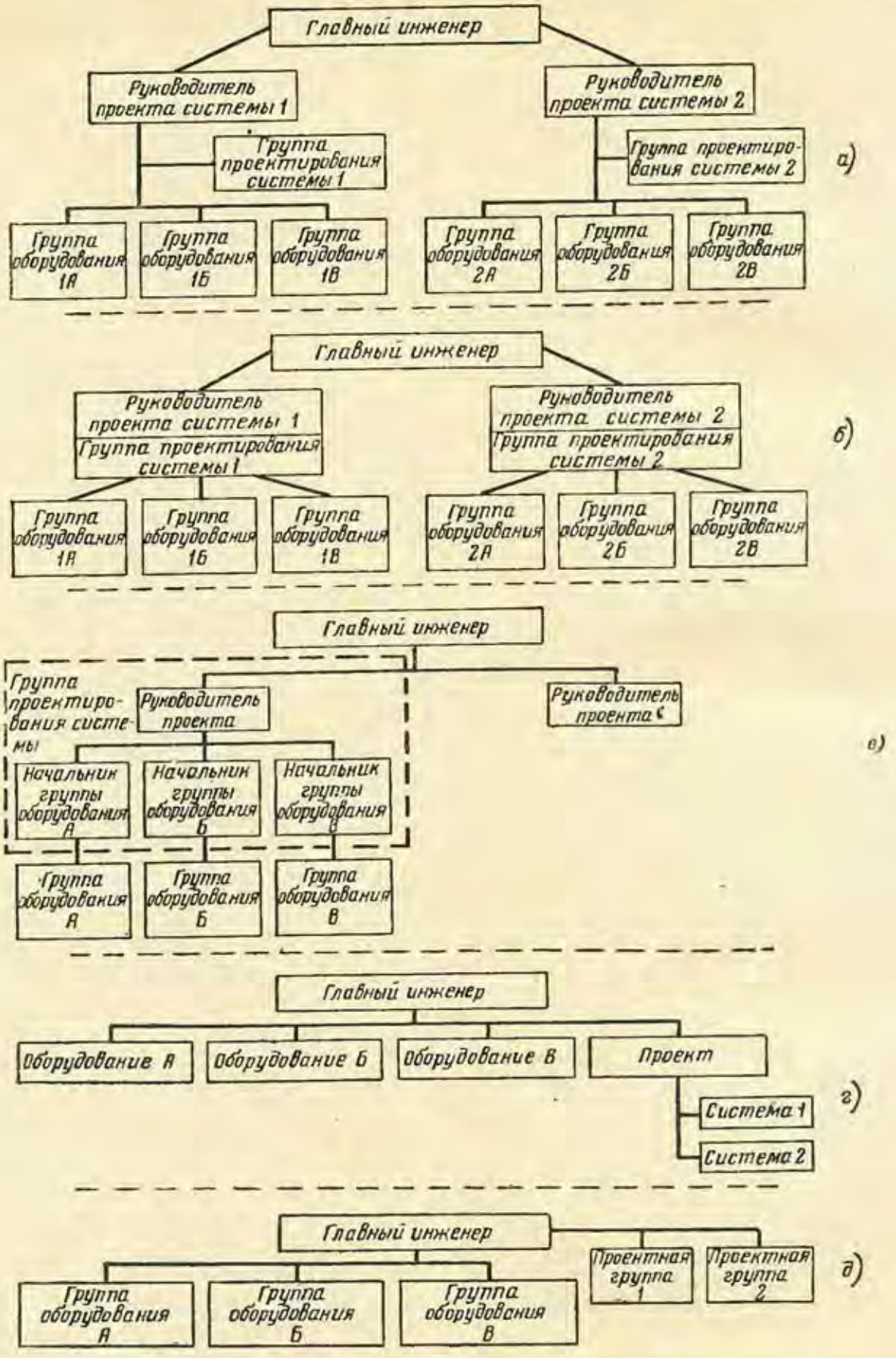


Рис. 31.1. Организация работы по проектированию систем.

роятность успеха. Отсюда ясно, что дублировать нужно исследования, а не разработки.

Организация. Руководители имеют свои проектные и рабочие обязанности, как и инженеры. Проектирование обычно производится силами организации. Следует отметить, что качество работы группы людей очень зависит от способностей индивидуумов и довольно слабо зависит от того, как они организованы. Другими словами, хорошие работники работают довольно хорошо почти при любой организации и несколько лучше при хорошей организации. Неспособные люди работают плохо при плохой организации и работают столь же плохо при любой организации. Повторные реорганизации наблюдаются в группах людей, мало пригодных к своим обязанностям, хотя никакая организация не даст здесь хорошей работы. Наилучший архитектурный проект провалится при плохих кирпичах и плохой извести. Но выигрыш от хорошей организации при хороших работниках стоит потраченного времени.

Исследования о личном составе в несистемных областях промышленности перечисляют некоторые методы отбора талантливых людей, но редко упоминают о выработке постоянного заряда энергии и инициативы, что составляет первое требование для хорошей разработки системы. Приобретение такой энергичности зависит опять-таки от личных качеств, но в организации могут быть созданы хорошие условия для ее роста.

Группа проектирования системы может быть организована следующим образом:

1) как штабная группа при руководителе проекта: последний наблюдает за всеми разделами программы, тогда как системная группа обеспечивает планы и общее ведение программы (рис. 31.1,*а*);

2) как линейная группа во главе с руководителем проекта (который является непосредственным начальником системной группы), функционирующая над всеми частями проектной организации, привлекаемыми к выполнению задания, за исключением обычных инженерных служб, цехов, изготовления образца и т. д. (рис. 31.1,*б*);

3) как расчлененная группа, состоящая из начальников групп оборудования, которые регулярно встречаются для выполнения задач проектирования системы в целом (рис. 31.1,*в*);

4) как отдельная линейная организация на равных правах с группами оборудования, быстро переключающаяся с одного оборудования на другое (рис. 31.1,*г*);

5) как модификация метода № 4, когда каждый проект системы получает права отдельной линии, — идея проектного бюро (рис. 31.1,*д*).

Выбор зависит от назначения, размеров, окружения и количества различных разрабатываемых систем. Существуют также важные факторы, не показанные на организационной схеме, например: кто дает средства, кто имеет право найма и увольнения и является ли власть полной или имеются другие органы, обладающие правом инициативы и запрета. Основные различия между этими формами организации сводятся к различию между функциональной организацией и организацией по конкретным заданиям, с одной стороны, и к различию между упором на систему и упором на оборудование, с другой стороны.

С учетом зависимости от персонала, методы № 1 и 2 дают одинаково хорошие результаты; метод № 3 довольно плох из-за наплыва повседневной работы и недостаточного упора на системные аспекты, если только каждый начальник по оборудованию не является специалистом по проектированию систем, что встречается редко; методы № 4 и 5 различаются лишь тем, перед кем отчитывается системная группа, но на практике это ведет к большому различию. При небольшом количестве больших проектов наилучшим, вероятно, является метод № 1. При большом количестве малых проектов лучше всего, вероятно, метод № 4. Но опять-таки в основном это зависит от людей, а не от организационной схемы.

Поток информации. Вероятно, лучшей организацией является такая, которая больше всего поддерживает поток информации. Осуществить действенный поток полезной информации нелегко. Эта информация может быть двух видов: технические данные и информация, касающаяся судеб организации в целом.

Необходимая техническая информация должна передаваться быстро как извне, так и изнутри. Контроль за тем, чтобы она двигалась достаточно быстро, является главным делом технического руководства. Что касается потока информации извне, то, кроме книг и технических журналов, жизненно необходимых для технических профессий, инженеру-системотехнику должны быть предоставлены возможности для участия в заседаниях технических обществ и для посещения других технических учреждений. Одна из наиболее существенных льгот инженеру-системотехнику — это предоставление средств на поездки.

Другим источником внешней информации является консультант. Консультанты бывают двух видов: работник и советник. Первый по существу является временным прибавлением

штата. Второй является своего рода носителем «перекрестного опыления» и должен использоваться именно как таковой. Настоящий консультант знаком со многими разработками, к которым инженеры-системотехники, особенно те, кто работает в промышленных организациях (в отличие от работающих в правительственных организациях и в университетах), не имеют доступа.

Для внутренней информации неоценимыми средствами являются отчеты и семинары. Отчет отражает существующее понимание решенной и нерешенной части задачи и весьма ценен не только для читателя, который знакомится с тем, что сделано, и с тем, что задумано, но также и для автора, повышающего уровень своего понимания задачи благодаря упорядочению, необходимому для представления выводов на бумаге. Семинар дает возможность «перемолоть» еще не вполне оформленные идеи. Семинар часто плохо используется, когда требуют, чтобы он проходил как административное заседание, или когда требуют формальных сообщений, как на заседании технического общества.

Большую пользу, особенно в организациях, не стремящихся к прибыли, приносят симпозиумы. Для сбора информации по одному вопросу, наспех, когда расходы делятся между участниками и выигрывают почти все, хорошо организованный симпозиум лучше всего.

Организации по проектированию систем являются неконсервативными системами в смысле Винера (§ 25.3), и информация служит энергией, движущей проектирование системы; если нет притока информации, проектирование засыхает. Во всякой организации нужно обеспечить беспрепятственное течение информации. Кроме того, как следует из рассмотрения групповой динамики в § 25.2, существуют некоторые доводы в пользу того, что задачи системотехники, как и другие трудные задачи, лучше всего решаются в организациях, построенных по принципу окружности, когда глава организации не является руководителем на стадии решения задачи. После того как задача понята, т. е. на стадии исполнения, глава снова становится руководителем.

Члены здоровой организации не тратят времени на предположения об организационных событиях, будущей работе и общем положении организации. Единственный способ свести подобные занятия к минимуму — это давать членам организации такую же полную информацию по этим вопросам, как по техническим вопросам. Такая откровенность приводит к новым затруднениям, когда организация попадает в состояние временной депрессии, но в конце концов хорошо информированная организация окажется в лучшем положении. Плохие новости так или иначе просачиваются в виде слухов.

Контроль расходов. В § 31.1 мы упоминали об оценке расходов на разработку перед началом разработки. По мере продвижения разработки возникают еще две проблемы в связи с расходами: расходы должны удерживаться в разумных границах и руководство должно знать их величину. Сокращением расходов, конечно, нельзя пренебрегать, но проведение разработок требует значительной свободы и может лишиться всякого смысла из-за скупости. Вероятно, лучшим решением этой проблемы является система выделения лимитов, когда на более высоких уровнях сохраняется лишь право вето; иначе говоря, каждый ведущий инженер может делать без специального разрешения любые расходы, которые он считает необходимыми (до некоторой максимальной цифры), но все его расходы просматриваются (либо его начальником, либо — в организации, изображенной на рис. 31.1,*д*,— проектным бюро), и ему задаются вопросы или налагается вето, если это нужно.

Необходимо также вести запись расходов, чтобы справедливо распределять средства между проектами, выполнять договорные условия и осуществлять действенный административный контроль. Для этой цели нужно иметь систему учета и вести точные записи всех трат. Однако эту нудную работу нужно свести к минимуму, приспособив методы учета к нуждам инженеров, а не к нуждам бухгалтеров. Действительно [123], «нет никаких причин, по которым принципы учета *должны* обязательно оказывать значительное влияние на практику контроля, применяемую в лаборатории. Иногда какой-нибудь метод рекомендуется на том единственном основании, что он согласуется с общепринятыми принципами учета. Такой довод лишь в редких случаях бывает справедливым. Общепринятых принципов учета немного, и их формулировка настолько широка, что обычно они не дают основы для ответа на вопросы, относящиеся к стоимости исследования».

Принятие решений. Важной обязанностью бригады проектирования системы является принятие критических решений, подобных перечисленным в § 3.3 (аналоговая или цифровая система и т. д.). Не следует принимать решение до того, как это становится необходимым. До последнего момента имеется возможность получения добавочной информации.

которую можно положить в основу выбора, или открытия новой альтернативы, или появления новой разработки, которая приведет к тому, что решение потеряет смысл — либо вследствие изменения требований, либо вследствие нового изобретения. Но задержка решения сверх необходимого времени губительна. Это приводит к тому, что решение откладывается со дня на день, и к неопределенности. Если решение можно немного задержать, то почему его не задержать еще дольше? Зачем вообще принимать решение? И задание не будет никогда выполнено!

Людям, а возможно, и животным биологически свойственно бояться решений. Но это отвращение необоснованно. Во-первых, принимающий решение должен утешаться тем, что задача проектирования системы допускает много решений. Будет ли данное решение жизненным, зависит в значительной степени от выполнения сделанного выбора и в меньшей степени — от выбранной альтернативы, так как предложенные альтернативы, по-видимому, были достаточно разумны, если они рассматривались прежде всего. Отсюда принимающий решение приходит к другой утешительной мысли. Если всей имеющейся информации выбор был труден, то альтернативы, видимо, не очень далеки друг от друга по своей выгодности, так как если бы одна из них была значительно хуже, то выбор был бы очевиден. Ударение на словах «при всей имеющейся информации» напоминает, что плохой выбор может вызываться пренебрежением какими-то доступными фактами.

Число необходимых выборов можно уменьшить на всех уровнях инженерного проектирования, если с самого начала потратить некоторое время на то, чтобы оформить и сделать ясной всем участникам идею выбранного проекта. Все последующие выборы при проектировании будут либо соответствовать выбранной идее, либо дополнять ее (если выбор не указывался), либо изменять ее — что уже связано с вопросами более высокой технической политики. Нерешительность и колебания при проектировании — признак того, что системная группа не имеет определенной ориентации. Такая группа никогда не разберется в системе.

Окончание. Истинное понимание системы будет достигнуто тогда, когда может быть предсказано влияние изменения одной переменной на все другие переменные системы. Это можно осуществить, когда полностью разработана математическая модель системы и известны все функциональные зависимости и значения всех параметров.

Например, в системах зенитных управляемых реактивных снарядов в число используемых переменных могут входить: дальность обнаружения цели; время ответа системы вплоть до выпуска снаряда; ошибки при управлении движением снаряда до дальности самонаведения; дальность самонаведения; радиус поражения боевой головки. Во всякой конкретной системе можно вывести точную зависимость между изменением любой одной из этих переменных и всеми остальными. Увеличение радиуса поражения приводит к менее строгим требованиям к самонаведению или к менее строгим требованиям к наведению на маршевом участке, к увеличению допустимого времени ответа системы или к тому, что можно будет допустить более позднее обнаружение цели,— все это при неизменных боевых качествах системы.

С другой стороны, в области уличного движения система настолько плохо изучена, что нельзя предсказать влияние, которое окажет на движение по перекрестку изменение цикла светофора, отстоящего на три квартала. Как было сказано раньше, неясно даже, приведет ли улучшение движения на одном перекрестке к улучшению всей системы. Когда можно делать такие предсказания, система является вполне изученной.

Когда система изучена, иногда быстро проводят «ударную» программу для достижения какой-либо ограниченной цели, например для завершения фазы предварительного проектирования. Хотя, возможно, проектная группа будет склонна считать, что ей нехватает времени для обдумывания и что решения принимаются экспромтом, в излишней спешке, тем не менее общий результат часто оказывается превосходным. Ударная программа после периода повседневного проектирования иногда способствует консолидации идей группы. Решения в действительности были негласно приняты, группа закончила обдумывание, и ударная программа приводит к письменному воплощению идей группы.

Процесс проектирования состоит в выборе групп компонентов, анализе этих групп и последующей их оценке в свете выбранных критериев эффективности. Затем выбор оборудования видоизменяется и цепь опять замыкается через те же этапы. По-видимому, при наличии у проектировщиков опыта проверка эффективности будет показывать, что с каждым изменением система совершенствуется. Очевидно, этому совершенствованию по существу нет конца, так как всегда можно будет внести какое-нибудь небольшое усовершенствование. Как обычно бывает в таких слу-

чаях, решение о моменте, когда нужно остановиться, зависит от того, как велика разница между последней оценкой и предыдущей. Когда эта разница становится достаточно мала, достигается «оптимум». Это весьма сходно с процессом итерации в вычислениях, когда исследуют ошибку на каждом этапе и прекращают вычисления, коль скоро ошибка становится меньше некоторого заранее заданного значения.

Какая разница задается как «достаточно малая», зависит от срочности выполнения системы. Но во всяком случае инженер-системотехник должен реалистически учитывать стоимость ошибки II рода — продолжать вводить усовершенствования и добавления после того, как работу следовало бы закончить. То же относится и к книге о системотехнике.

ЛИТЕРАТУРА

В книгах по «экономике» рассматриваются главным образом модели экономических систем с целью предсказания результатов изменения таких входных параметров, как *спрос* и *полезность*. Книги по «инженерной экономике» (хорошим образцом которых является Терборг [116]) посвящены минимизации стоимостей, но по своему изложению они обычно мало пригодны для инженера-системотехника. Книги по «техническому управлению предприятиями» (хороший образец — Кэнфилд и Боумен [124]) охватывают промежуточную область между техникой, с одной стороны, и вопросами права, трудовых отношений, правительственного контроля, учета, страхования и банковского дела, с другой стороны. Антони [123] дает тонкий анализ методов, которые применялись для руководства исследовательскими и проектными работами (хотя и без специального упора на проектирование систем). Келли [119] служит хорошим описанием того, как осуществляется руководство большой исследовательской и проектной организацией, участвующей в проектировании нескольких больших систем.

ЦИТИРОВАННАЯ ЛИТЕРАТУРА

1. *Agriculture Outlook Charts.* U. S. Department of Agriculture, 1955.

2. van Gorder H. F. et al. A new approach to office mechanization: Integrated data processing through common language machines. American Management Association, New York, 1954.

3. Haselton M. L. and Schmidt E. L. Automatic inventory system for air travel reservations. *Elec. Eng.*, 1954, July, p. 641—646.

4. Ewing D. H. et al., Teleran. *RCA Rev.*, 1946, v. 7, Dec., p. 601—621; 1947, v. 8, Dec., p. 612—632.

5. Bruce J. A. and Rudden J. B. Denver's traffic control system electronically moves auto flow. *Western City*, 1953, Nov.

6. Korn F. A. and Ferguson J. G. Number 5 Crossbar dial telephone switching system. *Trans. AIEE*, 1950, v. 69, p. 244—254.

7. Meszar J. Fundamentals of the AMA system. *Trans. AIEE*, 1950, v. 69, p. 255—269.

8. Reports on Signal Corps contract № DA-36-039-sc-30250 with General Electric Company.

9. Project Tinkertoy. *Natl. Bur. Standards (U. S.) Tech. News Bull.*, 1953, v. 37, № 11, p. 161—170.

10. Osborn R. F. G. E. and UNIVAC: Harnessing the high-speed computer, *Harvard Bus. Rev.*, 1954, v. 32, № 4, July, p. 99—107.

11. Morse P. M. and Kimball G. E. Methods of operations research. John Willey and Sons, Inc., New York, 1951.

Русск. пер.: Морз Ф. М. и Кимбелл Д. Е. Методы исследования операций», «Советское радио», 1956.

12. Ackoff R. L., Production scheduling: operations research case study. *Advanced Management*, 1955, v. 20, March, p. 21—28.

13. Zipf G. K. The hypothesis of the «Minimum Equation» as a unifying social principle: with attempted synthesis. *Am. Sociol. Rev.*, 1947, p. 627—650.

14. Stewart J. Q. Empirical mathematical rules concerning the distribution and equlibrium of population. *Geog. Rev.*, 1947, p. 461—485.

15. *Operations Research Group Study* 1. Office of Naval Research. December, 1953.

16. Trautman D. L. et al., Analysis and simulation of traffic flow. *University of California, Institute of Transportation and Traffic Engineering Research Rept.* 20, 1954.

17. Kolmogoroff A. Grundbegriffe der Wahrscheinlichkeitsrechnung, Berlin, 1933.

Рус. пер.: Колмогоров А. Н. Основные понятия теории вероятностей, ОНТИ, 1936.

18. Reichenbach H. The Theory of Probability, Berkeley, Calif., 1949.

Нем. изд.: Wahrscheinlichkeitslehre. Sijthoff, Leiden, Niederlande, 1934.

19. von Mises R. Probability, statistics and truth, New, York, 1939.

Нем. изд.: Wahrscheinlichkeitslehre, Statistik und Wahrheit, Springer, Wien, 1928.

Рус. пер.: Мизес Р. Вероятность и статистика, Госиздат, 1930.

20. Carnap R. Logical foundations of probability, Chicago, 1950.

21. Jeffreys. H. Theory of probability. Oxford, 1939.

22. Keynes J. M. A. treatise on probability, London, 1921.

23. Koopman B. O. The basis of probability. *Bull. Am. Math Soc.*, 1950, v. 46, p. 763—774.

24. Wald A. Sequential analysis. John Wiley and Sons, Inc., New York, 1947.

Рус. пер.: Вальд А. Последовательный анализ, Физматгиз, М., 1960.

25. Dodge H. F. and Romig H. G. A method of sampling inspection. *Bell System Tech. J.*, 1929, v. 8, p. 613—631.

26. Wald A. Sequential tests of statistical hypotheses. *Ann. Math. Statist.*, 1945, v. 16, June.

27. Wald A., Sequential method of sampling for deciding between two courses of action. *J. Am. Statist. Assoc.*, 1945, v. 40, p. 277—306.

28. Studies of methods for achieving reliability of guided missiles, *Naval Air Missile Test Center, Pt. Mugu Tech. Rept.* 75, 1950.

29. Kendall M. G., The advanced theory of statistics, 2d ed., vols I and II. Charles Griffin and Co., Ltd., London, 1945, 1948.

29a. Kendall M. G. The advanced theory of statistics, v. I, p. 380. Charles Griffin and Co., Ltd., London, 1945.

29b. Kendall M. G., The advanced theory of statistics, v. I, p. 246, Charles Griffin and Co., Ltd., London, 1945.

30. Whittaker E. T. and Watson G. N. A course of modern analysis, p. 117. The Macmillan Company, New York, 1946.

31. The Computer Issue. *Proc. IRE*, 1953, v. 41, № 10.

31a. Rubinoff M. Analog vs. digital computers—a comparison. *Proc. IRE*, 1953, v. 41, p. 1254—1262.

32. Forrester J. and Vance A. Institute of Radio Engineers metting, New York, 1951.

33. Fry T. C. Probability and its engineering uses. D. Van Nostrand Company, Inc., Princeton, N. J., 1928. Рус. пер.: Фрай. Теория вероятностей для инженеров, Гостехиздат, 1934.

34. Salveson M. E. The assembly line balancing problem. *J. Ind. Eng.*, 1955, v. 6, № 3, May—June, p. 18—25.

35. Morse P. M. *Phys. Today*, 1955, v. 8, № 9, September, p. 14.

36. McCloskey J. F. and Trefethen F. N. Operations research for management. Johns Hopkins Press, Baltimore, 1954.

37. Kempthorne O. The design and analysis of experiments. John Wiley and Sons, Inc. New York, 1952.

38. Von Neumann J. and Morgenstern O. Theory of games and economic bechavior. Princeton University Press, Princeton, N. J., 1947.

39. Aitken A. C. On the graduation of data by the orthogonal polynomials of least squares. *Proc. Roy. Soc. Edinburgh*, 1933, v. 53, p. 54.

40. Shannon C. The mathematical theory of communication. *Bell System Tech. J.*, 1948, July and October. reprinted in book form, University of Illinois Press, Urbana, Ill., 1949. Рус. пер. (частичный): Шэннон К., Статистическая теория передачи сигналов. В сб. «Теория передачи электрических сигналов при наличии помехи». Изд-во иностранной литературы, 1953.

41. An outline plan for modernizing USAF logistics, p. 81. United States Air Force Air Material Command. Research and Planning Office, Logistical Research Systems, 1955.

42. Green L. Probability confidence belts in weight estimation. *University of Michigan, Engineering Research Institute Rept.* 2063-5-J,1953.

43. Cramer H. Mathematical methods of statistics. Princeton University Press, Princeton, N. J., 1946. Рус. пер.: Крамер Г. «Математические методы статистики». Изд-во иностранной литературы, 1947.

44. Uspensky J. V. Introduction to mathematical probability. McGraw-Hill Book Company, Inc., New York, 1937.

45. Luce R. D. et al. Information flow in task oriented groups. *Massachusetts Institute of Technology, Lincoln Laboratories Tech. Rept.* 264, 1953.

46. Wilson E. B., Jr. An introduction to scientific research. McGraw-Hill Book Company, Inc., New York, 1952.

47. Fisher R. A. The design of experiments, 6th ed. Hafner Publishing Company, New York, 1951.

48. Mood A. McF. Introduction to the theory of statistics. McGraw-Hill Book Company, Inc., New York, 1957.

49. Dixon W. J. and Massey F. J., Jr. Introduction to statistical analysis. 2d McGraw-Hill Book Company, Inc., New York, 1957.

50. Korn G. A. and Korn T. M. Electronic analog computers. 2d ed., McGraw-Hill Book Company, Inc., New York, 1956. Рус. пер.: Корн Г. А. и Корн Т. М. Электронные моделирующие устройства. Под ред. Б. Я. Когана. Изд-во иностранной литературы, 1955.

51. Soroka W. W. Analog methods in computation and simulation. McGraw-Hill Book Company, Inc., New York, 1954.

52. Shannon C. J. *Math. and Phys.*, 1941, v. 20, p. 337—354.

53. Wiener N. Cybernetics. John Wiley and Sons, Inc., New York, 1948. Рус. пер.: Винер Н. Кибернетика. «Советское радио», 1953.

54. Hunt M. M. Bell Labs. 230 long-range planners. *Fortune*, 1954, May, p. 120.

55. Drucker P. F. The promise of automation. *Harper's Magazine*, 1955, April.

56. Engineering Research Associates. High-speed computing devices. McGraw-Hill Book Company, Inc. New York, 1950. Рус. пер.: «Быстродействующие вычислительные машины». Пер. под ред. Ю. Л. Панова, Изд-во иностранной литературы, 1952.

57. Richards R. K. Arithmetic operations in digital computers. D. Van Nostrand Company, Inc., Princeton, N. J., 1955. Рус. пер.: Ричардс Р. К. Арифметические операции на цифровых вычислительных машинах. Изд-во иностранной литературы, 1957.

58. Martin N. M., On completeness of decision element sets. *J. Comput. Systems*, 1953, v. 1, p. 150—154.

59. De Hartog J. P. Mechanical vibrations, 3d ed., sec. 24, McGraw-Hill Book Company, Inc., New York, 1947.

60. Levy M., Automation in post offices, *Proc. 11th Nat. Electronics Conf.* Chicago, Oct. 3—5, 1955.

61. Final report on indoor warning devices for civil defense *University of Michigan, Engineering Research Institute Rept.* 2369—4—F, 1955.

62. Willams J. D. The compleat strategyst. A Rand Corporation research study, McGraw—Hill Book Company, Inc., New York, 1954. Рус. пер.: Вильямс Дж. Д. Совершенный стратег. «Советское радио», 1960.

63. McKinsey J. Introduction to the theory of games. A Rand Corporation research study. McGraw—Hill Book Company, Inc., New York., 1954. Руск. пер.: Мак-Кинси Дж. Введение в теорию игр. Физматгиз, 1960.

64. Charnes A., Cooper W. W. and Henderson A. An introduction to linear programming. John Wiley and Sons, Inc., New York, 1953.

65. Isaacs R. Optimal horse race bets. *Am. Math. Monthly*, 1953, v. 60, p. 310.

66. Lewin K. Principles of topological psychology. McGraw-Hill Book Company, Inc., New York, 1936.

67. Bavelas A., A mathematical model for group structures. *Appl. Anthrop.*, 1948, v. 7, p. 16—30.

68. Leavitt H. J., Some effects of certain communication patterns on group performance. *J. Abnormal Social Psychol.*, 1951, v. 46, p. 38—50.

69a. Dinneen G. P. Programming pattern recognition.* *Proc. Western Joint Computed Conf.*, Los Angeles, March, 1955.

69b. Selfridge O. G. Pattern recognition and modern computers*. *Proc. Western Joint Computer Conf.*, Los Angels, March, 1955.

70. Weil H. Reduction of runs in multiparameter computations. *J. Assoc. Comput; Mach.*, 1955, v. 2, № 2, April, p. 99—110.

71. Goode H. H. Simulation—its place in system design. *Proc. IRE*, 1951, v. 39, p. 1501—1506.

72. Project Cyclone Symposium I on Reac Techniques, Port Washington, N. Y., Mar. 15—16, 1951, SDC, USN.

73. Project Cyclone Symposium II on simulation and computing techniques, Port Washington, N. Y., Apr. 28 May 2, 1952, SDC and BuAer, USN.

74. Symposium III on simulation and computing techniques, Port Washington, N. Y., Oct. 12—14, 1953, BuAer and NADC, USN.

75. *Proc. Natl. Simulation Conf.*, sponsored by Institute of Radio Engineers, Dallas, Jan. 19—21, 1956.

76. Chapanis A., Garner W. R. and Morgan C. T. Applied experimental psychology. John Wiley and Sons, Inc., New York, 1949.

77. Handbook of human engineering data. 2d ed., Tufts College, Institute of Applied Experimental Psychology; *U. S. Navy, Special Devices Center (Navexos* P-643) *Tech. Rept.* SDC-199-1-2.

78. Feller W. Probability theory and its applications, vol. 1, chap. 17. John Wiley and Sons. Inc., New York, 1950. Рус. пер.: Феллер В. Введение в теорию

* Исследование по этим документам финансировалось совместно армией, ВМФ и ВВС по контракту с Массачусетсским технологическим институтом. — *Прим. авт.*

вероятностей и ее приложение, Изд-во иностранной литературы, 1952.

79. Bartlett M. S. An introduction to stochastic processes, sec. 4. 21. Cambridge University Press, London, 1955. Рус. пер.: Бартлетт М. С. Введение в теорию случайных процессов. Изд-во иностранной литературы, 1958.

80. Kendall D. G. Some problems in the theory of queues. *J. Roy. Statist. Soc.*, 1951, v. 13, № 2, p. 151—185.

81. Lindley D. V. Theory of Queues with a Single Server. *Proc. Cambridge Phil. Soc.* 1952, v. 48, № 2, April, p. 277—289.

82. Gilchrist B., Pomerene J. H. and Wong S. Y. Fast carry logic for digital computers, *Trans. IRE, Professional Group on Electronic Computers*, 1955, v. EC-4, № 4, December.

83. Alder R. B. and Fricker S. J. The flow of scheduled air traffic, parts 1 and 2, *Massachusetts Institute of Technology, Research Laboratory for Electronics Tech. Repts.* 198, 199, May, 1951.

84. Goode H. H., Wright J. and Polmar C. Use of a digital computer to model a signalized intersection. Highway Research Board, Washington, D. C., January, 1956.

85. Johnson M. H. *J. Roy Statist. Soc.*, 1951, v. 13, p. 178.

86. Jackson W. (ed.) Communication theory. Academic Press, Inc., New York, 1953.

87. Shannon C. Communication in the presence of noise. *Proc. IRE*, 1949, v. 37, pp. 10—21, January, 1949.

88. Ridenour L. N. (ed.). Modern physics for the engineer. McGraw-Hill Book Company, Inc., New York, 1954.

89. Woodward P. M. Probability and information theory, with applications to radar. McGraw-Hill Book Company, Inc., New York, 1953.

Рус. пер.: Вудворд П. М., Теория вероятностей и теория информации с применениями в радиолокации. «Советское радио», 1953.

90. Goldman S. Information theory. Prentice—Hall, Inc., Englewood Cliffs, N. J., 1953.

Рус. пер.: Голдман С. Теория информации. Изд-во иностранной литературы, 1957.

91. Blackwell D. and Girshick M. A. Theory of games and statistical decisions. John Wiley and Sons, Inc., New York, 1954.

Рус. пер.: Блекуэлл Д. и Гиршик М. А. Теория игр и статистических решений. Изд-во иностранной литературы, 1958.

92. Koopmans T. J. (ed.) Activity analysis of production and allocation. John Wiley and Sons, Inc., New York, 1951.

93. Visual presentation of information. *Wright Air Development Center, Aero Medical Laboratory, WADC Tech. Rept.* 54-160, ASTIA № AD 43064.

94. Thaler G. J. Elements of servomechanism theory. McGraw-Hill Book Company, Inc., New York, 1955.

95. Brown G. S. and Campbell D. P. Principles of servomechanisms. John Wiley & Sons, Inc., New York, 1948.

96. Cage J. M. Theory and application of industrial electronics. McGraw-Hill Book Company, Inc., New York, 1951.

97. Truxal J. G. Automatic feedback control system synthesis. McGraw-Hill Book Company, Inc., New York, 1955.

Рус. пер.: Траксел Дж. Синтез систем автоматического регулирования. Машгиз, 1959.

98. Lawson J. L. and Uhlenbeck G. E. Threshold signals, vol. 24. Massachusetts Institute of Technology Radiation Laboratory Series. McGraw-Hill Book Company, Inc., New York, 1950.

Рус. пер.: Пороговые сигналы, «Советское радио», 1952.

99. Yule G. U. and Kendall M. G. An introduction to the theory of statictics. Charles Griffin and Co. Ltd, London, 1940.

Рус. пер.: Юл. Дж. Э. и Кендэл М. Дж. Теория статистики, Госстатиздат, 1960.

100. Scheffé H. Theory of probability. *National Defense Research Committee, Arm. and Ord. Rept.* A-224 (OSRD 1918), Div. 2.

101. Fisher R. A. Statistical methods for research workers. Oliver and Boyd, Ltd., Edinburg and London, 1938.

Рус. пер.: Фишер Р. А. Статистические методы для исследователей. Госстатиздат, 1958.

102. Turner W. O. Estimation of requirements in dial telephone central offices. *Proc. Operations Research Conf.* Case Institute of Technology, Cleveland, January, 1958.

103. Forrester J. W. Digital information storage in three dimensions using magnetic cores. *J. Appl. Phys.*, 1951, v. 22, January, p. 44—48.

104. Tocher K. D. *J. Roy. Statist. Soc.*, 1951, v. 13, № 2, p. 181.

105. Ridenour L. N. (ed.) Radar system engineering, vol. I. Massachusetts Institute of Technology Radiation Laboratory Series, McGraw-Hill Book Company, Inc., New York, 1947.

Рус. пер.: «Радиолокационная техника», «Советское радио», 1949.

106. Hopper G. M. Automatic coding for digital computers, talk at Louisiana State University, Feb. 16, 1955; published by Remington-Rand.

107. Encyclopaedia Britannica, Statistical Abstracts, United Statistical Yearbook, etc.

108. Sloan L. L. Rate of dark adaptation and reginal threshold gradient of the dark adapted eye: Psychologie and chemical studies. *Am. J. Ophthalmol.*, 1947, v. 30, p. 705—720.

109. Luckiesch M. Light, vision and seeing, in O. Glasser (ed.). Medical Physics. Year Book Publishers, Inc., Chicago, 1944.

110. Sleight R. B. The effect of instrument dial shape on legibility. *J. Appl. Psychol.*, 1948, v. 32, p. 170—188.

111. Fletcher H. and Munson W. A. Loudness, its definition, measurement and calculation. *J. Acoust. Soc. Amer.*, 1953, v. 5, p. 91—108.

112. Fletcher H. Speech and hearing. D. Van Nostrand Company, Inc., Princeton, N. J., 1929.

113. James H., Nichols N. and Phillips R. Theory of servomechanisms, vol. 25. Massachusetts Institute of Technology Radiation Laboratory Series, McGraw-Hill Book Company, Inc., New York, 1947.

Рус. пер.: Джеймс Х., Никольс Н., Филлипс Р. Теория следящих систем. Изд-во иностранной литературы, 1951.

114. Terman F. E. Electronic and radio engineering. McGraw-Hill Book Company, Inc., New York, 1955.

115. Thaler G. J. and Brown R. G. Servomechanism analysis. McGraw-Hill Book Company, Inc., New York, 1953.

116. Terborgh G. Dynamic equipment policy. McGraw-Hill Book Company, Inc., New York, 1949.

117. Industrial Bulletin. Arthur D. Little Company, Cambridge, Mass., May, 1954.

118. Scientific research and development in American industry, *U. S. Dept. Labor. Bull.* 1148, 1953.

119. Kelly M. J. Research and development in en-

gineering management in the electronic industry, talk at Institute of Radio Engineers, meeting, March, 1953.

120. Brillouin L. Maxwell's demon cannot operate; information and entropy I. *J. Appl. Phys.*, 1951, v. 222, p. 334—337.

121. Hawkins J. E. and Stevens S. S. The masking of pure tones and speech by white noise. *J. Acoust. Soc. Amer.*, 1950, v. 22, p. 6—13.

122. Porter F. J. Computers in basic business applications. *Proc. Eastern Joint Computer Conf.*, Boston, Nov. 7—9, 1955.

123. Anthony R. N. Management controls in industrial research organisations. Harvard University Graduate School of Business Administration, 1952.

124. Canfield D. T. and Bowman J. M. Business, legal and ethical phases of engineering. 2d ed. McGraw-Hill Book Company, Inc., New York, 1954.

125. Larrowe V. L. Direct simulation. *Control Eng.*, November, 1954, p. 24—31.

126. Carr J. W. and Scott N. R. Digital computers and data processing. University of Michigan, College of Engineering, 1955.

127. Greenshields B. D. Traffic performance at urban street intersections. Yale University, Bureau of Highway Traffic, 1947.

128. Malcolm D. G. Queueing theory in organisation design. *J. Ind. Eng.*, November—December, 1955, p. 19—26.

129. Thorndike F. Applications of Poisson's probability summation. *Bell System Tech. J.*, 1926, v. 5, p. 610ff.

130. Kendall D. G. Stochastic processes occurring in the theory of queues and their analysis by the method of the imbedded Markov chain. *Ann. Math. Statist.*, 1953, v. 24, № 3, September, p. 338.

131. Merz W. J. and Anderson J. R. Ferroelectric storage devices. *Bell Lab. Record*, 1955, September.

132. Paynter H. M. (ed.) A palimpsest on the electronic analog art. G. A. Philbrick, Boston, 1955.

133. Shannon C. E. Prediction and entropy of printed English. *Bell System Tech. J.*, 1951, v. 30, p. 50—64.

134. Clarke A. B. A waiting line process of Markov type. *Ann. Math. Statist.*, 1956, v. 27, June, p. 452—459.

135. Walquist R. L. Analysis and design of a digital-to-analog decoder, thesis for M. S. in electrical engineering. Massachusetts Institute of Technology, 1951.

136a. «What is GCA?». Bendix Radio Division of Bendix Aviation Corporation. Baltimore, 1955.

136b. «Bendix GCA System, ASR-3 (PAR-2)». Bendix International. Division of Bendix Aviation Corporation, New York, 1955.

137. The air traffic story. *Radio Tech. Commission for Aeronautics Paper* 194-52/DO-47, Washington, D. C., December, 1952.

138. Benaglio R. V. Description of CP-74 computer-recorder. Bendix Aviation Corporation, Detroit, February, 1956 (unpublished).

139. Stevens S. S. (ed.) Handbook of experimental psychology. John Wiley and Sons, Inc., New York, 1951.

140. Clark A. B. et al. Automatic switching for nation-wide telephone service, *Bell Telephone System Monograph* 2015, 1952.

141. Thrall R. M., Coombs C. H. and Davis R. L. Decision processes. John Wiley and Sons, Inc., New York, 1954.

142. Keister W., Ritchie A. E. and Washburn S. H. Design of switching circuits. D. Van Nostrand Company, Inc., Princeton, N. J., 1951.

143. Kenney J. F. Mathematics of statistics. D. Van Nostrand Company, Inc., Princeton, N. J., 1939.

144. Laning J. H., Jr. and Battin R. H. Random processes in automatic control. McGraw-Hill Book Company, Inc., New York, 1956.

Рус. пер.: Лэнинг Д. Х. и Бэттин Р. Г. Случайные процессы в задачах автоматического управления. Изд-во иностранной литературы, 1958.

145. Carhart R. R. A survey of the current status of the electronic reliability problem. *Rand Memo 1131*, United States Air Force Project Rand, Santa Monica, Calif., Aug. 14, 1953.

ДОПОЛНИТЕЛЬНАЯ ЛИТЕРАТУРА

(Добавлено при переводе)

Д.1. Осипов Г. В. Техника и общественный прогресс. Изд-во АН СССР, 1959.

Д.2. Лилли С. Автоматизация и социальный прогресс. Изд-во иностранной литературы, 1958.

Д.3. Эшби У. Р. Введение в кибернетику. Изд-во иностранной литературы, 1959.

Д.4. Ивахненко А. Г. Техническая кибернетика. Гостехиздат УССР, Киев, 1959.

Д.5. «Кибернетику на службу коммунизму». Сб. статей под ред. А. И. Берга, Госэнергоиздат, 1961.

Д.6. Философские вопросы кибернетики. Соцэкгиз, 1961.

Д.7. Кельзон А. С. Динамические задачи кибернетики. Судпромгиз, 1959.

Д.8. Поваров Г. Н. Логика на службе автоматизации и технического прогресса. «Вопросы философии», 1959, № 10, стр. 48—62.

Д.9. Гнеденко Б. В. Курс теории вероятностей. Физматгиз, 1962.

Д.10. Гливенко В. И. Курс теории вероятностей. ГОНТИ, 1939.

Д.11. Ван дер Варден Б. Л. Математическая статистика. Изд-во иностранной литературы, 1960.

Д.12. Митропольский А. К. Техника статистических вычислений. Физматгиз, 1961.

Д.13. Хольд А. Математическая статистика с техническими приложениями. Изд-во иностранной литературы, 1956, гл. XX, § 7.

Д.14. Хинчин А. Я. Математические методы теории массового обслуживания. Труды математического института им. В. А. Стеклова АН СССР, 1955, т. 49.

Д.15. «Применение математики в экономических исследованиях». Сборник под ред. В. С. Немчинова, Соцэкгиз, 1959.

Д.16. Китов А. И., Криницкий Н. А. Электронные цифровые машины и программирование. Физматгиз, 1959.

Д.17. Автоматизация программирования. Под ред. А. П. Ершова, Физматгиз, 1961.

Д.18. Реймон Ф. Автоматика переработки информации. Физматгиз, 1961.

Д.19. Ричардс Р. К. Элементы и схемы цифровых вычислительных машин. Изд-во иностранной литературы, 1961.

Д.20. Ликлайдер. Симбиоз человека и машины. «Зарубежная радиоэлектроника», 1960, № 9, стр. 84—96.

Д.21. Клейн М. Л., Морган Г. С., Аронсон М. Г. Цифровая техника для вычислений и управления. Изд-во иностранной литературы, 1960.

Д.22. «Применение вычислительной техники для автоматизации производства». Сборник под ред. В. В. Солодовникова, Машгиз, 1961.

Д.23. Begun C. J. Magnetic memory device for

business machines. *Electrical Engineering*, 1955, v. 74, № 6, p. 466—488.

24. Иньков Ю. И. Радиоэлектроника на службе военных монополий США. Соцэкгиз, 1962.

Д.25. Харкевич А. А. Очерки общей теории связи. Гостехиздат, 1955.

Д.26. «Надежность наземного радиоэлектронного оборудования». Сборник переводов под ред. Н. М. Шулейкина, «Советское радио», 1957.

Д.27. Кармазов М. Г. АТС системы Кроссбар. Инф. сборник «Техника связи». Связьиздат, 1947, июль, стр. 3—17.

Д.28. Метельский Г. Б. Координатные АТС. Связьиздат, 1961.

Д.29. Панкратов А. Г., Етрухин Н. Н. Оргасвязь. Связьиздат, 1960 (серия «Техника связи за рубежом»).

Д.30. Михайлов А. В. Метод гармонического анализа в теории регулирования. «Автоматика и телемеханика», 1938, № 3.

Д.31. Розенберг В. Я., Прохоров А. И. Что такое теория массового обслуживания. «Советское радио», 1962.

Д.32. «Применение теории игр в военном деле». Под ред. В. О. Ашкенази, «Советское радио», 1961.

РЕШЕНИЕ ЗАДАЧ

8.1. Расписания автобусов не случайные и в этой задаче имеют смещение. В частности, например, автобусы северного направления останавливаются каждый час точно в момент наступления этого часа; автобусы южного направления останавливаются каждый час в 18 мин. после наступления этого часа.

8.2.

$$P(\infty)=\frac{\beta}{1-\alpha+\beta}P(n)=$$

$$=(\alpha-\beta)^n P(0)+\frac{1-(\alpha-\beta)^n}{1-\alpha+\beta}\beta.$$

8.3. а) Вероятность отказа в момент T (т. е. между T и $T+dT$) определяется равенством (6.34). Вероятность того, что отказ произойдет в промежутке времени от 0 до T, определяется путем интегрирования этого выражения от 0 до T:

$$\int_0^T me^{-mt}\,dt = 1-e^{-mT},$$

а искомую вероятность находим, вычитая этот результат из единицы:

$$P = e^{-mT}.$$

б) Вероятность того, что данная лампа выйдет из строя в течение промежутка dt, равна mdt, а вероятность того, что в течение этого времени выйдет из строя хотя бы одна лампа, равна

$$1-(1-mdt)^k = 1-\left[1-kmdt+\right.$$

$$\left.+\frac{k(k-1)}{2}(mdt)^2-\ldots\right]\approx kmdt,$$

так как мы можем отбросить члены с $(dt)^2$ и более высокими степенями. Поэтому вероятность того, что приемник выйдет из строя в течение промежутка времени от t до $t+dt$, определяется следующим вариантом формулы (6.34):

$$p(t)\,dt = kme^{-mkt}\,dt.$$

Решая как в предыдущем случае, находим $P=e^{-kmt}$.

в) Вероятность того, что в выбранном случайно приемнике стоят лампы первого изготовителя, равна d_1; условная вероятность того, что такой приемник будет действовать в момент T, равна e^{-km_1T}; совместная вероятность этих двух событий равна $d_1e^{-km_1T}$, а общая вероятность того, что приемник будет действовать в момент T, равна $P=d_1e^{-km_1T}+d_2e^{-km_2T}$.

г) Искомая условная вероятность равна $e^{-rm_1T-(k-r)m_2T}$. Вероятность того, что приемник будет иметь r ламп первого типа и $k-r$ ламп второго типа, равна по формуле (5.2)

$$\frac{k}{r!(k-r)!}d_1^r d_2^{k-r}.$$

Совместная вероятность равна произведению этих двух вероятностей, а искомая общая вероятность получается путем суммирования по r от $r=0$ до $r=k$, так что r исключается:

$$P=\sum_{r=0}^{k}\frac{k!}{r!(k-r)!}d_1^r d_2^{k-r}e^{-rm_1T}e^{-(k-r)m_2T}=$$

$$=\sum_{r=0}^{k}\frac{k!}{r!(k-r)!}(d_1e^{-m_1T})^r(d_2e^{-m_2T})^{k-r}.$$

Это выражение, как можно видеть, есть биномиальное разложение выражения $(d_1e^{-m_1T}+d_2e^{-m_2T})^k$, которое и представляет искомую вероятность.

д) Вероятность того, что одна случайно выбранная лампа продолжает работать в данный момент, равна $d_1e^{-m_1T}+d_2e^{-m_2T}$; вероятность того, что продолжают работать k ламп, равна этой величине, возведенной в k-ю степень.

е) При ответе на такой вопрос всегда полезно сначала рассмотреть вырожденные случаи. Здесь интересны два вырожденных случая. Если $m_1=m_2$, то интуитивно ясно, что при перемешивании и неперемешивании получаются одинаковые результаты, и, действительно, оба выражения приводят к формуле случая б). Если m_2 равно нулю, а m_1 равно бесконечности, то, очевидно, при отсутствии перемешивания d_1 приемников все еще действуют после любого момента, а при перемешивании число все еще действующих приемников значительно меньше (а именно, оно равно d_1^k—числу приемников, в которых все лампы будут 1-го класса). Вообще же

составляют отношение двух вероятностей [находимых соответственно из формул для случаев в) и г)]

$$\frac{d_1e^{-km_1T}+d_2e^{-km_2T}}{(d_1e^{-m_1T}+d_2e^{-m_2T})^k}=\frac{e^{-km_2T}(d_1e^{-k\delta T}+d_2)}{e^{-km_2T}(d_1e^{-\delta T}+d_2)^k}$$

и дифференцируют его по $\delta=m_1-m_2$. Если положить эту производную равной нулю, то получается уравнение, имеющее лишь одно решение: $\delta=0$, или $m_1=m_2$. Следовательно, это должен быть минимум или максимум, и отношение не имеет других экстремальных значений. Но при $\delta=0$ отношение равно 1, а при $\delta=\infty$ отношение больше 1; следовательно, это есть минимум. Таким образом, мы видим, что, разделив лампы, мы получим бо́льшую вероятность продолжения работы приемника, чем при перемешивании ламп,— для всех значений k ($k>1$), для всех значений T ($T>0$) и для всех значений m ($m_1\neq m_2$).

ж) Этот вопрос имеет ряд сторон, связанных с дополнительными расходами на сырье и производство при отсутствии перемешивания ламп и с вероятным доходом. Вообще говоря, m_1 и m_2 не будут в точности известны, но тем не менее могут существовать основания к предположению, что они заметно различаются. Записи качества работы приемников, изготовленных из ламп разных заводов, могут доставить ценные данные, которые можно использовать при организации снабжения.

8.4.

а) Из задачи 8.3б находим: $Pe^{-kmt}=e^{-0,6}=0,549$, $E(n)=np=549$.

б) $\sigma=(npq)^{1/2}=15,75$.

в) $P(500)=\dfrac{1000!}{500!500!}(0,549)^{500}(0,451)^{500}=2,1\times10^{-4}$.

г) При нормальном распределении 500 отличается от 549 на $49/15,75=3,11\sigma$. Вероятность отклонения $\pm3,11\sigma$ равна 0,0019. Следовательно, вероятность того, что работают менее 500 приемников, равна около 0,001, а вероятность того, что работают 500 или более приемников, равна 0,999.

8.5. $P(10)=0,383$, $P(9)=0,242$, $P(<9)=0,067$.

8.6.

$$P=\frac{10!}{1!2!3!4!}(0,067)(0,242)^2(0,383)^3(0,242)^4=0,00957.$$

8.7. Совместное распределение x и y определяется выражением

$$P(x,y)\,dxdy=\frac{1}{2\pi\sigma^2}e^{-(x^2+y^2)/2\sigma^2}\,dxdy,$$

причем

$$P(r,\theta)=P(x,y)\frac{\partial(x,y)}{\partial(r,\theta)}=rP(x,y).$$

Отсюда

$$P(r,\theta)\,drd\theta=\frac{r}{2\pi\sigma^2}e^{-r^2/2\sigma^2}\,drd\theta$$

и

$$P(r)\,dr=\int_0^{2\pi}P(r,\theta)\,dr\,d\theta=\frac{r}{\sigma^2}e^{-r^2/2\sigma^2}\,dr.$$

В нашем примере $\sigma=100$ футам. Указанное распределение переменной r называется *распределением Рэлея*; распределение переменной r^2 есть распределение отношения хи-квадрат с двумя степенями свободы. Распределение хи-квадрата с n степенями свободы можно вывести таким же образом, найдя распределение переменной

$$r^2=\sum_{i=1}^{n}x_i^2.$$

8.8.

$$\left(1+\frac{t^2}{\nu}\right)^{(\nu+1)/2}\to e^{t^2/2};$$

$$\frac{\Gamma(\nu+1)/2}{\Gamma(\nu/2)}\to\sqrt{\frac{\nu}{2}}.$$

8.9.

$$n=145.$$

8.10. Подставив в уравнение (7.27) $\nu=1$ и используя преобразование $d(\chi^2)=2\chi d\chi$, находим $P(\chi)=2N(0,1)$. Коэффициент 2 компенсирует в нулевой момент отсутствующие отрицательные значения.

8.11. Нам нужно найти вероятность того, что абсолютная величина разности x_1-x_2 превосходит $k\sigma$, где x_1 и x_2 — независимые наблюдения функции $N(\mu,\sigma)$. Распределение разности x_1-x_2 равно на основании (7.22) $N(0,\sqrt{2}\sigma)$.

Отсюда

$$P(|x_1-x_2|\geqslant k\sigma)=2\int_{k\sigma}^{\infty}N(0,\sqrt{2}\sigma)\,dx=$$

$$=2\int_{k/\sqrt{2}}^{\infty}N(0,1)\,dy.$$

Эту величину можно найти в таблицах. При $k=1$, $P=0,480$.

8.12. Нет. Вероятность того, что будет выбран n-й прибор аппаратуры, равна 1/100, так что математическое ожидание числа совпадений равно в точности 1. (Вероятность точно n совпадений равна $1/n!\,e$. Читатель, возможно, захочет это доказать. Вывод, хотя и трудный, не связан с новыми принципами; он приводится в [44]).

12.1.

а) $P(A)=1/3$, $P(B)=16/27$, $P(C)=2/27$.

б) $P(A)=1/3$, $P(b)=16/27$, $P(C)=2/27$.

Отметим, что $P(i)$ должно равняться $P(j)$ если брать любые пары из бесконечной последовательности.

в)

i \ j	A	B	C
A	0	4/5	1/5
B	1/2	1/2	0
C	1/2	2/5	1/10

г)

i \ j	A	B	C
A	0	9/20	9/10
B	8/9	1/2	0
C	1/9	1/20	1/10

12.3.

а) Приблизительно 1 000 000 испытаний; б) $\pm6^2/_3\%$. Подробности см. на последней странице гл. 23.

12.4.

а) $m=113,1$.

б) $s=16,65$.

в) Конечно, на такой вопрос нельзя дать простой ответ. Однако мы можем применить к этим данным критерий χ^2. Так как мы оценили и математическое ожидание, и стандартное отклонение по данным, то число степеней свободы на три меньше, чем число классов, и еще раз объединять эти классы нецелесообразно. Вычисленный хи-квадрат равен 1,99, а при четырех степенях свободы и хи-квадрате, меньшем чем 7,78, мы не отвергли бы гипотезу о нормальном распределении даже при уровне 10%. Поэтому мы можем сказать, что эти данные нельзя считать несовместимыми с гипотезой о нормальном распределении.

12.5. Нет. Мы допустили за отсутствием других сведений, что обе генеральные совокупности нормальны. Вычисляя, находим, что F равно $s_1^2/s_2^2 = 17/29 = 0{,}586$. Мы ищем F с 14 и 11 степенями свободы в таблице для 5% и видим, что гипотезу можно отвергнуть только в случае, если $F > 2{,}74$ или $F < 0{,}391$.

12.6. В качестве независимой переменной удобно взять $Z = (x - 15)/5$. Тогда Z_i принимает значения —3, —2, —1, 0, 1, 2 и 3, а члены ΣZ_i и ΣZ_i^3 исчезают. Нормальные уравнения получаются из (12.24) при подстановке $a_0 = a$, $a_1 = b$, $a_2 = c$, $x_{i0} = 1$, $x_{i1} = Z_i$, $x_{i2} = Z_i^2$, когда k принимает значения 0, 1 и 2. Тогда получаются уравнения

$$a\Sigma 1 + b\Sigma Z + c\Sigma Z^2 = \Sigma Y = 7a + c\Sigma Z^2,$$

$$a\Sigma Z + b\Sigma Z^2 + c\Sigma Z^3 = \Sigma YZ = b\Sigma Z^2,$$

$$a\Sigma Z^2 + b\Sigma Z^3 + c\Sigma Z^4 = \Sigma YZ^2 = a\Sigma Z^2 + c\Sigma Z^4,$$

в которых все суммы берутся по $i = 1, 2, 3, 4, 5, 6, 7$.

Решение. $a = 322/3$, $b = 441/28$ и $c = 179/84$. Чтобы восстановить первоначальное уравнение, мы применяем подстановку $Z = (x - 15)/5$ и получаем

$$Y = 78{,}62 + 0{,}68x + 0{,}0824x^2.$$

12.7. а) Назовем нулевой гипотезой предположение, что напряжение равно 1000 *в*, а альтернативной — предположение, что напряжение равно 1020 *в*. Ошибка I рода (принятие плохой партии) может очень дорого стоить, если, скажем, плохие 60-центовые конденсаторы вызовут неисправность в 300-долларовых телевизорах. Ошибка II рода (браковка хорошей партии) может стоить очень дорого, если она вызывает задержку производства, раздражение поставщика и уплату за конденсаторы, испорченные при испытании. Увеличение n может стоить очень дорого, если мы должны уничтожить в испытании большую часть партии.

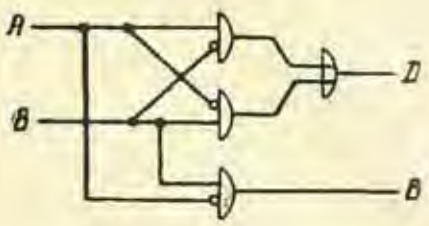
б) Принять, если наблюденное среднее значение $> \theta$ отвергнуть, если оно меньше θ. Значение θ зависит от выбранных значений α и β. Если $\alpha = \beta$, то $\theta = 1\,010$.

в) Задача этого типа не имеет подходящего ответа, за исключением уменьшения σ (путем более тщательного изготовления), уменьшения нулевой гипотезы (путем изменения схемы) или увеличения альтернативной гипотезы (путем покупки конденсаторов на большее напряжение). Например, если $\alpha = \beta = 0{,}026$ (точка 2σ), то $n = 400$. Читатель при желании может исследовать эту ситуацию, чтобы узнать, почему область неопределенности (20 *в*) так узка сравнительно со стандартным отклонением (100 *в*).

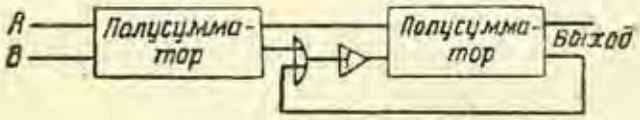
12.8. Отвергнуть после 21-го наблюдения.

15.1.

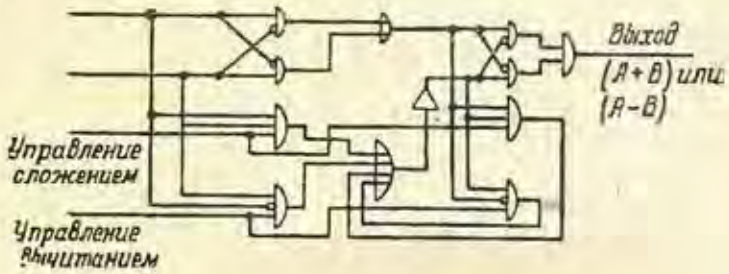
а) Полувычитатель:



б) Вычитатель:

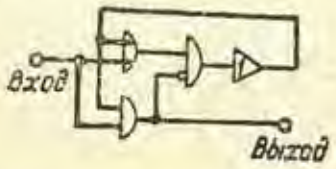


в) Комбинированный сумматор-вычитатель:

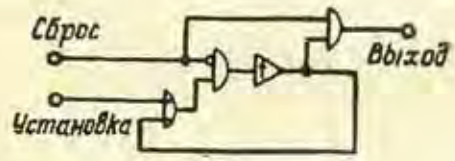


15.2.

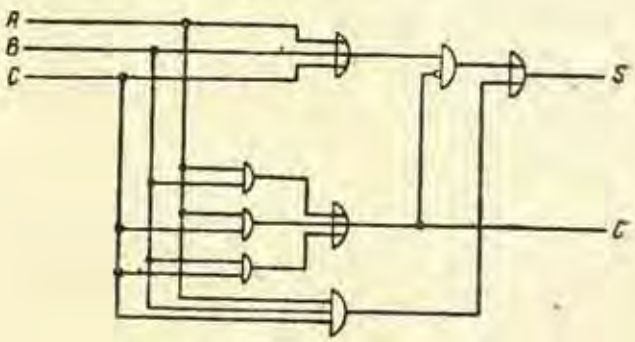
а) Триггер с одним входом:



б) Триггер с двумя входами:



15.3. Сумматор с тремя входами:

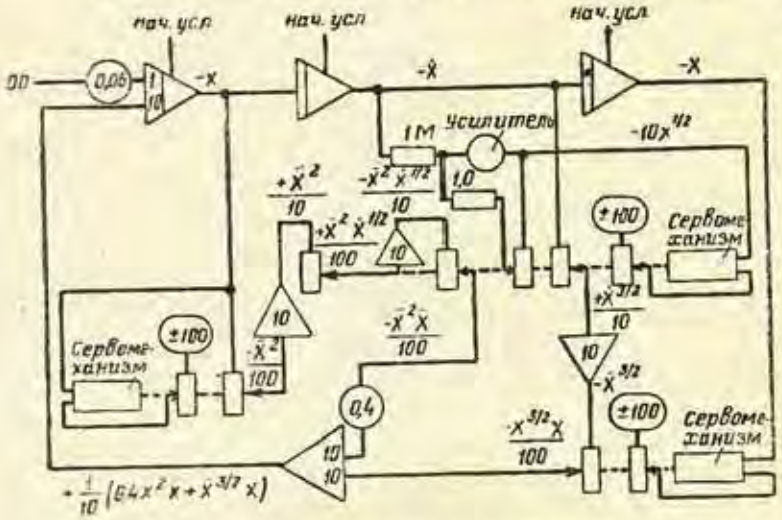


Заметим, что если между двумя проводами, обозначенными буквой C, включить линию задержки, то эту схему можно использовать как обычный сумматор с двумя входами.

16.3

Команда	α	β	γ	Операция	Примечание	Ячейка памяти	Содержимое
001	009	011	010	01	Фиксация y_1	009	x
002	009	010	014	06	Вычисление y_{k+1}; в 004 можно применить деление на 2 или умножение на 1/2, но при сдвиге тратится меньше машинного времени	010	y_k
003	014	010	014	01		011	Нуль
004	014	012	014	08		012	—1
005	014	010	015	02	Подготовка сравнения	013	2^{-30}
006	014	011	010	01	Фиксация y_{k+1}	014	Промежуточный результат
007	015	013	002	12	При использовании операции № 12 вместо № 15 устраняются вопросы о знаке числа в 015	015	$y_{k+1}-y_k$
008					Команда 008 — возврат к основной программе: искомый квадратный корень хранится в ячейке 010.		

18.1. Решение для $\ddot{x}=0{,}4\ddot{x}^2x+\dot{x}^{3/2}x+6$. Масштабные коэффициенты не указаны; указанные усиления просто компенсируют коэффициент 100, вводимый в каждое умножение. Присоединения концов потенциометров, если они не указаны, зависят от знаков числовых величин. Умножение $\ddot{x}^2$ на $\dot{x}$ было выполнено путем умножения дважды на $-10\ddot{x}^{1/2}$, так как эта величина имеется на выходе одного из сервомеханизмов.



23.1.

$$p(T)\,dT=\frac{1}{\sqrt{2\pi}}\,e^{-T/2}T^{-1/2}\,dT,$$

$$\mu_T=\int_{-\infty}^{\infty}Tp(T)\,dT=\frac{1}{\sqrt{2\pi}}\int_0^{\infty}T^{1/2}e^{-T/2}\,dt$$

(пределы интегрирования изменены, потому что T не может принимать отрицательных значений). Для вычисления интеграла подставляем $T/2=x$, $dT=2dx$:

$$\mu_T=\frac{2^{3/2}}{\sqrt{2\pi}}\int_0^{\infty}x^{1/2}e^{-x}dx=\frac{2}{\sqrt{\pi}}\,\Gamma\left(\frac{3}{2}\right)=1,$$

$$\sigma_T^2=\int_{-\infty}^{\infty}T^2p(T)\,dt-\mu^2=\frac{1}{\sqrt{2\pi}}\int_0^{\infty}T^{3/2}e^{-T/2}dT-1=$$

$$=\frac{2^{5/2}}{\sqrt{2\pi}}\int_0^{\infty}x^{3/2}e^{-x}dx-1=\frac{4}{\sqrt{\pi}}\,\Gamma\left(\frac{5}{2}\right)-1=3-1=2.$$

Отсюда $M=1/\mu_T=1$, $m=\rho M=1/2$.

Для экспоненциального распределения времени занятия получается $M=1$, $m=1/2$, $\sigma_T^2=1$. Из формулы (23.5) находим соответственно $E(n)=5/4$ и 1, из формулы (23.7) находим соответственно $E(w)=3/2$ и 1.

23.2.

а) Оценка по формулам (23.18) и (23.17) при $\rho=5/6$ дает следующие значения:

ν	$\frac{\rho^\nu}{\nu!\,(\nu-\rho)}$	$\sum_{n=0}^{\nu-1}\frac{\rho^n}{n!}$	$p(0)$	$p(1)$	$p(2)$	$p(3)$	$p(4)$	$\sum_{n=0}^{\nu}p(n)$
1	5	1	0,167	0,139	—	—	—	0,306
2	25/42	11/6	0,412	0,343	0,113	—	—	0,874
3	125/936	157/72	0,432	0,360	0,150	0,042	—	0,984
4	625/24 624	2951/1296	0,434	0,362	0,151	0,042	0,009	0,998

Следовательно, два канала достаточны, чтобы вероятность образования линии ожидания уменьшить до 0,126.

б) $\rho=2{,}08>\nu$. Следовательно, линия ожидания будет бесконечной длины.

24.1. Матрица игры имеет вид:

		Бандит	
		Врач	Ветеринар
Полиция	Врач	0	1/4
	Ветеринар	1/2	1/8

Стратегии таковы: полиция к врачу — с вероятностью 0,6; арестант к врачу — с вероятностью 0,2; вероятность спасения 0,2.

24.2. Цена игры, очевидно (по соображениям симметрии), равна 400 долларам. Матрица — порядка 6×9 и имеет четыре седловые точки; таким образом, каждый игрок имеет две чистые оптимальные стратегии и может предлагать 300 долларов или 400 долларов. Имеются некоторые соображения за то, что предложить 300 долларов является «лучшей» стратегией (§ 24.6).

24.3. Точные решения: $p_1 = {}^{6}/_{27}$, $p_2 = {}^{1}/_{9}$, $p_3 = {}^{7}/_{27}$, $p_4 = {}^{1}/_{3}$, $q_1 = {}^{5}/_{18}$, $q_2 = {}^{7}/_{18}$, $q_3 = {}^{1}/_{6}$, $q_4 = {}^{1}/_{6}$, $v = {}^{22}/_{9}$.

24.4. Оба животных имеют по восьми различных стратегий. Например, кошка может пойти к точкам пересечения *4*, *7* и *8* или *4*, *5* и *8* и т. д. Получается матрица 8×8, которую можно составить с платежами 1 и —1. После вычеркивания подчиненных стратегий она сводится к матрице 6×2 и после дальнейших вычеркиваний приводится к матрице 2×2, приведенной в табл. 24.4. В результате кошка должна итти к точкам пересечения *4*, *5*, *2* одну половину времени и к точкам *2*, *5*, *4* другую половину времени; мышь должна итти к точкам пересечения *6*, *3*, *2* одну половину времени и к точкам *8*, *7*, *4* другую половину времени. Вероятность выживания мыши равна 0,5.

28.1. По уравнению (28.3) энтропия термометра равна $-100 \times 1/100 \times \log_2 1/100 = 6{,}64$ бита на символ = = 132,8 бита в секунду, а энтропия барометра равна $-10 \times 1/10 \times \log_2 1/10 = 3{,}32$ бита на символ = 332 битам в секунду.

Следовательно, общая необходимая пропускная способность равна 133 + 332 = 465 битам в секунду.

28.2. Используя (28.3), получаем

$[6 \times 0{,}05 \times \log_2 0{,}05 + 2 \times 0{,}1 \times \log_2 0{,}1 + 0{,}2 \times \log_2 0{,}2 + 0{,}3 \times \log_2 0{,}3] = 2{,}95$ бита на символ, или 295 битам в секунду для барометра, или 428 битам в секунду всего.

28.3. Из равенства (28.30)

$$465 = W \log_2 (1 + 10) \text{ и } W = 134 \text{ } гц.$$

28.4. Практически некоторые температуры и давления будут более вероятны, а другие менее вероятны, как в задаче 28.2. Более существенно то, что здесь, бесспорно, будет большая степень автокорреляции, т. е. будет значительная вероятность того, что температура не будет изменяться, например, каждую двадцатую долю секунды, а если изменится, то почти наверное не будет меняться больше, чем на 1—2°. Следовательно, если на аэростате можно разместить соответствующую кодирующую аппаратуру и буферный накопитель, то требования к пропускной способности канала можно, вероятно, снизить до величины порядка двух или трех битов в секунду. С другой стороны, эта кодирующая и буферная аппаратура, вероятно, будет стоить дороже, чем канал шириной 134 *гц*, и на практике, вероятно, будет применяться очень простое кодирование, с небольшим накопителем или совсем без накопителя.

28.5. $H(x) = -16({}^{1}/_{16} \log_2 {}^{1}/_{16}) = 4$.

Если посылают или принимают согласную, то ненадежность не получает никакого приращения. Следовательно, равенство (28.13) нужно вычислить только для i и j, принимающих восемь различных значений гласных.

Если, в частности, передана гласная V_i, то условная вероятность того, что она принята, равна ${}^{1}/_{2}$, так что

$$-p(V_i, V_i) \log_2 p_{V_i}(V_i) = -{}^{1}/_{16} \times {}^{1}/_{2} \log_2 {}^{1}/_{2} = {}^{1}/_{32}.$$

Всего получится восемь таких членов, вносящих в сумме ${}^{1}/_{4}$ в $H_y(x)$.

Если передана V_j, то вероятность того, что принята данная V_j, равна ${}^{1}/_{2} \times {}^{1}/_{7} = {}^{1}/_{14}$. Совместная вероятность того, что передана V_i, а принята V_j, равна ${}^{1}/_{16} \times {}^{1}/_{14}$, так что

$$-p(V_i, V_j) \log_2 p_{V_i}(V_j) = -\frac{1}{16 \times 14} \log_2 \frac{1}{14}.$$

Всего имеется $7 \times 8 = 56$ таких членов, и их пай в $H_y(x)$ равен

$$-\frac{56}{16 \times 14} \log_2 \frac{1}{14} = -\frac{1}{4} \log_2 \frac{1}{14}.$$

$$R = H(x) - H_y(x) = 4 - \left(\frac{1}{4} + \frac{3{,}81}{4}\right) =$$

$$= 2{,}80 \text{ бита в секунду.}$$

28.6.

а) $\log_2 3 = 1{,}585$ бита на символ; следовательно, получается 3,17 бита в секунду.

б) $\log_2 2 = 1{,}00$ бита в секунду.

в) Каждые 2 *сек* можно передавать пару символов канала, переносящих $\log_2 3$ битов информации; следовательно, скорость передачи равна 0,79 бита в секунду.

г) 0,79 бита в секунду.

29.1. а) $\alpha = \frac{f}{2J} = \xi\omega_n = \frac{30}{2(1{,}5)} = 10$, $\omega_n = \frac{10}{0{,}5} = 20$ *рад/сек*;

б) $\omega = \omega_n \sqrt{1 - \xi^2} = 20\sqrt{1 - 0{,}5^2} = 17{,}32$ *рад/сек*;

в) $k = J\omega_n^2 = 1{,}5 \times 400 = 600$ футов на радиан;

г) $\varepsilon_{ст} = \frac{f \times 0{,}3}{k} = \frac{30 \times 0{,}3}{600} = 0{,}015$ *рад*;

д) $\frac{T}{k} = 0{,}015$, $T = 9$ футо-фунтов.

29.2. а) $L[6(1 - e^{-0.2t})] = 6L(1) - 6Le^{-0.2t} = \frac{6}{s} - \frac{6}{s + 0{,}2}$;

б) $L(\theta) = \frac{10 \times \omega}{\omega^2 + s^2}$;

в) $10 \int_0^\infty \cos(t + 30°)\, e^{-st} dt = 10 \int_0^\infty [\sqrt{3}/2 \cos t - {}^{1}/_{2} \sin t] \times e^{-st} dt = 5\sqrt{3} L(\cos t) - 5L(\sin t) = \frac{5\sqrt{3}s}{s^2 + 1} - \frac{5}{s^2 + 1} = 5\left(\frac{\sqrt{3}s - 1}{s^2 + 1}\right)$.

ПРЕДМЕТНЫЙ УКАЗАТЕЛЬ

ОГЛАВЛЕНИЕ

ЧАСТЬ 4

ТЕОРИЯ ВЫЧИСЛИТЕЛЬНЫХ МАШИН — ОСНОВНОЕ ОРУДИЕ ВНУТРЕННЕГО ПРОЕКТИРОВАНИЯ СИСТЕМ

ЧАСТЬ 5

ВНУТРЕННЕЕ ПРОЕКТИРОВАНИЕ СИСТЕМ

Г. Х. ГУД, Р. Э. МАКОЛ
СИСТЕМОТЕХНИКА
ВВЕДЕНИЕ В ПРОЕКТИРОВАНИЕ БОЛЬШИХ СИСТЕМ
Редактор *И. М. Болкова*
Техн. редактор *В. В. Беляева*
Обложка художника *В. Т. Сидоренко*

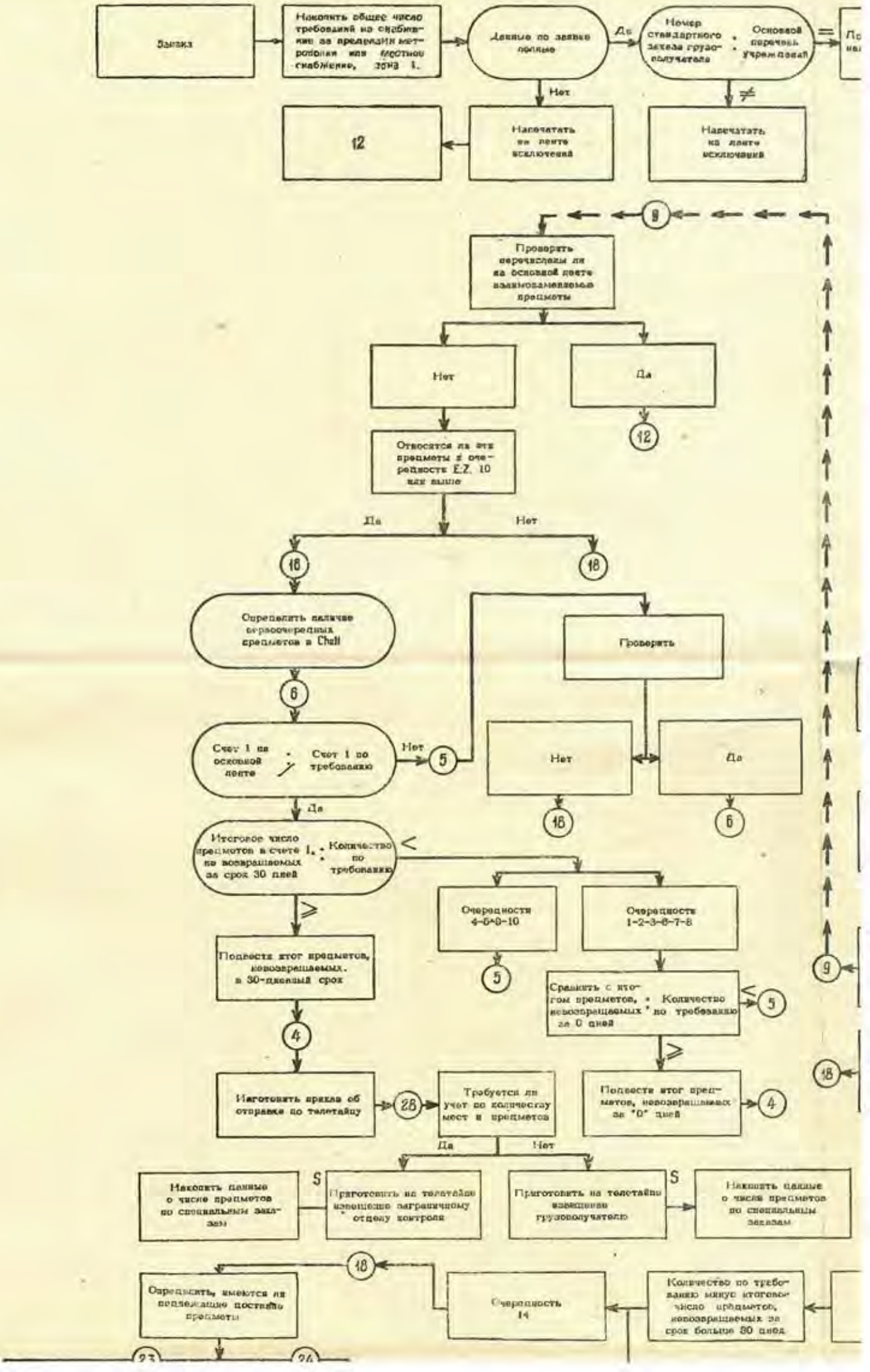
Сдано в набор 30.I.62. Подписано к печати 11.VII.62. Формат 84×108¹/₁₆
Уч.-изд. л. 43,775. Объем 40,18 п. л. Тираж 10 000.
Цена в пер. № 5—3 р. 21 коп. Заказ 2066

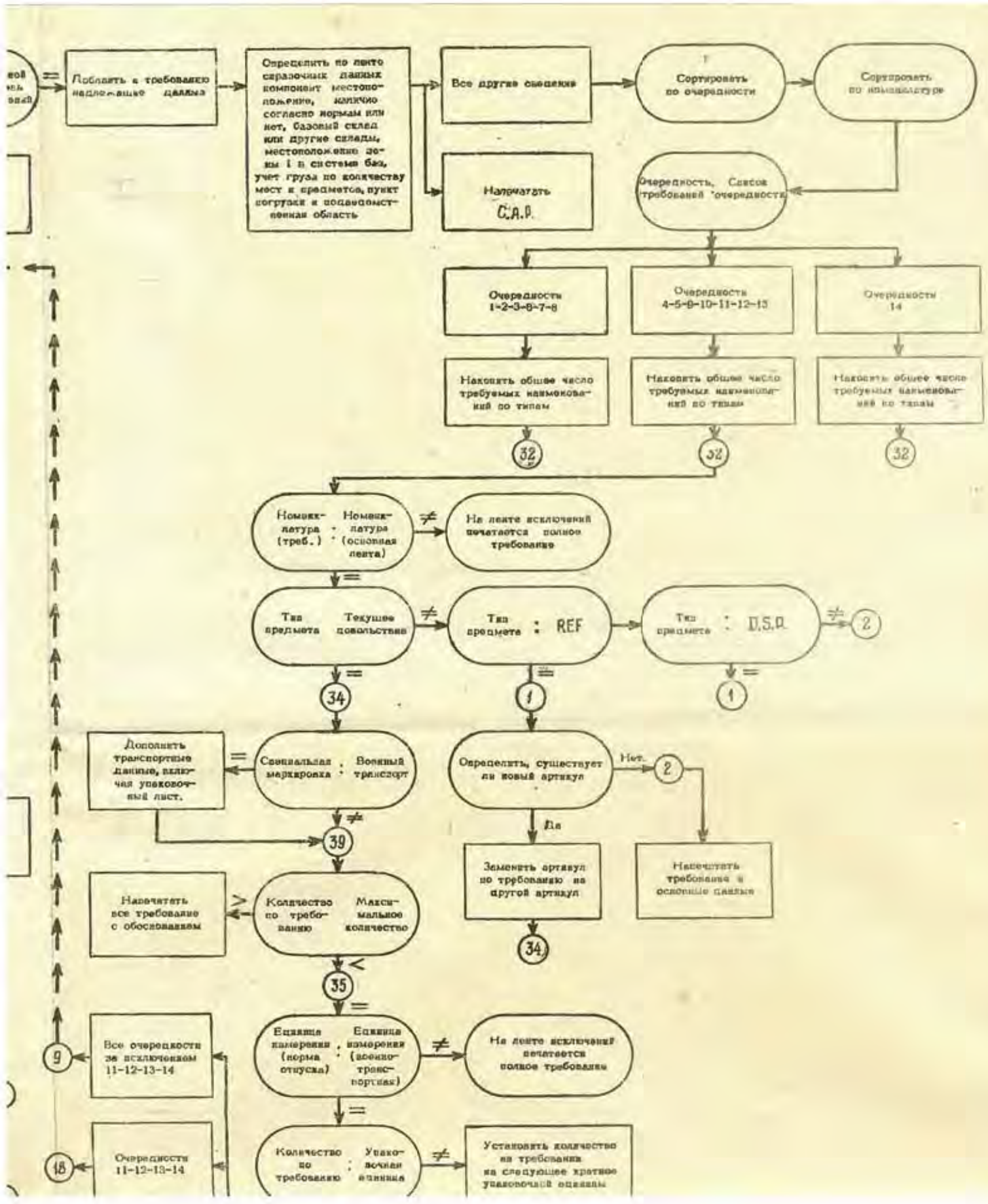
Типография Госэнергоиздата. Москва, Шлюзовая наб., 10.

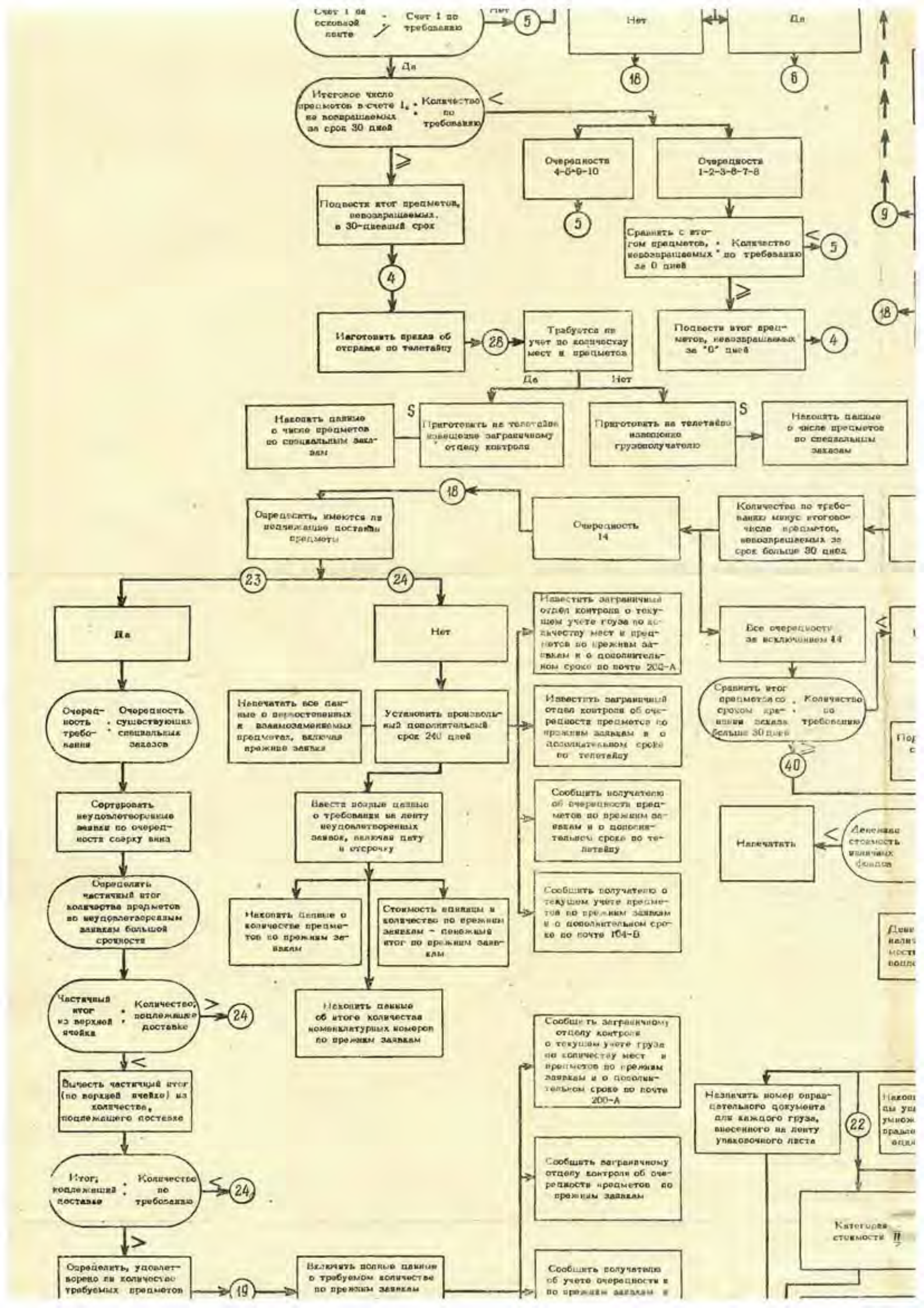
ЗАМЕЧЕННЫЕ ОПЕЧАТКИ

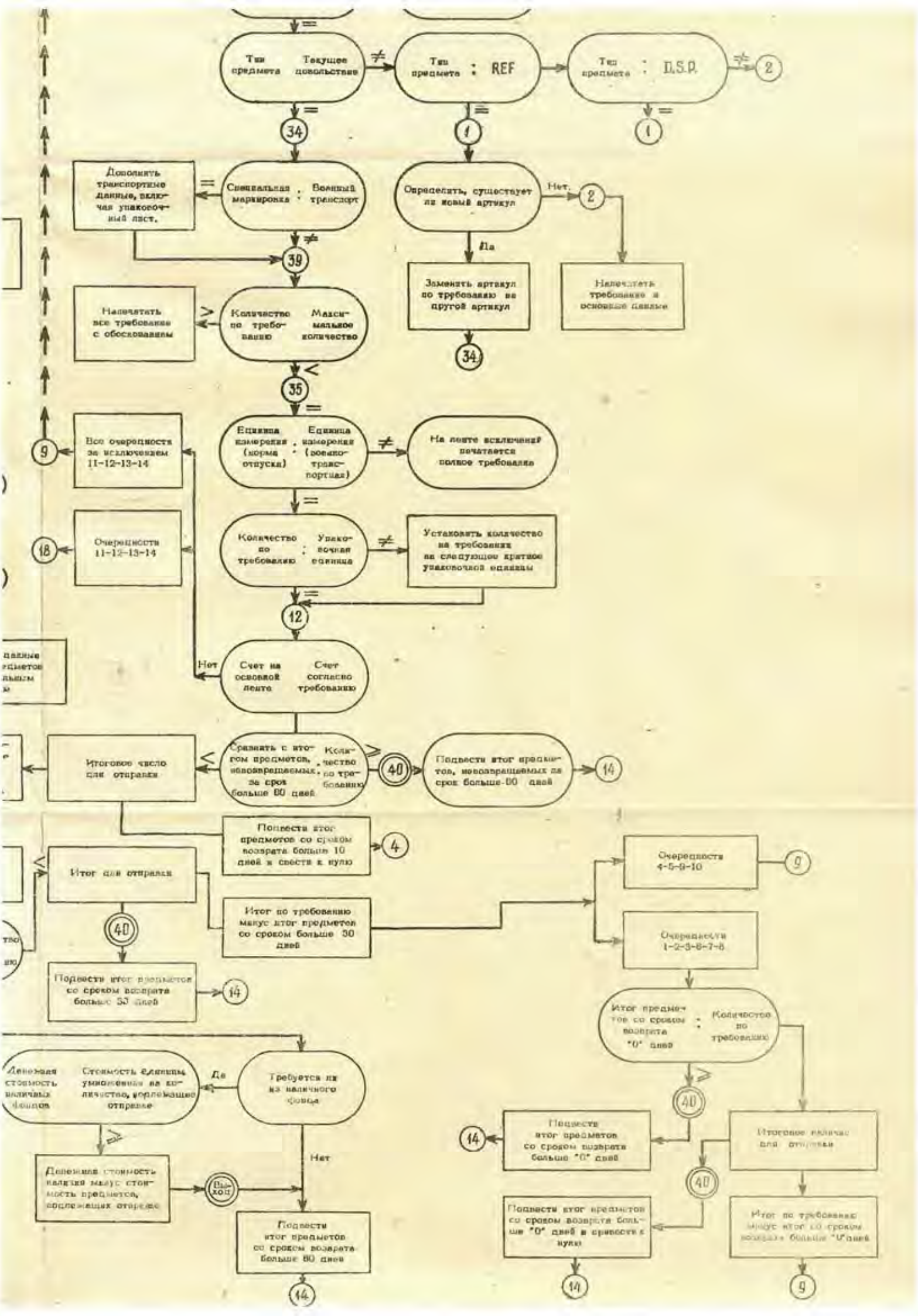
Стр.	Строка	Напечатано	Читать
29 лев.	2 сн.	опробывании	опробовании
120 лев.	1 св.	служащие статистики	служащие-статистики
140 пр.	8 сн.	$F = \frac{36}{2} \div \frac{58}{9} = 2{,}79.$	$F = \frac{36}{2} : \frac{58}{9} = 2{,}79.$
217, табл. 21.1		Плотность населения:	Плотность населения
242 пр.	22 св.	σ_{2T}	σ_T^2
308 пр.	8 св.	$t=0, \pm \frac{1}{2}W, \pm \frac{2}{2}W, \ldots, \pm \frac{n}{2}W.$	$t=0, \pm \frac{1}{2W}, \pm \frac{2}{2W}, \ldots, \pm \frac{n}{2W}.$
308 пр.	9 сн.	$t = \frac{n}{2w}$	$t = \frac{n}{2W}$
319 пр.	18 св.	$\xi = \frac{f}{f_C}$	$\xi = \frac{f}{f_{кр}}$
365, табл. к задаче 23.2		$\frac{\rho^{\nu!}}{\nu!(\nu-\rho)}$	$\frac{\rho^{\nu}\nu}{\nu!(\nu-\rho)}$
366 лев.	29 сн.	$[6 \times 0{,}05 \times \log_2 0{,}05 \ldots$	$-[6 \times 0{,}05 \times \log_2 0{,}05 \ldots$
367 лев.	10—11 сн.	analog-to-digital converter	analog-to-time-to-digital converter

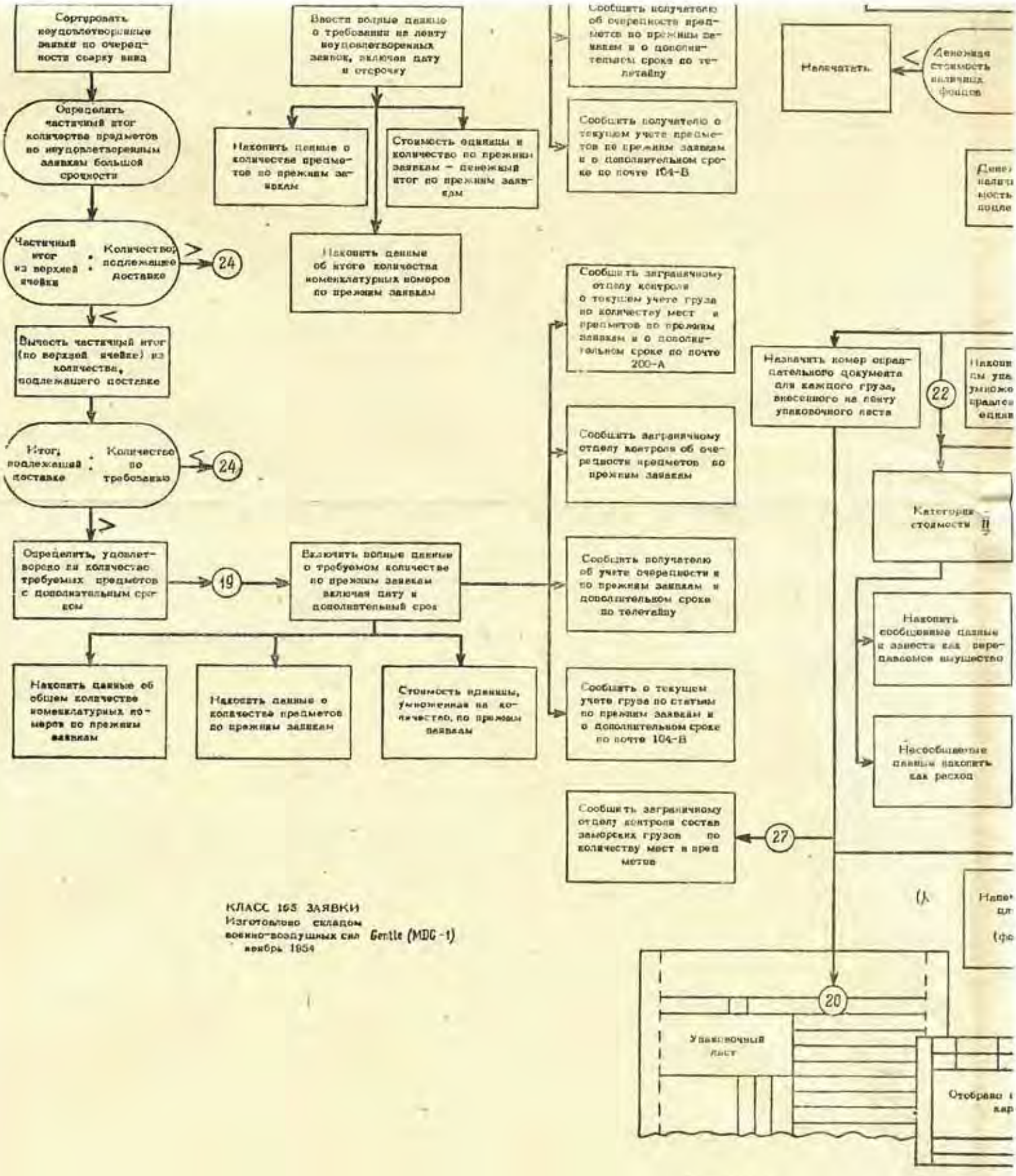
К зак. 2066.


Выход
Накопить общее число требований на снабжение за предприятия метрополии или простое снабжение, зона 1.
Данные по заявке полные
Да
Номер стандартного заказа грузополучателя
Основной перечень учреждений
Нет
12
Напечатать на ленте исключения
Напечатать на ленте исключения
9
Проверить перечислены ли на основной ленте взаимозаменяемые предметы
Нет
Да
12
Относятся ли эти предметы к очередности E.Z. 10 или выше
Да
Нет
16
18
Определить наличие первоочередных предметов в Chall
Проверить
6
Счет 1 на основной ленте
Счет 1 по требованию
Нет
5
Нет
Да
Да
16
6
Итоговое число предметов в счете 1 на возвращаемых за срок 30 дней
Количество по требованию
Очередности 4-5-9-10
Очередности 1-2-3-6-7-8
Подвести итог предметов, невозвращаемых, в 30-дневный срок
5
Сравнить с итогом предметов, невозвращаемых за 0 дней
Количество по требованию
5
9
4
18
Изготовить приказ об отправке по телетайпу
28
Требуется ли учет по количеству мест и предметов
Подвести итог предметов, невозвращаемых за "0" дней
4
Да
Нет
Накопить данные о числе предметов по специальным заказам
S
Приготовить на телетайпе извещение заграничному отделу контроля
Приготовить на телетайпе извещение грузополучателю
S
Накопить данные о числе предметов по специальным заказам
18
Определить, имеются ли подлежащие постановке предметы
Очередность 14
Количество по требованию минус итоговое число предметов, невозвращаемых за срок больше 30 дней
23
24


Добавить к требованию надлежащие данные
Определить по ленте справочник данных компонент местоположение, наличие согласно нормам или нет, базовый склад или другие склады, местоположение зоны I в системе баз, учет груза по количеству мест и предметов, пункт погрузки и подчиненная область
Все другие сведения
Напечатать С.А.Р.
Сортировать по очередности
Сортировать по номенклатуре
Очередность, Список требований "очередность"
Очередность 1-2-3-6-7-8
Очередность 4-5-9-10-11-12-13
Очередность 14
Находить общее число требуемых наименований по типам
Номенклатура (треб.) = Номенклатура (основная лента)
На ленте исключений печатается полное требование
Тип предмета = Текущее довольствие
Тип предмета = REF
Тип предмета = D.S.R.
Дополнить транспортные данные, включая упаковочный лист.
Специальная маркировка = Военный транспорт
Определить, существует ли новый артикул
Нет
Да
Заменить артикул по требованию на другой артикул
Напечатать требование с основными данными
Напечатать все требование с обоснованием
Количество по требованию ≥ Максимальное количество
Единица измерения (норма отпуска) = Единица измерения (военно-транспортная)
На ленте исключений печатается полное требование
Все очередности за исключением 11-12-13-14
Очередность 11-12-13-14
Количество по требованию = Упаковочная единица
Установить количество на требование на следующее кратное упаковочной единицы

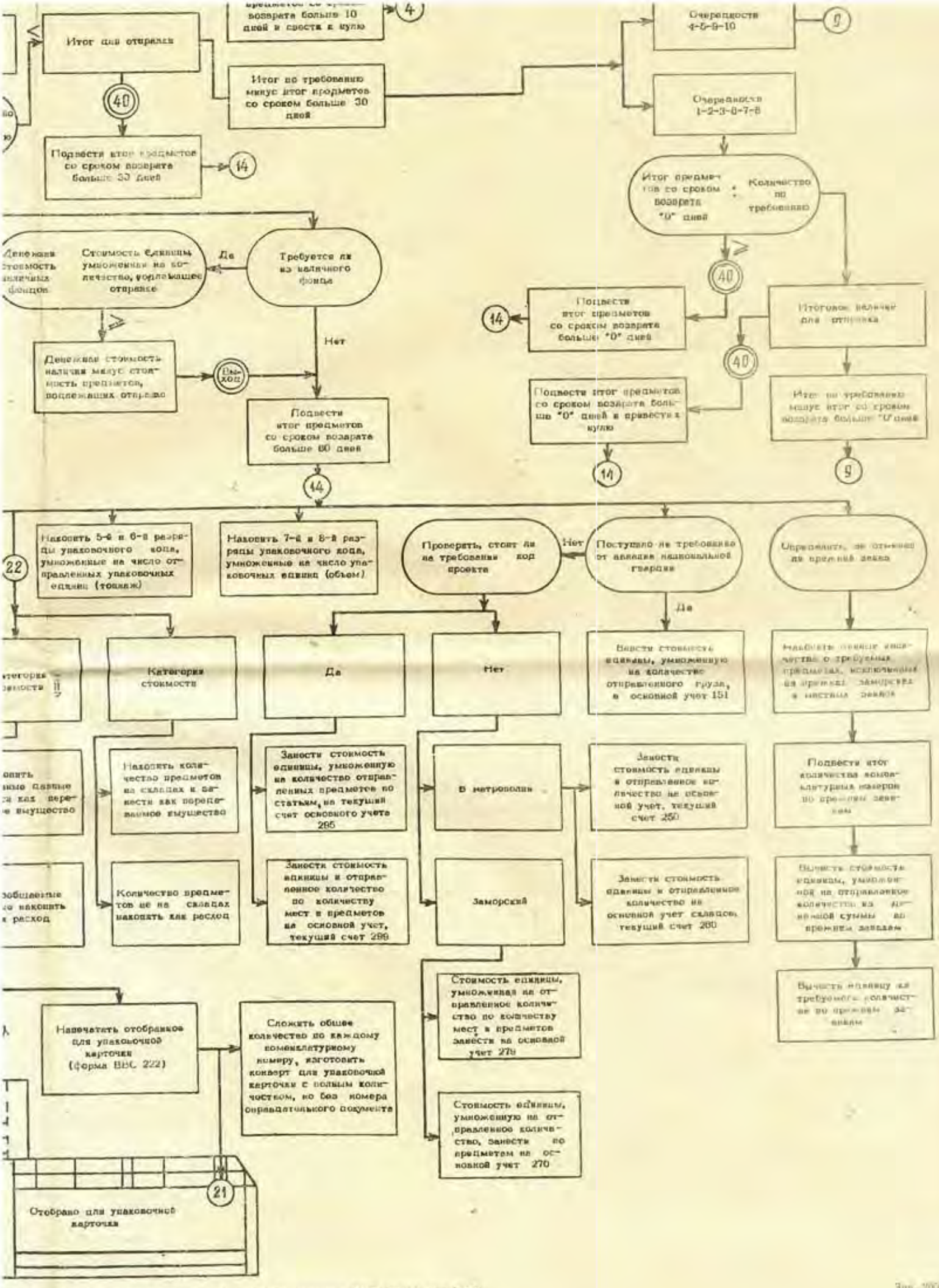

Счет 1 по основной ленте
Счет 1 по требованию
Да
Нет
Итоговое число предметов в счете 1 на невозвращаемых за срок 30 дней
Количество по требованию
Очередность 4-5-9-10
Очередность 1-2-3-6-7-8
Подвести итог предметов, невозвращаемых, в 30-дневный срок
Сравнить с итогом предметов, невозвращаемых за 0 дней
Количество по требованию
Подвести итог предметов, невозвращаемых за "0" дней
Изготовить приказ об отправке по телетайпу
Требуется ли учет по количеству мест и предметов
Да
Нет
Накопить данные о числе предметов по специальным заказам
Приготовить на телетайпе извещение заграничному отделу контроля
Приготовить на телетайпе извещение грузополучателю
Накопить данные о числе предметов по специальным заказам
Определить, имеются ли подлежащие поставке предметы
Очередность 14
Количество по требованию минус итоговое число предметов, невозвращаемых за срок больше 30 дней
Да
Нет
Очередность требования
Очередность специальных заказов
Напечатать все данные о приостановленных к взаимозаменяемых предметах, включая прежние заявки
Установить произвольный дополнительный срок 240 дней
Известить заграничный отдел контроля о текущем учете груза по количеству мест и предметов по прежним заявкам и о дополнительном сроке по почте 200-А
Известить заграничный отдел контроля об очередности предметов по прежним заявкам и о дополнительном сроке по телетайпу
Сообщить получателю об очередности предметов по прежним заявкам и о дополнительном сроке по телетайпу
Сообщить получателю о текущем учете предметов по прежним заявкам и о дополнительном сроке по почте 164-В
Все очередности за исключением 40
Сравнить итог предметов со сроком прежних заявок больше 30 дней
Количество по требованию
Напечатать
Денежная стоимость наличных фондов
Сортировать неудовлетворенные заявки по очередности спереду вниз
Определить частичный итог количества предметов по неудовлетворенным заявкам большой срочности
Ввести полные данные о требованиях на ленту неудовлетворенных заявок, включая дату и отсрочку
Накопить данные о количестве предметов по прежним заявкам
Стоимость единицы и количество по прежним заявкам - денежный итог по прежним заявкам
Накопить данные об итоге количества номенклатурных номеров по прежним заявкам
Частичный итог из верхней ячейки
Количество, подлежащее доставке
Вычесть частичный итог (по верхней ячейке) из количества, подлежащего доставке
Итог, подлежащий доставке
Количество по требованию
Определить, удовлетворено ли количество требуемых предметов
Включить полные данные о требуемом количестве по прежним заявкам
Сообщить заграничному отделу контроля о текущем учете груза по количеству мест и предметов по прежним заявкам и о дополнительном сроке по почте 200-А
Сообщить заграничному отделу контроля об очередности предметов по прежним заявкам
Сообщить получателю об учете очередности и по прежним заявкам и
Назначить номер определительного документа для каждого груза, внесенного на ленту упаковочного листа
Категория стоимости II


Тип предмета : Текущее довольствие
Тип предмета : REF
Тип предмета : D.S.P.
2
34
1
1
Дополнить транспортные данные, включая упаковочный лист.
Специальная маркировка . Военный транспорт
Определить, существует ли новый артикул
Нет.
2
Да
39
Заменить артикул по требованию на другой артикул
Напечатать требование с основными данными
Напечатать все требования с обоснованием
Количество по требованию : Максимальное количество
34
35
9
Все очередности за исключением 11-12-13-14
Единица измерения (норма отпуска) . Единица измерения (военно-транспортная)
На ленте вспомогательной печатается полное требование
18
Очередности 11-12-13-14
Количество по требованию : Упаковочная единица
Установить количество на требованиях на следующее кратное упаковочной единицы
12
Нет
Счет на основной ленте : Счет согласно требованию
Итоговое число для отправки
Сравнить с итогом предметов, невозвращаемых за срок больше 60 дней . Количество по требованию
40
Подвести итог предметов, невозвращаемых за срок больше 60 дней
14
Подвести итог предметов со сроком возврата больше 10 дней к свести к нулю
4
Итог для отправки
Итог по требованию минус итог предметов со сроком больше 30 дней
Очередность 4-5-9-10
9
Очередность 1-2-3-6-7-8
40
Подвести итог предметов со сроком возврата больше 30 дней
14
Итог предметов со сроком возврата "0" дней : Количество по требованию
Денежная стоимость наличных фондов
Стоимость единицы умноженная на количество, подлежащее отправке
Да
Требуется ли из наличного фонда
Нет
Денежная стоимость наличных минус стоимость предметов, подлежащих отправке
Выход
Подвести итог предметов со сроком возврата больше 60 дней
14
40
Подвести итог предметов со сроком возврата больше "0" дней
14
Итоговое число для отправки
40
Подвести итог предметов со сроком возврата больше "0" дней и свести к нулю
14
Итог по требованию минус итог со сроком возврата больше "0" дней
9



КЛАСС 165 ЗАЯВКИ
Изготовлено складом военно-воздушных сил Gentle (MDC-1)
ноябрь 1954

Рис. 21.1. Блок-схема для части системы



ги системы материально-технического обеспечения ВВС США (из [41]).

Зак. 2006