

МАТЕМАТИЧЕСКИЕ ОСНОВЫ в.г. лапа КИБЕРНЕТИКИ



МАТЕМАТИЧЕСКИЕ ОСНОВЫ в. г. лапа КИБЕРНЕТИКИ

ИЗДАНИЕ ВТОРОЕ, ПЕРЕРАБОТАННОЕ
И ДОПОЛНЕННОЕ

Издательское объединение «Вища школа»
Головное издательство
Киев — 1974

517.2 + 6 Ф0.1
Л24

УДК 519.5:62—50(07)

Математические основы кибернетики. Лапа В. Г.
Издательское объединение «Вища школа», 1974, 452 с.

В основу книги положен курс лекций по программе «Математические основы кибернетики». Главная задача курса — ознакомить студентов с теоретическими направлениями и математическими методами, составляющими фундамент современной науки об управлении и связи.

В книге излагаются основные вопросы, являющиеся теоретической базой кибернетики и обычно не входящие в традиционный курс математики втузов, а именно: теория множеств и основы математической логики, теории алгоритмов, вероятностей и случайных функций, корреляции, спектров, а также элементы теории информации и игр. Предлагаемое учебное пособие предназначено для студентов электроприборостроительных специальностей втузов. Изложение основного материала и примеров подчинено специфике подготовки специалистов широкого профиля. Однако книга несомненно будет полезна и специалистам других направлений. Ею можно пользоваться при подготовке специалистов по программам экономических вузов и биологических факультетов с кибернетическим уклоном. Книга может быть полезна также аспирантам и научным работникам в качестве своеобразного справочного пособия по отдельным теоретическим вопросам.

Табл. 33. Ил. 113. Библиогр. 49.

Редакция литературы по радиоэлектронике, кибернетике
и связи
Зав. редакцией А. В. Дьячков

Л 30501 — 179
М211 (04) — 74 53 — 40 — 18 — 73

© Издательское объединение «Вища школа», 1974

ПРЕДИСЛОВИЕ

Обеспечение значительного повышения материального и культурного уровня жизни народа на основе высоких темпов развития социалистического производства, повышение его эффективности, научно-технического прогресса и ускорение роста производительных сил — вот главная задача, поставленная КПСС перед трудящимися СССР.

Успешное решение этой задачи в значительной степени определяется уровнем подготовки, квалификацией кадров и, в том числе специалистов, выпускаемых высшими учебными заведениями. Особого внимания требует подготовка кадров для науки и производства, обеспечивающая существенное сокращение цикла «исследование — производство». В учебных планах многих специальностей появились новые специальные дисциплины: математические основы кибернетики, основы построения АСУ, теория вычислительных комплексов, отображение информации, исследование операций и многие другие.

Предлагаемая книга является учебным пособием нового типа. В ней сделана попытка по возможности охватить те теоретические направления, которые являются основой современной кибернетики. При этом учитывалась специфика подготовки кадров в высших технических учебных заведениях.

В книге десять глав.

Первые две главы посвящены теории множеств и основам математической логики. Основное внимание здесь уделено определениям и наиболее важным соотношениям. При изложении доказательства большинства теорем опущены. В главе 2, посвященной математической логике и исчислению предикатов, примеры подобраны из области электроники и приборостроения. Дальнейшее углубление знаний студентов по специальным вопросам осуществляется в курсах

по вычислительной технике и логическим элементам (электромагнитным, пневмо- и гидроавтоматики и т. д.).

В главе 3 кратко изложены основы теории алгоритмов и некоторых ее приложений. Эта глава является введением для последующих курсов алгоритмизации, управляющих машин, программирования.

Основам теории вероятностей и случайных функций посвящена четвертая глава. Материал здесь подобран в соответствии с требованиями практических приложений и использования в последующих специальных курсах.

Значительное место в главе занимает раздел, в котором излагаются основы математического описания динамических систем. Сформулированы основные задачи анализа систем под воздействием различного рода сигналов, в том числе и случайных. Введено понятие оператора системы, рассмотрены различные классы операторов и систем. Эта глава служит основой для последующих курсов по автоматическому управлению, следящим системам, статистической динамике и другим.

В главе 5 изложены методы статистического анализа и теории корреляции (дисперсионный анализ). Задачи обработки результатов наблюдений, определения статистических параметров, статистического измерения связи иллюстрируются примерами из области производства и надежности аппаратуры.

Следующие две главы — 6 и 7 — посвящены теории спектров и элементам теории информации. Здесь изложены такие важные вопросы, как теория рядов Фурье и интеграл Фурье, модуляция, детектирование, основы общей теории связи, элементы теории прохождения сигналов по каналам связи и т. д. Все это является теоретической базой для многих специальных курсов и в том числе для курсов по телемеханике.

В 8 главе изложены элементы теории игр.

Глава 9 посвящена теории графов и некоторым важным приложениям этой теории. В частности, в терминах теории графов сформулированы задачи распределения потоков (в том числе и транспортные), задачи составления графиков и расписаний и некоторые другие.

В 10 главе представлены теоретические основы тензорного исчисления. Этот аппарат широко используется в механике, гидравлике, электродинамике, а в теоретической и технической кибернетике используется пока мало. А ведь проблемы передачи и переработки информации, управления

и связи можно весьма эффективно описывать и изучать в терминах тензорной алгебры и тензорного анализа. Это перспективное направление еще ожидает своих исследователей.

Чтобы не загромождать текст, ссылки на литературу по ходу изложения не делались. Различные вопросы математических основ кибернетики освещаются во многих десятках монографий и статей, поэтому число ссылок должно быть весьма большим.

Вместо этого в приложении помещен довольно обширный аннотированный список литературы по главам.

Автор выражает благодарность чл.-кор. АН СССР А. Е. Алексееву и проф. И. В. Кузьмину, которые прочитали рукопись и сделали ряд полезных замечаний. Автор с признательностью примет пожелания и предложения читателей, направленные по адресу: 252054, Киев, 54, ул. Гоголевская, 7, Головное издательство издательского объединения «Вища школа».

ВВЕДЕНИЕ

Традиционный курс математики для студентов высших технических учебных заведений включает дифференциальное и интегральное исчисление и примыкающие дисциплины. По мере развития различных научных и технических направлений и особенно в связи с быстрым ростом автоматизации производственных процессов возникла необходимость в изучении новых математических дисциплин. Этим объясняется введение в разных вузах для различных специальностей дополнительных разделов математики. Студенты — гидравлики, акустики слушают курс уравнений математической физики. Студенты приборостроительных и радиотехнических специальностей изучают вопросы теории случайных функций и гармонический анализ. Для будущих специалистов по конструированию и эксплуатации вычислительных машин необходимы основы теории множеств и математическая логика.

Границы практических приложений знаний современного специалиста с высшим образованием трудно очертить. Физикам приходится изучать проблемы биологии, биологи участвуют в создании аппаратуры для автоматизации и управления, математики заняты анализом лингвистических законов, а лингвисты изучают математику.

Существуют задачи, исследуемые с разных сторон математикой, статистикой, электроникой, нейрофизиологией. Одно и то же понятие у каждой группы специалистов получает особое название. Многие важные исследования продвигаются по несколько раз без учета результатов, полученных в смежных областях. Другие важные работы затягиваются из-за того, что в одной области неизвестны результаты, ставшие уже классическими в смежной.

Нужны объединенные усилия специалистов различных областей, чтобы преодолеть языковой барьер и подойти к

решению задачи с принципиальной точки зрения, не ставя во главу угла специфику только одной конкретной отрасли. В начале 40-х годов сформировалась группа ученых, известных специалистов различных отраслей науки, которых объединял интерес к одной общей проблеме — проблеме управления.

Инженеры разрабатывали и создавали электронную аппаратуру для управления. Математики исследовали свойства сигналов в различных системах и описывали их аналитически. Параллельно специалисты развивали теорию кодирования информации. Они пытались дать ответ на вопрос: как можно измерить содержание информации в сообщении и как точно выразить эту меру. Специалисты по статистике рассматривали поток информации в живом организме как основу физиологического регулирования его функций.

Постепенно некоторые исследователи, несмотря на разделявший их языковой барьер, пришли к выводу, что их исследования привели к формированию новой области научной мысли. Эта новая наука «об общих законах управления и связи в живых организмах» была названа кибернетикой.

Слово «кибернетика» в значении «наука о кораблевождении» применял еще Платон. В 1843 году оно было использовано французским физиком и математиком Ампером.

Ампер заимствовал это слово из греческого языка, в котором «кибернетес» означает — кормчий, рулевой. Но этот термин не получил распространения и по существу был забыт.

Вновь появился этот термин в 1948 году в связи с изданием книги профессора Массачусетского технологического института Норберта Винера «Кибернетика». В книге были обобщены результаты исследований американских ученых в области управления, которые проводились в 40-х годах. По сути, исследования общих законов переработки информации в системах различной природы начались с задачи определения будущих значений траектории движущегося объекта. Для решения этой задачи необходимо было выполнять определенные операции над прошлыми наблюдениями.

Первые математические исследования по предсказанию случайных процессов принадлежат академику А. Н. Колмогорову. В отчетах и публикациях по разработке искусственных систем для сложных вычислений и предсказаний Н. Винер и его коллеги ссылаются также на работы А. Н. Крылова и Н. Н. Боголюбова, посвященные вопросам управления.

Всемирно известны исследования отечественной школы кибернетиков (работы А. И. Берга, В. М. Глушкова, А. Г. Ивахненко, В. В. Солодовникова). Интенсивно развиваясь, кибернетика в настоящее время объединяет уже комплекс таких самостоятельных наук, как теоретическая кибернетика, биологическая кибернетика, техническая кибернетика, промышленная кибернетика, экономическая кибернетика.

Все перечисленные направления кибернетики базируются на широкой теоретической основе, включающей множество разделов, технических наук, биологии.

Основными в кибернетике являются понятия, широко используемые во многих областях знаний: «система», «структура», «информация», «сигнал», «управление», «обратная связь» и другие.

Одной из центральных проблем в развитии науки и человеческого общества в настоящее время стала проблема создания больших систем и управления ими. Промышленные предприятия, планирующие организации, все отрасли производства и сфера потребления, все народное хозяйство в целом можно рассматривать как большие системы. При изучении больших систем необходимо анализировать огромное количество связей между элементами и явлениями, учитывать взаимодействие частей и целого, неопределенность в поведении системы, связи с окружающей средой и взаимодействие с ней.

В промышленности, в научных и технических исследованиях благодаря применению автоматических систем решение многих сложных задач управления стало возможным без непосредственного вмешательства человека. По мере усложнения структуры управляемых объектов, увеличения объема информации о протекающих в них процессах человек часто не в состоянии наилучшим образом осуществлять функции управления. Это объясняется недостатком времени, в течение которого должно быть принято оптимальное решение, невозможностью мобилизовать в короткий срок значительный объем памяти, свойством забывания информации и рядом других факторов.

Сложные системы автоматического управления обладают большим быстродействием и достаточным объемом запоминающих устройств.

Кроме того, они должны осуществлять многие функции «интеллектуального» характера, такие как сопоставление различных вариантов решения задачи, выбор наилучше-

го варианта в соответствии с определенными критериями, учет изменения внешних воздействий и изменения в связи с этим характера решения и критериев.

Поскольку характер моделируемых в автоматических системах мыслительных способностей постоянно усложняется, следует при создании подобных систем учитывать одно из важных качеств, присущих человеческому мышлению,— способность обучаться предсказанию.

Ни одно действие не совершается человеком без того, чтобы он в достаточно определенной форме не предвидел результатов этого действия.

При постановке задачи предсказания в технике, очевидно, невозможно обойтись без исследования того, как выполняются соответствующие функции в живых организмах.

Многие из этих задач еще каких-нибудь тридцать лет назад даже не могли быть поставлены, поскольку решение их с привлечением наличных средств и в разумные сроки не представлялось возможным.

С появлением универсальных цифровых вычислительных машин расширились возможности решения сложных задач обработки информации и автоматизации управления. При этом центр тяжести приложений математики сместился в сторону конечной, дискретной математики. Это повлекло за собой глубокие изменения в математической науке. Сформировались и нашли практическое приложение новые большие направления теоретической математики. Некоторые из них (теория игр, теория информации) стали на сегодняшний день самостоятельными математическими науками.

Вот как определяет А. Н. Колмогоров важнейшие цели современных математических исследований:

«1. Привести общие логические основы современной математики в такое состояние, чтобы их можно было излагать в школе подросткам 14—15 лет.

2. Уничтожить расхождение между «строгими» методами чистых математиков и «нестрогими» приемами математических рассуждений, применяемых прикладными математиками, физиками и техниками.

Две сформулированные задачи тесно связаны между собой. По поводу второй замечу, что в отличие от времен создания Ньютоном и Лейбницем дифференциального и интегрального исчисления, математики умеют сейчас без большого промедления подводить фундамент логически безукоризненных математических построений под любые методы

расчета, родившиеся из живой физической и технической интуиции и оправдывающие себя на практике. Но фундамент этот иногда оказывается столь хитро построенным, что молодые математики, гордые пониманием его устройства, принимают фундамент за все здание. Физики же и инженеры, будучи не в силах в нем разобраться, изготавливают для себя вместо него временные шаткие мостки» (А. Н. Колмогоров. Простоту сложному. «Известия» от 31 декабря 1962 г.).

Язык, на котором говорит кибернетика,— это математический язык, причем математическими основами кибернетики являются направления, далеко выходящие за рамки традиционной «инженерной» высшей математики. Специалисты по созданию и исследованию сложных систем переработки информации и управления должны владеть аппаратом теории вероятностей, статистическим анализом, теорией информации и т. д.

Глава 1

ЭЛЕМЕНТЫ ТЕОРИИ МНОЖЕСТВ

§ 1. ОСНОВНЫЕ ОПРЕДЕЛЕНИЯ. СПОСОБЫ ЗАДАНИЯ МНОЖЕСТВ

При изучении окружающего мира мы сталкиваемся с большим многообразием явлений, событий, объектов. Все существующее и происходящее в мире, как правило, характеризуется некоторыми индивидуальными свойствами или особенностями, которые позволяют выделить тот или иной объект, то или иное явление.

В математике совокупности объектов обычно называются *множествами*.

Множество задано (или, что то же, определено), если обо всяком объекте можно сказать, принадлежит он данному множеству или нет. Самый простой способ задания множества состоит в перечислении всех входящих в него объектов. Эти объекты называют *элементами* множества. Так, например, множество цифр десятичной системы исчисления состоит из десяти знаков

0, 1, 2, 3, 4, 5, 6, 7, 8, 9.

Значок $\sqrt{}$ в этот список не входит, следовательно, значок $\sqrt{}$ не принадлежит множеству цифр десятичной системы исчисления.

Множество букв русского алфавита состоит из тридцати двух знаков

а, б, в, г, д, е, ж, з, и, й, к, л, м, н, о, п, р, с, т, у, ф, х,
ц, ч, ш, щ, ъ, ы, ь, э, ю, я.

Знаки «?» и «,» в этом списке отсутствуют, следовательно, они не принадлежат множеству букв русского алфавита.

Рассмотрим множество диагоналей правильного пятиугольника. Это множество обозначим D_5 . Оно состоит из отрезков

$A_1A_3; A_2A_4; A_3A_5; A_4A_1; A_5A_2.$

Рассмотрим теперь множество диагоналей правильного выпуклого стоугольника. Обозначим это множество D_{100} . Множество D_{100} принципиально ничем не отличается от множества D_5 . Однако составить список элементов множества D_{100} не так уж просто. В нем 4850 элементов. Но нужен ли нам такой список? Предположим, что задан некоторый отрезок и нужно определить, принадлежит ли он множеству D_{100} . Стали бы мы данный отрезок искать в кем-то составленном списке из 4850 элементов? Безусловно, нет; мы бы даже не поинтересовались, существует ли такой список. Посмотрим на концы отрезка; если оба его конца являются какими-нибудь двумя несоседними вершинами данного стоугольника, то можно ответить: да, отрезок принадлежит множеству D_{100} . Если же хотя бы один из концов отрезка не является вершиной стоугольника или если концы этого отрезка совпадают с двумя соседними вершинами стоугольника, мы ответили бы отрицательно.

Таким образом, в последнем примере множество задавалось не списком своих элементов, а свойством, которым должны обладать эти элементы — диагонали многоугольника.

Рассмотренные нами в виде примеров множества называются *конечными*, хотя некоторые из них состоят из большого количества элементов. Во всех этих случаях можно, хотя бы принципиально, составить списки элементов множества. В математике чаще всего имеют дело не с конечными, с бесконечными множествами. Например, множество целых чисел, множество квадратов, множество сфер. Ясно, что перечислить все элементы этих множеств, составить списки всех элементов невозможно: *Бесконечное множество* определяется обычно *свойствами его элементов*.

Множество A задано (определено, установлено), если указано свойство α , которым обладают все элементы, принадлежащие множеству A , и которым объекты, не принадлежащие множеству A , не обладают.

Множества мы будем обозначать заглавными буквами латинского алфавита, элементы множества — строчными буквами латинского алфавита, а свойства этих элементов — буквами греческого алфавита. Знаком \in обозначается вхождение, включение элемента в множество: $a \in A$ — a входит в A , или $A \ni a$ — множество A содержит a .

Рассмотрим примеры бесконечных множеств.

Элементы множества B чётных чисел обладают свойством делиться на 2 без остатка. Значит, $24 \in B$. Множество

квадратов состоит из четырехугольников, обладающих следующими свойствами: все стороны каждого из этих четырехугольников равны между собой и, по крайней мере, один из углов такого четырехугольника прямой. Множество сфер состоит из замкнутых поверхностей, каждая из которых обладает следующими свойствами: все точки этой поверхности находятся на одном и том же расстоянии от точки, называемой центром сферы.

Определяя бесконечные множества (множество четных чисел, множество квадратов, множество сфер), мы, по существу, выделяли их из других, более общих множеств (множества целых чисел, множества четырехугольников, множества поверхностей).

Понятия «множество» и «свойство» тесно связаны между собой. Например, свойство α определяет множество A объектов, обладающих этим свойством; при этом предполагается, что в множество A входят *все* объекты, обладающие свойством α . Справедливо и обратное утверждение. Если определено множество A , то определено также и свойство «принадлежать множеству A ». Так, свойство быть знаком арифметического действия означает принадлежать к совокупности (или быть похожим на один из значков) $+$, $-$, \times , $:$, называемых элементами множества знаков арифметических операций.

По-видимому нетрудно придумать такое свойство, которым не обладал бы ни один объект. К примеру, не существует четырехугольников, у которых сумма внутренних углов равна 500° . Подобные свойства никаких множеств не определяют. Однако с математической точки зрения в целях общности говорят, что такое свойство определяет множество, не содержащее ни одного элемента. Такие, не содержащие ни одного элемента, множества, называются *пустыми*. Пустое множество будем обозначать O .

Если все элементы некоторого множества A можно перенумеровать в виде бесконечной последовательности $a_1, a_2, \dots, a_n, \dots$, причем каждый элемент имеет только один номер, то такое множество называют *счетным*.

Бесконечные множества, элементы которых нельзя перенумеровать, называются *несчетными множествами*. Пример несчетного множества — множество точек отрезка прямой между 0 и 1.

Часто при рассмотрении некоторого множества приходится иметь дело не только с его отдельными элементами, но и с упорядоченными парами элементов этого множества. При этом допускается, что оба элемента пары могут совпадать.

Например, $\{3, 3\}$, $\{0, 8\}$, $\{8, 0\}$ — упорядоченные пары, составленные из действительных чисел. Точно так же можно рассматривать упорядоченные тройки, четверки, и, в общем случае, упорядоченные n -ки элементов множества. Причем, как и ранее, допускается совпадение отдельных элементов. Например, $\{1, 2, 3, 4\}$, $\{5, 5, 10, 10\}$ — примеры упорядоченных четверок, составленных из элементов натурального ряда.

Если обобщить понятия упорядоченной пары, упорядоченной тройки, ..., упорядоченной n -ки, мы приходим к понятию упорядоченного набора элементов множества. В различных областях математики упорядоченные наборы элементов называются по-разному. В комбинаторном анализе их называют *размещениями*, в теории вероятностей и статистическом анализе — *выборками*, в алгебре — *векторами*. В теории множеств для упорядоченного множества введен термин *кортеж*. Кортеж, составленный из элементов множества D , называют обычно кратко *кортежем над D* .

Кортеж над D , составленный из элементов d_1, d_2, \dots, d_s , взятых в этом же порядке, обозначается

$$\{d_1, d_2, \dots, d_s\}.$$

При этом говорят, что d_i — i -я *координата*, или *компонента*, кортежа.

Число s координат называется длиной кортежа. Кроме кортежей длины 2, 3, ..., n для общности рассматриваются также кортежи $\{d\}$ длины 1 и пустые кортежи $\{\}$. Для пустого кортежа введем обозначение то же, что и для пустого множества — O . Длину кортежа O будем считать равной 0.

Примерами множеств, важных с точки зрения приложений, являются *линейные точечные множества*.

§ 2. ЛИНЕЙНЫЕ ТОЧЕЧНЫЕ МНОЖЕСТВА

Линейными точечными называются множества, элементами которых являются точки прямой линии или, точнее говоря, точки числовой оси.

Простейшие примеры линейных точечных множеств — *отрезок* и *интервал*.

Отрезком (иногда *сегментом*) называют множество точек числовой оси, удовлетворяющих неравенствам

$$a \leq x \leq b.$$

Отрезок обозначают символом $[a, b]$.

Интервалом (или *промежутком*) называют множество точек x числовой оси, удовлетворяющих неравенствам

$$a < x < b.$$

Интервал принято обозначать символом (a, b) .

Из определения видно, что отрезок $[a, b]$ отличается от интервала (a, b) тем, что точки $x = a$ и $x = b$ принадлежат отрезку и не принадлежат интервалу. Таким образом, между этими множествами установлено определенное соотношение. Рассмотрим некоторые соотношения, которые могут существовать между множествами.

§ 3. СООТНОШЕНИЯ МЕЖДУ МНОЖЕСТВАМИ

Включение. *Множество A входит (включено) в множество B* , если каждый элемент множества A входит также в множество B . Это соотношение будем записывать $A \subset B$. Часто его формулируют иначе: «Множество A составляет часть множества B », «множество A является *подмножеством* множества B ».

Если в множестве B есть элементы b , не входящие в множество A , то говорят, что множество A составляет *правильную часть* множества B . Например, множество прямоугольников входит в множество параллелограммов, которое в свою очередь входит в множество четырехугольников.

Если множество A определяется свойством α , множество B — свойством β и $A \subset B$, то любой объект, обладающий свойством α , должен обладать также и свойством β .

Если всякий объект, обладающий свойством α , обладает также и свойством β , то говорят, что свойство α включает свойство β : $\alpha \supset \beta$. Например, свойство числа делиться на 4 включает свойство числа быть четным. Таким образом, соотношению $A \subset B$ между множествами соответствует соотношение $\alpha \supset \beta$ между свойствами, определяющими эти множества.

Сумма множеств. Прибавить к множеству A множество B — значит образовать новое множество C , включающее как все элементы множества A , так и все элементы множества B , не входящие в множество A . Объект c входит в множество A и B , если он входит в множество A или в множество B , т. е. $c \in C$ в том и только в том случае, когда $c \in A$ или $c \in B$.

Сумма множеств записывается: $A \cup B$. Множество чисел C , делимых на 2 или на 3, есть сумма множества P чисел,

делящихся на 2, и множества Q чисел, делящихся на 3. В эту сумму войдут числа 2, 3, 4, 6, 8, 9, ... 10, 12, ..., из них числа 6, 12, ... входят как в множество P , так и в множество Q ; в множество C они входят только один раз. Отсюда следует, например, что $A \cup A = A$.

Сумма множества A , элементы которого обладают свойством α , и множества B , элементы которого обладают свойством β , есть множество, элементы которого обладают свойством α или свойством β . Новое свойство обладать свойством α или свойством β будем называть произведением свойств α и β и обозначать: $\alpha \vee \beta$ (иногда просто $\alpha\beta$). Например, элементы указанного выше множества C обладают произведением свойств: 1) быть четными, 2) делиться на 3.

Таким образом, *сумме множеств соответствует произведение свойств*, определяющих эти множества.

Перечисление множеств. Если рассмотреть множество людей, населяющих европейский материк, и множество людей, населяющих Советский Союз, то можно заметить, что среди них есть множество людей, которые по делению, принятому в физической географии, относятся к населению Европы, а по политико-административному делению — к населению Советского Союза. «Пересечение» этих множеств (европейцев и граждан СССР) предстает собой множество людей, населяющих Европейскую часть территории СССР. Образовать пересечение множества A с множеством B — значит образовать множество C из всех элементов множества A , входящих в множество B . Объект c входит в множество C , называемое *пересечением* (или *произведением*) множеств A и B , если он входит как в множество A , так и в множество B ; т. е. $c \in C$ в том и только в том случае, когда $c \in A$ и $c \in B$.

Пересечение множеств A и B запишем в виде $A \cap B$ (можно встретить также обозначения $A \cdot B$ или просто AB). Из определения пересечения множеств, в частности, следует, что

$$A \cap A = A.$$

Пересечение множества чисел, делящихся на 2, и множества чисел, делящихся на 3, есть множество чисел, делящихся на 6.

Если множества A и B не имеют общих элементов, то их пересечение представляет собой пустое множество: $A \cap B = O$, где O — пустое множество.

Если элементы множества A обладают свойством α , а элементы множества B обладают свойством β , то элементы

пересечения $A \cap B$ должны обладать как свойством α , так и свойством β . Свойство обладать как свойством α , так и свойством β , мы будем называть *суммой свойств* α и β и обозначим так: $\alpha \wedge \beta$ (или $\alpha \& \beta$).

Таким образом, *пересечению (произведению) множеств соответствует сумма свойств, определяющих эти множества*.

Разность множеств. Разностью двух множеств M и A называют множество B тех элементов множества M , которые не входят в множество A .

Разность множеств обозначается $M - A$ или $M \setminus A$. Пусть элементы множества A обладают свойством α . Об объектах, не обладающих свойством α , мы будем говорить, что они обладают свойством $\bar{\alpha}$ («не α »). Если множество M обладает свойством μ , то разность $M - A$ обладает суммой свойств $\mu \wedge \bar{\alpha}$. Обозначим множество объектов, обладающих свойством $\bar{\alpha}$, так: \bar{A} . Тогда разность $M - A$, обладая свойством $\mu \wedge \bar{\alpha}$, будет пересечением MA . Например, множество иррациональных чисел представляет собой разность между множествами действительных и рациональных чисел. А рассмотренное нами множество людей, населяющих Европейскую часть Союза ССР, можно получить в виде разности множества людей, населяющих СССР, и множества людей, населяющих Азию.

§ 4. ЭКВИВАЛЕНТНЫЕ МНОЖЕСТВА. МОЩНОСТЬ МНОЖЕСТВА

Два конечных множества можно сравнивать по числу элементов. Если множество A имеет m , а множество B — n элементов, то справедливо только одно из трех соотношений

$$m = n; \quad m < n; \quad m > n.$$

Если множества бесконечные, то сравнивать их по числу элементов нельзя, так как тогда нет смысла говорить о числе их элементов. Но конечные множества можно сравнивать и другим способом.

Пусть A и B — конечные множества и пусть по некоторому правилу удалось каждому элементу $a_i \in A$ *поставить в соответствие* один и только один элемент $b_i \in B$. Такое попарное соответствие между элементами двух множеств называется взаимно-однозначным или 1 — 1-соответствием.

В этом случае говорят, что между множествами A и B удалось установить взаимно-однозначное соответствие, а сами множества A и B называются *эквивалентными*.

Эквивалентность множеств обозначается

$$A \sim B.$$

Взаимно-однозначное соответствие между элементами конечных множеств можно установить только тогда, когда число элементов A и B одно и то же. Поэтому для конечных множеств понятие «эквивалентность» совпадает с понятием «равнозначность».

Рассмотрим свойства эквивалентных множеств.

1. *Свойство симметрии (или взаимности)*. Если множества A и B эквивалентны, то эквивалентны также B и A , т. е. если $A \sim B$, то $B \sim A$.

2. *Свойство транзитивности (или переходности)*. Если эквивалентны множества A и B , а также множества B и C , то множества A и C тоже эквивалентны. Иначе, два множества A и B , эквивалентные третьему множеству C , эквивалентны между собой, т. е. если $A \sim C$ и $B \sim C$, то $A \sim B$.

3. *Свойство рефлексивности*. Каждое множество A эквивалентно самому себе: $A \sim A$.

Очевидно, что два эквивалентных конечных множества A и B содержат одинаковое число элементов. Число элементов конечного множества есть то общее, что присуще всем эквивалентным друг другу конечным множествам. Учитывая это, сформулируем следующее определение.

Мощностью произвольного множества A называется то общее, что есть у всех множеств, эквивалентных данному.

Эквивалентные множества равномощны. Все множества можно разделить на классы равномощных множеств.

Тогда выражение «мощность множества» будет означать принадлежность множества к тому или иному классу.

У конечных эквивалентных множеств общим является число элементов множества. Неэквивалентные между собой конечные множества состоят из различного числа элементов. Таким образом, под мощностью конечного множества обычно подразумевается число элементов этого множества. В этом случае классу конечных эквивалентных множеств, состоящих из n элементов, ставится в соответствие число n , называемое *кардинальным числом* данного класса.

Мощность всех бесконечных счетных множеств одинакова, так как каждое из них, по определению, эквивалентно одному и тому же множеству всех чисел натурального ряда.

В силу свойства транзитивности они, следовательно, эквивалентны между собой. К таким множествам относятся множества всех четных чисел, множество кубов целых чисел и т. д. Их называют еще *перечислимыми*, но не потому, что элементы бесконечного множества можно перечислить исчерпывающим образом, а потому, что их элементы всегда перечисляемы в порядке своих номеров, причем этот процесс можно продолжать сколь угодно далеко.

Из всех бесконечных множеств счетные множества имеют наименьшую мощность, если, конечно, существуют бесконечные множества, не эквивалентные счетному. Существование таких множеств доказывает следующая теорема.

Теорема Кантора. *Множество P точек отрезка $[0, 1]$ неэквивалентно множеству N натуральных чисел.*

Доказательство. Допустим, что множество $P = [0, 1]$ счетно, т. е. точки этого отрезка можно представить занумерованной последовательностью

$$x_1, x_2, \dots, x_n, \dots \quad (1.1)$$

Разделим отрезок $[0, 1]$ на три равных отрезка. Тогда по крайней мере один из этих отрезков не содержит точки x_1 . Точка x_1 может принадлежать либо одному частному отрезку, либо двум, если это их пограничная точка. Отрезок Δ_1 , не содержащий точки x_1 , снова разделим на три равных отрезка. По крайней мере один из них, Δ_2 , не содержит точки x_2 . Отрезок Δ_2 второго деления, не содержащий точки x_2 , снова разделим на три равных отрезка и т. д. В результате получается последовательность отрезков $\Delta_1, \Delta_2, \dots, \Delta_n, \dots$, вложенных друг в друга. Пусть x_0 — точка, принадлежащая всем этим отрезкам. Тогда, с одной стороны, $x_0 \in [0, 1]$ и, следовательно, совпадает с одной из точек x_n -последовательности (1.1). С другой стороны, точка x_0 не может совпадать ни с одной точкой x_n -последовательности (1.1), так как точка x_n не принадлежит отрезку Δ_n , а точка x_0 входит в этот отрезок.

Это противоречие доказывает неверность предположения об эквивалентности P и N .

Все множества, эквивалентные множеству точек отрезка $[0, 1]$ (так называемые множества элементов непрерывной протяженности), называются *множествами мощности континуума*.

Существуют множества, мощность которых больше мощности континуума. Например, множество всех математических функций.

Важным в теории множеств является утверждение, что множества с наибольшей мощностью не существует, точно так же, как не существует самого большого натурального числа. Это утверждение доказывает следующая теорема.

Множество всех частей данного множества имеет мощность большую, чем мощность данного множества.

Д о к а з а т е л ь с т в о. Пусть A — произвольное множество и B — множество всех его подмножеств. В число подмножеств множества A входит также само множество A и пустое множество O .

Очевидно, в B есть часть B_0 , эквивалентная A . Это совокупность всех одноэлементных подмножеств множества A . Необходимо, следовательно, доказать, что A и B неэквивалентны.

Допустим обратное: $A \sim B$. Тогда каждому подмножеству $b \in B$ взаимно однозначно соответствует элемент $a \in A$. Разобьем все элементы множества A на два класса. К первому отнесем элементы, входящие в соответствующие им подмножества; так, элемент a' , соответствующий $b' = A$, входит в первый класс. Ко второму классу отнесем элементы, не входящие в соответствующее подмножество; так, элемент a'' , соответствующий $b'' = O$, входит во второй класс. Рассмотрим совокупность всех элементов второго класса. Это некоторое подмножество b_0 множества A , т. е. некоторый элемент множества B . Так как мы допустили, что $A \sim B$, элементу $b_0 \in B$ соответствует элемент $a_0 \in A$. Попытаемся определить, какому классу принадлежит элемент a_0 . Пусть a_0 принадлежит первому классу, т. е. входит в соответствующее ему подмножество b_0 . Но это невозможно, так как b_0 представляет собой совокупность элементов второго класса.

Предположим, что a_0 принадлежит второму классу, т. е. не входит в соответствующее ему подмножество b_0 . Но это опять невозможно, так как в b_0 собраны все элементы второго класса.

Таким образом, a_0 не может принадлежать ни к первому, ни ко второму классу. Но ведь любой элемент множества A , в том числе и a_0 , должен принадлежать либо первому, либо второму классу. Это противоречие показывает несправедливость допущения об эквивалентности A и B и тем самым доказывает сформулированную теорему.

Подсчитаем, сколько подмножеств содержится в конечном множестве A , состоящем из n элементов: пустое подмножество O , C_n^1 — одноэлементных подмножеств, C_n^2 — двух-

элементных подмножеств, ..., C_n^k — k -элементных подмножеств, ..., $1 = C_n^n$ — само множество.

Итого — $C_n^0 + C_n^1 + C_n^2 + \dots + C_n^k + \dots + C_n^n = 2^n$ подмножеств.

Если A бесконечное множество мощности α , то по аналогии со случаем конечного множества мощность множества B всех его подмножеств обозначают 2^α .

Таким образом, доказанная выше теорема утверждает, что

$$2^\alpha > \alpha.$$

В теории множеств, в частности, доказывается, что

$$2^a = c,$$

где a — мощность бесконечного счетного множества, а c — мощность континуума.

§ 5. ОСНОВНЫЕ ТЕОРЕМЫ

Рассмотрим некоторые важные теоремы о счетных множествах и множествах мощности континуума.

Теорема 1. *Сумма конечного или счетного множества конечных или счетных множеств является конечным или счетным множеством.*

Следует различать несколько случаев. Сумма конечного числа конечных множеств, очевидно, является конечным множеством. При рассмотрении счетного множества конечных множеств могут быть два случая. Если число отличающихся друг от друга элементов среди всех элементов множеств-слагаемых конечно, то сумма этих множеств, очевидно, конечна. Если число отличающихся друг от друга элементов в множествах-слагаемых бесконечно, то сумма будет счетной. Пронумеруем элементы первого множества-слагаемого в каком-либо порядке, например,

$$a_1, a_2, \dots, a_{n_1}.$$

Добавим отличающиеся от них элементы второго множества (если они есть) и пронумеруем их. Имеем

$$a_1, a_2, \dots, a_{n_1}, a_{n_1+1}, a_{n_1+2}, \dots, a_{n_2} \text{ и т. д.}$$

Поскольку различных элементов бесконечное множество, этот процесс не может оборваться после конечного числа шагов. Мы получаем бесконечную последовательность, т. е. счетное множество

$$a_1, a_2, \dots, a_m, \dots$$

Очевидно, что любой элемент любого множества — слагаемого A_k на некотором шаге попадает в эту последовательность, т. е. она действительно будет суммой данных множеств A_1, A_2, \dots, A_n .

Докажем, что *сумма счетного множества счетных множеств есть счетное множество*.

Пусть

$$B = \bigcup_{n=1}^{\infty} A_n,$$

где

$$A_n = \{a_1^{(n)}, a_2^{(n)}, \dots, a_k^{(n)}, \dots\}.$$

Назовем рангом элемента $a_i^{(j)}$ сумму его верхнего и нижнего индекса. Будем нумеровать $a_i^{(j)}$ в порядке возрастания рангов. Если элементы имеют равные ранги, нумерацию будем производить в порядке возрастания нижнего индекса, пропуская уже занумерованные элементы. Имеем

$$b_1 = a_1^{(1)}; \quad b_2 = a_1^{(2)}; \quad b_3 = a_2^{(1)}; \quad b_4 = a_1^{(3)}, \dots,$$

если все элементы различны. Если $a_2^{(2)}$ совпадает, например, с $a_2^{(1)}$, то $a_2^{(2)}$ пропускаем, поэтому $b_5 = a_3^{(1)}$, $b_6 = a_1^{(4)}$, $b_7 = a_2^{(3)}$, если каждый из этих элементов отличается от уже занумерованных, и т. д. Таким образом, мы занумеруем элементы множества B в последовательность, т. е. B будет счетным множеством.

Лемма. *Всякое бесконечное подмножество счетного множества есть счетное множество.*

Доказательство. Пусть $A = \{a_i\}$ — счетное множество и A' — бесконечное подмножество A . Пусть a_{n_1} — элемент из A' с наименьшим номером. Обозначим его b_1 . Далее, a_{n_2} — элемент из A' с наименьшим номером, большим, чем n_1 , обозначим b_2 . Элемент a_{n_3} из A' с наименьшим номером, большим, чем n_2 , обозначим b_3 и т. д. Ясно, что таким образом можно занумеровать A' в последовательность $b_1, b_2, \dots, b_k, \dots$.

Из доказанной теоремы получается ряд важных следствий. Приведем их без доказательства.

1. *Если элементы множества A снабжены конечным числом индексов, каждый из которых независимо от других принимает счетное множество значений, то A — счетное множество.*

2. *Множество рациональных чисел есть счетное множество.* Действительно, любое рациональное число $r = \frac{p}{q}$

определяется двумя индексами p и q . Каждый из них принимает счетное множество значений. Поэтому на основании следствия 1 множество рациональных чисел счетно.

3. Множество рациональных точек (т. е. точек с рациональными координатами) n -мерного пространства счетно.

Рассмотрим еще несколько теорем.

Теорема 2. Из всякого бесконечного множества можно выделить конечное или счетное подмножество так, что оставшееся множество будет эквивалентно первоначальному.

Теорема 3. Если к бесконечному множеству прибавить конечное или счетное множество, то мощность полученного множества равна мощности данного множества.

Теперь перейдем к множествам мощности континуума.

Очевидно, всякий отрезок имеет мощность континуума. Равенства

$$y = a + (b - a)x, \quad 0 \leq x \leq 1,$$

$$x = \frac{y - a}{b - a}, \quad a \leq y \leq b$$

устанавливают взаимно однозначное соответствие между точками отрезков $[a, b]$ и $[0, 1]$, следовательно, отрезки эквивалентны.

В соответствии с теоремой 3 добавление конечного числа точек не меняет мощности множества. Так что любой конечный интервал и полуинтервал имеет мощность континуума. Любой бесконечный интервал также имеет мощность континуума. Например, формулы

$$y = \operatorname{tg} x, \quad \operatorname{arctg} a < x < \frac{\pi}{2},$$

$$x = \operatorname{arctg} y, \quad a < y < \infty$$

устанавливают взаимно однозначное соответствие между точками интервалов

$$\left(\operatorname{arctg} a, \frac{\pi}{2}\right) \text{ и } (a, \infty), \quad a \geq 0.$$

Из счетности множества всех рациональных чисел и теоремы 3 следует, что множество иррациональных точек отрезка $[0, 1]$ (а следовательно, и любого отрезка или интервала) имеет мощность континуума.

Теорема 4. Сумма конечного или счетного множества множеств мощности континуума имеет мощность континуума.

Прежде чем переходить к следующей теореме, остановимся на представлении вещественных чисел отрезка $[0, 1]$ в виде двоичных дробей.

Любое вещественное число α , лежащее на отрезке $[0, 1]$, можно записать в виде

$$\alpha = \frac{\alpha_1}{2} + \frac{\alpha_2}{2^2} + \dots + \frac{\alpha_k}{2^k} + \dots, \quad (1.2)$$

где $\alpha_k = 0$ или 1, подобно тому как, представляя это число в виде десятичной дроби, мы имеем

$$\alpha = \frac{a_1}{10} + \frac{a_2}{10^2} + \dots + \frac{a_k}{10^k} + \dots,$$

где $a_k = 0$, или 1, или 2, ..., или 9.

Например,

$$\frac{5}{6} = \frac{1}{2} + \frac{1}{2^2} + \frac{1}{2^3} + \frac{0}{2^4} + \frac{1}{2^5} + \frac{0}{2^6} + \frac{1}{2^7} + \dots$$

Равенство (1.2) можно сокращенно записать в виде

$$\alpha = 0, \alpha_1, \alpha_2, \dots, \alpha_k, \dots \quad (1.3)$$

Следует отметить, что рациональное число вида $\frac{p}{2^k}$ можно записать в виде двоичной дроби двояко: либо с нулем, либо с единицей в периоде. Например, для дроби $\frac{3}{4}$ имеем

$$\frac{3}{4} = \frac{1}{2} + \frac{1}{2^2} + \frac{0}{2^3} + \frac{0}{2^4} + \dots = 0,1100 \dots,$$

$$\frac{3}{4} = \frac{1}{2} + \frac{0}{2^2} + \frac{1}{2^3} + \frac{1}{2^4} + \dots = 0,1011 \dots$$

Если не пользоваться записью, содержащей единицу в периоде, то каждое число $\alpha \in [0, 1]$ однозначно представимо в виде дроби (1.3).

Теорема 5. Пусть $A = \{a_{i_1, i_2, i_3, \dots, i_k}\}$ множество элементов, определяемых счетным числом параметров $i_1, i_2, \dots, i_k, \dots$, каждый из которых, независимо от других, может принимать два значения, a и b . Тогда множество A имеет мощность континуума.

Доказательство. Каждому элементу $a_{i_1, i_2, \dots, i_k, \dots}$ ставится в соответствие двоичная дробь $0, \alpha_1, \alpha_2, \dots, \alpha_k, \dots$ по следующему правилу:

$$\alpha_k = \begin{cases} 1, & \text{при } i_k = a, \\ 0, & \text{при } i_k = b. \end{cases}$$

Обратно, каждой двоичной дроби $0, \alpha_1, \alpha_2, \dots, \alpha_k \dots$ по этому же закону ставится в соответствие элемент $a_{l_1, l_2, \dots, l_k} \in A$. Таким образом, получается, что множество A и множество всех двоичных дробей, в том числе и с единицей в периоде, эквивалентны. Множество всех возможных двоичных дробей отличается от множества всех вещественных чисел отрезка $[0, 1]$ на множество двоичных дробей с единицей в периоде, которые *изображают* двоично рациональные числа и которых, следовательно, счетное множество. Поскольку прибавление счетного множества к множеству мощности континуума или удаление из него счетного множества не меняет мощности этого множества, то совокупность всех двоичных дробей имеет ту же мощность, что и отрезок $[0, 1]$, т. е. мощность континуума. Но тогда и множество A имеет мощность континуума.

Теорема 6. *Множество всех последовательностей, составленных из чисел натурального ряда, имеет мощность континуума.*

Для доказательства прежде всего покажем, что *множества всех возможных последовательностей и всех возрастающих последовательностей натуральных чисел эквивалентны*. Действительно, любой произвольной последовательности

$$m_1, m_2, \dots, m_k, \dots$$

можно поставить в соответствие возрастающую последовательность

$$n_1 = m_1, n_2 = m_1 + m_2, n_3 = m_1 + m_2 + m_3, \dots$$

С другой стороны, любой возрастающей последовательности

$$n_1, n_2, \dots, n_k, \dots$$

можно поставить в соответствие некоторую последовательность

$$m_1 = n_1, m_2 = n_2 - n_1, m_3 = n_3 - n_2, \dots,$$

и требуемая эквивалентность доказана.

Рассмотрим множество всех возрастающих последовательностей натуральных чисел. Возрастающей последовательности $n_1, n_2, \dots, n_k, \dots$ ставим в соответствие двоичную дробь

$$\alpha = 0, \alpha_1, \alpha_2, \dots, \alpha_k.$$

При этом полагаем $\alpha_i = 1$, если i совпадает с каким-либо членом n_k последовательности, и $\alpha_i = 0$ для i , не

совпадающих ни с одним членом последовательности. С другой стороны, дроби

$$\beta = 0, \beta_1, \beta_2, \beta_3, \dots, \beta_k \dots$$

ставим в соответствие возрастающую последовательность натуральных чисел, составленную из индексов тех двоичных знаков, которые равны 1. Таким образом, множество всех возрастающих последовательностей чисел натурального ряда эквивалентно множеству всех двоичных дробей и, следовательно, имеет мощность континуума.

§ 6. ФУНКЦИИ. ОТНОШЕНИЯ. СПОСОБЫ ЗАДАНИЯ ФУНКЦИЙ

При изучении явлений природы математическими методами вместо физических величин рассматривают измеряющие их числа. Таким образом, соответствие между величинами заменяется соответствием между числами.

Если при рассматриваемых условиях некоторая величина может принимать различные числовые значения, то эта величина называется *переменной*. Если же некоторая величина при рассматриваемых условиях имеет вполне определенное, неизменное значение, то она называется *постоянной величиной*, или *константой*.

Переменные величины обозначаются буквами x, y, z, u, \dots ; постоянные — буквами a, b, c . Ни в одной задаче переменные величины не встречаются изолированно друг от друга. Изменение значений некоторых величин влечет за собой изменение значений других величин.

Если в силу некоторого закона или свойства каждому значению переменной x отвечает одно или несколько определенных значений переменной y , то переменную y называют функцией переменной x . Записывают это в виде

$$y = f(x), y = F(x), y = \varphi(x) \dots$$

По своему содержанию понятие *функции* совпадает с понятием *соответствия*. В самом деле, если есть два множества $X = \{x\}$ и $Y = \{y\}$ какой угодно мощности, составленные из каких угодно элементов x и y , и если каждому элементу x множества X соответствует один или несколько элементов y множества Y , то говорят, что существует функция, определенная на множестве X , и пишут символическое равенство

$$y = f(x),$$

где y — тот самый элемент множества Y , который соответствует элементу x множества X .

Элементы x множества X — значения аргумента, а все множество X — множество значений аргумента или область задания (определения) данной функции.

Переменная x — независимая, переменная y — зависимая.

Частное значение функции находят подстановкой в выражение для функции вместо независимой переменной конкретного числового значения. Если частное значение функции $y = f(x)$ при $x = a$ равно b , то записывают $b = f(a)$.

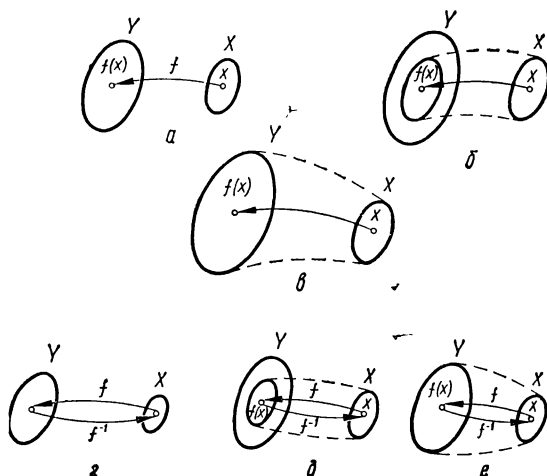


Рис. 1.1. Соотношения.

Всякое конкретное значение аргумента принято также называть *прообразом*, а соответствующее ему значение функции $f(a)$ — *образом*. Следовательно, каждому элементу множества X соответствует его образ в множестве Y (рис. 1.1, а). Вместо слова «функция» употребляется термин *отображение*.

Функции, значения которых определяются по переменной, принадлежащей к одному множеству, называются *функциями одной переменной*, или *одноместными*.

Если некоторая величина зависит не от одной, а от двух переменных, то математическая функция, описывающая эту величину, определяется не одним, а двумя аргументами. Это *функция двух переменных*, или *двуместная функция*:

$$z = f(x, y).$$

Областью определения двуместной функции являются множества $X = \{x\}$ и $Y = \{y\}$.

Область значений функции находится в множестве $Z = \{z\}$.

В общем случае n -местная функция применима к упорядоченной системе n аргументов и дает некоторое значение при условии, что упорядоченная система n аргументов принадлежит к области определения функции

$$v = f(x, y, z, u, \dots).$$

Если функция $f(x)$ всюду определена, т. е. если область определения совпадает с X , то $f(x)$ называется *отображением множества X в Y* (рис. 1.1, б). Иногда кратко это записывается в виде $\Gamma_Y(X)$.

Если образ всего множества X равен Y , т. е. если каждый элемент из Y есть образ по крайней мере одного элемента из X , то говорят, что имеет место *отображение X на Y* .

В этом случае $f(x)$ называется также *сюръективным отображением* (рис. 1.1, в).

Если X и Y совпадают, то $y = f(x)$ есть *отображение X в X* . Элемент x , удовлетворяющий отношению $x = f(x)$, называется *неподвижной точкой отображения $f(x)$* .

Если между x и $f(x)$ установлено взаимно однозначное соответствие, то f называется *инъективной* (или *однозначной*) функцией, а f^{-1} *обратной* (рис. 1.1, г).

Если, кроме того, это соответствие всюду определено, то $f(x)$ называется *инъективным отображением, или инъекцией* (рис. 1.1, д).

Отображения, которые одновременно являются сюръективными и инъективными, называют *биективными отображениями, или биекциями* (рис. 1.1, е).

Функция считается *заданной, или определенной*, если указана совокупность всех значений, которые принимают независимые переменные, и способ определения значений функции по данным значениям независимой переменной. Основными способами задания функций являются: табличный, графический и аналитический.

1. Если приведена таблица, в которой указаны числовые значения аргументов и соответствующие значения функции, то говорят, что функция задана *таблично*. Например, таблицы логарифмов чисел, таблицы значений тригонометрических функций и т. п.

В виде таблиц обычно представляют результаты экспериментальных исследований, например, изменение какого-

либо физического параметра во времени. В таком виде получаются зависимости при автоматической регистрации, например, при использовании информационных цифровых машин.

Табличное задание функций удобно прежде всего потому, что значения функции для конкретных значений аргументов можно определять без дополнительных вычислений. К недостаткам табличного задания функции следует отнести малую наглядность. Трудно бывает судить об общем характере изменения функции при изменении аргументов, особенно при большом количестве данных. Кроме того, таблицы дают значения функций лишь для определенных дискретных значений аргументов.

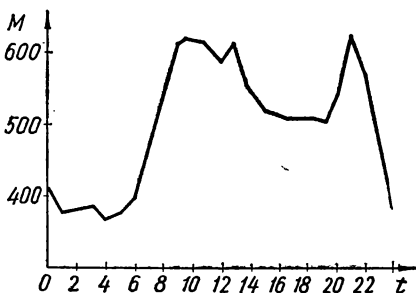


Рис. 1.2. Суточные измерения нагрузки энергосистемы.

2. Функции одной и двух переменных могут быть заданы *графически*. Функции одной независимой переменной представляют собой графики в плоскости xOy . Каждая точка такого графика характеризуется значением абсциссы — величиной независимой переменной и соответствующей ординатой — значением функции. На рис. 1.2 изображена за-



Рис. 1.3. Автоматическая запись изменения качества проката.

висимость изменения нагрузки энергосистемы в течение суток $P(t)$. Здесь независимой переменной является время t . Другой пример — графики тригонометрических функций $y = \sin x$; $y = \operatorname{tg} x$ и т. д. Это типичный пример графического задания функции.

Очень часто мы получаем графически заданные зависимости различных физических переменных при помощи самопишущих приборов, установленных на промышленных

объектах. На рис. 1.3 показаны записанные самопишущим прибором изменения качества проката на металлургическом комбинате.

Графическое задание функции очень наглядно. Часто лишь беглый взгляд на график позволяет определить такие закономерности и тенденции в изменении переменной, которые не являются очевидными при других способах задания функций.

Функция двух независимых переменных $z = f(x, y)$ может быть задана в пространстве трех измерений. Значения аргументов при этом откладывают по осям x и y горизонтальной плоскости xOy , а значения функции $f(x, y)$ отсчитывают по вертикальной оси z .

На практике графический способ задания применяется обычно для функций одной переменной, реже для функций двух переменных. Графическое задание n -местной функции ($n > 2$) в трехмерном пространстве лишено всякой наглядности.

3. Если закон соответствия между переменными задается математической формулой, то говорят, что функция задана *аналитически*. Преимущества аналитического способа задания функции — компактность выражения, возможность определения значений функции многих переменных и вычисления этих значений.

К недостаткам аналитического задания следует отнести недостаточную наглядность, а иногда и громоздкость вычислений для сложных аналитических зависимостей.

Все указанные способы задания функций могут встречаться изолированно. Но практически чаще всего используются в некоторых комбинациях. Например, почти всегда полученную экспериментально в виде таблицы зависимость целесообразно для наглядности представить в графической форме. Во многих случаях ставится задача описать полученную зависимость некоторым аналитическим выражением. Возникает необходимость аппроксимации, или аналитического продолжения функций.

§ 7. ИЗОМОРФИЗМ

Часто приходится рассматривать множества физически разнородных элементов, между которыми можно, однако, установить взаимно однозначное соответствие.

Рассмотрим в качестве примера какой-либо прибор или аппарат — радиоприемник, трансформатор, металлорежу-

щий станок и т. д. Любой из них — это совокупность элементов, обладающих определенными свойствами. Кроме того, между элементами существуют определенные взаимосвязи. Такие совокупности (множества) обычно называют *системами*.

Однако тот же радиоприемник или трансформатор можно представить в виде чертежа, принципиальной или монтажной схемы, на которых все элементы и взаимосвязи между ними выражены условными обозначениями и линиями. Кроме того, можно описать системы элементов и связей математическими уравнениями, выражающими взаимодействие частей. Например, металлорежущий станок можно представить графически в виде кинематической схемы, или аналитически, системой уравнений движения элементов и частей.

В любом из этих представлений реальных физических систем можно дать исчерпывающую характеристику элементов и связей между ними.

Если при решении задачи отвлечься от качественной физической природы элементов, то каждая система из конкретной превращается в абстрактную. В этом случае устанавливается только структура системы, а природа ее элементов остается неопределенной во всех отношениях. Говорят, что *все представители одной и той же абстрактной системы или схемы изоморфны*.

В теории множеств два множества X и Y называются изоморфными, если выполняются следующие условия:

1. Каждый элемент $x \in X$ может быть взаимно однозначно сопоставлен с элементом $y \in Y$, т. е. $x \rightarrow y$ и $y \rightarrow x$.

2. Каждая операция f (из некоторого класса операций, выражающих отношения между элементами множества X) может быть взаимно однозначно сопоставлена с операцией φ в множестве Y .

Если операция f преобразует $x_1 \in X$ в $x_2 \in X$ в множестве X (например, изменение состояния электронной лампы под воздействием входного сигнала), то ей должна соответствовать операция φ , преобразующая элемент $y_1 \in Y$ в $y_2 \in Y$ (математическая операция преобразования сигналов в уравнении, описывающем процессы в электронной лампе).

3. Если $x_1 \in X$ соответствует $y_1 \in Y$ и $x_2 \in X$ соответствует $y_2 \in Y$, а также $f(x_1) = x_2$ и $f \rightarrow \varphi$, то для всех x , y , f будет иметь место $\varphi(y_1) = y_2$.

Два множества изоморфны, если их элементы попарно взаимно-однозначно соответствуют друг другу и преобразо-

вания элементов одного множества соответствуют преобразованиям соответствующих элементов другого множества.

Понятие изоморфизма является одним из фундаментальных в кибернетике. В этом понятии теоретически обосновывается очень важное в практическом отношении положение. В любом исследовании изучаемое явление на отдельных этапах можно заменить изоморфной ему *моделью* и дальнейшее изучение свести к изучению модели.

Особое значение приобрели математические описания (математические модели) физических процессов и объектов в связи с созданием быстродействующих универсальных цифровых вычислительных машин (ЦВМ). С помощью вычислительной техники и математических моделей можно исследовать многие сложные системы и процессы, которые ранее были недоступны детальному анализу из-за сложности экспериментирования в реальных условиях и ограниченного времени.

Контрольные вопросы и задания

1. Перечислите способы задания множеств. Придумайте примеры.
2. Охарактеризуйте линейные точечные множества.
3. В каких соотношениях могут находиться множества элементов? Приведите примеры всех видов соотношений.
4. Сформулируйте основные свойства эквивалентных множеств. Что такое мощность множества?
5. Назовите основные виды отображений. Дайте их характеристику.
6. Какие вы знаете виды задания функций? Приведите конкретные примеры.
7. Сформулируйте условия изоморфности. Приведите примеры изоморфных множеств.

Глава 2

ЭЛЕМЕНТЫ МАТЕМАТИЧЕСКОЙ ЛОГИКИ

В алгебре действия над числами заменяются действиями над буквами. Буквы в алгебре заменяют или любые или некоторые вполне определенные числа. Слова «сложить», «вычесть», «разделить» и т. п. заменяются в алгебре значками-символами: $+$, $-$, $:$ и т. п. Пользуясь алгебраическими обозначениями, кубическое уравнение можно кратко записать так: $x^3 + ax = b$. Но до такой краткости записи ученые дошли только в XVIII в. Итальянский математик Тарталья, нашедший решение кубического уравнения (1535), записывал его примерно так: «куб» плюс некоторое количество «вещей» равно известному «числу». При указании решения неизвестное (x) именовали «вещью», «количеством» вещей называли коэффициент (a) при x , а b именовали «числом». Виета в конце XVI в. несколько усовершенствовал алгебраическую символику, но и у него это уравнение выглядит довольно сложно:

X cubus A planum X aequatur B solido.

Наши утверждения и отрицания об объектах и их свойствах напоминают записи Тартальи. Даже краткие формулировки вроде: «если объект a обладает свойством α , то он обладает свойством β » — недалеко ушли от записи Виеты.

Внедрение алгебраических обозначений в исследования в области логики началось со второй половины XVII в. Как самостоятельная научная область математическая логика возникла в середине XIX в. Выдающийся вклад в математико-логические исследования внесла советская школа математической логики (И. П. Жегалкин, В. И. Гливленко, А. Н. Колмогоров, П. С. Новиков, А. А. Марков).

В последние десятилетия математическая логика получила разнообразные технические приложения. Современная

математическая логика связана с автоматикой, вычислительной математикой, проблемой автоматического перевода с одного языка на другой, с теорией связи и передачи информации.

§ 1. ВЫСКАЗЫВАНИЯ. ИХ ИСТИННОСТЬ И ЛОЖНОСТЬ

Одним из разделов математической логики является исчисление высказываний. Предмет изучения исчисления высказываний *суждения (предложения, высказывания)*.

Для примера рассмотрим такие фразы:

дважды два — семь, $\sin x$ не больше единицы, число 4 — нечетное, снег белый, июль — летний месяц.

Поставим вопрос, какие из этих суждений верны, или истинны, а какие неверны, ложны. Из приведенных высказываний 1 и 3 ложны, 2 и 4 истинны. Что же касается 5-го высказывания, то оно истинно для северного полушария и ложно для южного. Существуют и такие высказывания, относительно которых мы не можем сказать, истинны они или ложны. Например, « $\sqrt{2}$ — число нечетное».

Нам известно, что такое «целое число нечетно», но что значит «иррациональное число нечетно» — этого мы не знаем.

В дальнейшем мы будем употреблять только такие высказывания, относительно которых известно, что они либо истинны, либо ложны, причем непременно одно из двух.

Существуют суждения, состоящие из нескольких простых соединенных между собой союзами: и, или, если — то и т. п. Одну и ту же мысль можно выразить как простым, так и сложным высказыванием.

Например, простое суждение « $\sin x$ не больше 1» можно заменить суждением « $\sin x$ меньше 1 или равен 1».

Вопрос об истинности или ложности суждений является основным вопросом, которым мы будем заниматься. Изучая суждения, мы абстрагируемся от их содержания, происхождения и других важных характеристик и сосредоточим все внимание на двух вопросах:

1. Как зависит истинность того или иного сложного высказывания от истинности входящих в его состав более простых суждений?

2. Как зависит истинность некоторых (или всех) простых высказываний, входящих в состав сложного, от истинности этого сложного высказывания и остальных простых, входящих в его состав?

Высказывания будем кратко обозначать большими буквами латинского алфавита

$$X, Y, Z, U, V, \dots$$

При этом различные буквы соответствуют различным высказываниям, а одни и те же буквы — одинаковым высказываниям. Сложные высказывания обозначим либо одной из букв латинского алфавита, либо буквами, соответствующими составляющим высказываниям, связанными особыми знаками.

§ 2. СВЯЗЬ ВЫСКАЗЫВАНИЙ. СИМВОЛЫ ЛОГИЧЕСКИХ СВЯЗЕЙ

Для выражения логической связи высказываний вводятся следующие 5 знаков.

1. \bar{X} (читается «не x ») обозначает противоположность X . \bar{X} обозначает высказывание, которое истинно, если X ложно, и ложно, если X истинно.

2. $X \wedge Y$ (читается « X и Y ») обозначает высказывание, которое истинно в том и только в том случае, когда X и Y истинны.

3. $X \vee Y$ (читается « X или Y ») обозначает высказывание, которое истинно в том и только в том случае, когда по крайней мере одно из двух высказываний X , Y является истинным.

4. $X \rightarrow Y$ (читается «если X , то Y ») обозначает высказывание, которое ложно в том и только в том случае, когда X истинно, а Y ложно.

5. $X \sim Y$ (читается « X равнозначно Y ») пишут также $X \rightleftharpoons Y$ или $X \leftrightarrow Y$) обозначает высказывание, которое истинно тогда и только тогда, когда X и Y оба истинны или X и Y оба ложны. Таким образом, $X \sim Y$ означает, что X и Y имеют одно и то же значение истинности или ложности.

Относительно третьего определения следует заметить, что сложное высказывание типа « X или Y » может принимать два различных значения. Например: «объект r входит в множество A или в множество B » можно понимать так: 1) объект r входит в одно и только в одно из двух множеств, 2) или множество A или множество B , т. е. что в оба эти множества объект r не может входить; это утверждение можно понимать и так, что объект r входит по крайней мере в одно из двух множеств, т. е. что не исключена возможность вхождения объекта r в оба множества. Мы будем рассматривать союз

«или» и соотношение \sim только в этом втором смысле. Исключающее «или — или» (в первом смысле) может быть выражено при помощи некоторой комбинации основных знаков. «Или X или Y » является отрицанием $X \sim Y$ и выражается так: $\overline{X \sim Y}$.

Соотношение 4 («если X , то Y ») не следует понимать как выражение для отношения основания и следствия. Напротив, высказывание $X \rightarrow Y$ истинно всегда уже в том случае, когда X ложно или же Y истинно. Например, истинными высказываниями являются: если « $3 \times 3 = 9$ », то «слон — животное», если $3 \times 3 = 7$, то «слон — животное», если $3 \times 3 = 7$, то «слон — птица», ложным же было высказывание: если « $3 \times 3 = 9$ », то «слон — птица».

Соотношение $X \rightarrow Y$ имеет общее с соотношением основания и следствия то, что в случае истинности $X \rightarrow Y$ из истинности X можно заключить об истинности Y .

Соотношение 5 ($X \sim Y$) не понимается как равносильность по смыслу X с Y ; оно имеет место между любыми двумя истинными, а также между любыми двумя ложными высказываниями.

Например, высказывания « $3 \times 3 = 9$ » \sim «слон — животное», « $3 \times 3 = 7$ » \sim «слон — птица» истинны.

Особо отметим то, что, в силу нашего определения основных логических связей, истинность или ложность сложного высказывания зависит только от истинности и ложности составляющих высказываний, а не от их содержания.

Введем еще два обозначения. Будем обозначать истинное высказывание буквой I , ложное — буквой L . Тогда, например, связь « \rightarrow » характеризуется тем, что высказывания $I \rightarrow I$, $L \rightarrow I$ и $L \rightarrow L$ являются истинными, а высказывание $I \rightarrow L$ — ложным.

Для связи « \wedge » высказывание $I \wedge I$ является истинным, а все остальные: $I \wedge L$, $L \wedge I$, $L \wedge L$ — ложными. Далее $I \vee I$, $I \vee L$, $L \vee I$ — истинны, а $L \vee L$ — ложны.

Связь « \sim » характеризуется тем, что $I \sim I$, $L \sim L$ истинны, между тем как $I \sim L$ и $L \sim I$ ложны. Наконец, I ложно, L истинно.

Таким образом, мы будем рассматривать основные связи как функции истинности, т. е. как определение функций, для которых в качестве аргументов и значений функций рассматриваются только I и L .

§ 3. ЭКВИВАЛЕНТНОСТЬ. ЗАМЕНЯЕМОСТЬ ОСНОВНЫХ СВЯЗЕЙ

Каждое сложное высказывание так же, как и простые связи высказываний, представляет собой определенную функцию истинности. Рассмотрим

$$[X \rightarrow X] \wedge (Y \rightarrow Z) \wedge (X \vee Z).$$

Для X, Y, Z возможны восемь троек значений

$I, I, I; I, I, Л; I, Л, I; I, Л, Л; Л, I, I;$

$Л, I, Л; Л, Л, I; Л, Л, Л.$

При подстановке в исходную формулу любой тройки получаем значения I или $Л$. Например, комбинации $Л, I, Л$ соответствует значение $Л$

$$[(Л \rightarrow I) \wedge (I \rightarrow Л)] \wedge (Л \vee Л)$$

$$I \wedge Л \wedge Л$$

$$Л \wedge Л$$

$$Л$$

Отметим, что некоторые различные из этих связей равнозначны, т. е. выражают ту же самую функцию истинности. Так, X равнозначно X . Подобные равнозначные связи будем называть эквивалентными

$$X \text{ экв } X, \quad (2.1)$$

$$X \wedge Y \text{ экв } Y \wedge X, \quad (2.2)$$

$$X \wedge (Y \wedge Z) \text{ экв } (X \wedge Y) \wedge Z, \quad (2.3)$$

$$X \vee Y \text{ экв } Y \vee X, \quad (2.4)$$

$$X \vee (Y \vee Z) \text{ экв } (X \vee Y) \vee Z, \quad (2.5)$$

$$X \vee (Y \wedge Z) \text{ экв } (X \vee Y) \wedge (X \vee Z). \quad (2.6)$$

Из эквивалентностей (2.2) — (2.6) следует перестановочный, сочетательный и распределительный законы (коммутативный, ассоциативный, дистрибутивный). Из приведенных законов видно, что в логических выражениях можно, как в алгебре, перемножать или выносить за скобки общий множитель. По аналогии с алгеброй $X \wedge Y$ называют логической суммой, а $X \vee Y$ логическим произведением. Но в отличие от алгебры, в исчислении высказываний действует второй дистрибутивный закон

$$X \wedge (Y \vee Z) \text{ экв } (X \wedge Y) \vee (X \wedge Z). \quad (2.7)$$

Таким образом, мы с успехом могли бы назвать $X \wedge Y$ логическим произведением, а $X \vee Y$ логической суммой.

Поскольку в логике относительно употребления слов «сумма» и «произведение» существует неопределенность, мы по возможности будем избегать этих выражений

$X \wedge Y$ будем называть конъюнкцией,

$X \vee Y$ » » дизъюнкцией,

$X \rightarrow Y$ » » импликацией.

В соответствии с законом коммутативности и ассоциативности многочисленные конъюнкции, дизъюнкции можно писать без скобок. Кроме того, для уменьшения количества скобок установим, что \vee связывает теснее, чем \wedge , а \wedge , в свою очередь теснее, чем \rightarrow и \sim . Знак \vee можно не ставить точно так же, как в алгебре не ставят знак умножения. Рассмотрим еще ряд эквивалентностей

$$X \wedge X \text{ экв } X, \quad (2.8)$$

$$X \vee X \text{ экв } X, \quad (2.9)$$

$$X \wedge I \text{ экв } X, \quad (2.10)$$

$$X \wedge L \text{ экв } L, \quad (2.11)$$

$$X \vee I \text{ экв } I, \quad (2.12)$$

$$X \vee L \text{ экв } X, \quad (2.13)$$

$$I \rightarrow X \text{ экв } X, \quad (2.14)$$

$$L \rightarrow X \text{ экв } I, \quad (2.15)$$

$$X \sim I \text{ экв } X, \quad (2.16)$$

$$X \sim L \text{ экв } \bar{X} \quad (2.17)$$

и несколько более сложных эквивалентностей

$$\overline{X \wedge Y} \text{ экв } \bar{X} \vee \bar{Y}, \quad (2.18)$$

$$\overline{X \vee Y} \text{ экв } \bar{X} \wedge \bar{Y}, \quad (2.19)$$

$$X \rightarrow Y \text{ экв } \bar{X} \vee Y. \quad (2.20)$$

Используя (2.18), можно (2.20) записать в виде

$$X \rightarrow Y \text{ экв } \bar{X} \vee Y$$

и далее в соответствии с (2.1)

$$X \rightarrow Y \text{ экв } \bar{X} \vee Y, \quad (2.21)$$

$$X \vee Y \text{ экв } \overline{X} \rightarrow Y, \quad (2.22)$$

$$X \rightarrow Y \text{ экв } \overline{Y} \rightarrow \overline{X}, \quad (2.23)$$

$$X \sim Y \text{ экв } (X \rightarrow Y) \wedge (Y \rightarrow X), \quad (2.24)$$

Из определения связи \sim непосредственно получаем, что

$$X \sim Y \text{ экв } Y \sim X, \quad (2.25)$$

$$X \sim Y \text{ экв } X \sim Y. \quad (2.26)$$

Из (2.18) и (2.19) получаем

$$X \wedge Y \text{ экв } \overline{\overline{X} \vee \overline{Y}}, \quad (2.27)$$

$$X \vee Y \text{ экв } \overline{\overline{X} \wedge \overline{Y}}. \quad (2.28)$$

Очевидно, что некоторые из основных логических связей излишни. Из (2.24) видно, что можно обойтись без знака \sim . Затем из (2.20) и (2.27) следует, что знаки \rightarrow и \vee также заменимы и можно обойтись только знаками \wedge и $\overline{}$. Из (2.21) и (2.28) следует, что можно ограничиться знаками \vee и $\overline{}$.

Можно обойтись также одним единственным логическим знаком, как это показал Шеффер. Знак / называется штрих Шеффера. X/Y означает: « X и Y несовместны». X/X тогда равнозначно \overline{X} . $X/X/Y/Y$ эквивалентно X/\overline{Y} , т. е. $X \vee Y$. А раз знаки \vee и $\overline{}$ можно выразить при помощи штриха Шеффера, то можно выразить и другие основные связи.

Приведем эквивалентности, важные для представления отношения равнозначности

$$X \sim Y \text{ экв } \overline{X} \vee Y \wedge \overline{Y} \vee X, \quad (2.29)$$

$$Y \sim Y \text{ экв } (X \wedge Y) \vee (\overline{X} \wedge \overline{Y}). \quad (2.30)$$

Наиболее целесообразно применять три знака \wedge , \vee , $\overline{}$, так как в силу эквивалентностей (2.2) — (2.6) при этом получается особенно простая вычислительная трактовка логических выражений.

П р и м е р. Найдем дизъюнктивную нормальную форму сложного высказывания

$$X \wedge (X \rightarrow Y).$$

Воспользовавшись эквивалентностью (2.21), получаем

$$X \wedge (\overline{X} \vee Y).$$

Раскрывая скобки, имеем

$$(X \wedge \overline{X}) \vee (X \wedge Y).$$

§ 4. НОРМАЛЬНАЯ ФОРМА ЛОГИЧЕСКИХ ВЫРАЖЕНИЙ

Каждое сложное высказывание можно привести к известной нормальной форме путем эквивалентного преобразования. Эта нормальная форма состоит из некоторой конъюнкции дизъюнкций, в которой каждый дизъюнктивный член является либо основным высказыванием, либо его отрицанием.

Правила преобразования. 1. Со знаками \wedge и \vee можно оперировать, как в алгебре, пользуясь ассоциативным, коммутативным и дистрибутивным законами.

2. $\overline{\overline{X}}$ можно заменить на X .

3. $\overline{X \wedge Y}$ можно заменить выражением $\overline{X} \vee \overline{Y}$, а $\overline{X \vee Y}$ — выражением $\overline{X} \wedge \overline{Y}$.

4. $X \rightarrow Y$ можно заменить выражением $\overline{X} \vee Y$, а $X \sim \sim Y$ — выражением $\overline{X}Y \wedge \overline{Y}X$.

Пример. Найдем конъюнктивную нормальную форму высказывания

$$(XY \wedge \overline{Y}) \vee (Z \wedge Y).$$

Последовательно получаем

$$\begin{aligned} & \overline{(XY \wedge \overline{Y}) \wedge (\overline{Z} \wedge \overline{Y})}, \\ & \overline{XY \wedge \overline{Y} \wedge \overline{Z} \vee \overline{Y}}, \\ & (\overline{X} \wedge \overline{Y}) \overline{\overline{Y}} \wedge \overline{\overline{Z} \overline{Y}}. \end{aligned}$$

Теперь применим закон дистрибутивности

$$\begin{aligned} & \overline{X} \overline{Y} \wedge \overline{Y} \overline{\overline{Z}} \wedge \overline{\overline{Z} \overline{Y}}, \\ & \overline{X}Y \wedge \overline{Y}Y \wedge \overline{Z} \overline{Y}. \end{aligned}$$

Это конъюнктивная нормальная форма.

Наряду с ней существует еще вторая нормальная форма — дизъюнктивная. Она представляет собой некоторую дизъюнкцию конъюнкций, в которой каждый конъюнктивный член является отрицаемым или неотрицаемым основным высказыванием.

§ 5. ВСЕГДА ИСТИННЫЕ И ВСЕГДА ЛОЖНЫЕ ВЫСКАЗЫВАНИЯ

Первой задачей логики является нахождение таких связей, которые всегда истинны, независимо от того, представляют ли основные высказывания истинные или ложные утверждения.

Поскольку каждому логическому условию мы можем привести в соответствие эквивалентное ему выражение в нормальной форме, то для ответа на поставленный вопрос нужно решить, когда выражение в нормальной форме — всегда истинное высказывание.

Всегда истинными высказываниями оказываются выражения, которые в конъюнктивной нормальной форме характеризуются тем, что в каждой дизъюнкции по меньшей мере одно основное высказывание встречается одновременно с его отрицанием.

П р и м е р:

$$\begin{aligned} X \wedge Y \rightarrow X, \\ \overline{X \wedge Y} \vee X, \text{ (по правилу 4)} \\ \overline{X} \overline{Y} X. \text{ (по правилу 3)} \end{aligned}$$

Последняя дизъюнкция содержит X и \overline{X} , следовательно, она истинна.

С помощью дизъюнктивной нормальной формы можно установить, является ли высказывание всегда ложным.

Это бывает в том и только в том случае, когда каждый дизъюнктивный член одновременно с основным высказыванием содержит и противоположное ему высказывание.

П р и м е р:

$$\overline{X}Y \wedge \overline{Y}Z \wedge X \wedge \overline{Z}. \quad (2.31)$$

Применим второй закон дистрибутивности

$$\begin{aligned} (\overline{X} \wedge \overline{Y} \wedge X \wedge \overline{Z}) \vee (\overline{X} \wedge Z \wedge X \wedge \overline{Z}) \vee \\ \vee (Y \wedge \overline{Y} \wedge X \wedge \overline{Z}) \vee (Y \wedge Z \wedge X \wedge \overline{Z}). \end{aligned}$$

Здесь каждый дизъюнктивный член содержит какое-нибудь основное высказывание вместе с его отрицанием. Таким образом, выражение (2.31) всегда ложно.

§ 6. ПРЕДИКАТЫ

Дальнейшим развитием исчисления высказываний является и исчисление предикатов. Оно содержит в себе все исчисления высказываний, т. е. простые высказывания, принимающие два значения I и L , плюс все операции исчисления высказываний. Но, кроме того, в исчислении предикатов рассматриваются высказывания, отнесенные к предметам. Здесь уже высказывания расчленяются на субъект и предикат.

Рассмотрим некоторое множество предметов M , и пусть a, b, c, d — определенные предметы из этого множества. Высказывания об этих предметах обозначим так:

$$P(a), Q(a, b), S(b, c, d).$$

Пусть, например, множество M — натуральный ряд чисел, а буквы a, b, c, d соответственно числа 2, 6, 7, 9. Тогда $P(b)$ может быть высказыванием «6 — простое число», $Q(a, b)$ — «2 меньше 6», $S(b, c, d)$ — «6, 7, 9 есть числа четные».

Подобные высказывания могут быть как истинными, так и ложными. Мы будем рассматривать их так же, как и в исчислении высказываний, лишь с той точки зрения, что они представляют собой истину либо ложь (*И* или *Л*). Однако теперь мы будем считать, что значения *И* и *Л* ставятся в соответствие определенным предметам или группам предметов.

В рассмотренных примерах высказывание $P(b)$ является ложью, поставленной в соответствие 6, а $Q(a, b)$ — истина, поставленная в соответствие паре 2 и 6.

Пусть M — произвольное непустое множество, x — произвольный элемент этого множества.

Тогда выражение $F(x)$ обозначает высказывание, которое становится определенным, когда x замещено конкретным элементом из M . $F(a), F(b)$... — уже вполне определенные высказывания.

Например, если множество M — натуральный ряд чисел, то высказывание $F(x)$ может означать: « x есть простое число». Это неопределенное высказывание. Оно становится определенным, если заменить x некоторым числом, «3 есть простое число», «6 есть простое число» и т. д.

Пусть $S(x, y)$ обозначает: « x меньше y ». Это высказывание становится определенным, если x и y заменить конкретными числами: «2 меньше 6», «9 больше 2» и т. д.

Поскольку каждое высказывание представляет собой *И* или *Л*, то $F(x)$ означает, что каждому элементу из M поставлен в соответствие один из двух символов *И* или *Л*. Другими словами, $F(x)$ является функцией, определенной на множестве M и принимающей только два значения — *И* и *Л*. Точно также неопределенные высказывания о двух и более предметах $S(x, y)$, $G(x, y, z)$ и т. д. являются функциями двух, трех и т. д. переменных. Переменные x, y, z ... пробегают множество M , а функция принимает значения только *И* и *Л*.

Такие неопределенные высказывания, функции одной или нескольких переменных, называются *логическими функциями, или предикатами*.

Используя ранее введенную терминологию при сопоставлении функций и отношений между элементами множества, соответствующие логические функции часто называют *одноместными предикатами, двухместными предикатами и т. д.*

Одноместный предикат может выразить свойство предмета, например, « x есть простое число», « y — выходной сигнал» и т. д. Предикаты нескольких переменных позволяют выразить различные отношения между предметами. Пусть, например, M — множество материалов. Тогда предикатами можно выразить их сравнительные характеристики: « x тверже y », или « x и y — металлы» и т. д.

Все вводимые понятия всегда относятся к некоторому произвольному множеству M , называемому *полем*. Элементы поля обозначаются малыми буквами латинского алфавита (иногда с индексами).

Неопределенные элементы поля обозначаются буквами конца алфавита

$$x, y, z, u, v, x_1, x_2, \dots$$

Их называют *свободными (предметными) переменными*. Буквами начала алфавита

$$a, b, c, d_1, d_2, \dots$$

обозначаются определенные элементы поля.

Их называют *индивидуальными предметами или предметными постоянными*.

Как и в исчислении высказываний, большими буквами

$$A, B, \dots, X, A_1, A_2, \dots$$

обозначаются переменные, принимающие значения I и L . Их называют *переменными высказываниями*.

Будем называть *элементарными формулами* высказывания, выражаемые большими латинскими буквами, как переменные, так и постоянные, а также выражения

$$P(a), S(a, b), \dots,$$

где P и S — предикаты, а a и b — индивидуальные предметы.

Этот термин употребляется для того, чтобы отличить такие формулы от сложных, составляемых из элементарных.

Так как элементарные формулы (и высказывания, и предикаты) всегда принимают только значения *И* или *Л*, их можно связывать операциями

$$\wedge, \vee, \rightarrow, -, \sim,$$

причем эти операции определяются так же, как и в исчислении высказываний.

Получаемые таким образом сложные формулы в свою очередь могут определять высказывания или предикаты. **П р и м е р,**

$$\begin{aligned} & A \vee F(x); \\ & A(x, y) \rightarrow (B \wedge \overline{A}(x_1, x_2)); \\ & G(x, y) \rightarrow G(x_1, x_2); \\ & L(x) \sim L(y) \\ & \text{и т. д.} \end{aligned}$$

Первая формула при фиксированных *A* и *F* (*x*) определяет некоторый предикат. Четвертая формула при всяком *L* является предикатом от двух переменных *x* и *y*. Этот предикат принимает значение *И* при *x* = *y*.

§ 7. ОПЕРАЦИИ НАВЕШИВАНИЯ КВАНТОРОВ

Кроме рассмотренных операций алгебры логики, в исчислении предикатов вводятся две специфические операции — навешивание кванторов.

Пусть *Q* (*x*) — определенный предикат, принимающий значение *И* или *Л* для каждого элемента *x* ∈ *M*. Тогда выражение

$$\forall x Q(x)$$

будем называть истинным, если *Q* (*x*) истинно для каждого элемента *x* ∈ *M*. Символ \forall называется *квантором всеобщности*. Выражение $\forall x$ — *квантор всеобщности по переменной x*. Переход от предиката *Q* (*x*) к предикату $\forall x Q(x)$ называется *навешиванием* на предикат *Q* (*x*) *квантора всеобщности по переменной x*.

П р и м е р. Пусть *M* — множество процессов. Предикат *Q* (*x*) для *x* ∈ *M* имеет интерпретацию: «процесс *x* протекает во времени». В результате навешивания квантора всеобщности имеем формулу $\forall x Q(x)$, которая интерпретируется так: «все процессы *x* протекают во времени», что, очевидно, истинно.

Пусть опять $Q(x)$ — предикат с единственной переменной x . Свяжем с ним формулу

$$\exists xQ(x),$$

принимаящую значение $И$, если существует элемент поля M , для которого $Q(x)$ истинно, и значение $Л$, если таких элементов не существует.

Символ \exists называется *квантором существования*, а выражение $\exists x$ — *квантором существования по переменной x* . Переход от $Q(x)$ к $\exists xQ(x)$ называется *навешиванием на предикат $Q(x)$ квантора существования по переменной x* .

П р и м е р. Пусть M — множество измерительных приборов. Предикат $Q(x)$ для $x \in M$ можно интерпретировать так: « x имеет цифровую индикацию». Тогда при навешивании квантора \exists получим формулу $\exists xQ(x)$, которая будет интерпретироваться так: «существует x , имеющий цифровую индикацию», что, очевидно, истинно.

Контрольные вопросы и задания

1. Назовите и объясните все виды логических связей. Приведите примеры.
2. Назовите основные виды эквивалентных преобразований для конъюнктивных и дизъюнктивных связей.
3. Приведите эквивалентные преобразования для представления импликации и равнозначности.
4. Сформулируйте правила приведения высказываний к нормальной форме.
5. Для чего применяется конъюнктивная нормальная форма?
6. В каком случае сложное высказывание оказывается всегда ложным?
7. Что такое логические функции?
8. Приведите примеры записей сложных логических предикатов.
9. Приведите примеры записи предикатов с использованием операций навешивания кванторов.

Глава 3

АЛГОРИТМЫ

Понятие алгоритма в современной математике является одним из основных. Особое место оно занимает среди фундаментальных понятий кибернетики.

Любая система управления собирает, передает и перерабатывает информацию. Переработка информации состоит в выполнении некоторых операций в определенной последовательности. В итоге выполнение последовательности операций приводит к получению результата или решения. Термин «алгоритм», или «алгорифм», как часто пишут и произносят математики, встречается довольно часто в специальной литературе. А. А. Марков определяет этот термин следующим образом: «В математике принято понимать под «алгорифмом» точное предписание, определяющее вычислительный процесс, ведущий от варьируемых исходных данных к искомому результату». С математической точки зрения это скорее описание, чем строгое определение. Понятие алгоритма относится к классу первоначальных математических понятий, таких, как «соответствие», «множество», «натуральное число». Первоначальные понятия нельзя свести к более простым. Поэтому понятие алгоритма считают обычно неопределимым. Оно абстрагируется из опыта и усваивается на конкретных примерах.

Обычно в качестве классического примера алгоритма в математике приводят пример алгоритма Эвклида для нахождения наибольшего общего делителя. Более простыми являются алгоритмы сложения чисел столбиком и другие алгоритмы арифметических действий. Умение складывать или делить числа по существу означает знание некоторых алгоритмов.

Каждый конкретный алгоритм характеризуется некоторой совокупностью возможных исходных данных — объектов, к которым имеет смысл применять этот алгоритм. Например, для алгоритма нахождения наибольшего общего

делителя такой совокупностью является набор всех пар положительных целых чисел.

Операции по переработке информации в системах управления можно разделить на два класса: количественные и логические.

Количественные операции производятся над числовыми характеристиками различных величин. Логические операции основаны на качественных особенностях соответствующих величин. При переработке информации количественные и логические операции чередуются в строго определенной последовательности.

В кибернетике алгоритмом называют совокупность правил или ограничений, которые определяют порядок чередования отдельных операций для получения некоторого результата. Другими словами, алгоритм состоит из последовательности совершенно определенных простых шагов и точных правил. Эти правила указывают, когда и какой шаг должен быть сделан и когда должен быть прекращен выполняемый процесс.

Если в результате применения алгоритма к какому-либо объекту из совокупности исходных данных получается решение, то говорят, что алгоритм применим к этому объекту. Далеко не к каждому объекту из рассматриваемой совокупности можно применить определенный алгоритм. Более того, применение алгоритма к какому-либо объекту из совокупности исходных данных не гарантирует получение результата, то есть в общем случае заранее неизвестно, применим ли алгоритм к данному объекту. Этот факт приводит к определению области применимости алгоритма в множестве всех возможных исходных данных. Основой современных прикладных аспектов теории алгоритмов являются работы по абстрактной теории алгоритмов А. А. Маркова, С. К. Клини, В. А. Успенского, В. М. Глушкова.

§ 1. ЧИСЛЕННЫЕ И ЛОГИЧЕСКИЕ АЛГОРИТМЫ

Рассмотрим типичные примеры и сформулируем на их основе общие свойства алгоритмов.

Алгоритм Эвклида. Это алгоритм для нахождения наибольшего общего делителя двух заданных положительных целых чисел a и b . Алгоритм запишем в виде последовательности указаний:

1. Обозревай данные числа a и b . Переходи к следующему указанию.

2. Сравни обозреваемые числа ($a = b$, $a < b$ или $a > b$). Переходи к следующему указанию.

3. Если обозреваемые числа равны, то каждое из них дает искомый результат. Остановка. Если нет, переходи к следующему указанию.

4. Если первое из обозреваемых чисел меньше второго, то переставь их местами. Переходи к следующему указанию.

5. Вычитай второе число из первого и обозревай два числа: вычитаемое и остаток. Переходи к указанию 2.

После выполнения пятого указания следует вновь возвращаться ко второму, и т. д. до тех пор, пока не окажется

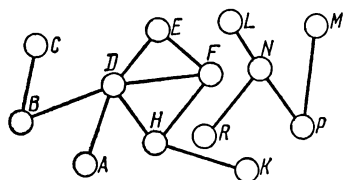


Рис. 3.1. Конечный лабиринт.

выполненным условие, содержащееся в третьем указании. При этом процесс прекращается. В приведенной записи алгоритм Эвклида представляет собой достаточно детализированное предписание, в котором элементарными операциями служат простейшие арифметические действия. Эти действия (вычитание, сравнение) можно было бы разложить на еще более элементарные операции, но этого не делают, так как арифметические правила просты и привычны.

Если решение какой-либо задачи сводится к арифметическим действиям, то соответствующие алгоритмы называются численными. Алгоритм Эвклида является численным. К численным алгоритмам относятся также любые формулы для решения некоторого класса задач, если эти формулы полностью определяют состав действий (умножение, сложение, деление) и порядок их выполнения.

Кроме численных существуют алгоритмы, в которых предписание о способе действия относится не к цифровым объектам. Это так называемые логические алгоритмы. Рассмотрим пример.

Алгоритм поиска пути в конечном лабиринте. Лабиринт состоит из конечного числа площадок, от которых расходятся коридоры. Каждый коридор соединяет две площадки, называемые смежными. Площадки, из которых выходит только один коридор, называются тупиками. Геометрически лабиринт имеет вид системы кружков A, B, C, D, \dots , изображающих площадки и соединенных отрезками прямых линий, изображающих коридоры (рис. 3.1.).

Площадка Y достижима из площадки X , если существует путь, ведущий через промежуточные коридоры и площадки от X и Y .

При этом X и Y могут быть либо смежными площадками, либо существует последовательность смежных площадок $X, X_1, X_2, \dots, X_n, Y$.

Если площадка Y вообще достижима с площадки X , то она достижима и по простому пути (без петель). На простом пути каждая площадка приходится только один раз. Например, одним из простых путей от площадки B к площадке K будет $BDFHK$. Площадка N недостижима из площадки B . Пусть требуется определить, достижима ли площадка K из площадки C . Если да, то нужно найти путь из C в K . Если же площадка K окажется недостижимой, то требуется после поиска вернуться в C . Вначале устройство лабиринта предполагается неизвестным. Неизвестно также, в каком месте лабиринта находятся площадки C и K . Решение задачи, поставленной в таком виде, должно быть общим методом поиска, пригодного в любом конечном лабиринте. Один из таких методов заключается в том, что «разведчик» имеет нить, конец которой закреплен на исходной площадке C . Двигаясь по лабиринту, «разведчик» может разматывать клубок или, наоборот, наматывать на него нить. Кроме того, пройденные коридоры можно отмечать. При этом различаются коридоры трех видов: зеленые — ни разу не пройденные, желтые — пройденные один раз, красные — пройденные дважды.

Находясь на какой-либо из площадок, на смежную площадку можно попасть одним из двух видов.

1. Разматывание нити. От данной площадки совершается проход к смежной по любому зеленому коридору. По этому коридору теперь протянута нить, и он считается желтым.

2. Наматывание нити. Возвращение по последнему пройденному желтому коридору до смежной площадки. Нить наматывается, и коридор объявляется красным.

Строго говоря, необходимо различать только зеленые и красные коридоры, так как по желтым протянута нить.

Остановка на каждой площадке характеризуется одним из пяти признаков:

1. Площадка K — цель достигнута.

2. Петля. От данной площадки расходится по крайней мере два желтых коридора, т. е. через площадку уже протянута нить.

3. Зеленая улица. От данной площадки отходит по крайней мере один зеленый коридор.

4. Исходная площадка C .

5. Пятый случай. Все указанные ранее признаки отсутствуют. Метод поиска задается следующей схемой:

Признак	Ход
1. Площадка K	Остановка
2. Петля	Наматывание нити
3. Зеленая улица	Разматывание нити
4. Площадка C	Остановка
5. Пятый случай	Наматывание нити

Очередной ход с любой площадки осуществляется так:

1. Проверяем по порядку номеров, какой из признаков имеет место (в соответствии с левым столбцом).

2. При обнаружении признака следует, не проверяя последующих признаков, сделать соответствующий ход, указанный в правой колонке. Ходы повторяются, пока не наступит остановка.

Для изложенного метода справедливы следующие три утверждения:

1. При любом взаимном расположении площадок C и K после конечного числа ходов наступит остановка либо на площадке K , либо на площадке C .

2. Остановка на площадке K соответствует достижению цели. При этом нить протянута по простому пути, ведущему от C к K .

3. Остановка на исходной площадке означает, что площадка K недостижима.

Рассмотрим действие метода на примере лабиринта (рис. 3.1). Процесс поиска представлен в табл. 3.1.

Площадка K оказалась достижимой. Выделяя в предпоследнем столбце таблицы коридоры, оставшиеся желтыми, получаем простой путь (без петель) от C к K .

Описанный метод поиска в конечном лабиринте характеризуется на некоторых этапах произвольным выбором. Если от некоторой площадки отходят несколько зеленых коридоров, «разведчик» может выбирать любой из них. Поэтому различные разведчики могут прийти от площадки C к площадке K разными путями.

В рассмотренном ранее алгоритме Эвклида никакого произвола не было. Операции по нахождению наибольшего общего делителя, выполненные разными вычислителями в соответствии с указаниями 1—5 этого алгоритма, совпадут

во всех деталях. Невозможна такая ситуация, чтобы вычислитель выбирал что-либо по своему усмотрению.

Следует заметить, что традиционно алгоритмом называют лишь строго детерминированную последовательность правил, где все возможные случаи предусмотрены и однозначно

Таблица 3.1

Признак	Ход	Пройденный коридор	Цвет коридора после прохождения
зеленая улица	разматывание	<i>CB</i>	желтый
» »	»	<i>BD</i>	желтый
» »	»	<i>DE</i>	желтый
» »	»	<i>EF</i>	желтый
» »	»	<i>FD</i>	желтый
петля	наматывание	<i>DF</i>	красный
зеленая улица	разматывание	<i>FH</i>	желтый
» »	»	<i>HD</i>	желтый
петля	наматывание	<i>DH</i>	красный
зеленая улица	разматывание	<i>HK</i>	желтый
площадка <i>K</i>	остановка		

определены соответствующие ходы. Поэтому, для того чтобы описанный метод поиска можно было назвать алгоритмом, необходимо, кроме указаний 1—5, четко определить, какой коридор выбирать из нескольких зеленых (например, первый по часовой стрелке).

§ 2. ЭМПИРИЧЕСКИЕ СВОЙСТВА АЛГОРИТМОВ

Рассмотренные примеры дают возможность сформулировать некоторые общие свойства, присущие любому алгоритму.

Свойство детерминированности. Метод вычислений (логических действий) должен быть точен и общепонятен. Никакой произвольный выбор недопустим. Сущность метода можно сообщить любому лицу в виде конечного числа указаний. Действия, предпринимаемые в соответствии с этими указаниями, представляют собой детерминированный процесс и не зависят от произвола действующего лица. Этот процесс может быть в любое время повторен другим лицом.

Свойство массовости. Алгоритм дает возможность решать целый класс задач. Указания, составляющие сущность алгоритма, применимы к начальным данным, которые могут варьироваться.

Алгоритм Эвклида, например, применим к любой паре целых чисел $a > 0$ и $b > 0$; формулы для решения системы уравнений дают решение при любых коэффициентах системы; метод поиска применим к любому как угодно сложному конечному лабиринту и т. д.

Задачи некоторого класса считаются решенными, если для этого класса найден алгоритм решения.

Если же для решения всех задач данного класса нет алгоритма, то приходится в частных случаях искать методы решения, которые хотя и дают решение в конкретном случае, но оказываются неприменимыми для других случаев. Например, нет алгоритма, позволяющего для любых $n = 1, 2, 3, \dots$ определить, имеет ли уравнение

$$x^n + y^n = r^n \quad (3.1)$$

целочисленное решение. Для конкретных значений n эта задача может быть решена. При $n = 2$ можно подобрать тройку чисел ($x = 3, y = 4, z = 5$), удовлетворяющую приведенному уравнению. Для $n = 3$ доказано, что уравнение (3.1) не имеет целочисленных решений. Но это доказательство оказывается непригодным для других n . Последовательность операций для решения конкретной задачи не называют алгоритмом.

Свойство результативности. Алгоритм, примененный к любой задаче определенного класса, через конечное число шагов должен привести к остановке операций. После остановки мы получаем результат. Это свойство называют иногда *направленностью алгоритма*.

Например, если применить алгоритм Эвклида к любым двум числам $a \geq 1, b \geq 1$, то рано или поздно наступит остановка и будет найден наибольший общий делитель.

Алгоритм поиска в любом как угодно сложном конечном лабиринте обязательно приведет к остановке. По тому, на какой площадке наступила остановка, можно сделать вывод, достижима ли искомая площадка из исходной.

Формально, конечно, можно применить алгоритм Эвклида к любым целым числам $a \geq 0, b \geq 0$, а также к отрицательным. Но при этом может случиться, что остановка алгоритмической процедуры никогда не наступит. Например, при $a = 0, b = 6$ (наибольший общий делитель равен 6), применяя указания 1—5, получаем следующие пары: 0,6; 0,6; 0,6; 6,0; 0,6; 0,6, ... и так до бесконечности.

Из свойства результативности алгоритма вытекает понятие *области применимости алгоритма*. Это наибольшая об-

ласть (множество) начальных данных, для которой алгоритм результативен. Если условия взяты из области применимости, то алгоритм перерабатывает условия в решение задачи, и наступает остановка с выдачей результатов. Если же начальные условия не относятся к области применимости, то либо алгоритмическая процедура длится бесконечно, либо остановка наступает, но результат мы не получаем.

Для рассмотренных примеров областями применимости являются: множество целых положительных чисел 1, 2, 3, ... (для алгоритма Эвклида), множество всех конечных лабиринтов.

Количество операций, необходимое для той или иной алгоритмической процедуры, заранее неизвестно. Оно определяется выбором исходных данных. Осуществимость алгоритма в общем смысле следует понимать как потенциально возможный процесс. В самом деле, для некоторых конкретных задач из области применимости алгоритма на практике может не хватить времени вычислителю или памяти при реализации алгоритма в вычислительной машине.

Указанные три общих свойства алгоритмов — свойства детерминированности, массовости и результативности (направленности) — являются эмпирическими свойствами. Они сформулированы на основании опыта, анализа всех существующих на сегодняшний день алгоритмов.

Естественно, что эти свойства нельзя считать строгой математической формулировкой понятия «алгоритм».

§ 3. ЭЛЕМЕНТЫ ТЕОРИИ АЛГОРИТМОВ.

АЛФАВИТНЫЕ ОПЕРАТОРЫ И АЛГОРИТМЫ

Абстрактным алфавитом принято называть конечную совокупность объектов, называемых буквами данного алфавита. Объекты эти могут быть любой природы. В качестве букв абстрактных алфавитов можно рассматривать отдельные значения множества значений параметра технологических процессов, определенные состояния объектов управления, конкретные производственные ситуации, буквы алфавита какого-либо языка, цифры, рисунки и т. д. Можно рассматривать абстрактный алфавит, буквами которого являются целые слова того или иного языка. Основное ограничение, накладываемое при определении, заключается в том, чтобы алфавит был конечным, т. е. состоял из конечного числа букв.

Словом в абстрактном алфавите будем называть любую конечную упорядоченную последовательность букв. Например, в алфавите $A = A(x, y)$, состоящем из двух букв x и y , любые из последовательностей $x, xy, yux, xxxuy, \dots$ являются словами. Длина слова определяется числом букв. Приведенные выше слова, например, имеют длины соответственно 1, 2, 3, 5, ...

Кроме слов положительной длины (содержащих не менее одной буквы), для общности целесообразно рассматривать так называемое пустое слово, не содержащее ни одной буквы. Для обозначения пустого слова будем употреблять букву O . Иногда пустое слово никак не обозначается — на соответствующем ему месте не выписывается ни одна буква.

Алфавит расширяется при включении в его состав новых букв. При этом понятие слова может существенно измениться. Например, выражение $32 + 45$ представляет собой два слова (32 и 45) в алфавите A из 10 цифр (0, 1, 2, 3, 4, 5, 6, 7, 8, 9), соединенных знаком $+$. Но это же выражение можно рассматривать как одно слово в расширении алфавита A , которое получится с присоединением к нему новой буквы « $+$ ».

Алфавитным оператором, или отображением, называется всякое соответствие, сопоставляющее слова в том или ином алфавите словам в том же самом или некотором другом алфавите. Первый алфавит называется при этом входным, а второй — выходным алфавитом данного оператора. Если входной и выходной алфавиты совпадают, говорят, что алфавитный оператор задан в соответствующем алфавите.

Мы будем рассматривать однозначные алфавитные операторы, т. е. такие, которые сопоставляют каждое слово во входном алфавите (входное слово) не более чем с одним словом в выходном алфавите (выходным словом). Если алфавитный оператор не сопоставляет данное входное слово P с никаким выходным словом (в том числе и пустым), то говорят, что он не определен на этом слове.

Областью определения алфавитного оператора называется совокупность всех слов, на которых он определен. Таким образом, под алфавитным оператором впредь будем подразумевать однозначное, частично определенное отображение множества слов во входном алфавите оператора в множество слов в его выходном алфавите. Алфавитные операторы можно задавать не на всех словах. Поэтому можно всегда считать, что входной и выходной алфавиты оператора совпадают. Для этого достаточно объединить входной и выходной алфавиты данного оператора Φ в общий алфавит A .

Простейшими из алфавитных операторов являются операторы, осуществляющие *побуквенные отображения*. Каждая буква x входного слова P заменяется соответствующей буквой y выходного алфавита. Такое отображение не зависит от входного слова P и полностью определяется заданием соответствия между буквами входного и выходного алфавитов.

Важным видом отображений являются так называемые *кодирующие изображения*. При простейшем кодировании слова в алфавите A кодируются словами в другом алфавите B следующим образом. Каждая буква a_i алфавита A сопоставляется с некоторой конечной последовательностью $b_{i_1}, b_{i_2}, \dots, b_{i_k}$ букв в алфавите B . Эту последовательность назовем *кодом соответствующей буквы*. Различным буквам алфавита A должны соответствовать различные коды.

Кодирующее изображение слова P осуществляется при замене всех его букв соответствующими кодами. При этом получим некоторое слово в алфавите B , называемое кодом исходного слова P .

Одно из основных условий состоит в том, что кодирующее изображение должно быть обязательно обратимым. Другими словами, должно выполняться условие взаимной однозначности кодирующего отображения.

Очевидно, что выполнение требования о том, чтобы различным буквам соответствовали различные коды, еще не обеспечивает обратимости кодирования. Пусть, например, букве a_1 соответствует код b , а букве a_2 — код bb . Тогда код bbb может в равной степени соответствовать как слову a_2a_1 , так и словам $a_1a_1a_1$ и a_2a_1 .

Для того чтобы кодирование было обратимым, необходимо выполнить следующие два условия:

- 1) коды различных букв входного алфавита A должны быть различными;
- 2) код любой буквы алфавита A не должен совпадать ни с каким из начальных отрезков кодов других букв этого алфавита.

Начальным отрезком слова $q = pr$ называется слово p , причем r — любое слово (в том числе и пустое).

Предположим, что слово $q = b_{i_1}, b_{i_2}, \dots, b_{i_n}$ является кодом некоторого слова $p = a_{j_1}, a_{j_2}, \dots, a_{j_m}$ в алфавите A . Если условия 1) и 2) выполнены, то можно показать, что по коду q слово p восстанавливается однозначно. В самом деле, из 2) следует, что только один начальный отрезок слова q

может совпадать с кодом какой-либо буквы алфавита A . Ясно, что таким отрезком является код буквы a_{j_1} . Отбрасывая этот отрезок, получаем код q_1 слова $p = a_{j_2}, a_{j_3}, \dots, a_{j_m}$. Используя те же рассуждения, однозначно восстанавливаем следующую букву — a_{j_2} слова p и т. д. Точно так же однозначно восстанавливаются одна за другой все буквы слова p .

Если коды всех букв исходного алфавита A имеют одинаковую длину, кодирование называется нормальным. Используя кодирование, можно сводить изучение произвольных алфавитных отображений к алфавитным отображениям в выбранном определенным образом стандартном алфавите.

Чаще всего таким стандартным алфавитом бывает двоичный алфавит, состоящий из двух букв, например 0 и 1.

Пусть A произвольный, а B — стандартный алфавиты, состоящие более чем из одной буквы. Если в алфавите A число букв равно n , а в алфавите B — m , то всегда можно выбрать число k так, чтобы удовлетворялось неравенство

$$m^k \geq n. \quad (3.2)$$

Число различных слов длины k в m -буквенном алфавите равно m^k . Поэтому из (3.2) следует, что все буквы в алфавите A можно закодировать словами длины k в алфавите B так, чтобы коды различных букв были различными.

Понятие алфавитного оператора достаточно общее. Он объединяет фактически любые процессы преобразования информации.

§ 4. СЛОВА В АССОЦИАТИВНОМ ИСЧИСЛЕНИИ

Описанный ранее пример поиска был рассмотрен для произвольного конечного лабиринта. Проблему слов в абстрактном алфавите в определенном смысле можно рассматривать как обобщение задачи поиска в бесконечном лабиринте.

Рассмотрим два слова L и M в некотором абстрактном алфавите A . Если L является частью M , то говорят, что слово L входит в слово M , иначе, есть вхождение слова L в M . Например слова ab и $cdaba$.

Вхождение в общем случае может быть многократным

$$dad \ bdbdadadb.$$

Преобразование слов в абстрактном алфавите описывается некоторой процедурой, которая дает возможность из заданного слова получать новые слова.

Пусть в некотором алфавите задана конечная система

допустимых подстановок

$$L - M; S - R; \dots U - V,$$

где $L, M, R, S, \dots U, V$ — слова в том же алфавите.

Любую из заданных подстановок можно применить к некоторому слову P этого алфавита. Если, например, в слово P один или несколько раз входит слово S , то любое из этих вхождений можно заменить словом R и наоборот, если есть вхождение слова R , то его можно заменить словом S . Для слова $abcbcbab$ подстановку $ab - bcb$ можно осуществить четырьмя способами. Замена каждого из двух вхождений bcb дает слова $aabcbab$, $abscabab$, а замена каждого из двух вхождений ab дает слова $bcbcbcbab$, $abcbcbcbcb$.

К слову $bacb$ подстановка $ab - bcb$ неприменима, так как в него не входят ни ab , ни bcb . Если к полученным новым словам применять другие заданные подстановки, мы будем получать другие новые слова и т. д.

Ассоциативным исчислением называется совокупность всех слов в данном алфавите вместе с системой допустимых подстановок. Задать ассоциативное исчисление — значит задать алфавит и систему подстановок.

Смежными называются два слова P_1 и P_2 в некотором ассоциативном исчислении, если одно из них может быть преобразовано в другое однократным применением некоторой допустимой подстановки.

Дедуктивной цепочкой, ведущей от слова P к слову Q , называется последовательность слов P, P_1, P_2, \dots, Q , если каждые из двух рядом стоящих слов этой цепочки являются смежными.

Эквивалентными называются два слова P и Q , если существует дедуктивная цепочка, ведущая от слова P к слову Q . Для обозначения отношения эквивалентности применяется тот же знак, что и в исчислении высказываний: $P \sim Q$. Поскольку допустимые подстановки применимы в обе стороны, справедливо, что $(P \sim Q) (Q \sim P)$.

Пример. Задано ассоциативное исчисление:

$$\{a, b, c, d, e\} \text{ — алфавит,}$$

$ac - ca$	}	допустимые подстановки
$ad - da$		
$bc - cb$		
$bd - db$		
$abac - dbacc$		
$eca - ae$		
$edb - be$		

Слова $abcde$ и $acbde$ смежные в этом исчислении, так как слово $abcde$ преобразуется в $acbde$ одной подстановкой $bc \rightarrow cb$. Слово $aaabbb$ не имеет смежных слов — к нему неприменима ни одна подстановка.

В силу наличия дедуктивной цепочки $abcde, acbde, cabde, cadbe, cadedb$ слова $abcde$ и $cadedb$ эквивалентны. При построении дедуктивной цепочки последовательно применены 3-я, 1-я, 4-я и 5-я подстановки.

Ассоциативному исчислению можно поставить в соответствие некоторый бесконечный лабиринт. Каждому слову некоторого алфавита ставится в соответствие определенная площадка лабиринта. Лабиринт будет бесконечен, так как из букв данного абстрактного алфавита можно составить бесконечное множество слов. Любые две площадки этого лабиринта, соответствующие смежным словам, соединяются коридором.

Эквивалентность слов P и Q означает, что в построенном таким образом лабиринте площадка, соответствующая слову Q , достижима с площадки, соответствующей слову P .

В некоторых случаях целесообразно рассматривать специальный вид ассоциативного исчисления, которое задается алфавитом и системой ориентированных подстановок вида $R \rightarrow S$. Стрелка означает, что подстановка возможна в данном случае лишь слева направо, т. е. можно заменять R на S , но не наоборот. Подобное ассоциативное исчисление соответствует бесконечному лабиринту, в котором каждый коридор можно проходить только один раз, в одном направлении.

Очевидно, что в таком ассоциативном исчислении из эквивалентности $P \sim Q$ не следует, что $Q \sim P$.

Проблема слов для любого ассоциативного исчисления определяется следующим образом. для любых двух слов в данном исчислении требуется узнать, эквивалентны они или нет. В такой постановке проблема эквивалентна проблеме достижимости в случае лабиринта. Однако теперь мы имеем бесконечный лабиринт, поэтому описанный ранее метод не применим. Обследовать бесконечный лабиринт в конечное время невозможно.

Проблема эквивалентности в любом ассоциативном исчислении — это бесконечное множество однотипных задач. Решение ее должно представлять собой алгоритм для установления эквивалентности или неэквивалентности любой пары слов.

Таким образом, логическую задачу поиска пути в лабиринте можно сформулировать в терминах ассоциативного исчисления. На языке ассоциативного исчисления можно трактовать и другие логические процессы. Например, любую логическую формулу математической логики можно рассматривать, как запись слова в некотором алфавите, буквами которого являются логические символы \neg , \vee , \wedge , \rightarrow и т. д., логические переменные (высказывания) и логические функции (предикаты).

Эквивалентности, используемые в математической логике для преобразования различных высказаний, рассматриваются как заданные подстановки, например, $\overline{X \vee Y}$ можно заменить на $\overline{X} \wedge \overline{Y}$ и наоборот.

§ 5. ЭКВИВАЛЕНТНЫЕ АЛГОРИТМЫ. НОРМАЛЬНЫЙ АЛГОРИТМ МАРКОВА

Два алгоритма α_1 и α_2 в некотором алфавите называются эквивалентными, если области их применимости совпадают и результаты переработки ими любого слова из их общей области применимости также совпадают. Если алгоритм α_1 применим к некоторому слову P , то α_2 также должен быть применим к этому слову и наоборот. Причем, оба алгоритма должны перерабатывать слово P в одно и то же слово Q .

Дальнейшее уточнение понятия алгоритм в смысле его точного математического определения было сделано А. А. Марковым. Им построен нормальный алгоритм в терминах системы подстановок. Но вместо расплывчатого понятия «общепонятное указание» использования подстановок А. А. Марков сформулировал точные правила о порядке использования подстановок. Нормальный алгоритм Маркова определяется так.

Задается алфавит A и определяется система подстановок. Отправляясь от произвольного слова P в алфавите A , необходимо просмотреть формулы подстановок в том порядке, в каком они заданы в схеме. При этом разыскивается формула с левой частью, входящей в P . Если такой формулы нет, алгоритмическая процедура останавливается. Если формула есть, берем первую по порядку проверки и делаем подстановку ее правой части вместо первого вхождения ее левой части в P . Получаем новое слово P_1 в алфавите A . На этом заканчивается первый шаг. Второй шаг точно такой же, но теперь исходным словом является

P_1 вместо P . Аналогичные шаги делаются до тех пор, пока не наступает остановка. А это может произойти в двух случаях:

1) при получении такого слова P_n , в которое не входит ни одна из левых частей формул подстановок;

2) когда при получении слова P_n применяется последняя подстановка.

И в том, и в другом случае говорят, что данный алгоритм перерабатывает слово P в слово P_n .

Различные нормальные алгоритмы отличаются только алфавитами и системами допустимых подстановок. Задать нормальный алгоритм — значит задать алфавит и систему подстановок.

Пример. Зададим алфавит и систему подстановок в следующем виде:

$$A = \{1, +\} \quad \begin{array}{l} 1 + \rightarrow + 1 \\ + 1 \rightarrow 1 \\ 1 \rightarrow 1. \end{array}$$

Стрелками принято обозначать систему подстановок в нормальном алгоритме Маркова (в отличие от произвольного ассоциативного исчисления).

Исходное слово имеет вид $1111 + 11 + 111$. Применяя нормальный алгоритм Маркова последовательно, получаем слова

$$\begin{array}{l} 1111 + 11 + 111 \\ 111 + 111 + 111 \\ 11 + 1111 + 111 \\ 1 + 11111 + 111 \\ + 111111 + 111 \\ + 1111 + 11111 \\ + 111 + 111111 \\ + 11 + 1111111 \\ + 1 + 11111111 \\ + + 11111111 \\ + 11111111 \\ 11111111 \\ 11111111 \end{array}$$

Алгоритмический процесс останавливается с применением последней подстановки $1 \rightarrow 1$, которая перерабатывает слово 111111111 в себя.

Уточнение понятия алгоритма, сделанное А. А. Марковым, заключается в предположении, что всякий алгоритм в алфавите A эквивалентен некоторому нормальному алгоритму в том же алфавите.

Строго доказать это утверждение не представляется возможным, так как в нем сопоставляются расплывчатое понятие «всякий алгоритм» и «нормальный алгоритм». Однако до настоящего времени не был построен такой алгоритм, для которого нельзя было бы построить эквивалентный ему нормальный алгоритм в том же алфавите. Так что предположение Маркова можно считать законом, который не доказан, но проверен и подтвержден всем предыдущим опытом.

§ 6. АЛГОРИТМИЧЕСКИ НЕРАЗРЕШИМЫЕ ПРОБЛЕМЫ

Нормальный алгоритм Маркова можно рассматривать как удобную «стандартную форму» для задания любого алгоритма. Другими словами, можно предположить, что вообще любой алгоритм может быть задан в форме нормального алгоритма Маркова.

Это утверждение является, конечно, гипотетическим. Но если принять подобную гипотезу, можно определить, как строго доказать алгоритмическую неразрешимость определенного круга проблем.

Например, доказательство алгоритмической неразрешимости проблемы слов сводится к доказательству того, что в соответствующем ассоциативном исчислении не существует нормального алгоритма, распознающего эквивалентность данных слов P и Q .

Подобные примеры были впервые построены А. А. Марковым в 1946 г. После этого стало ясно, что не существует алгоритма для распознавания эквивалентности слов в любом ассоциативном исчислении.

Таким образом, при решении какой-либо задачи следует иметь в виду, что алгоритм для ее решения может и не существовать. Поэтому одновременно с поиском нужного алгоритма следует работать и над доказательством того, что он не существует. Короче говоря, решение задачи и доказательство невозможности алгоритма — результаты эквивалентные.

То, что алгоритм для решения какого-либо класса задач не существует, вовсе не означает неразрешимости вообще. Рассматриваемый класс задач может быть настолько широк, что нет единого эффективного метода для их решения. Однако для конкретных задач не исключается возможность нахождения частных приемов решения.

Для уточнения понятия «алгоритм» в математике были распространены две точки зрения.

1. *Все проблемы являются алгоритмически разрешимыми.* Просто для решения некоторых из них не найден алгоритм. Не хватает средств в современной математике для его построения.

Сторонники этой точки зрения считали, что для решения проблем, называемых алгоритмически неразрешимыми, просто не хватает средств современной математики, построение искоемых алгоритмов — дело будущего.

2. *Есть классы задач, для решения которых вообще не существует алгоритма,* т. е. некоторые проблемы нельзя решать механически, с помощью формальных рассуждений и вычислений. Эти проблемы требуют творческого мышления.

Это очень сильное утверждение. Ведь оно распространяется на все будущие времена и средства.

Пока в определении алгоритма фигурировало расплывчатое понятие «общепонятное точное предписание», о доказательстве правоты второй точки зрения не могло быть и речи.

Благодаря наличию гипотез о существовании «стандартных форм» (например нормальный алгоритм), в которых могут быть выражены любые алгоритмы, стало возможным сформулировать понятие «алгоритм» и «алгоритмически неразрешимая проблема» в точных терминах.

§ 7. СВЕДЕНИЕ ЛЮБОГО АЛГОРИТМА К ЧИСЛЕННОМУ. МЕТОД ГЕДЕЛЯ

Как видно из рассмотренных примеров, определение для численных и логических алгоритмов одно и то же. И в том, и в другом случае алгоритмом названа система правил для решения некоторого класса задач. И в том, и в другом случае указанная система правил характеризуется свойствами детерминированности, массовости и результативности (направленности).

Более строгим становится понятие алгоритма при введении в рассмотрение нормального алгоритма Маркова. Оказывается, любой логический алгоритм можно достаточно простыми методами свести к численному. Поэтому теорию численных алгоритмов можно считать универсальным аппаратом для исследования всех алгоритмических проблем.

Покажем, как любую алгоритмическую проблему можно свести к вычислению значений некоторой целочисленной функции при целочисленных значениях аргументов.

Обозначим все условия задачи, перерабатываемые данным алгоритмом α , в виде последовательности с целыми неотрицательными индексами-номерами

$$A_0, A_1, A_2, \dots, A_n, \dots$$

Решения можно представить занумерованной последовательностью

$$B_0, B_1, B_2, \dots, B_m, \dots$$

После введения нумерации будем оперировать не с самими записями условий и решений, а с их номерами. Теперь можно представить алгоритм, который перерабатывает номер записи условий в номер записи решения. Этот алгоритм осуществляет вычисление значений числовой функции

$$m = \varphi(n),$$

т. е. он является численным алгоритмом.

Если существует алгоритм, решающий исходную задачу, то существует алгоритм, вычисляющий значения соответствующей функции. В самом деле, для нахождения значения $\varphi(n)$ при $n = n^*$ можно выбрать запись условия для n^* , далее с помощью имеющегося алгоритма найти запись решения и по ней определить соответствующий номер m^* . Таким образом,

$$\varphi(n^*) = m^*.$$

Справедливо и обратное утверждение: если существует алгоритм вычисления функции $\varphi(n)$, то существует и алгоритм решения исходной задачи.

Рассмотрим широко применяющийся для нумерации метод Гёделя.

Представим некоторое число n в виде

$$n = 2^{a_1} \cdot 3^{a_2} \cdot 5^{a_3} \cdot 7^{a_4} \cdot \dots \cdot p_{m-1}^{a_m},$$

где $p_0 = 2$; $p_1 = 3$; $p_2 = 5$ и т. д., то есть p_m — простое число.

Учитывая, что любое число можно разложить на простые множители единственным образом, можно утверждать, что каждому числу однозначно соответствует набор a_1, a_2, \dots, a_m и, наоборот, каждому набору a_1, a_2, \dots, a_m однозначно соответствует число n .

Например, если $n = 60$, то

$$60 = 2^2 3^1 5^2, \text{ т. е. } a_1 = 2, a_2 = 1, a_3 = 1.$$

Этим способом можно нумеровать любые упорядоченные последовательности m чисел.

П р и м е р ы: 1. Каждой паре чисел a_1 и a_2 , для которой пишется наибольший общий делитель q , можно поставить в соответствие гёделевский номер этой пары

$$n = 2^{a_1} 3^{a_2}.$$

Тогда алгоритм Эвклида сводится к вычислению функции $q = \varphi(n)$.

2. Припишем номер n уравнению n -й степени, записанному в общем виде

$$x^n + b_1 x^{n-1} + b_2 x^{n-2} + \dots + b_n = 0.$$

По n можно легко восстановить запись уравнения. При $n = 2$ имеем

$$x^2 + b_1 x + b_2 = 0.$$

Решение выражается через коэффициенты при помощи формулы

$$x = -\frac{b_1}{2} \pm \sqrt{\left(\frac{b_1}{2}\right)^2 - b_2}. \quad (3.3)$$

Запишем (3.3) в строку

$$x = -b_1 : 2 + - \sqrt{(b_1 \times b_1 : 4 - b_2)}.$$

В приведенной записи считается, что радикал действует на скобки, записанные непосредственно после него.

Пусть нужно найти выражение для решения уравнения n -й степени в радикалах. При любом виде этого решения оно может состоять только из символов

$$+, -, \times, :, (,), 1, b_1, b_2, \dots, b_n, \sqrt{}, \sqrt[3]{}, \dots, \sqrt[n]{}.$$

Обозначим эти символы некоторыми числами:

$\sqrt[3]{}$	соответствует	число 2)	соответствует	число 1,3
+	—	»	—	3 1	— » — 17
—	—	»	—	5 b_1	— » — 19
\times	—	»	—	7 b_2	— » — 23
:	—	»	—	9	
(—	»	—	11	
				b_n	— » — $(2_n + 15)$.

Любому выражению, составленному из перечисленных символов, соответствует определенный набор чисел. Например, для выражения

$$\sqrt[3]{(b_1 + b_2)}$$

получаем набор чисел

$$6, 11, 17, 3, 19, 13.$$

А набору чисел может быть поставлен в соответствие его гёделевский номер:

$$2^6 \cdot 3^{11} \cdot 5^{17} \cdot 7^3 \cdot 11^{19} \cdot 13^{13}.$$

С другой стороны, по заданному гёделевскому номеру можно восстановить набор чисел, затем по каждому из них определить соответствующий символ и восстановить запись любой формулы. Таким образом можно нумеровать любые выражения, составленные как из цифр, так и из иных символов — букв, значков различных операций и т.п.

3. С помощью метода Гёделя можно перенумеровать все слова в некотором алфавите A . Каждой букве ставится в соответствие некоторое число. Тогда любому слову B алфавита A соответствует последовательность чисел, от которой легко перейти к гёделевскому номеру, зависящему от выбранной системы соответствий букв и чисел. Далее можно перенумеровать все последовательности слов (например, все дедуктивные цепочки).

Другими словами, не только арифметические алгоритмы сводятся к вычислению значений целочисленных функций. *Любой нормальный алгоритм Маркова с применением метода Гёделя также можно свести к вычислению значений целочисленных функций.* Так что алгоритм вычисления значений целочисленных функций можно считать универсальной формой алгоритма.

Следует заметить, что при решении алгоритмических задач множество исходных данных предполагается счетным, хотя может быть и бесконечным.

Контрольные вопросы и задания

1. Как определяется понятие алгоритма в кибернетике?
2. Приведите примеры численных и логических алгоритмов.
3. Сформулируйте эмпирические свойства алгоритмов. Приведите примеры.
4. Что такое абстрактный алфавит?
5. Что такое слово и отображение в абстрактном алфавите?
6. Что значит задать ассоциативное исчисление?
7. Какие алгоритмы называются эквивалентными?
8. Покажите на контрольном примере сведение логического алгоритма к численному.

Глава 4

ВЕРОЯТНОСТИ. СЛУЧАЙНЫЕ ВЕЛИЧИНЫ. СЛУЧАЙНЫЕ ФУНКЦИИ

В подавляющем большинстве случаев будущие события и характеристики случайных процессов можно предсказать на основе накопленного опыта лишь приближенно. Точное предсказание нельзя сделать по многим причинам. Прежде всего нам могут быть известны не все причинные факторы, участвующие в анализируемом явлении или процессе. Кроме того может быть неопределенность в основе самого физического явления, сложность и многочисленность причин не позволяет вычислить их суммарный эффект, не все условия задачи четко определены и т. д.

Однако каковы бы ни были причины случайности при большом «опыте», т. е. при многократных наблюдениях исследуемого явления или процесса, наблюдается так называемая статистическая устойчивость. Здесь вступает в силу закон больших чисел, представляющий собой, по определению академика А. Н. Колмогорова, общий принцип, в силу которого совокупное действие большого числа случайных факторов приводит при некоторых весьма общих условиях к результату, почти не зависящему от случая.

Изучением случайных событий, величин и процессов занимается теория вероятностей и статистический анализ.

§ 1. ЭЛЕМЕНТЫ ТЕОРИИ ВЕРОЯТНОСТЕЙ. ОСНОВНЫЕ ПОНЯТИЯ И ОПРЕДЕЛЕНИЯ

Теория вероятностей, как и любая наука, опирается на ряд основных понятий. Одним из них является событие. Понятие «события» следует относить ко всему, что происходит или не происходит в результате опыта. При этом под «опытом» понимается не специально организованное исследование. Опытом или испытанием в теории вероятностей считается любая совокупность условий и воздействий,

при которых происходят интересующие нас события. С теоретико-вероятностной точки зрения опытом является любое наблюдение окружающего нас мира, а событием — любой результат, получаемый при наблюдениях. Примерами событий могут служить:

А — получение кандидатом определенного количества голосов на выборах;

Б — выход из строя прибора при включении его под определенное напряжение;

С — завоевание спортивной командой звания чемпиона;

Д — отсутствие осадков в некотором районе в течение определенного отрезка времени;

Е — получение продукции заданного качества.

Каждое из таких событий обладает той или иной степенью возможности. Чтобы количественно сравнить между собой события по степени их возможности, нужно с каждым событием связать определенное число, которое должно быть тем больше, чем более возможно событие. Это число называют вероятностью события.

Заметим, что уже при самом введении понятия вероятности с ним связывается определенный практический смысл. Более вероятными событиями считаются те, которые чаще происходят при испытании. И наоборот, события, которые происходят реже, считаются менее вероятными. Таким образом, понятие вероятности связывается с понятием частоты события.

Для того чтобы количественно оценить степень возможности события, необходимо ввести какую-то единицу измерения.

В теории вероятностей такой мерой является вероятность достоверного события — события, которое обязательно происходит в результате опыта.

Например, выпадение не более шести очков при одном бросании игральной кости. Если приписать достоверному событию вероятность, равную единице, то все другие события — возможные, но недостоверные — будут характеризоваться вероятностями, меньшими единицы.

Противоположным достоверному является невозможное событие (появление 12 очков при одном бросании игральной кости). Вероятность его — нуль.

Таким образом, диапазон изменения вероятностей любых событий представляет собой отрезок числовой оси $[0, 1]$. Вероятность некоторого события A обозначается символом $P(A)$.

Соотношения между событиями. Рассмотрим множество N событий A, B, C, \dots . Могут существовать различные соотношения.

1. Если при каждом испытании, в результате которого происходит событие A , происходит также и событие B , то говорят, что событие A влечет за собой событие B . Символически это можно записать

$$A \subset B \text{ или } B \supset A.$$

2. Если при каждом испытании оба события A и B происходят или не происходят, то говорят, что такие события равносильны. Равносильные события считаются тождественными и могут заменять друг друга

$$A = B.$$

3. Событие, заключающееся в наступлении хотя бы одного из событий A и B , называется суммой событий

$$A + B.$$

В общем случае суммой нескольких событий A, B, C, \dots называют событие, состоящее в появлении хотя бы одного из этих событий

$$A + B + C + \dots$$

4. Событие, состоящее в совместном наступлении событий A и B в одном опыте, называется *произведением событий*

$$A \cdot B.$$

Произведением нескольких событий A, B, C, \dots называется событие, состоящее в совместном появлении всех этих событий

$$A \cdot B \cdot C \cdot \dots$$

5. Обозначим *достоверное* событие U . *Невозможное* событие будем обозначать V .

6. Два события A и B называются *несовместными*, если в одном опыте их совместное появление невозможно, т. е. если

$$A \cdot B = V.$$

7. События A_1, A_2, \dots, A_n образуют *полную группу*, если в результате опыта обязательно произойдет по крайней мере одно из них, т. е. если

$$A_1 + A_2 + \dots + A_n = U.$$

8. Если два события A и \bar{A} образуют полную группу несовместных событий, то такие события называются *противоположными*. Для них одновременно удовлетворяются соотношения

$$A + \bar{A} = U; \quad A \cdot \bar{A} = V.$$

9. Несколько событий называются *равновозможными*, если в силу тех или иных причин (например, симметрии) ни одно из этих событий не является объективно более возможным, чем другое.

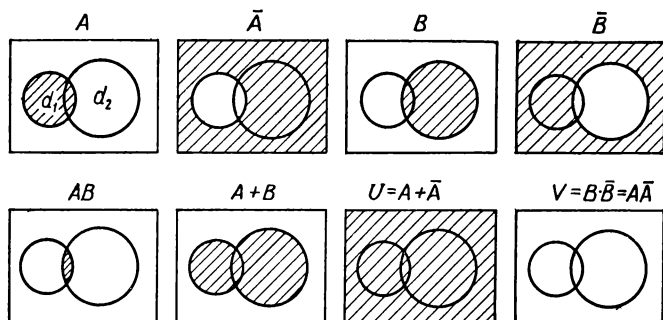


Рис. 4.1. Геометрическая интерпретация основных соотношений между событиями.

10. События A_1, A_2, \dots, A_n , образующие полную группу попарно несовместных равновозможных событий, называются *случаями*.

При пользовании введенными понятиями и соотношениями удобна наглядная геометрическая интерпретация.

Пусть, например, достоверное событие U состоит в том, что некоторая точка a окажется внутри области D . Невозможным событием V будет по условию такое, при котором точка не окажется внутри области D .

Обозначим через A событие, состоящее в том, что точка окажется внутри окружности диаметра d_1 (рис. 4.1), и через B событие, состоящее в том, что точка a окажется внутри окружности диаметра d_2 . Штриховкой внутри области D можно показать площади, соответствующие следующим событиям: 1) событие A ; 2) событие \bar{A} ; 3) событие B ; 4) событие \bar{B} ; 5) событие $A \cdot B$; 6) событие $A + B$; 7) событие $U = \bar{A} + A$; 8) событие $V = A \cdot \bar{A} = B \cdot \bar{B}$.

Непосредственный подсчет вероятностей. Рассмотрим, при каких условиях опытов можно определить $P(A)$, не прибегая с самого начала к испытаниям.

Если какое-нибудь испытание характеризуется симметрией возможных исходов, то говорят, что такое испытание *сводится к схеме случаев*. Классическое определение вероятностей опирается как раз на схему случаев. При этом становится возможным непосредственный подсчет вероятностей, основанный на определении доли «благоприятствующих» случаев в общем числе случаев.

Случай называют *благоприятствующим* событием, если появление его влечет за собой появление события.

Пусть, например, на девяти карточках написаны числа от единицы до девяти. Предположим, что событие A заключается в том, что на одной из карточек, выбранной наугад, окажется четное число. Всего здесь возможно девять случаев. Из них событию A благоприятствуют четыре случая — 2, 4, 6, 8. Если опыт сводится к схеме случая, то вероятность события A можно определить как отношение числа m случаев, благоприятствующих наступлению этого события, к общему числу n всех возможных случаев

$$P(A) = \frac{m}{n}. \quad (4.1)$$

Вероятность любого случайного события A заключена между вероятностями невозможного события V и достоверного события U

$$0 = P(V) \leq P(A) \leq P(U) = 1. \quad (4.2)$$

Формула (4.1) пригодна тогда и только тогда, когда опыт сводится к схеме случаев. В большинстве практических задач, связанных с реальными явлениями, вероятность непосредственно связывают с эмпирическим понятием частоты.

Частота или статистическая вероятность Пусть выполнено n опытов. В каждом из них могло произойти или не произойти некоторое событие A .

Частотой, или статистической вероятностью, события A в данной серии опытов называется отношение числа опытов, в которых появилось событие A , к общему числу произведенных опытов

$$P^*(A) = \frac{m}{n}. \quad (4.3)$$

В отличие от классической формулы, в выражении (4.3) m — число появлений события A ; n — общее число произведенных опытов.

Если число опытов невелико, то частота события носит случайный характер. При увеличении числа опытов частота события стабилизируется и приближается к некоторой постоянной величине. Эта закономерность характерна для случайных явлений. Математически это свойство было впервые сформулировано в теореме Я. Бернулли. При неограниченном увеличении числа опытов частота события будет сколь угодно мало отличаться от его вероятности в отдельном опыте.

Рассмотренные непосредственные способы определения вероятностей не всегда удобны и часто не применимы для расчетов. Во многих практических задачах требуется определять вероятности таких событий, которые трудно или бессмысленно воспроизводить экспериментально. Опыт может оказаться громоздким, дорогостоящим, а то и вообще невозможным (например, при оценке функционирования не существующего, а проектируемого оборудования).

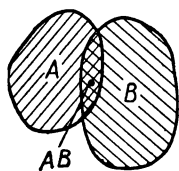
В большинстве случаев для определения вероятностей применяются косвенные методы. Они позволяют по известным вероятностям одних событий определять вероятности связанных с ними других событий. Применяя косвенные методы, пользуются основными теоремами теории вероятностей.

§ 2. ОСНОВНЫЕ ТЕОРЕМЫ

Вероятность суммы двух несовместных событий равна сумме вероятностей этих событий:

$$P(A + B) = P(A) + P(B). \quad (4.4)$$

Для суммы любого числа несовместных событий эту теорему можно записать в виде



$$P\left(\sum_{i=1}^n A_i\right) = \sum_{i=1}^n P(A_i). \quad (4.5)$$

Если события A и B совместны, вероятность суммы их можно выражать как

$$P(A + B) = P(A) + P(B) - P(A \cdot B). \quad (4.6)$$

Рис. 4.2. Геометрическая интерпретация вероятности суммы совместных событий.

Выражение (4.6) можно достаточно наглядно интерпретировать геометрически (рис. 4.2).

При произвольном числе событий A_i , ($i = 1, 2, \dots, n$), формула (4.6) обобщается выражением

$$P\left(\sum_{i=1}^n A_i\right) = \sum_i P(A_i) - \sum_i \sum_{j \neq i} P(A_i A_j) + \\ + \sum_i \sum_{j \neq i} \sum_{k \neq i, j} P(A_i A_j A_k) - \dots + (-1)^{n-1} P\left(\prod_{i=1}^n A_i\right). \quad (4.7)$$

Теорема сложения вероятностей имеет важные следствия.

С л е д с т в и е 1. Если события A_1, A_2, \dots, A_n образуют полную группу несовместных событий, то сумма их вероятностей

$$\sum_{i=1}^n P(A_i) = 1.$$

С л е д с т в и е 2. Сумма вероятностей противоположных событий

$$P(A) + P(\bar{A}) = 1.$$

Следствие 2 является частным случаем следствия 1. При решении практических задач часто легче вычислить вероятности противоположных событий, чем вероятности прямых событий. В этих случаях вычисляют $Q(A) = P(\bar{A})$ и находят $P(A) = 1 - Q(A)$.

Как видно, в формулы (4.6) и (4.7), кроме вероятностей отдельных слагаемых событий, входят вероятности произведений событий в различных сочетаниях. Вероятность произведения событий определяется теоремой умножения вероятностей. Прежде чем формулировать эту теорему, нам необходимо ввести понятие о зависимых и независимых событиях.

Событие A считается *не зависимым от события B* , если вероятность события A не зависит от того, произошло событие B или нет. Если же вероятность события A зависит от того, произошло событие B или нет, то говорят, что A *зависимо от B* . Вероятность события A , вычисленная в предположении, что совершилось событие B , называется *условной вероятностью события A* . Условная вероятность обозначается $P(A/B)$. Применяя понятие условной вероятности, можно записать условие независимости события A от события B в виде

$$P(A/B) = P(A). \quad (4.8)$$

Если же A от B не зависит, то

$$P(A/B) \neq P(A). \quad (4.9)$$

Сформулируем теперь теорему умножения вероятностей.

Вероятность произведения двух событий равна произведению вероятности одного из них на условную вероятность другого, вычисленную в предположении, что первое произошло

$$P(AB) = P(A)P(B/A). \quad (4.10)$$

Выражение (4.10) можно обобщить на произведение n событий

$$P\left(\prod_{i=1}^n A_i\right) = P(A_1)P(A_2/A_1)P(A_3/A_1A_2) \dots \\ \dots P\left(A_n/\sum_{i=1}^{n-1} A_i\right).$$

С л е д с т в и е 1. Если событие A не зависит от события B , то и событие B не зависит от события A , т. е. если

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

то

$$P(B/A) = P(B). \quad (4.11)$$

Другими словами, свойства зависимости или независимости событий всегда взаимны.

С л е д с т в и е 2. Вероятность произведения независимых событий равна произведению их вероятностей

$$P\left(\prod_{i=1}^n A_i\right) = \prod_{i=1}^n P(A_i). \quad (4.12)$$

Рассмотрим пример применения теорем сложения и умножения вероятностей. На рис. 4.3 изображена электрическая цепь с элементами k_1, k_2, k_3, k_4 . Выход из строя различных элементов цепи за время t — независимые события с вероятностями

$$P(K_1) = 0,4; \quad P(K_2) = P(K_3) = 0,6; \quad P(K_4) = 0,4.$$

Цепь разрывается при выходе из строя элемента k_1 либо k_4 , либо при совместном выходе из строя элементов k_2 и k_3 . Определим вероятность события C — разрыва цепи за время t .

Пусть событие A — выход из строя одного из элементов k_1 или k_4 . Событие B — выход из строя элементов k_2 и k_3 . Искомая вероятность

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

Поскольку

$$P(A) = P(K_1) + P(K_4) - P(K_1)P(K_4) = 0,64$$

и

$$P(B) = P(K_2)P(K_3) = 0,36,$$

то

$$P(C) \approx 0,77.$$

Важным следствием теорем сложения и умножения вероятностей является *формула полной вероятности*. Эта формула позволяет определить вероятность некоторого события A , которое происходит вместе с одним из событий H_1, H_2, \dots, H_n , образующих полную группу несовместных событий. События H_i ($i = 1, 2, \dots, n$) обычно называют гипотезами, при них возможно наступление события A . При такой постановке задачи вероятность события A равна сумме произведений вероятности каждой гипотезы на условную вероятность события A при этой гипотезе

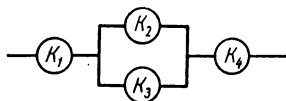


Рис. 4.3. К задаче определения вероятности разрыва электроцепи.

$$P(A) = \sum_{i=1}^n P(H_i)P(A/H_i). \quad (4.13)$$

Рассмотрим пример применения формулы полной вероятности. Двадцать экзаменационных билетов содержат по два вопроса, которые не повторяются. Экзаменующийся может ответить только на 35 вопросов. Экзамен считается сданным, если экзаменующийся ответит на оба вопроса одного билета или на один вопрос из первого билета и на указанный дополнительный вопрос из другого билета.

Определим вероятность того, что экзамен будет сдан. Рассмотрим две гипотезы:

H_1 — в первом билете известен один вопрос;

H_2 — в билете известны оба вопроса.

Определим вероятности этих гипотез, пользуясь теоремами сложения и умножения вероятностей. В первом билете известным может оказаться или первый или второй вопрос. Учитывая, что общее число вопросов равно сорока, получим

$$P(H_1) = \frac{35}{40} \cdot \frac{5}{39} + \frac{5}{40} \cdot \frac{35}{39} = \frac{35}{156};$$

$$P(H_2) = \frac{35}{40} \cdot \frac{34}{39} = \frac{1191}{1560}.$$

Событие, заключающееся в сдаче экзамена, обозначим A . Условные вероятности события A при рассматриваемых гипотезах равны

$$P(A/H_1) = \frac{34}{38}; \quad P(A/H_2) = 1.$$

По формуле полной вероятности получим

$$P(A) = P(H_1) P(A/H_1) + P(H_2) P(A/H_2) = \frac{35}{156} \cdot \frac{34}{38} + \\ + \frac{1191}{1560} \cdot 1 \approx 0,89.$$

Строго говоря, в этой задаче нам необходимо было рассматривать не две гипотезы H_1 и H_2 , а три. По условиям задачи возможной является также гипотеза H_0 — в билете оба вопроса неизвестны. Действительно, только в этом случае мы будем иметь дело с полной группой H_0, H_1, H_2 . Но условная вероятность события A при этой гипотезе равна нулю, поэтому в формуле полной вероятности член $P(H_0) P(A/H_0)$ отсутствует. При решении практических задач обычно нет смысла рассматривать гипотезы, при которых интересующее нас событие является невозможным. Это только усложняет вычисления.

Очень часто на практике приходится определять вероятности гипотез после того, как опыт произведен. При этом пользуются *теоремой гипотез* или *формулой Байеса*, которая представляет собой следствие теоремы умножения и формулы полной вероятности. Задача заключается в следующем.

Пусть H_1, H_2, \dots, H_n полная группа несовместных гипотез. Априорно известны вероятности этих гипотез. В результате испытания произошло некоторое событие A . Требуется определить апостериорные вероятности гипотез H_1, H_2, \dots, H_n , определить условную вероятность $P(H_i/A)$ с учетом того, что событие A совершилось.

Эта вероятность определяется выражением

$$P(H_i/A) = \frac{P(H_i) P(A/H_i)}{P(A)}, \quad i = 1, 2, \dots, n, \quad (4.14)$$

или, учитывая (4.13),

$$P(H_i/A) = \frac{P(H_i) P(A/H_i)}{\sum_j P(H_j) P(A/H_j)}, \quad i = 1, 2, \dots, n. \quad (4.15)$$

В правых частях формул (4.14) и (4.15) $P(H_i)$ — априорные вероятности гипотез H_i , а $P(A/H_i)$ — условные вероятности события A при гипотезах H_i , ($i = 1, 2, \dots, n$).

В качестве примера рассмотрим задачу о контроле качества продукции.

В одной из двух партий все изделия качественные, а в другой 45% изделий бракованные. Из любой партии наудачу взято качественное изделие. Определить вероятность того, что изделие выбрано из партии, в которой все изделия качественные.

Рассмотрим две гипотезы:

H_1 — деталь выбрана из партии с браком;

H_2 — деталь выбрана из партии, в которой все детали качественные.

Поскольку первоначальный выбор чисто случайный, то априорные вероятности $P(H_1) = P(H_2) = 0,5$. Событие A — выбрана качественная деталь. Условные вероятности этого события при H_1 и H_2 равны соответственно $P(A/H_1) = 0,55$; $P(A/H_2) = 1$.

Согласно формуле полной вероятности $P(A) = 0,5(0,55 + 1) = 0,775$. Искомая вероятность того, что партия состоит только из качественных деталей

$$P(H_2/A) = \frac{P(H_2) \cdot P(A/H_2)}{P(A)} = \frac{0,5 \cdot 1}{0,775} = 0,63.$$

Таким образом, новая информация, полученная нами при испытании, дает возможность пересмотреть первоначальное распределение вероятностей гипотез.

Контрольные вопросы и задания

1. Перечислите основные соотношения между событиями. Дайте их геометрическое представление.
 2. Что такое полная группа событий?
 3. Охарактеризуйте схему случая.
 4. Дайте определения классической и статистической вероятностей.
- В каком соотношении находятся эти две величины?
5. Приведите пример применения теории сложения вероятностей.
 6. То же для теоремы умножения вероятностей.
 7. То же для формулы гипотез Байеса.

§ 3. СЛУЧАЙНЫЕ ВЕЛИЧИНЫ

Реальные опыты или испытания, в том числе и наблюдения, характеризуются обычно различными переменными величинами. Их называют параметрами, характеристиками, координатами и т. д. Если в результате испытания эти переменные величины могут принимать те или иные значения,

причем, заранее неизвестно, какие именно, то такие переменные называют случайными величинами. Предположим, мы выбираем наугад какую-нибудь деталь из большой партии однотипных деталей. Размеры выбранной детали — случайные величины. Поскольку результаты исследований и измерений обычно оцениваются числами, случайные величины могут принимать различные числовые значения. Если случайные величины принимают отдельные, изолированные друг от друга значения, которые можно перечислить, то эти величины называются дискретными. Случайные величины, которые непрерывно переходят в функции некоторого параметра, причем значения их нельзя заранее перечислить, называются непрерывными.

Понятие случайной величины является одним из основных в теории вероятностей. Классическая теория вероятностей оперирует преимущественно событиями. В большинстве задач современной теории вероятностей основными объектами исследования являются случайные величины.

Примерами дискретных случайных величин могут служить: 1) количество сбоев вычислительной машины за некоторый отрезок времени; 2) число повторений одной буквы в сообщении определенной длины; 3) количество сердечных сокращений в минуту.

К непрерывным случайным величинам относятся: 1) напряжение в энергосистеме; 2) температура окружающей среды; 3) скорость движущегося объекта.

Для обозначения случайных величин обычно используют большие буквы латинского алфавита X, Y, Z, \dots . Возможные отдельные значения дискретных величин и текущие значения непрерывных величин будем обозначать соответствующими малыми буквами.

Пусть дискретная случайная величина X может принять одно из значений x_1, x_2, \dots, x_n . Так как каждое из этих возможных значений не является достоверным, то естественно говорить о вероятности, с которой случайная величина X принимает то или иное значение x_i ($i = 1, 2, \dots, n$). В результате опыта, например измерения параметра, величина X примет одно из значений x_i ; произойдет одно из полной группы несовместных событий

$$\left. \begin{aligned} X &= x_1, \\ X &= x_2, \\ &\vdots \\ X &= x_n. \end{aligned} \right\} \quad (4.16)$$

Вероятности этих событий обозначим

$$P(X = x_1) = P_1; \quad P(X = x_2) = P_2; \quad \dots; \quad P(X = x_n) = P_n.$$

Сумма этих вероятностей

$$\sum_{i=1}^n P_i = 1.$$

С вероятностной точки зрения случайную величину X можно полностью описать, если точно указано, какой вероятностью характеризуется каждое возможное значение x_i .

Законом распределения случайной величины называют соотношение, связывающее возможные значения случайной величины с их вероятностями. Простейшей формой задания закона распределения дискретной случайной величины является ряд распределения. Эта таблица, в которой перечислены все возможные значения случайной величины и соответствующие вероятности

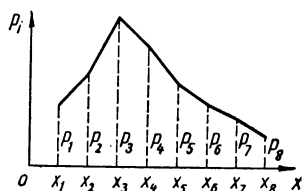


Рис. 4.4. Полигон распределения дискретной случайной величины.

x_i	x_1	x_2	\dots	x_n
P_i	P_1	P_2	\dots	P_n

Для большей наглядности ряды распределения представляют в графическом изображении (рис. 4.4). По оси абсцисс отложены возможные значения случайной величины, а по оси ординат — их вероятности. Более подробно мы остановимся на формах представления дискретных случайных величин в разделе, связанном со статистическим анализом.

Функция распределения

Ряд распределения представляет собой исчерпывающую вероятностную характеристику дискретной случайной величины. Непрерывную случайную величину, очевидно, нельзя охарактеризовать таким образом. Непрерывная случайная величина может принимать бесчисленное множество значений, и составить таблицу всех предполагаемых значений невозможно.

Для количественной характеристики распределения непрерывной случайной величины вводится функция распределения

$$\Phi(x) = P(X < x). \quad (4.17)$$

Это вероятность события, которое заключается в том, что случайная величина X примет значение, меньшее некоторой текущей переменной x . Иногда эту характеристику называют интегральной функцией распределения или интегральным законом распределения.

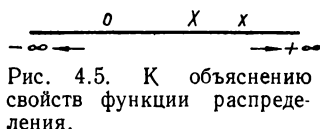


Рис. 4.5. К объяснению свойств функции распределения.

Определим общие свойства

функции распределения.

1. Функция распределения $\Phi(x)$ — неубывающая,

$$\Phi(x_2) \geq \Phi(x_1) \text{ при } x_2 > x_1.$$

$$2. \Phi(-\infty) = 0.$$

$$3. \Phi(+\infty) = 1.$$

Эти свойства достаточно очевидны из геометрического представления (рис. 4.5). Пусть случайная величина X представляется случайной точкой на числовой оси. Функция распределения $\Phi(x)$ — это вероятность того, что случайная точка X в результате опыта попадет левее x . Конечно же вероятность этого события не может уменьшиться с увеличением x , т. е. при сдвиге x вправо. Значит, $\Phi(x)$ при росте x не может убывать.

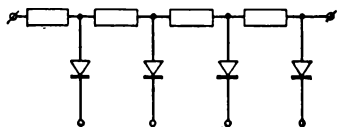


Рис. 4.6. Диодная логическая схема.

Если значение x устремить к $-\infty$, а точку x неограниченно сдвигать влево, то в пределе событие $X < x$ становится невозможным: $\Phi(-\infty) = P(X < -\infty) = 0$. При неограниченном перемещении точки x вправо событие $X < x$ в пределе становится достоверным

$$\Phi(+\infty) = P(X < +\infty) = 1.$$

Функцию распределения можно построить графически. В общем случае это будет график неубывающей функции, изменяющейся в пределах $[0, 1]$ при изменении аргумента от $-\infty$ до $+\infty$.

Рассмотрим пример построения функции распределения для дискретных величин. Логическая схема (рис. 4.6) испытывает резкий бросок напряжения. Пробой любого

диода равновероятен. Случайная величина X — количество пробитых диодов — характеризуется рядом распределения

x_i	0	1	2	3	4
P_i	0,2	0,4	0,3	0,08	0,02

Функцию распределения можно определить в виде

$$\Phi(x) = P(X < x) = \sum_{x_i < x} P(X = x_i).$$

Неравенство $x_i < x$ указывает, что суммирование распространяется на все те значения x_i , которые меньше x .



Рис. 4.7. Функция распределения дискретной случайной величины.

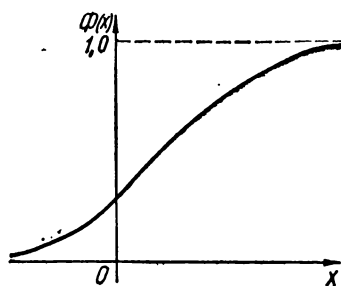


Рис. 4.8. Функция распределения непрерывной случайной величины.

Когда текущая переменная x проходит через какое-нибудь из возможных значений дискретной величины X , функция распределения меняется скачкообразно, причем величина скачка равна вероятности этого значения.

1. при $x \leq 0$ $\Phi(x) = 0$,
2. » $0 < x \leq 1$ $\Phi(x) = 0,2$,
3. » $1 < x \leq 2$ $\Phi(x) = 0,6$,
4. » $2 < x \leq 3$ $\Phi(x) = 0,9$,
5. » $3 < x \leq 4$ $\Phi(x) = 0,98$,
6. » $4 < x$ $\Phi(x) = 1$.

Функция распределения (рис. 4.7) представляет собой разрывную ступенчатую функцию. Скачки функции соответствуют возможным значениям случайной величины и равны вероятностям этих значений.

По мере увеличения числа возможных значений число скачков увеличивается, сами скачки уменьшаются. В пределе для бесконечного числа возможных значений (непрерывная случайная величина) функция распределения становится непрерывной (рис. 4.8).

Функция распределения является универсальной вероятностной характеристикой. Она существует как для непрерывных, так и для дискретных случайных величин. Для непрерывных случайных величин можно представить закон распределения в другой форме.

Функция плотности вероятности

Пусть функция распределения $\Phi(x)$ случайной величины непрерывна и дифференцируема. Вероятность события

$$x \leq X < x + \Delta x$$

определяется как приращение $\Phi(x)$ на участке с границами x и $x + \Delta x$:

$$P(x \leq X < x + \Delta x) = \Phi(x + \Delta x) - \Phi(x). \quad (4.18)$$

Рассмотрим среднюю вероятность этого события на участке $(x, x + \Delta x)$ и перейдем к пределу при $\Delta x \rightarrow 0$:

$$\lim_{\Delta x \rightarrow 0} \frac{P(x \leq X < x + \Delta x)}{\Delta x} = \lim_{\Delta x \rightarrow 0} \frac{\Phi(x + \Delta x) - \Phi(x)}{\Delta x} = \Phi'(x). \quad (4.19)$$

Полученную производную функцию распределения обозначим

$$\Phi(x) = f(x). \quad (4.20)$$

Функцию $f(x)$ называют плотностью распределения или плотностью вероятности случайной величины X .

В отличие от функции распределения — интегрального закона — плотность вероятности называют дифференциальным законом распределения. Графически плотность вероятности изображается кривой распределения (рис. 4.9).

Вероятность попадания случайной величины X на элементарный участок dx равна $f(x) dx$ и называется элементом вероятности. Это площадь элементарного прямоугольника с основанием dx и высотой $f(x)$.

Вероятность попадания случайной величины на некоторый участок (a, b) можно выразить через плотность вероятности в виде

$$P(a < x < b) = \int_a^b f(x) dx. \quad (4.21)$$

Функция распределения может быть выражена через плотность вероятности как

$$\Phi(x) = \int_{-\infty}^x f(x) dx. \quad (4.22)$$

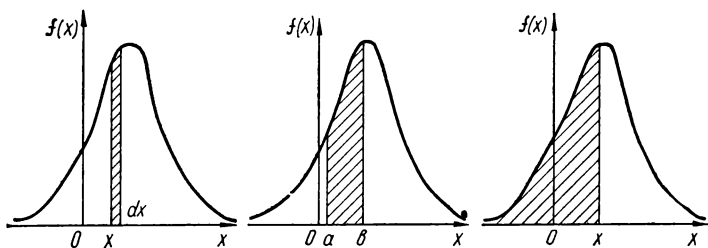


Рис. 4.9. Кривая плотности вероятности непрерывной случайной величины.

Общие свойства плотности вероятности следующие:

1. Плотность вероятности — функция неотрицательная

$$f(x) \geq 0.$$

2. Интеграл в бесконечных пределах от плотности вероятности

$$\int_{-\infty}^{\infty} f(x) dx = 1.$$

Геометрически это означает, что кривая распределения лежит не ниже оси абсцисс и полная площадь, ограниченная кривой распределения и осью абсцисс, равна единице.

Указанные свойства со всей очередностью вытекают из общих свойств функции распределения случайной величины.

Числовые характеристики

Рассмотренные законы распределения в интегральной и дифференциальной форме являются исчерпывающими вероятностными характеристиками случайной величины. Однако при решении многих практических задач полное

определение этих законов не является необходимым. Иногда достаточно ограничиться знанием некоторых основных, типичных показателей закона распределения. К ним относятся числовые характеристики случайных величин.

На практике обычно применяют числовые характеристики двух типов: 1) характеристика положения; 2) моменты.

Характеристики положения определяют средние, типичные значения случайной величины. К этим характеристикам относятся: математическое ожидание, мода и медиана.

Для дискретной случайной величины математическое ожидание определяется как сумма произведений всех возможных значений случайной величины на вероятности этих значений

$$m_x = M[X] = \sum_{i=1}^n x_i' P_i. \quad (4.23)$$

Математическое ожидание иногда называют средним значением. Между математическим ожиданием и средним арифметическим существует соответствие, аналогичное соответствию между частотой и вероятностью. При увеличении числа опытов среднее арифметическое полученных при наблюдении значений случайной величины приближается к ее математическому ожиданию. Как мы увидим далее, при решении практических задач различия между этими понятиями не делают.

Для непрерывной случайной величины математическое ожидание определяется как

$$m_x = M[X] = \int_{-\infty}^{\infty} x f(x) dx. \quad (4.24)$$

Часто вместо самой случайной величины X рассматривают величину $\hat{X} = X - m_x$, которая называется центрированной случайной величиной. Определяется она как разность действительной случайной величины и ее математического ожидания. Операция центрирования, по сути, представляет собой перенесение начала координат в точку, соответствующую математическому ожиданию.

Значение случайной величины, имеющее наибольшую вероятность, называется модой

$$Mo_x = x_i' / P_i = P_{\max}. \quad (4.25)$$

На полигоне распределения, или кривой распределения, значению Mo_x соответствует наивысшая точка (рис. 4.10).

Если полигон распределения (кривая распределения) имеет два и больше максимумов, то распределение называется полимодальным.

Значение случайной величины, удовлетворяющее равенству

$$P(X < Me_x) = P(X > Me_x), \quad (4.26)$$

называется медианой случайной величины — Me_x (рис. 4.11). Равенство (4.26) показывает, что события $X < Me_x$ и $X > Me_x$ — равновероятны. Геометрически, это означает,

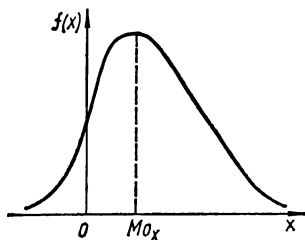


Рис. 4.10. Мода.

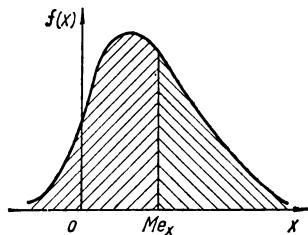


Рис. 4.11. Медиана.

что перпендикуляр, восстановленный из точки Me_x , делит площадь, ограниченную кривой распределения, на две равные части.

В общем случае значения m_x , Mo_x и Me_x не совпадают. Если закон распределения — симметричный модальный закон, то все три характеристики положения определены одним значением случайной величины.

Как уже было сказано, кроме характеристик положения для описания свойств распределения случайной величины применяются моменты. Обычно пользуются моментами двух видов: начальными и центральными.

Начальные моменты порядка s определяются как

$$\alpha_s = [X] = \sum_{i=1}^n x_i^s P_i, \quad (4.27)$$

$$\alpha_s [X] = \int_{-\infty}^{\infty} x^s f(x) dx. \quad (4.28)$$

Выражение (4.27) определяет начальные моменты для дискретной случайной величины, а (4.28) — для непрерывной. При $s=1$ получим математическое ожидание

случайной величины X . Используя знак математического ожидания, запишем (4.27) и (4.28) в виде

$$\alpha_s = [X] = M[X^s]. \quad (4.29)$$

Формула (4.29), очевидно, справедлива как для непрерывных, так и для дискретных случайных величин.

Моменты центрированной случайной величины называются центральными. Для вычисления центральных моментов порядка s дискретной и непрерывной случайных величин соответственно применяются формулы

$$\mu_s = \sum_{i=1}^n (x_i - m_x)^s P_i; \quad (4.30)$$

$$\mu_s = \int_{-\infty}^{\infty} (x - m_x)^s f(x) dx. \quad (4.31)$$

Для любой случайной величины центральный момент первого порядка ($s = 1$) равен нулю

$$\mu_1 = M[X - m_x] = 0. \quad (4.32)$$

В общем случае можно рассматривать моменты относительно произвольной точки a

$$\gamma_s = M[(X - a)^s]. \quad (4.33)$$

Практически же из всех моментов в качестве характеристик случайной величины чаще всего применяют первый начальный (математическое ожидание) и второй центральный моменты.

Второй центральный момент характеризует рассеивание значений случайной величины X относительно ее среднего значения m_x . Эта характеристика называется дисперсией случайной величины. Для вычисления дисперсии используют формулы

$$\mu_2 = D[X] = \sum_{i=1}^n (x_i - m_x)^2 P_i \quad (4.34)$$

для дискретной случайной величины и

$$\mu_2 = D[X] = \int_{-\infty}^{\infty} (x - m_x)^2 f(x) dx \quad (4.35)$$

для непрерывной.

Часто рассеивание характеризуют средним квадратическим отклонением («стандартом») случайной величины X

$$\sigma[X] = \sqrt{D[X]}, \quad (4.36)$$

или

$$\sigma_x = \sqrt{D_x}. \quad (4.37)$$

Характеристика σ , более наглядна, так как имеет размерность самой случайной величины.

Если знания математического ожидания и дисперсии недостаточно, то для более полной характеристики случайной величины пользуются моментами высших порядков.

Например, может оказаться, что математические ожидания и дисперсии двух случайных величин совпадают, а законы распределения тем не менее различны. На рис. 4.12 показаны кривые распределения, имеющие различную асимметрию.

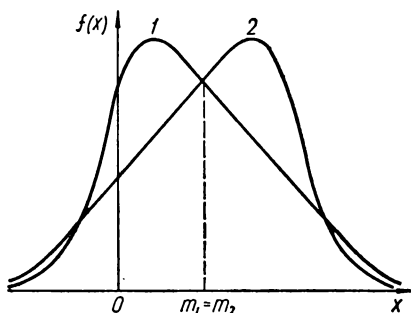


Рис. 4.12. Асимметрия распределения.

Асимметрию характеризуют при помощи третьего центрального момента μ_3 . Для симметричных распределений все нечетные центральные моменты равны нулю. Поэтому естественно принять в качестве характеристики асимметрии какой-либо из них. Самый простой — третий. Для того чтобы характеристика была безразмерной, μ_3 делят на среднее квадратическое отклонение в третьей степени

$$S_k = \frac{\mu_3}{\sigma^3}. \quad (4.38)$$

Характеристику (4.38) называют коэффициентом асимметрии. Асимметрия может быть положительной $S_k > 0$ (кривая 1 рис. 4.12) и отрицательной $S_k < 0$ (кривая 2 рис. 4.12).

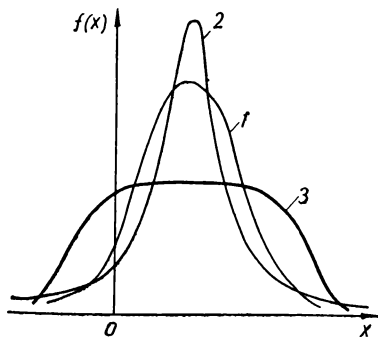


Рис. 4.13. Эксцесс распределения.

Кривые распределения могут быть более или менее «крутыми» (рис. 4.13). Для

характеристики островершинности или плосковершинности применяют четвертый центральный момент. С его помощью

определяется эксцесс распределения

$$E_x = \frac{\mu_4}{\sigma^4} - 3. \quad (4.39)$$

Для наиболее распространенного в природе нормального распределения значение эксцесса принимается равным нулю. Все другие законы распределения характеризуются определенным эксцессом по отношению к нормальному. Величина $\frac{\mu_4}{\sigma^4}$ для нормального закона равна 3. Поэтому в общей формуле (4.39) в правой части вычитается тройка. Кривые, более островершинные (кривая 2 рис. 4.13) по сравнению с нормальной (кривая 1), обладают положительным эксцессом, более плосковершинные (кривая 3) — отрицательным.

§ 4. НЕКОТОРЫЕ ЗАКОНЫ РАСПРЕДЕЛЕНИЯ

Случайные величины, характеризующие различные физические процессы и явления, могут иметь самые разнообразные законы распределения. Простейшим с математической точки зрения является закон равномерной плотности вероятности. В конкретных задачах часто приходит-

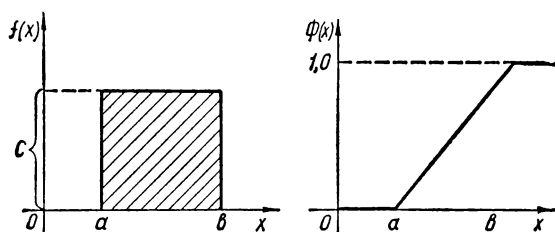


Рис. 4.14. Закон равномерной плотности.

ся рассматривать случайные величины, о которых заранее известно, что все их возможные значения лежат в пределах заданного интервала и все значения случайной величины внутри этого интервала равновероятны. Дифференциальный закон распределения в этом случае (рис. 4.14) можно записать в виде

$$f(x) = \begin{cases} C = \text{const}, & a < x < b, \\ 0, & x < a \text{ и } x > b. \end{cases}$$

Поскольку полная площадь, ограниченная осью x и кривой распределения, равна единице

$$(b-a)C = 1,$$

имеем

$$C = \frac{1}{b-a}.$$

Тогда плотность распределения

$$f(x) = \begin{cases} \frac{1}{b-a}, & a < x < b, \\ 0, & x < a \text{ и } x > b. \end{cases}$$

Функция распределения (см. рис. 4.14)

$$\Phi(x) = \begin{cases} 0, & x < a, \\ \frac{x-a}{b-a}, & a < x < b, \\ 1, & x > b. \end{cases}$$

Числовые характеристики:

1. Математическое ожидание

$$m_x = \int_a^b \frac{x}{b-a} dx = \frac{a+b}{2}.$$

2. Так как кривая плотности распределения симметрична,

$$Me_x = m_x.$$

3. Моды закон равномерной плотности не имеет.

4. Дисперсия

$$D_x = \frac{1}{b-a} \int_a^b \left(x - \frac{a+b}{2}\right)^2 dx = \frac{(b-a)^2}{12},$$

а среднее квадратическое отклонение

$$\sigma_x = \sqrt{D_x} = \frac{b-a}{2\sqrt{3}}.$$

5. Асимметрия равна нулю.

6. Четвертый центральный момент

$$\mu_4 = \frac{1}{b-a} \int_a^b \left(x - \frac{a+b}{2}\right)^4 dx = \frac{(b-a)^4}{80},$$

и, следовательно, эксцесс

$$E_x = \frac{\mu_4}{\sigma^4} - 3 = -1,2.$$

Большое практическое значение имеет закон распределения Пуассона или, как его иногда называют, закон редких явлений. Этому распределению подчиняются события, вероятность которых при каждом отдельном испытании очень мала, а число испытаний велико. Вероятность случайной величины X , распределенной по закону Пуассона,

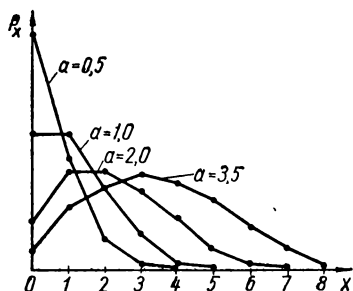


Рис. 4.15. Закон распределения Пуассона.

$$P_x = \frac{a^x}{x!} e^{-a},$$

где a — математическое ожидание случайной величины. На

рис. 4.15 приведены кривые распределения Пуассона для различных значений параметра a .

Замечательным свойством распределения Пуассона является то, что его дисперсия равна математическому ожиданию

$$D_x = m_x = a.$$

Это свойство позволяет практически проверять справедливость гипотезы о том, что изучаемая случайная величина распределена по закону Пуассона.

На практике наиболее часто встречается так называемый нормальный закон распределения (или закон Гаусса). Он характеризуется плотностью вероятности

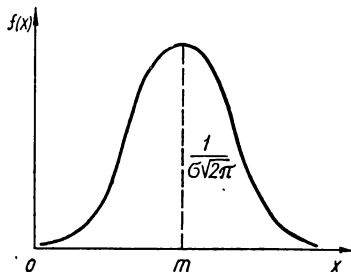


Рис. 4.16. Нормальный закон распределения.

$$f(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x-m)^2}{2\sigma^2}},$$

где m — математическое ожидание случайной величины X , σ — среднее квадратическое отклонение, так что

$$D_x = \sigma^2.$$

Кривая плотности вероятности нормального закона распределения имеет характерную колоколообразную форму (рис. 4.16). Нормальный закон распределения — это предельный закон, к которому приближаются другие законы

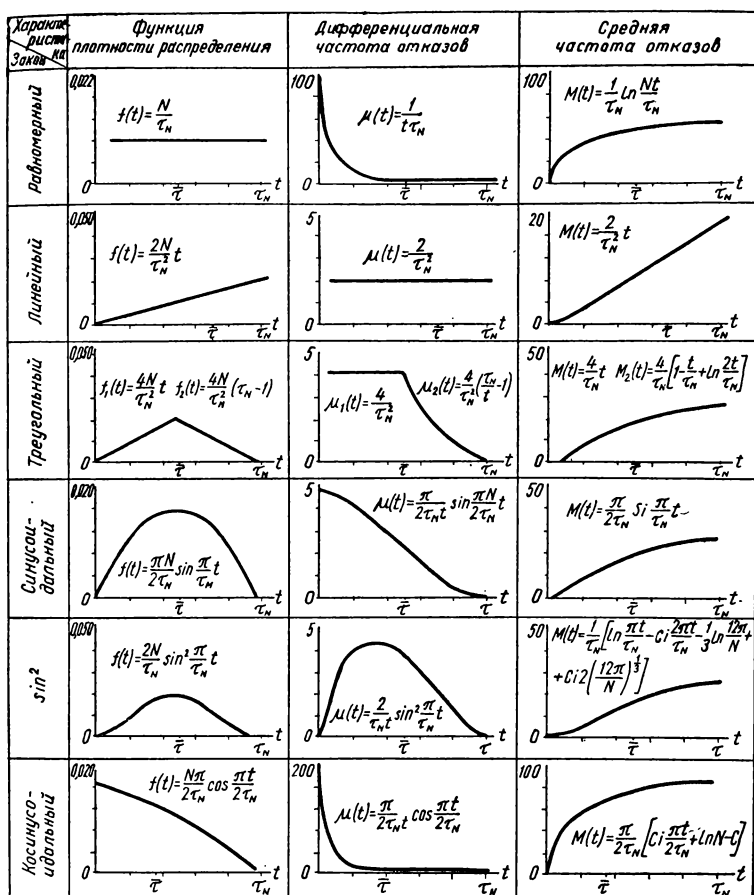


Рис. 4.17. Законы распределения и надежностные характеристики.

распределения при часто встречающихся типичных условиях. Сумма большого числа независимых или слабо зависящих случайных величин, распределенных по произвольным законам, приближенно подчиняется нормальному закону. Для этого должно лишь выполняться условие, состоящее в том, чтобы отдельные слагаемые в сумме имели

примерно одинаковую значимость. В противном случае в суммарном распределении будет доминировать распределение наиболее значимой составляющей.

Исследование распределений случайных величин — один из основных этапов в решении задач статистических предсказаний. Так, например, пуассоновское распределение играет важную роль при решении вопросов массового обслуживания. Большое практическое значение имеет анализ распределений при предсказании отказов в теории надежности. На рис. 4.17 приведены типичные законы распределения и соответствующие характеристики надежности.

Контрольные вопросы и задания

1. Что такое случайная величина? Приведите примеры непрерывных и дискретных случайных величин.
2. Что такое закон распределения случайной величины? Какие вы знаете формы задания закона распределения?
3. Сформулируйте основные свойства функции распределения случайной величины.
4. Что такое дифференциальный закон распределения случайной величины? Сформулируйте основные свойства.
5. Перечислите основные числовые характеристики случайной величины.
6. Изобразите геометрически представление характеристики положения.
7. Что такое операция центрирования случайной величины?
8. Что такое среднее квадратическое отклонение?
9. Чем характеризуется асимметрия распределения?
10. Что такое эксцесс?

§ 5. СИСТЕМЫ СЛУЧАЙНЫХ ВЕЛИЧИН

На практике результат опыта обычно описывается не одной, а двумя и более случайными величинами.

Например, на электростанции контролируется напряжение генераторов, частота, потребляемая мощность. Каждый съем данных дает в общем случае случайные значения этих переменных. Он зависит от нагрузки, температуры, метеорологических условий и многих других факторов.

Будем в таких случаях говорить, что случайные величины образуют комплекс или систему.

Систему нескольких случайных величин X, Y, \dots, W обозначают (X, Y, \dots, W) .

Свойства системы случайных величин не исчерпываются свойствами отдельных ее составляющих. Для полной характеристики системы необходимо учитывать зависимости между случайными величинами.

Геометрически систему двух случайных величин (X, Y) можно изобразить случайной точкой на плоскости с координатами X и Y (рис. 4.18). Система трех случайных величин изображается точкой в трехмерном пространстве. В общем случае может идти речь о системе n случайных величин как о случайной точке в n -мерном пространстве.

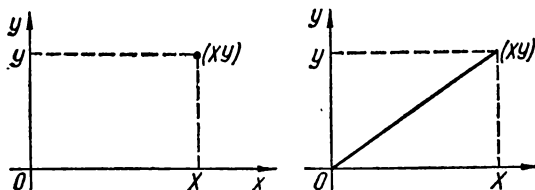


Рис. 4.18. Система двух случайных величин.

Другим подходом является представление системы случайных величин случайным вектором на плоскости xOy . Составляющие этого вектора по осям — это случайные величины X и Y (рис. 4.18). Систему n случайных величин можно интерпретировать n -мерным случайным вектором. При этом теория систем случайных величин рассматривается как теория случайных векторов.

Функция распределения

Функция распределения системы двух случайных величин — это вероятность совместного выполнения двух неравенств $X < x$ и $Y < y$

$$\Phi(x, y) = P((X < x)(Y < y)). \quad (4.40)$$

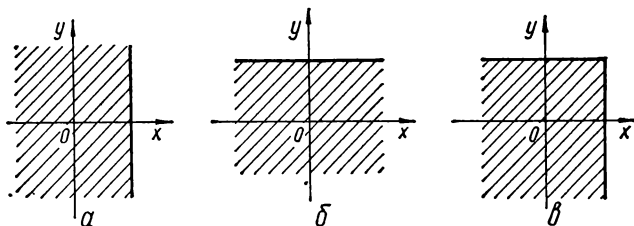


Рис. 4.19. К объяснению свойств функции распределения системы случайных величин

Геометрически — это вероятность попадания случайной точки (X, Y) в бесконечный квадрат с вершиной (x, y) , лежащий левее и ниже ее (рис. 4.19, а).

Функция распределения одной случайной величины $X — \Phi_1(x)$ — вероятность попадания случайной точки в полуплоскость левее абсциссы x (рис. 4.19, б). Функция распределения $Y — \Phi(y)$ — вероятность попадания случайной точки в полуплоскость ниже ординаты y (рис. 4.19, в).

Ранее мы приводили основные свойства функции распределения $\Phi(x)$ для одной случайной величины. Теперь сформулируем свойства функции распределения системы случайных величин. Определим общие свойства функции распределения системы двух случайных величин.

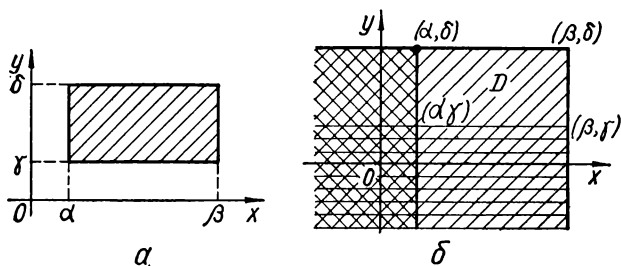


Рис. 4.20. К определению вероятности попадания в заданную область.

1. Функция распределения $\Phi(x, y)$ является неубывающей функцией своих аргументов, т. е.

$$\Phi(x_2, y) \geq \Phi(x_1, y) \text{ при } x_2 > x_1; \quad (4.41)$$

$$\Phi(x, y_2) \geq \Phi(x, y_1) \text{ при } y_2 > y_1.$$

2. Повсюду на $-\infty$ функция распределения

$$\Phi(x, -\infty) = \Phi(-\infty, y) = \Phi(-\infty, -\infty) = 0. \quad (4.42)$$

3. При одном из аргументов, равном $+\infty$, функция распределения системы превращается в функцию распределения случайной величины, соответствующей другому аргументу

$$\Phi(x, +\infty) = \Phi_1(x); \quad (4.43)$$

$$\Phi(+\infty, y) = \Phi_2(y).$$

4. Если оба аргумента равны $+\infty$, функция распределения системы

$$\Phi(+\infty, +\infty) = 1. \quad (4.44)$$

Эти свойства с достаточной очевидностью вытекают из геометрического представления (рис. 4.19).

Для системы двух случайных величин можно поставить задачу о вероятности попадания случайной точки (X, Y) в некоторую область D на плоскости xOy . Эта задача аналогична задаче о вероятности попадания случайной точки X , изображающей одну случайную величину, на заданный отрезок оси Ox .

Чтобы упростить задачу, будем рассматривать область D в виде прямоугольника (рис. 4.20), причем точки, лежащие на нижней и левой границах, включим в прямоугольник, а точки, лежащие на верхней и правой границах, не включим. Рассмотрим события

$$A_{\text{экр}} \alpha \leq X < \beta \text{ и } B_{\text{экр}} \gamma \leq Y < \delta.$$

Нас интересует вероятность совместного выполнения этих событий

$$P(AB) = P((\alpha \leq X < \beta)(\gamma \leq Y < \delta)). \quad (4.45)$$

Эту вероятность можно получить из рассмотрения четырех бесконечных квадрантов с вершинами в точках, соответствующих вершинам прямоугольника D с координатами (β, δ) , (α, δ) , (β, γ) и (α, γ) (рис. 4.20).

Вероятность попадания в прямоугольник D выражается через функцию распределения системы следующим образом

$$P((X, Y) \in D) = \Phi(\beta, \delta) - \Phi(\alpha, \delta) - \Phi(\beta, \gamma) + \Phi(\alpha, \gamma). \quad (4.46)$$

Вероятность попадания в область произвольной формы будет определена позднее через плотность распределения.

Функция плотности вероятности

Как и для одной случайной величины, функция распределения системы случайных величин является универсальной характеристикой. Она существует как для непрерывных, так и для дискретных случайных величин. Практически основное значение имеют непрерывные случайные величины.

Распределение системы непрерывных величин обычно характеризуют плотностью распределения.

Плотность распределения одной случайной величины определялась как предел отношения вероятности попадания на малый участок к длине этого участка при его неограниченном уменьшении.

Рассмотрим систему двух случайных величин (X, Y) . На плоскости xOy выделим малый прямоугольник D_Δ со

сторонами Δx и Δy , примыкающий к точке с координатами (x, y) (рис. 4.21).

Вероятность попадания в этот прямоугольник в соответствии с формулой (4.46)

$$P((X, Y) \in D_{\Delta}) = \Phi(x + \Delta x, y + \Delta y) - \Phi(x + \Delta x, y) - \Phi(x, y + \Delta y) + \Phi(x, y). \quad (4.47)$$

Разделим эту вероятность на площадь прямоугольника и перейдем к пределу при $\Delta x \rightarrow 0$ и $\Delta y \rightarrow 0$

$$\lim_{\substack{\Delta x \rightarrow 0 \\ \Delta y \rightarrow 0}} \frac{P((X, Y) \in D_{\Delta})}{\Delta x \Delta y} = \\ = \lim_{\substack{\Delta x \rightarrow 0 \\ \Delta y \rightarrow 0}} \frac{\Phi(x + \Delta x, y + \Delta y) - \Phi(x + \Delta x, y) - \Phi(x, y + \Delta y) + \Phi(x, y)}{\Delta x \Delta y}. \quad (4.48)$$

Если $\Phi(x, y)$ непрерывна и дифференцируема, то правая часть (4.48) представляет собой вторую смешанную част-

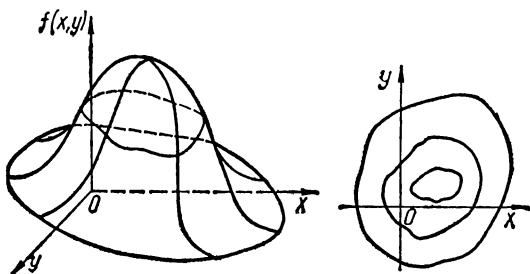


Рис. 4.22. Плотность вероятности системы двух случайных величин.

ную производную функции $\Phi(x, y)$ по x и по y

$$f(x, y) = \frac{\partial^2 \Phi(x, y)}{\partial x \partial y} = \Phi''_{xy}(x, y). \quad (4.49)$$

Функция $f(x, y)$ называется плотностью распределения системы случайных величин.

Геометрически — это некоторая поверхность (рис. 4.22), называемая поверхностью распределения. Если пересечь поверхность распределения $f(x, y)$ плоскостью, параллельной xOy , и спроектировать полученное сечение на

xOy , получим кривую, в каждой точке которой плотность распределения постоянна. Такие кривые называются *кривыми равной плотности*. Они представляют собой горизонтали поверхности распределения. Часто бывает удобно задавать распределение семейством кривых равной плотности (см. рис. 4.22).

При рассмотрении плотности распределения $f(x)$ для одной случайной величины было введено понятие «элемента вероятности» $f(x) dx$. Это вероятность попадания случайной величины X на элементарный участок dx , прилежащий к точке x .

Для системы случайных величин элемент вероятности $f(x, y) dxdy$ — вероятность попадания случайной точки в элементарный прямоугольник со сторонами dx, dy , примыкающий к точке (x, y) . Эта вероятность равна объему элементарного параллелепипеда, ограниченного сверху поверхностью $f(x, y)$ и опирающегося на элементарный прямоугольник $dxdy$ (рис. 4.23).

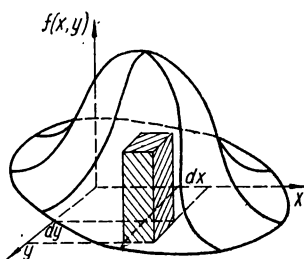


Рис. 4.23. Элемент вероятности.

Пользуясь понятием элемента вероятности, можно вывести выражение для вероятности попадания случайной точки в произвольную область D . Эта вероятность получается интегрированием элементов вероятности по всей области D

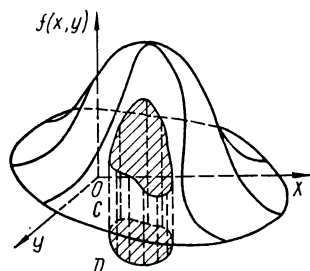


Рис. 4.24. Вероятность попадания в заданную область.

$$P((X, Y) \in D) = \iint_{(D)} f(x, y) dxdy. \quad (4.50)$$

Геометрически эта вероятность изображается объемом цилиндрического тела C , ограниченного сверху поверхностью распределения и опирающегося на область D (рис. 4.24).

Из общей формулы следует формула для вероятности попадания в прямоугольник D , ограниченный абсциссами α и β и ординатами γ и δ

$$P((X, Y) \in D) = \int_{\alpha}^{\beta} \int_{\gamma}^{\delta} f(x, y) dxdy. \quad (4.51)$$

Выразим функцию распределения $\Phi(x, y)$ через плотность распределения $f(x, y)$. $\Phi(x, y)$ определялась как вероятность попадания в бесконечный квадрант. Этот квадрант можно рассматривать как прямоугольник, ограниченный абсциссами $-\infty, x$ и ординатами $-\infty, y$.

Из (4.51) получаем

$$\Phi(x, y) = \int_{-\infty}^x \int_{-\infty}^y f(x, y) dx dy. \quad (4.52)$$

Рассмотрим основные свойства плотности распределения.

1. Плотность распределения системы — функция неотрицательная

$$f(x, y) \geq 0. \quad (4.53)$$

Это ясно из того, что плотность распределения есть предел отношения двух неотрицательных величин: вероятности попадания в прямоугольник и площади прямоугольника.

2. Двойной интеграл в бесконечных пределах от плотности распределения системы случайных величин

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dx dy = 1. \quad (4.54)$$

Этот интеграл — вероятность попадания случайной точки в плоскость xOy , т. е. вероятность достоверного события.

Геометрически это свойство означает, что полный объем тела, ограниченного поверхностью распределения и плоскостью xOy , равен единице.

Зависимые и независимые случайные величины

При изучении систем случайных величин всегда надо обращать внимание на степень и характер зависимости этих величин.

Случайная величина Y называется не зависимой от случайной величины X , если закон распределения величины Y не зависит от того, какое значение приняла случайная величина X .

Зависимость между величинами можно охарактеризовать с помощью условных законов распределения.

Условным законом распределения величины Y , входящей в систему (X, Y) , называется ее закон распределения при условии, что другая случайная величина X приняла определенное значение. Условный закон распределения

можно задавать как функцией распределения, так и плотностью вероятности. Условная функция распределения обозначается $\Phi(y/x)$. Условная плотность распределения $f(y/x)$.

Для непрерывных случайных величин условие независимости Y от X можно записать в виде

$$f(y/x) = f_2(y). \quad (4.55)$$

Если же Y зависит от X , то

$$f(y/x) \neq f_2(y). \quad (4.56)$$

Зависимость или независимость случайных величин всегда взаимны: если величина Y не зависит от X , то и величина X не зависит от Y . Исходя из этого определение независимых случайных величин можно дать так: случайные величины X и Y называются независимыми, если закон распределения каждой из них не зависит от того, какое значение приняла другая.

Числовые характеристики системы случайных величин

Ранее были рассмотрены числовые характеристики одной случайной величины — начальные и центральные моменты различных порядков. Было отмечено, что важнейшими являются математическое ожидание m_x и дисперсия D_x .

Введем аналогичные характеристики для системы двух случайных величин.

Начальным моментом порядка k, s системы (X, Y) называется математическое ожидание произведения X^k на Y^s

$$\alpha_{k,s} = M[X^k Y^s]. \quad (4.57)$$

Центральным моментом порядка k, s системы (X, Y) называется математическое ожидание произведения k -й и s -й степени соответствующих центрированных величин

$$\mu_{k,s} = M[X^k Y^s], \quad (4.58)$$

где

$$\dot{X} = X - m_x, \quad \dot{Y} = Y - m_y.$$

Запишем формулы для вычисления моментов. Для дискретных случайных величин

$$\alpha_{k,s} = \sum_i \sum_j x_i^k x_j^s P_{ij}, \quad (4.59)$$

где

$$P_{ij} = P[(X = x_i)(Y = y_j)] \text{ и} \\ \mu_{k,s} = \sum_i \sum_j (x_i - m_x)^k (y_j - m_y)^s P_{ij}. \quad (4.60)$$

Для непрерывных случайных величин

$$\alpha_{k,s} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x^k y^s f(x, y) dx dy, \quad (4.61)$$

$$\mu_{k,s} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - m_x)^k (y - m_y)^s f(x, y) dx dy. \quad (4.62)$$

Помимо значений k и s , характеризующих порядок момента по отношению к отдельным случайным величинам системы, рассматривается еще суммарный порядок момента $k + s$. Соответственно суммарному порядку моменты бывают первые, вторые и т. д.

На практике применяются обычно только первые и вторые моменты. Первые начальные моменты — это математические ожидания случайных величин X и Y , входящих в систему

$$m_x = \alpha_{1,0} = M[X^1 Y^0] = M[X], \\ m_y = \alpha_{0,1} = M[X^0 Y^1] = M[Y]. \quad (4.63)$$

Совокупность математических ожиданий m_x , m_y характеризует положение системы. Геометрически — это координаты средней точки на плоскости, вокруг которой происходит рассеивание точки (X, Y) .

Рассмотрим вторые центральные моменты системы. Два из них — это дисперсии

$$D_x = \mu_{2,0} = M[\check{X}^2 \check{Y}^0] = M[\check{X}^2] = D[X], \quad (4.64)$$

$$D_y = \mu_{0,2} = M[\check{X}^0 \check{Y}^2] = M[\check{Y}^2] = D[Y].$$

Они характеризуют рассеивание случайной точки в направлении осей O_x и O_y . Кроме дисперсий, существует еще второй смешанный центральный момент

$$\mu_{1,1} = M[\check{X} \check{Y}]. \quad (4.65)$$

Его обычно обозначают

$$K_{xy} = M[\check{X} \check{Y}] = M[(X - m_x)(Y - m_y)]. \quad (4.66)$$

Это корреляционный момент (момент связи) случайных величин. Для дискретных случайных величин

$$K_{xy} = \sum_i \sum_j (x_i - m_x)(y_j - m_y) P_{ij}. \quad (4.67)$$

Для непрерывных —

$$K_{xy} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - m_x)(y - m_y) f(x, y) dx dy. \quad (4.68)$$

Корреляционный момент характеризует связь между случайными величинами X и Y . Для независимых случайных величин корреляционный момент равен нулю.

Система произвольного числа случайных величин

При решении практических задач приходится рассматривать системы более чем двух случайных величин.

Полной характеристикой системы произвольного числа случайных величин является закон распределения, который может быть задан функцией распределения или плотностью распределения.

Функцией распределения системы n случайных величин (X_1, X_2, \dots, X_n) называется вероятность совместного выполнения n неравенств вида $X_i < x_i$

$$\Phi(x_1, x_2, \dots, x_n) = P((X_1 < x_1)(X_2 < x_2) \dots (X_n < x_n)). \quad (4.69)$$

Плотностью распределения системы будет n -я смешанная частная производная функции $\Phi(x_1, x_2, \dots, x_n)$, взятая один раз по каждому аргументу

$$f(x_1, x_2, \dots, x_n) = \frac{\partial^n \Phi(x_1, x_2, \dots, x_n)}{\partial x_1 \partial x_2 \dots \partial x_n}. \quad (4.70)$$

Систему k случайных величин, взятую из системы n случайных величин произвольным образом, называют частной системой (при $k < n$).

Случайные величины X_1, X_2, \dots, X_n называются независимыми, если закон распределения каждой частной системы, выделенной из системы (X_1, X_2, \dots, X_n) , не зависит от того, какие значения приняли остальные случайные величины.

Определим минимальное количество числовых характеристик, с помощью которых можно описать систему n случайных величин:

1. n математических ожиданий

$$m_1, m_2, \dots, m_n,$$

характеризующих средние значения величин;

2. n дисперсий

$$D_1, D_2, \dots, D_n,$$

характеризующих рассеивание;

3. $n(n-1)$ корреляционных моментов

$$K_{ij} = M[\dot{X}_i \dot{X}_j], \quad (i \neq j), \quad (4.71)$$

где

$$\dot{X}_i = X_i - m_i; \quad \dot{X}_j = X_j - m_j,$$

характеризующих попарную связь всех величин, входящих в систему.

Дисперсия каждой из случайных величин X_i является частным случаем корреляционного момента

$$D_i = K_{ii} M[\dot{X}_i^2]. \quad (4.72)$$

Все корреляционные моменты обычно представляют в виде прямоугольной матрицы

$$\begin{vmatrix} K_{11} & K_{12} & \dots & K_{1n} \\ K_{21} & K_{22} & \dots & K_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ K_{n1} & K_{n2} & \dots & K_{nn} \end{vmatrix}$$

Из определения корреляционного момента ясно, что $K_{ij} = K_{ji}$, т. е. симметричные относительно главной диагонали элементы корреляционной матрицы равны. Поэтому матрицу иногда пишут в виде

$$\begin{vmatrix} K_{11} & K_{12} & \dots & K_{1n} \\ & K_{22} & \dots & K_{2n} \\ & & \ddots & \vdots \\ & & & K_{nn} \end{vmatrix}$$

По главной диагонали — дисперсии.

Если случайные величины не коррелированы, то все элементы матрицы, кроме диагональных, равны нулю

$$\| K_{ij} \| = \begin{vmatrix} D_1 & & 0 \\ & D_2 & \\ & & \ddots \\ 0 & & & D_n \end{vmatrix}$$

Контрольные вопросы и задания

1. Дайте определение системы случайных величин.
2. Сформулируйте свойства функции распределения системы двух случайных величин.
3. Как связаны между собой дифференциальный и интегральный законы распределения случайных величин?
4. Как определяются зависимые и независимые случайные величины?
5. Дайте определение начальным и центральным моментам системы случайных величин.
6. Что такое корреляционный момент случайных величин?
7. Охарактеризуйте корреляционную матрицу системы случайных величин.

§ 6. СЛУЧАЙНЫЕ ФУНКЦИИ. СЛУЧАЙНЫЕ ПРОЦЕССЫ

Классическая теория вероятностей рассматривает «массовые» случайные явления. Массовое явление представляет собой совокупность многократных повторений явлениями действий «наудачу», рассматриваемых в целом и без учета хронологической последовательности.

В отличие от классической теории вероятностей, теория вероятностных, или стохастических процессов, развитая в основном А. Н. Колмогоровым и А. Я. Хинчиным, оперирует процессами и последовательностями (дискретными процессами) случайных явлений. Случайные процессы и последовательности представляют собой совокупности случайных величин в динамике их развития. Это те же массовые явления. Но они рассматриваются не в виде, например, однородного массива случайных чисел, а в виде последовательности чисел в хронологии появления величин, которым они соответствуют.

Примерами случайных процессов могут служить изменения координаты броуновской частицы, флуктуации в электрических цепях, вибрации узлов станка во время его работы, изменение температуры больного в ходе болезни, изменение биоэлектрической активности мозга и т. п.

Случайной называют функцию, которая в результате испытания может принять тот или иной конкретный вид, причем заранее неизвестно, какой именно.

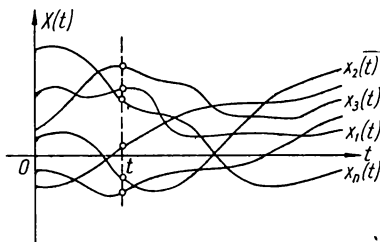


Рис. 4.25. Семейство реализаций случайной функции.

Конкретный вид, который принимает случайная функция в опыте, называют реализацией случайной функции.

Случайными процессами называются случайные функции времени. При этом различают непрерывные случайные процессы — функции непрерывного времени и случайные последовательности или цепи — функции дискретного времени.

Каждая реализация является неслучайной функцией. В результате группы испытаний получают «семейство» реализаций случайной функции (рис. 4.25). При определенном фиксированном значении аргумента (в общем случае — всех аргументов) случайная функция превращается в случайную величину. Эту случайную величину называют сечением случайной функции.

Случайные функции, как правило, обозначают $X(t)$, $Y(t)$, ... в отличие от неслучайных $x(t)$, $y(t)$, ...

Законы распределения

Рассмотрение случайной функции можно с некоторым приближением заменить рассмотрением системы случайных величин. В самом деле, в моменты t_1, t_2, \dots, t_m случайная функция $X(t)$ превращается в случайные величины

$$X(t_1), X(t_2), \dots, X(t_m).$$

С увеличением m точность такого представления будет повышаться. При $m \rightarrow \infty$ понятие случайной функции можно рассматривать как обобщение понятия системы бесконечного числа случайных величин. Учитывая это, дадим понятие закона распределения случайной функции.

Для одной случайной величины закон распределения является функцией одного аргумента, для системы двух случайных величин — функцией двух аргументов и т. д. Но использовать в качестве вероятностных характеристик функции нескольких аргументов чрезвычайно неудобно. Уже для системы трех-четырех случайных величин законом распределения не пользуются. Рассматриваются только числовые характеристики.

Если рассматривать случайную функцию как систему бесконечного числа случайных величин, то ее закон распределения будет функцией бесконечного числа аргументов. Формально этот закон можно записать, но практическое использование его исключено.

Пусть $X(t)$ — сечение случайной функции в момент t . Это случайная величина, которая характеризуется законом распределения. В общем случае закон распределения зависит от t . Можно записать его в виде: $f(x, t)$. Это одномерный закон распределения случайной функции $X(t)$. В самом деле, $f(x, y)$ характеризует закон распределения только для данного, хотя и произвольного, значения t . По такой характеристике нельзя судить о зависимости значений $X(t)$ при различных t . Но можно ввести в рассмотрение двумерный закон распределения

$$f(x_1, x_2; t_1, t_2). \quad (4.73)$$

Это характеристика системы двух случайных величин $X(t_1)$, $X(t_2)$ — двух сечений случайной функции $X(t)$ в произвольные моменты времени.

Очевидно, таким же образом можно вводить в рассмотрение все более и более исчерпывающие характеристики: трехмерный закон, четырехмерный и т. д. Однако оперировать с подобными громоздкими характеристиками, зависящими от многих аргументов, очень неудобно. При исследовании законов распределения случайных функций обычно рассматривают лишь некоторые частные случаи, когда для полной характеристики случайной функции достаточно знать функцию (4.73).

Большинство практических задач можно решить и без определения законов распределения, используя лишь некоторые характеристики случайных функций. Эти характеристики аналогичны числовым характеристикам случайных величин.

Характеристики случайных функций

В теории вероятностей большое значение имеют основные числовые характеристики случайных величин: математическое ожидание и дисперсия — для одной случайной величины, математические ожидания и корреляционная матрица для системы случайных величин. Числовые характеристики — очень эффективный аппарат. С его помощью удастся сравнительно легко решать многие практические задачи, не прибегая к определению законов распределения.

Для случайных функций также вводят простейшие основные характеристики, аналогичные числовым характеристикам случайных величин. Для случайных величин числовые характеристики представляют собой некоторые числа.

Характеристики случайных функций в общем случае — функции.

Определим математическое ожидание случайной функции $X(t)$. В любом сечении $X(t)$ представляет собой случайную величину, характеризуемую некоторым математическим ожиданием. Для различных значений t очевидно существует некоторая функция

$$m_x(t) = M[X(t)]. \quad (4.74)$$

Математическое ожидание случайной функции $X(t)$ — это неслучайная функция $m_x(t)$, которая при каждом значении аргумента совпадает со значением математического ожидания соответствующего сечения случайной функции.

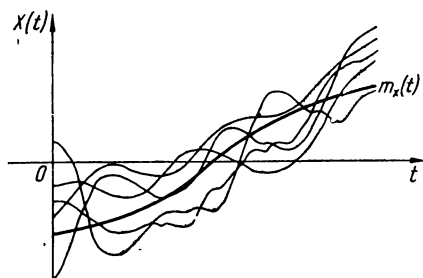


Рис. 4.26. Математическое ожидание случайной функции.

Другими словами, это некоторая средняя функция, около которой группируются и относительно которой колеблются все возможные реализации случайной функции (рис. 4.26).

В качестве второй важной характеристики определим дисперсию случайной функции.

Дисперсией случайной функции $X(t)$ называется неслучайная функция $D_x(t)$, значение которой для каждого значения аргумента t совпадает со значением дисперсии соответствующего сечения случайной функции

$$D_x(t) = D[X(t)]. \quad (4.75)$$

Эта зависимость характеризует разброс возможных реализаций случайной функции относительно математического ожидания. Функция $D_x(t)$, очевидно, не может быть отрицательной. Если извлечь из нее квадратный корень, получим характеристику, известную под названием среднего квадратического отклонения случайной функции

$$\sigma_x(t) = \sqrt{D_x(t)}. \quad (4.76)$$

Рассмотренных характеристик недостаточно для описания основных особенностей случайных функций. В этом легко убедиться на примере. Пусть две случайные функции

$X_1(t)$ и $X_2(t)$ заданы семействами реализаций (рис. 4.27). В результате расчетов оказалось, что случайные функции $X_1(t)$ и $X_2(t)$ имеют одинаковые математические ожидания и дисперсии. Но характер этих функций различен. Случайная функция $X_1(t)$ изменяется плавно. Для нее характерна ярко выраженная зависимость между значениями при различных t . Если в момент t случайная функция $X(t)$ приняла значение, заметно превышающее среднее, то, вероятно, и в точке t' она также примет значение больше среднего.

Совсем не так ведет себя функция $X_2(t)$. У нее резко колебательный беспорядочный характер. Связь между отдельными значениями случайной функции $X_2(t)$ резко уменьшается при увеличении расстояния между ними по t . То, что внутренняя структура обоих случайных процессов различна, очевидно. Но это различие не удастся уловить при помощи математического ожидания и дисперсии.

По-видимому, необходимо ввести некоторую специальную характеристику для описания тесноты связи между значениями случайной функции в различные моменты t . Эта характеристика носит название корреляционной функции.

Рассмотрим два сечения случайной функции $X(t)$ в моменты t и t' . При близких значениях t и t' величины $X(t)$ и $X(t')$ связаны более тесно. Если же интервал между t и t' увеличивается, то теснота связи, зависимость между $X(t)$ и $X(t')$, вообще говоря, должна убывать.

Для характеристики зависимости двух случайных величин $X(t)$ и $X(t')$ вводят второй смешанный центральный момент (корреляционный момент). В этом случае корреляционный момент будет функцией двух аргументов t и t' .

Корреляционной функцией случайной функции $X(t)$ называется неслучайная функция двух аргументов $K_x(t, t')$, которая при каждой паре значений t, t' равна корреляцион-

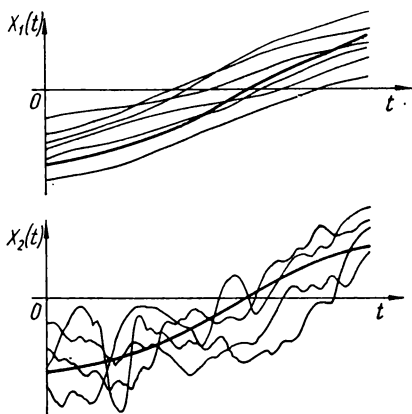


Рис. 4.27. К объяснению понятия автокорреляционной функции.

ному моменту соответствующих сечений случайной функции

$$K_x(t, t') = M[\dot{X}(t) \dot{X}(t')]. \quad (4.77)$$

В выражении (4.77)

$$\dot{X}(t) = X(t) - m_x(t); \quad \dot{X}(t') = X(t') - m_x(t').$$

Для рассмотренного примера (рис. 4.27) корреляционная функция случайной функции $X_1(t)$ медленно убывает с увеличением интервала (t, t') .

Корреляционная функция случайной функции $X_2(t)$ с увеличением этого интервала убывает быстро. Положим $t = t'$. Тогда

$$K_x(t, t) = M[(\dot{X}(t))^2] = D_x(t), \quad (4.78)$$

т. е. при равенстве аргументов корреляционная функция обращается в дисперсию случайной функции. А значит, отпадает

необходимость в дисперсии как отдельной самостоятельной характеристике, и в качестве основных характеристик случайной функции достаточно рассматривать ее математическое ожидание и корреляционную функцию.

Поскольку корреляционный момент двух случайных величин $X(t)$ и $X(t')$ не зависит от порядка, в котором рассматриваются эти величины, то корреляционная функция симметрична относительно своих аргументов

$$K_x(t, t') = K_x(t', t). \quad (4.79)$$

На рис. 4.28 изображена корреляционная функция в виде поверхности. Поверхность симметрична относительно вертикальной плоскости Q , проходящей через биссектрису угла tOt' .

Свойства корреляционной функции со всей очевидностью вытекают из свойств корреляционной матрицы системы случайных величин. Заменяем случайную функцию $X(t)$ системой m случайных величин.

Если уменьшать промежутки между аргументами, одновременно увеличивая m , двувходная корреляционная мат-

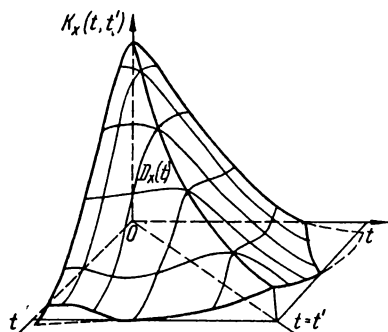


Рис. 4.28. Корреляционная функция.

рица системы в пределе переходит в функцию двух непрерывно изменяющихся аргументов. Свойство симметричности матрицы относительно главной диагонали преобразуется в свойство симметричности корреляционной функции (4.79). Главную диагональ корреляционной матрицы составляют дисперсии случайных величин. Аналогично $t' = t$ приводит к

$$K(t, t) = D_x(t).$$

Чтобы практически определить корреляционную функцию случайной функции $X(t)$, необходимо задаться рядом равноотстоящих значений аргумента и строить корреляционную матрицу полученной системы случайных величин. Эта матрица представляет собой таблицу значений корреляционной функции для прямоугольной сетки значений аргументов на плоскости tOt' . Затем строят функцию двух аргументов $K(t, t')$ путем интерполирования.

Иногда удобно пользоваться нормированной корреляционной функцией

$$r_x(t, t') = \frac{K_x(t, t')}{\sigma_x(t) \sigma_x(t')}. \quad (4.80)$$

При $t = t'$ нормированная корреляционная функция равна единице

$$r_x(t, t) = \frac{K_x(t, t)}{\sigma_x(t) \sigma_x(t)} = \frac{D_x(t)}{D_x(t)} = 1. \quad (4.81)$$

Элементарные операции над случайными функциями

Пусть к случайной функции $X(t)$ прибавляется неслучайное слагаемое $\varphi(t)$

$$Y(t) = X(t) + \varphi(t), \quad (4.82)$$

где $Y(t)$ — новая случайная функция.

Посмотрим, как изменяются характеристики случайной функции. В соответствии с теоремой сложения математических ожиданий имеем

$$m_y(t) = m_x(t) + \varphi(t), \quad (4.83)$$

т. е. при прибавлении к случайной функции неслучайного слагаемого к ее математическому ожиданию прибавляется то же неслучайное слагаемое.

Корреляционную функцию случайной функции $Y(t)$ можно определить в виде

$$K_y(t, t') = M[\dot{Y}(t) \dot{Y}(t')] = M[(Y(t) - m_y(t))(Y(t') -$$

$$\begin{aligned} -m_y(t')] &= M[X(t) + \varphi(t) - m_x(t) - \varphi(t)](X(t') + \varphi(t') - \\ &- m_x(t') - \varphi(t')) = M[(X(t) - m_x(t))(X(t') - m_x(t'))] = \\ &= K_x(t, t'). \end{aligned} \quad (4.84)$$

Таким образом, при прибавлении неслучайного слагаемого корреляционная функция случайной функции не меняется.

Пусть теперь случайная функция $X(t)$ умножается на неслучайный множитель $\varphi(t)$

$$Y(t) = \varphi(t) X(t). \quad (4.85)$$

Если вынести неслучайную величину за знак математического ожидания, получим

$$m_y(t) = M[\varphi(t) X(t)] = \varphi(t) m_x(t), \quad (4.86)$$

т. е. при умножении случайной функции на неслучайный множитель ее математическое ожидание умножается на тот же множитель.

Корреляционная функция равна

$$\begin{aligned} K_y(t, t') &= M[\dot{Y}(t) \dot{Y}(t')] = M[(Y(t) - m_y(t))(Y(t') - \\ &- m_y(t'))] = M[\varphi(t)(X(t) - m_x(t))\varphi(t')(X(t') - m_x(t'))] = \\ &= \varphi(t)\varphi(t') M[(X(t) - m_x(t))(X(t') - m_x(t'))] = \\ &= \varphi(t)\varphi(t') K_x(t, t'). \end{aligned} \quad (4.87)$$

Таким образом, при умножении случайной функции $X(t)$ на неслучайную функцию $\varphi(t)$ корреляционная функция случайной функции $X(t)$ умножается на $\varphi(t)\varphi(t')$. В частном случае, когда $\varphi(t)$ не зависит от времени

$$\varphi(t) = c,$$

где $c = \text{const}$, корреляционная функция умножается на c^2 .

Учитывая описанные свойства случайных функций, почти всегда можно значительно упростить операции с ними. Например, при исследовании случайной функции $X(t)$ можно заранее перейти к центрированной функции

$$\dot{X}(t) = X(t) - m_x(t). \quad (4.88)$$

Математическое ожидание центрированной функции тождественно равно нулю, а корреляционная функция совпадает с корреляционной функцией исходной случайной функции $X(t)$

$$K_{\dot{x}}(t, t') = M[\dot{X}(t) \dot{X}(t')] = K_x(t, t'). \quad (4.89)$$

Кроме центрирования широко применяется также операция нормирования случайных функций. Нормированную случайную функцию можно представить в виде

$$X_N(t) = \frac{\dot{X}(t)}{\sigma_x(t)}.$$

Корреляционная функция нормированной случайной функции $X_N(t)$ выражается как

$$r_{x_N}(t, t') = \frac{K_{x_N}(t, t')}{\sigma_x(t) \sigma_x(t')}. \quad (4.90)$$

Дисперсия ее равна единице.

Определение характеристик случайных функций из опыта

Допустим, произведено n независимых опытов, в результате чего получено n реализаций случайной функции $X(t)$ (рис. 4.29). Найдем оценки для характеристик случайной функции: математического ожидания $m_x(t)$, дисперсии $D_x(t)$, корреляционной функции $K_x(t, t')$.

Пусть выбран ряд сечений случайной функции в моменты времени

$$t_1, t_2, \dots, t_m.$$

Каждому из моментов t_1, t_2, \dots, t_m будет соответствовать n значений случайной величины. Значения t_1, t_2, \dots, t_m и величина интервала между соседними значениями выбирается в зависимости от вида экспериментальных кривых. Выбор этот делают так, чтобы по точкам можно было восстановить основной ход кривых (операция квантования по времени). Часто интервал между соседними значениями задается частотой работы регистрирующего прибора (например, в системах центрального контроля технологических процессов с выводом результатов на цифропечатающие устройства).

Зарегистрированные значения $X(t)$ заносят в таблицу, каждая строка которой соответствует определенной

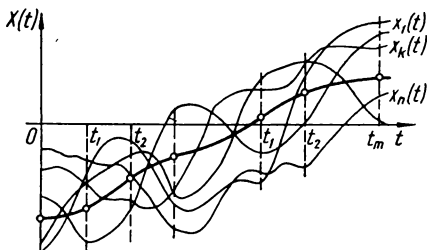


Рис. 4.29. К определению характеристик случайной функции из опыта.

реализации, а число столбцов равно числу сечений случай-
ной функции.

В таком виде данные представляют собой результаты n
опытов над системой m случайных величин

$$X(t_1), X(t_2), \dots, X(t_m).$$

Сначала определяют оценки математических ожиданий

$$\tilde{m}_x(t_k) = \frac{\sum_{i=1}^n x_i(t_k)}{n}. \quad (4.90)$$

Для дисперсий оценки имеют вид

$$D_x(t_k) = \frac{\sum_{i=1}^n [x_i(t_k) - \tilde{m}_x(t_k)]^2}{n-1}, \quad (4.91)$$

а для корреляционных моментов —

$$\tilde{K}_x(t_k, t_e) = \frac{\sum_{i=1}^n [x_i(t_k) - \tilde{m}_x(t_k)] [x_i(t_e) - \tilde{m}_x(t_e)]}{n-1}. \quad (4.92)$$

Вычислив эти характеристики, можно построить зави-
симости $\tilde{m}_x(t)$ и $\tilde{D}_x(t)$. Функция двух аргументов $\tilde{K}_x(t_1, t_2)$
строится в прямоугольной сетке точек по значениям аргу-
ментов. При необходимости все эти зависимости аппроксими-
руют какими-либо аналитическими выражениями. Ме-
тоды аппроксимации, или приближения функций, мы рас-
смотрим позже в разделах статистического анализа.

Контрольные вопросы и задания

1. Дайте определение случайной функции.
2. Что такое закон распределения случайных функций?
3. Перечислите характеристики случайной функции. Дайте им
определение.
4. Что такое операция нормирования?
5. Как изменяются характеристики случайной функции при при-
бавлении неслучайной? При умножении на неслучайную?
6. Как практически определить характеристики случайной функции?

§ 7. МЕТОДЫ ТЕОРИИ СЛУЧАЙНЫХ ФУНКЦИЙ В ИССЛЕДОВАНИИ СИСТЕМ

В теоретико-множественном смысле под системой пони-
мают множество элементов, которые характеризуются опре-
деленными свойствами и между которыми установлены
определенные связи.

В дальнейшем под термином система будем подразумевать все, что состоит из связанных друг с другом частей.

Мы уже познакомились с примером математической системы — системой случайных величин.

Прежде всего отметим, что определение любой конкретной системы — произвольно. Измерительный прибор можно назвать системой. Более сложная совокупность, включающая технологический объект, наш измерительный элемент, регулятор, также является системой. В свою очередь, объект с измерительным прибором и регулятором представляет часть более крупной системы — производственного предприятия и т. д.

Вся Вселенная состоит из множества систем, каждая из которых содержится в более крупной системе. Всегда можно определить более обширную систему, в которую входит данная, и выделить из данной системы более ограниченную.

Задача строгого научного исследования любой системы характеризуется рядом особенностей. Пусть есть система из n элементов. Если их не считать системой, то для выяснения природы элементов понадобится n отдельных исследований. Если же это множество элементов считать системой, потребуются исследования не только n элементов, но и $n(n - 1)$ связей между ними. Возьмем для примера систему из 8 элементов.

Внутри такой системы можно определить 56 связей. Число различных состояний, в которых может находиться система, составляет 2^{56} . Это фантастически большое число! Вот почему задача полного исследования систем является сложной и необычной. Система, находящаяся в динамическом режиме, может переходить из одного состояния в другое в течение любого интервала времени. Очевидно, что для оценки поведения такой системы необходим огромный объем исследований.

Входные, выходные сигналы, а также потоки информации внутри любой динамической системы в общем случае — случайные процессы. Ранее мы познакомились с методом непосредственного определения характеристик случайных функций из опыта. Подобный метод применяется далеко не всегда. С одной стороны, постановка специальных опытов для исследований может оказаться сложной и дорогостоящей. С другой стороны, очень часто требуется исследовать случайные функции, характеризующие объекты не существующие, а лишь проектируемые или разрабатываемые.

И сами эти исследования нужны для рационального выбора параметров системы.

В подобных случаях приходится пользоваться различными косвенными способами исследования случайных функций. В этом случае характеристики интересующих нас функций отыскивают косвенным путем, по характеристикам других случайных функций, связанных с интересующими нас. Прикладная теория случайных функций в основном занимается разработкой и развитием таких косвенных методов.

Представим себе некоторую динамическую систему A . Она может быть любой природы: электрической, механической, биологической. Это может быть измерительный прибор, вычислительная машина, система, автоматического управления, промышленное предприятие и т. д.

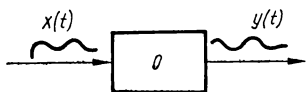


Рис. 4.30. Динамическая система.

На вход системы поступают некоторые входные данные. Система перерабатывает эти данные и выдает результаты. Сигналы, поступающие на вход системы, будем называть *воздействием*, а выдаваемый результат — *реакцией*. Причем, в общем случае, реакция соответствует не только «полезному» воздействию, но зависит также от неучитываемых помех (шумов) различного рода. Воздействиями могут быть переменные токи, напряжения, углы поворота, температура окружающей среды, сигналы и команды, подаваемые на систему управления. Реакциями также могут быть токи или напряжения, углы поворота стрелок приборов и исполнительных двигателей, качество продукции, результаты вычислений и т. п.

В простейшем случае на вход системы A подается одно воздействие, являющееся функцией времени $x(t)$. Реакцию системы на это воздействие можно представить другой функцией времени $y(t)$ (рис. 4.30).

Оператор системы

Говорят, что система A осуществляет некоторое преобразование входного воздействия. В результате этого функция $x(t)$ преобразуется в другую функцию $y(t)$. Это преобразование в символическом виде можно записать как

$$y(t) = A \{x(t)\}. \quad (4.93)$$

С математической точки зрения соответствие между функциями на входе и выходе системы является оператором. Оператор представляет собой совокупность математических и логических действий, в результате которых заданной функции приводится в соответствие некоторая другая функция.

Понятие оператора — это обобщение понятия функции.

Функцией, как известно, называется переменная величина, числовое значение которой определяется заданием числового значения другой переменной — аргумента.

Более общее понятие — функционал — переменная величина, числовое значение которой определяется заданием функции (например, площадь, ограниченная замкнутой кривой).

Понятие оператора является еще более широким, так как оператор приводит в соответствие каждой данной функции не число, а функцию.

Оператор системы — это ее полная исчерпывающая характеристика. Понятие оператора объединяет любые математические действия: все действия алгебры, дифференцирование, интегрирование, сдвиг во времени, решение дифференциальных, интегральных, интегродифференциальных, алгебраических и любых других функциональных уравнений. К понятию оператора относятся также любые логические действия. Задать оператор системы — это значит задать программу действий (операций), которые преобразуют входную функцию в выходную.

Линейные и нелинейные операторы и системы

Существуют различные типы операторов, применяемых к функциям. Одним из наиболее важных с практической точки зрения является класс так называемых *линейных операторов*.

Оператор называется линейным, если при любых числах n, c_1, c_2, \dots, c_n и при любых функциях $x_1(t), x_2(t), \dots, x_n(t)$

$$A \left\{ \sum_{k=1}^n c_k x_k(t) \right\} = \sum_{k=1}^n c_k A x_k(t), \quad (4.94)$$

т. е. результат действия этого оператора на любую линейную комбинацию данных функций является также линейной комбинацией от результатов его действия на каждую функцию в отдельности с теми же коэффициентами.

Динамическая система, которая описывается линейным оператором, называется линейной. Свойство линейных систем, выраженное формулой (4.94), известно под названием *принципа суперпозиции*.

Для того чтобы система была линейной, необходимо и достаточно выполнения двух условий:

- 1) сумме любых двух воздействий должна соответствовать сумма двух реакций;
- 2) при любом усилении входного воздействия без изменения его формы форма выходного сигнала также не должна изменяться.

Необходимость этих условий очевидна. Так как формула (4.94) справедлива для любого n и любых чисел c_1, c_2, \dots, c_n , то, полагая $n = 2$; $c_1 = c_2 = 1$, получаем

$$A \{x_1(t) + x_2(t)\} = Ax_1(t) + Ax_2(t). \quad (4.95)$$

Полагая $n = 1$ при произвольных c и $x(t)$, имеем

$$A \{cx(t)\} = cAx(t). \quad (4.96)$$

Для доказательства достаточности условий (4.95) и (4.96) заметим, что из этих условий можно получить зависимости

$$\begin{aligned} A \{c_1x_1(t) + c_2x_2(t)\} &= A \{c_1x_1(t)\} + A \{c_2x_2(t)\} = \\ &= c_1Ax_1(t) + c_2Ax_2(t), \end{aligned} \quad (4.97)$$

$$\begin{aligned} A \left\{ \sum_{k=1}^n c_k x_k(t) \right\} &= A \left\{ \sum_{k=1}^{n-1} c_k x_k(t) + c_n x_n(t) \right\} = \\ &= A \left\{ \sum_{k=1}^{n-1} c_k x_k(t) \right\} + A \{c_n x_n(t)\} = A \left\{ \sum_{k=1}^{n-1} c_k x_k(t) \right\} + \\ &\quad + c_n Ax_n(t). \end{aligned} \quad (4.98)$$

Формула (4.97) показывает, что из условий (4.95) и (4.96) следует справедливость принципа суперпозиции для двух слагаемых. Из формулы (4.98) видно, что принцип суперпозиции выполняется для n слагаемых, если он выполняется для $n - 1$ слагаемого.

Таким образом, принцип суперпозиции при любом числе n слагаемых является следствием условий (4.95) и (4.96), что и доказывает достаточность этих условий.

Принцип суперпозиции значительно облегчает исследование линейных систем по сравнению с исследованием нелинейных. Благодаря принципу суперпозиции теория линейных дифференциальных уравнений разработана в са-

мом общем виде для уравнений любого порядка, в то время как теории нелинейных дифференциальных уравнений по существу нет, и мы можем решать в аналитической форме только нелинейные дифференциальные уравнения частных видов невысокого порядка. Поэтому для решения всех математических вопросов, возникающих в приложениях, обращаются в первую очередь к линейным методам. При этом даже нелинейные системы стараются приближенно рассматривать как линейные. В результате появились различные методы линеаризации нелинейных систем, т. е. приближенной замены нелинейных систем равноценными линейными.

Из справедливости принципа суперпозиции для линейных систем при любом числе слагаемых n , любом выборе функций $x_k(t)$ и чисел c_k следует, что он применим не только к суммам, но и к интегралам. Другими словами, если входной сигнал системы представляет собой сумму бесконечно малых элементарных воздействий, то выходная переменная линейной системы есть сумма соответствующих бесконечно малых реакций на эти элементарные воздействия

$$A_t \left\{ \int_{\lambda_1}^{\lambda_2} c(\lambda) x(t, \lambda) d\lambda \right\} = \int_{\lambda_1}^{\lambda_2} c(\lambda) A_t x(t, \lambda) d\lambda. \quad (4.99)$$

Эта формула — принцип суперпозиции в интегральной форме.

Принцип суперпозиции дает возможность выразить реакцию линейной системы на любое возмущение через ее реакцию на определенный вид элементарных возмущений. Иными словами, любая линейная система полностью характеризуется ее реакцией на какой-нибудь стандартный тип возмущений.

Примерами линейных операторов могут служить:

1. Оператор дифференцирования

$$y(t) = \frac{dx(t)}{dt};$$

2. Оператор интегрирования

$$y(t) = \int_0^t x(\tau) d\tau;$$

3. Оператор умножения на некоторую функцию $\varphi(t)$

$$y(t) = \varphi(t) x(t);$$

4. Оператор интегрирования с заданным «весом»

$$y(t) = \int_0^t x(\tau) \varphi(\tau) d\tau$$

и т. д.

Рассмотренные операторы относятся к *линейным однородным*. Кроме них существуют еще *линейные неоднородные* операторы.

Линейным неоднородным называется оператор, состоящий из линейного однородного, к которому прибавлена некоторая определенная функция $\varphi(t)$. Примерами линейных неоднородных операторов могут служить

$$1. y(t) = \frac{dx(t)}{dt} + \varphi(t);$$

$$2. y(t) = \int_0^t x(\tau) \varphi(\tau) d\tau + \varphi_1(t);$$

$$3. y(t) = \varphi_1(t) x(t) + \varphi_2(t),$$

где $x(t)$ — функция, преобразуемая оператором, а $\varphi(t)$, $\varphi_2(t)$, $\varphi_3(t)$ — вполне определенные функции.

При математическом описании реальных объектов и процессов, например в теории автоматического регулирования и управления, применяется условная форма записи операторов, сходная с алгебраической символикой. Такая символика упрощает преобразования. Формулы становятся более простыми и удобными.

Например, оператор дифференцирования обозначается в виде

$$p = \frac{d}{dt}.$$

При этом выражение $y(t) = px(t)$ равносильно выражению

$$y(t) = \frac{dx(t)}{dt}.$$

Двойное дифференцирование записывается в виде множителя p^2

$$p^2 x(t) = \frac{d^2 x(t)}{dt^2}$$

и т. д.

При таких обозначениях значительно упрощается запись дифференциальных уравнений. Предположим, что функционирование некоторой динамической системы описывается

линейным дифференциальным уравнением с постоянными коэффициентами. Уравнение описывает работу системы относительно воздействия $x(t)$ и реакции $y(t)$

$$\begin{aligned} a_n \frac{d^n y(t)}{dt^n} + a_{n-1} \frac{d^{n-1} y(t)}{dt^{n-1}} + \dots + a_1 \frac{dy(t)}{dt} + a_0 y(t) = \\ = b_m \frac{d^m x(t)}{dt^m} + b_{m-1} \frac{d^{m-1} x(t)}{dt^{m-1}} + \dots + b_1 \frac{dx(t)}{dt} + b_0 x(t) \end{aligned} \quad (4.100)$$

В символической форме это же уравнение записывается как

$$\begin{aligned} (a_n p^n + a_{n-1} p^{n-1} + \dots + a_1 p + a_0) y(t) = \\ = (b_m p^m + b_{m-1} p^{m-1} + \dots + b_1 p + b_0) x(t), \end{aligned} \quad (4.101)$$

где $p = \frac{d}{dt}$ — оператор дифференцирования.

Обычно полиномы в левой и правой частях уравнений записывают для краткости в виде

$$\begin{aligned} A_n(p) &= a_n p^n + a_{n-1} p^{n-1} + \dots + a_1 p + a_0, \\ B_m(p) &= b_m p^m + b_{m-1} p^{m-1} + \dots + b_1 p + b_0. \end{aligned}$$

Тогда исходное уравнение становится еще более компактным

$$A_n(p) y(t) = B_m(p) x(t). \quad (4.102)$$

Если формально решить уравнение (4.102) относительно $y(t)$, то оператор решения линейного дифференциального уравнения в «явном виде» можно представить как

$$y(t) = \frac{B_m(p)}{A_n(p)} x(t).$$

Аналогично можно описать дифференциальное уравнение с переменными коэффициентами. Обычная математическая запись этого уравнения имеет вид

$$\begin{aligned} a_n(t) \frac{d^n y(t)}{dt^n} + a_{n-1}(t) \frac{d^{n-1} y(t)}{dt^{n-1}} + \dots + a_1(t) \frac{dy(t)}{dt} + \\ + a_0(t) y(t) = b_m(t) \frac{d^m x(t)}{dt^m} + b_{m-1}(t) \frac{d^{m-1} x(t)}{dt^{m-1}} + \\ + \dots + b_1(t) \frac{dx(t)}{dt} + b_0(t) x(t). \end{aligned} \quad (4.103)$$

Введем обозначения для многочленов по p , коэффициенты которых зависят от t

$$A_n(p, t) = a_n(t) p^n + a_{n-1}(t) p^{n-1} + \dots + a_1(t) p + a_0(t),$$

$$B_m(p, t) = b_m(t) p^m + b_{m-1}(t) p^{m-1} + \dots + b_1(t) p + b_0(t).$$

Тогда оператор дифференциального уравнения принимает вид

$$y(t) = \frac{B_m(p, t)}{A_n(p, t)} x(t). \quad (4.104)$$

Как уже указывалось, динамические системы, описываемые линейными операторами, называются линейными.

Кроме линейных операторов и систем рассматриваются системы и операторы нелинейные.

Нелинейным называется любой оператор, для которого принцип суперпозиции справедлив или справедлив только при некоторых вполне определенных функциях $x_1(t)$, $x_2(t)$, ..., $x_n(t)$ и числах c_1, c_2, \dots, c_n в (4.94).

В качестве примеров нелинейных операторов можно привести выражения

$$y(t) = x^2(t),$$

$$y(t) = \int_0^t x^2 \tau d\tau,$$

$$y(t) = \sin x(t)$$

или решение линейного дифференциального уравнения

$$y'(t) + \alpha \cos y(t) = x(t)$$

Уравнения, описывающие поведение линейной системы, всегда линейны. И наоборот, если среди уравнений, описывающих поведение системы, хотя бы одно — нелинейное, то система нелинейна.

Стационарные и нестационарные системы

Стационарной называется система, реакция которой на любой данный тип воздействия зависит только от интервала времени между данным моментом времени и моментом начала действия возмущения.

Если $x(t)$ произвольная функция, равная нулю при $t < t_0$, то согласно определению реакция системы на воздействие $x(t)$ будет зависеть только от интервала $t - t_0$.

Это будет некоторая функция $y(t - t_0)$ (рис. 4.31). Если это же воздействие на стационарную систему будет начинаться с момента $t_1 = t_0 + \tau$, то оно описывается функцией $x(t - \tau)$, а реакция системы $y(t - t_1) = y(t - t_0 - \tau)$.

Нестационарные системы характерны тем, что при сдвиге входного возмущения во времени без изменения формы их выходные переменные не только сдвигаются во времени, но и изменяют форму. И стационарные, и нестационарные системы могут быть как линейными, так и нелинейными.

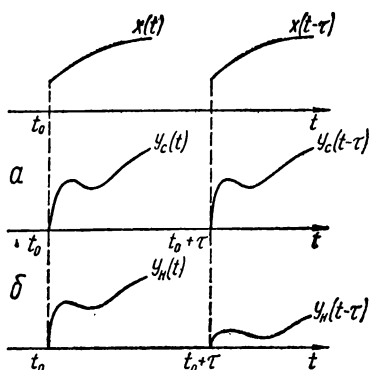


Рис. 4.31. Реакции стационарной (а) и нестационарной (б) системы.

Линейные преобразования случайных функций

На вход линейной системы, описываемой оператором L , воздействует случайная функция $X(t)$. Ее характеристики — математическое ожидание $m_x(t)$ и корреляционная функция $K_x(t, t')$ — известны. Реакцией системы будет также случайная функция

$$Y(t) = \{X(t)\}. \quad (4.105)$$

Необходимо по характеристикам случайной функции $X(t)$ на входе линейной системы определить характеристики случайной функции $Y(t)$ на выходе. Можно ограничиться решением этой задачи только для линейного однородного оператора L . В самом деле, пусть оператор L неоднороден и имеет вид

$$L\{X(t)\} = L_0\{X(t)\} + \varphi(t), \quad (4.106)$$

где L_0 — линейный однородный оператор; $\varphi(t)$ — определенная неслучайная функция. При этом

$$m_y(t) = M[L_0\{X(t)\}] + \varphi(t), \quad (4.107)$$

т. е. функция $\varphi(t)$ прибавляется к математическому ожиданию случайной функции на выходе линейной системы. Корреляционная функция, как было доказано, не меняется от прибавления к случайной функции неслучайного

слагаемого. Поэтому обычно понятие линейного оператора связано только с линейными однородными операторами.

Рассмотрим некоторые частные задачи определения характеристик функции на выходе системы.

Интегрирование случайной функции. Случайная функция $X(t)$ с математическим ожиданием $m_x(t)$ и корреляционной функцией $K_x(t, t')$ преобразуются при помощи линейного оператора интегрирования к виду

$$Y(t) = \int_0^t X(\tau) d\tau. \quad (4.108)$$

Найдем характеристики случайной функции $Y(t)$: $m_y(t)$ и $K_y(t, t')$.

Запишем (4.108) как предел суммы

$$Y(t) = \int_0^t X(\tau) d\tau = \lim_{\Delta\tau \rightarrow 0} \sum_i X(\tau_i) \Delta\tau. \quad (4.109)$$

Применим к выражению (4.109) операцию математического ожидания

$$\begin{aligned} m_y(t) &= M[Y(t)] = \lim_{\Delta\tau \rightarrow 0} \sum_i M[X(\tau_i)] \Delta\tau = \\ &= \lim_{\Delta\tau \rightarrow 0} \sum_i m_x(\tau_i) \Delta\tau = \int_0^t m_x(\tau) d\tau. \end{aligned} \quad (4.110)$$

Математическое ожидание интеграла от случайной функции равно интегралу от ее математического ожидания.

Будем искать корреляционную функцию $K_y(t, t')$. По определению корреляционной функции

$$K_y(t, t') = M[\dot{Y}(t) \dot{Y}(t')],$$

где

$$\dot{Y}(t) = \int_0^t \dot{X}(\tau) d\tau; \quad \dot{Y}(t') = \int_0^{t'} \dot{X}(\tau') d\tau'. \quad (4.111)$$

Перемножив выражения (4.111), получим

$$\dot{Y}(t) \dot{Y}(t') = \int_0^t \dot{X}(\tau) d\tau \int_0^{t'} \dot{X}(\tau') d\tau'. \quad (4.112)$$

Это можно переписать в виде двойного интеграла

$$\int_0^t \int_0^{t'} \dot{X}(\tau) \dot{X}(\tau') d\tau d\tau'. \quad (4.113)$$

Применив к (4.113) операцию математического ожидания и меняя ее в правой части с операцией интегрирования, получим

$$K_y(t, t') = M[\dot{Y}(t) \dot{Y}(t')] = \int_0^t \int_0^{t'} M[\dot{X}(\tau) \dot{X}(\tau')] d\tau d\tau',$$

что дает окончательно

$$K_y(t, t') = \int_0^t \int_0^{t'} K_x(\tau, \tau') d\tau d\tau'. \quad (4.114)$$

Следовательно, чтобы получить корреляционную функцию интеграла от случайной функции, необходимо дважды проинтегрировать корреляционную функцию исходной случайной функции. Первый раз — по одному аргументу, затем — по другому.

Дифференцирование случайной функции. Случайная функция с математическим ожиданием $X(t)$ и корреляционной функцией $K_x(t, t')$ преобразуется при помощи линейного оператора дифференцирования к виду

$$Y(t) = \frac{dX(t)}{dt}. \quad (4.115)$$

Найдем характеристики случайной функции $Y(t)$; $m_y(t)$ и $K_y(t, t')$.

Запишем (4.115) как предел отношения

$$Y(t) = \lim_{\Delta t \rightarrow 0} \frac{X(t + \Delta t) - X(t)}{\Delta t}. \quad (4.116)$$

Применяя операцию математического ожидания, получим

$$m_y(t) = M[Y(t)] = \lim_{\Delta t \rightarrow 0} \frac{m_x(t + \Delta t) - m_x(t)}{\Delta t} = \frac{dm_x(t)}{dt},$$

или

$$m_y(t) = \frac{dm_x(t)}{dt}. \quad (4.117)$$

Таким образом, математическое ожидание производной от случайной функции равно производной от ее математического ожидания. Операцию дифференцирования, как и операцию интегрирования, можно менять местами с операцией математического ожидания.

Будем искать корреляционную функцию $K_y(t, t')$. По определению

$$K_y(t, t') = M[\dot{Y}(t) \dot{Y}(t')].$$

Подставим выражения для $\dot{Y}(t)$ и $\dot{Y}(t')$

$$K_y(t, t') = M \left[\frac{d\dot{X}(t)}{dt} \frac{d\dot{X}(t')}{dt'} \right].$$

Выражение в квадратных скобках представим в виде второй смешанной частной производной

$$\frac{d\dot{X}(t)}{dt} \cdot \frac{d\dot{X}(t')}{dt'} = \frac{\partial^2 \dot{X}(t) \dot{X}(t')}{\partial t \partial t'}. \quad (4.118)$$

Так как математическое ожидание производной равно производной математического ожидания, получим

$$\begin{aligned} K_y(t, t') &= M \left[\frac{\partial^2 \dot{X}(t) \dot{X}(t')}{\partial t \partial t'} \right] = \frac{\partial^2}{\partial t \partial t'} M [\dot{X}(t) \dot{X}(t')] = \\ &= \frac{\partial^2}{\partial t \partial t'} K_x(t, t'). \end{aligned} \quad (4.119)$$

Следовательно, чтобы найти корреляционную функцию производной, необходимо дважды продифференцировать корреляционную функцию исходной случайной функции: сначала по одному аргументу, затем по другому.

Применяя формулы (4.117) и (4.119) многократно, можно получить выражения для математического ожидания $m_{y_s}(t)$ и корреляционной функции $K_{y_s}(t, t')$ производной порядка S случайной функции $X(t)$

$$m_{y_s}(t) = m_x^{(s)}(t),$$

$$K_{y_s}(t, t') = \frac{\partial^{2s} K_x(t, t')}{\partial t^s \partial t'^s}.$$

Если сравнить правила, по которым мы находили математические ожидания и корреляционные функции для операторов дифференцирования и интегрирования, то очевидно, что эти правила совершенно аналогичны. Чтобы найти математическое ожидание преобразованной случайной функции, тот же оператор нужно применить к математическому ожиданию исходной случайной функции; чтобы найти корреляционную функцию преобразованной случайной функции, тот же линейный оператор применяется дважды к корреляционной функции исходной случайной функции. В первом — частном случае это было двойное интегрирование, во втором — двойное дифференцирование. Это правило общее для всех линейных операторов.

Если случайная функция $X(t)$ с математическим ожиданием $m_x(t)$ и корреляционной функцией $K_x(t, t')$ преобразуется линейным однородным оператором L в случайную функцию

$$Y(t) = L\{X(t)\},$$

то для нахождения математического ожидания случайной функции $Y(t)$ нужно применить тот же оператор к математическому ожиданию случайной функции $X(t)$

$$m_y(t) = L\{m_x(t)\}, \quad (4.120)$$

корреляционную функцию $Y(t)$ определяют, применяя дважды тот же оператор к корреляционной функции случайной функции $X(t)$, сначала по одному аргументу, затем по другому

$$K_y(t, t') = L^{(t)}L^{(t')}\{K_x(t, t')\}. \quad (4.121)$$

Сложение случайных функций

Если на вход динамической системы поступает не одна случайная функция $X(t)$, а две или больше, то возникает задача сложения случайных функций. Точнее эту задачу можно определить как задачу определения характеристик суммы по характеристикам слагаемых.

Если складываемые случайные функции не коррелированы между собой, задача решается просто. В случае зависимости функций-слагаемых для решения задачи необходимо знать еще одну характеристику — взаимную корреляционную функцию или иначе — корреляционную функцию связи.

Взаимной корреляционной функцией двух случайных функций $X(t)$ и $Y(t)$ называется неслучайная функция двух аргументов t и t' , которая при каждой паре значений t, t' равна корреляционному моменту соответствующих сечений случайной функции $X(t)$ и случайной функции $Y(t)$

$$R_{xy}(t, t') = M[\dot{X}(t)\dot{Y}(t')]. \quad (4.121, a)$$

Взаимная корреляционная функция, как и автокорреляционная функция, не изменяется при прибавлении к случайным функциям любых неслучайных слагаемых, а значит, и при центрировании случайных функций.

Из определения следует важное свойство взаимной корреляционной функции

$$R_{xy}(t, t') = R_{yx}(t', t). \quad (4.122)$$

Иногда удобнее вместо взаимной корреляционной функции $R_{xy}(t, t')$ пользоваться нормированной взаимной корреляционной функцией

$$r_{xy}(t, t') = \frac{R_{xy}(t, t')}{\sigma_x(t) \sigma_y(t')} . \quad (4.123)$$

Случайные функции $X(t)$ и $Y(t)$ называются некоррелированными, если взаимная корреляционная функция равна нулю при всех значениях t, t' .

При решении практических задач о некоррелированности случайных функций судят в большинстве случаев не по равенству взаимной корреляционной функции нулю. Наоборот, взаимную корреляционную функцию полагают равной нулю, если на основании соображений о физических свойствах процессов их можно считать независимыми.

Если известны математические ожидания и корреляционные функции двух случайных функций $X(t)$ и $Y(t)$, а также их взаимная корреляционная функция, то можно определить характеристики суммы этих двух случайных функций

$$Z(t) = X(t) + Y(t). \quad (4.124)$$

В соответствии с теоремой сложения математических ожиданий

$$m_z(t) = m_x(t) + m_y(t), \quad (4.125)$$

т. е. математическое ожидание суммы двух случайных функций равно сумме их математических ожиданий.

Найдем корреляционную функцию $K_z(t, t')$. По определению корреляционной функции

$$K_z(t, t') = M[\dot{Z}(t) \dot{Z}(t')]. \quad (4.126)$$

Учитывая (4.124), после подстановки получим

$$\begin{aligned} K_z(t, t') &= M[(\dot{X}(t) + \dot{Y}(t))(\dot{X}(t') + \dot{Y}(t'))] = \\ &= M[\dot{X}(t) \dot{X}(t')] + M[\dot{Y}(t) \dot{Y}(t')] + \\ &\quad + M[\dot{X}(t) \dot{Y}(t')] + M[\dot{Y}(t) \dot{X}(t')], \end{aligned}$$

или

$$K_z(t, t') = K_x(t, t') + K_y(t, t') + R_{xy}(t, t') + R_{yx}(t, t'). \quad (4.127)$$

Если случайные функции $X(t)$ и $Y(t)$ не коррелированы,

$$R_{xy}(t, t') \equiv 0; \quad R_{yx}(t', t) \equiv 0,$$

выражение (4.126) принимает вид

$$K_z(t, t') = K_x(t, t') + K_u(t, t'). \quad (4.128)$$

Корреляционная функция суммы двух некоррелированных случайных функций равна сумме их корреляционных функций.

Формулы (4.125) и (4.128) можно обобщить при произвольном числе слагаемых. Если случайная функция $X(t)$ представляет собой сумму n случайных функций

$$X(t) = \sum_{i=1}^n X_i(t), \quad (4.129)$$

то ее математическое ожидание

$$m_x(t) = \sum_{i=1}^n m_{x_i}(t). \quad (4.130)$$

Корреляционная функция суммы n случайных функций равна

$$K_x(t, t') = \sum_{i=1}^n K_{x_i}(t, t') + \sum_{i \neq j} R_{x_i x_j}(t, t'). \quad (4.131)$$

Если же все случайные функции $X_i(t)$ не коррелированы, выражение (4.131) обращается в

$$K_x(t, t') = \sum_{i=1}^n K_{x_i}(t, t'). \quad (4.132)$$

Формула (4.132) представляет собой теорему сложения корреляционных функций:

корреляционная функция суммы взаимно некоррелированных случайных функций равна сумме корреляционных функций слагаемых.

Иногда приходится определять характеристики суммы случайных функций и случайных величин. Пусть, например, к случайной функции $X(t)$ с математическим ожиданием $m_x(t)$ и корреляционной функцией $K_x(t, t')$ прибавляется случайная величина Y с математическим ожиданием m_y и дисперсией D_y . Предположим, что случайная функция $X(t)$ и случайная величина Y не коррелированы

$$M[\dot{X}(t) \dot{Y}] = 0.$$

Математическое ожидание суммы $X(t)$ и Y

$$Z(t) = X(t) + Y, \quad (4.133)$$

очевидно, равно:

$$m_z(t) = m_x(t) + m_y. \quad (4.134)$$

Случайную величину Y будем рассматривать как частный случай случайной функции, не зависящей от времени. Ее корреляционная функция

$$K_y(t, t') = M[\dot{Y}(t) \dot{Y}(t')] = M[\dot{Y}^2] = D_y. \quad (4.135)$$

Найдем корреляционную функцию суммы. Пользуясь теоремой сложения корреляционных функций, получаем

$$K_z(t, t') = K_x(t, t') + D_y. \quad (4.136)$$

Если к случайной функции прибавить некоррелированную с нею случайную величину, то к корреляционной функции $K_x(t, t')$ необходимо прибавить постоянное слагаемое, равное дисперсии этой случайной величины.

Контрольные вопросы и задания

1. Дайте определение оператора системы.
2. Сформулируйте принцип суперпозиции для линейных операторов.
3. Приведите примеры линейных и нелинейных операторов.
4. Дайте определения и сравнительную характеристику стационарных и нестационарных систем.
5. Как изменяются характеристики случайной функции при ее интегрировании? При дифференцировании?
6. Определите взаимную корреляционную функцию случайных функций.
7. Как определяются характеристики суммы двух случайных функций?

§ 8. КОМПЛЕКСНЫЕ СЛУЧАЙНЫЕ ФУНКЦИИ

Часто при решении практических задач методами теории случайных функций эти функции и их характеристики удобно записывать в комплексной форме. Определим понятия комплексной случайной величины и комплексной случайной функции.

Комплексной случайной величиной будем называть случайную величину Z вида

$$Z = X + jY, \quad (4.137)$$

где X и Y — действительные случайные величины; $j = \sqrt{-1}$.

Обобщим основные понятия математического ожидания, дисперсии и корреляционного момента для комплексных случайных величин. Эти обобщения необходимо провести

так, чтобы в частном случае, при $Y = 0$ и действительном Z , они были обычными определениями характеристик действительных случайных величин.

Математическим ожиданием комплексной случайной величины $Z = X + jY$ будем называть комплексное число

$$m_z = m_x + jm_y. \quad (4.138)$$

Дисперсией комплексной случайной величины будем называть математическое ожидание квадрата модуля соответствующей центрированной величины

$$D_z = M[\dot{Z}^2], \quad (4.139)$$

где

$$\dot{Z} = Z - m_z.$$

Дисперсию комплексной случайной величины можно выразить через дисперсии ее действительной и мнимой частей. Запишем

$$\dot{Z} = Z - m_z = X + jY - m_x - jm_y = \dot{X} + j\dot{Y},$$

откуда

$$D_z = M[\dot{Z}^2] = M[\dot{X}^2 + \dot{Y}^2] = M[\dot{X}^2] + M[\dot{Y}^2],$$

или

$$D_z = D_x + D_y. \quad (4.140)$$

Дисперсия комплексной случайной величины равна сумме дисперсий ее действительной и мнимой частей.

Из определения ясно, что дисперсия комплексной случайной величины всегда действительна и положительна. Она обращается в нуль только в случае, когда величина Z не случайна.

Определим корреляционный момент двух комплексных случайных величин Z_1 и Z_2

$$Z_1 = X_1 + jY_1; \quad Z_2 = X_2 + jY_2. \quad (4.141)$$

Это определение должно быть сформулировано так, чтобы при $Z_1 = Z_2 = Z$ корреляционный момент превращался в дисперсию случайной величины Z .

Корреляционным моментом двух комплексных случайных величин Z_1 и Z_2 называется математическое ожидание произведения одной центрированной случайной величины на комплексную сопряженную другой

$$K_{z_1 z_2} = M[\dot{Z}_1 \overline{\dot{Z}_2}]. \quad (4.142)$$

При этом для $Z_1 = Z_2 = Z$ корреляционный момент обращается в дисперсию случайной величины Z

$$K_{zz} = M[(\dot{X} + j\dot{Y})(\dot{X} - j\dot{Y})] = M[\dot{X}^2] + M[\dot{Y}^2] = D_z. \quad (4.143)$$

Корреляционный момент двух комплексных случайных величин можно выразить через корреляционные моменты их действительных и мнимых частей. Получаем

$$\begin{aligned} K_{z_1 z_2} &= M[\overline{\dot{Z}_1} \dot{Z}_2] = M[(\dot{X}_1 + j\dot{Y}_1)(\dot{X}_2 - j\dot{Y}_2)] = \\ &= K_{x_1 x_2} + K_{y_1 y_2} + j(K_{y_1 x_2} - K_{x_1 y_2}), \end{aligned} \quad (4.144)$$

где $K_{x_1 x_2}$, $K_{y_1 y_2}$, $K_{y_1 x_2}$, $K_{x_1 y_2}$ — корреляционные моменты величин (X_1, X_2) , (Y_1, Y_2) , (Y_1, X_2) , (X_1, Y_2) .

Если эти величины некоррелированы между собой, корреляционный момент величин Z_1, Z_2 равен нулю.

Теперь определим *комплексную случайную функцию* и ее характеристики.

Комплексной случайной будем называть функцию вида

$$Z(t) = X(t) + jY(t). \quad (4.145)$$

В формуле (4.145) $X(t)$, $Y(t)$ — действительные случайные функции.

Математическое ожидание комплексной случайной функции равно:

$$m_z(t) = m_x(t) + jm_y(t). \quad (4.146)$$

Дисперсией комплексной случайной функции $Z(t)$ называется математическое ожидание квадрата модуля соответствующей центрированной функции

$$D_z(t) = M[|\tilde{Z}(t)|^2], \quad (4.147)$$

где

$$\tilde{Z}(t) = Z(t) - m_z(t) = \dot{X}(t) + j\dot{Y}(t). \quad (4.148)$$

Как видно из определения, дисперсия комплексной случайной функции действительна и неотрицательна. Дисперсия комплексной случайной функции равна сумме дисперсий ее действительной и мнимой частей

$$D_z(t) = D_x(t) + D_y(t). \quad (4.149)$$

Корреляционную функцию комплексной случайной функции можно определить как корреляционный момент ее сечений t и t'

$$K_z(t, t') = M[\tilde{Z}(t) \tilde{Z}^*(t')]. \quad (4.150)$$

Для $t = t'$ из (4.150) получаем дисперсию

$$K_z(t, t) = D_z(t). \quad (4.151)$$

Корреляционную функцию комплексной случайной функции можно выразить через характеристики ее действительной и мнимой частей.

Если в формулу (4.144) вместо случайных величин Z_1 и Z_2 подставить сечения случайной функции $Z(t)$ и $Z(t')$, получим:

$$K_z(t, t') = K_x(t, t') + K_y(t, t') + j \{R_{xy}(t', t) - R_{xy}(t, t')\}. \quad (4.152)$$

В выражении (4.152) $R_{xy}(t, t')$ — взаимная корреляционная функция случайных функций $X(t)$ и $Y(t)$. Если действительная и мнимая части некоррелированы ($R_{xy}(t, t') \equiv 0$), (4.152) превращается в

$$K_z(t, t') = K_x(t, t') + K_y(t, t'). \quad (4.153)$$

Контрольные вопросы и задания

1. Дайте определение комплексной случайной величины и ее числовых характеристик.
2. Определите комплексную случайную функцию.
3. Приведите выражения для характеристик комплексной случайной функции.

§ 9. КАНОНИЧЕСКИЕ РАЗЛОЖЕНИЯ СЛУЧАЙНЫХ ФУНКЦИЙ

Определение канонического представления случайных функций

Случайная функция — это очень сложный математический объект. В общем случае ее можно рассмотреть как систему бесконечного множества случайных величин.

При решении практических задач характеристики случайных функций определяются на основании опытных данных. Чаще всего они задаются не аналитически, а таблично. Поэтому находить характеристики преобразованных функций весьма затруднительно. Даже при простейшей форме оператора преобразования, например

$$K_y(t, t') = \int_0^t \int_0^{t'} K_x(\tau, \tau') d\tau d\tau', \quad (4.154)$$

интеграл приходится находить численными методами, определяя его как функцию обоих пределов. Это трудоемкая и громоздкая задача. Если же подынтегральную функцию аппроксимировать аналитически, то зачастую интеграл через известные функции не выражается. Таким образом, возникают существенные трудности даже при простейших преобразованиях. Тем более сложно определять характеристики преобразованных случайных функций, когда функционирование динамической системы описывается дифференциальными уравнениями, решение которых не выражается в явной форме. При решении подобных задач требуется интегрировать дифференциальные уравнения с частными производными.

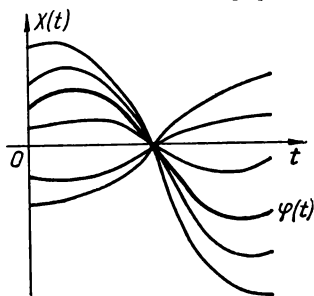


Рис. 4.32 Реализации элементарной случайной функции.

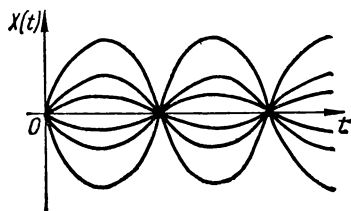


Рис. 4.33. Элементарная случайная функция $V \sin \varphi$.

Поэтому естественно попытаться выразить случайную функцию через более простые случайные объекты, например, через обычные случайные величины.

Практически простейшей формой выражения случайной функции через случайные величины является ее представление в виде линейной комбинации некоррелированных случайных величин с нулевыми математическими ожиданиями. Таким образом, любую случайную функцию можно представить в виде

$$X(t) = m_x(t) + \sum_{i=1}^n V_i \varphi_i(t), \quad (4.155)$$

где V_i — некоррелированные случайные величины с нулевыми математическими ожиданиями; $\varphi_i(t)$ — неслучайные функции. Каждое произведение $V_i \varphi_i$ называется *элементарной случайной функцией*. Это наиболее простой тип случайной функции. Все возможные реализации элементарной функции можно получить из графика неслучайной функции $\varphi(t)$ простым изменением масштаба по оси ординат

(рис. 4.32). На рис. 4.33, 4.34 приведены примеры элементарных случайных функций. Основное свойство элементарной случайной функции состоит в том, что вся ее случайность определяется коэффициентом V , а зависимость от времени — неслучайной функцией $\varphi(t)$.

Найдем характеристики элементарной случайной функции. Математическое ожидание

$$m_x(t) = M[V\varphi(t)] = m_v \varphi(t),$$

где m_v — математическое ожидание случайной величины V . При $m_v = 0$ математическое ожидание элементарной случайной функции тождественно равно нулю

$$m_x(t) \equiv 0.$$

Вместо случайной функции $\dot{X}(t)$ можно рассматривать соответствующую центрированную $\ddot{X}(t)$ с нулевым математическим ожиданием. В дальнейшем будем рассматривать центрированные элементарные случайные функции, для которых $m_v = 0$; $V = \dot{V}$; $m_x(t) \equiv 0$.

Корреляционная функция элементарной случайной функции определяется как

$$K_x(t, t') = M[\dot{X}(t) \dot{X}(t')] = \varphi(t) \varphi(t') M[V^2] = \varphi(t) \varphi(t') D,$$

где D — дисперсия случайной величины V .

Рассмотрим линейные преобразования элементарной случайной функции. При дифференцировании случайная величина V выйдет за знак производной, так как она не зависит от t

$$\frac{dX(t)}{dt} = V \frac{d\varphi(t)}{dt}.$$

Точно так же

$$\int_0^t X(\tau) d\tau = V \int_0^t \varphi(\tau) d\tau.$$

Обобщая, можно сказать, что если элементарная случайная функция преобразуется линейным оператором L , то случайный множитель V выносится за знак оператора

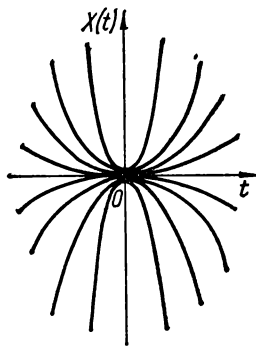


Рис. 4.34. Элементарная случайная функция Vt^2 .

ра, а неслучайная функция $\varphi(t)$ преобразуется тем же оператором L

$$L\{X(t)\} = VL\{\varphi(t)\}. \quad (4.156)$$

Если на вход динамической системы поступает случайная функция общего вида, то эту случайную функцию можно представить суммой элементарных случайных функций, после чего задача преобразования существенно упрощается. Эта идея лежит в основе метода канонических разложений, разработанного В. С. Пугачевым.

Представление случайной функции в виде

$$X(t) = m_x(t) + \sum_{i=1}^n V_i \varphi_i(t) \quad (4.157)$$

известно под названием разложения случайной функции. Случайные величины V_i ($i = 1, 2, \dots, n$) называются коэффициентами разложения, а функции $\varphi_i(t)$ — координатными функциями разложения. В общем случае разложение случайной функции представляется в виде бесконечного ряда. В частных случаях это может быть конечная сумма.

Определим корреляционную функцию и дисперсию случайной функции $X(t)$, заданной в виде (4.157). По определению

$$K_x(t, t') = M[\dot{X}(t) \dot{X}(t')]. \quad (4.158)$$

В выражении (4.158)

$$\dot{X}(t) = \sum_{i=1}^n V_i \dot{\varphi}_i(t); \quad (4.159)$$

$$\dot{X}(t') = \sum_{j=1}^n V_j \dot{\varphi}_j(t'). \quad (4.160)$$

Подставляя (4.159) и (4.160) в (4.158), получим

$$\begin{aligned} K_x(t, t') &= M \left[\sum_{i=1}^n \sum_{j=1}^n V_i V_j \dot{\varphi}_i(t) \dot{\varphi}_j(t') \right] = \\ &= \sum_{i=1}^n \sum_{j=1}^n M[V_i V_j] \dot{\varphi}_i(t) \dot{\varphi}_j(t'). \end{aligned} \quad (4.161)$$

При $i = j$ имеем

$$M[V_i V_i] = M[V_i^2] = K_{ii} = D_i, \quad (4.162)$$

где D_i — дисперсия случайной величины V_i . При $i \neq j$

$$M[V_i V_j] = K_{ij}, \quad (4.163)$$

где K_{ij} — корреляционный момент случайных величин V_i , V_j . Если учесть (4.162) и (4.163), то формулу (4.161) можно переписать в виде

$$K_x(t, t') = \sum_{i=1}^n \varphi_i(t) \varphi_i(t') D_i + \sum_{i \neq j} \varphi_i(t) \varphi_j(t') K_{ij}. \quad (4.164)$$

Если $t' = t$, то из (4.164) получаем выражение для дисперсии случайной функции $X(t)$, представленной разложением (4.157)

$$D_x(t) = \sum_{i=1}^n [\varphi_i(t)]^2 D_i + \sum_{i \neq j} \varphi_i(t) \varphi_j(t) K_{ij}. \quad (4.165)$$

Если все коэффициенты V_i разложения (4.157) не коррелированы, т. е. для $i \neq j$ $K_{ij} = 0$, то выражения (4.164), (4.165) существенно упрощаются. Само разложение (4.157) в этом случае называется *каноническим разложением* случайной функции $X(t)$ по координатным функциям $\varphi_i(t)$.

Из выражения (4.164), полагая $K_{ij} = 0$ при $i \neq j$, получим каноническое разложение корреляционной функции

$$K_x(t, t') = \sum_{i=1}^n \varphi_i(t) \varphi_i(t') D_i. \quad (4.166)$$

Если положить $t' = t$, получим дисперсию случайной функции в виде

$$D_x(t) = \sum_{i=1}^n [\varphi_i(t)]^2 D_i. \quad (4.167)$$

По известному каноническому разложению случайной функции можно определить каноническое разложение ее корреляционной функции. Обратное утверждение также справедливо. Если каноническое разложение корреляционной функции задано в виде (4.166), то для случайной функции $X(t)$ справедливо каноническое разложение вида (4.157) по координатным функциям $\varphi_i(t)$ с коэффициентами V_i . Доказательство этого утверждения можно найти в специальной литературе.

Каноническое разложение комплексной случайной функции

Элементарной комплексной случайной функцией будем называть функцию вида

$$X(t) = V \varphi(t), \quad (4.168)$$

где случайная величина V и функция $\varphi(t)$ — комплексны.

Корреляционная функция комплексной случайной функции определяется как

$$K_x(t, t') = M[V\varphi(t) \overline{V\varphi(t')}].$$

Учитывая, что

$$\overline{V\varphi(t)} = \overline{V\varphi(t')},$$

получаем

$$K_x(t, t') = \varphi(t) \overline{\varphi(t')} M[|V|^2].$$

По определению

$$M[|V|^2] = D.$$

Тогда

$$K_x(t, t') = \varphi(t) \overline{\varphi(t')} D. \quad (4.169)$$

Каноническим разложением комплексной случайной функции называется ее представление в виде

$$X(t) = m_x(t) + \sum_{i=1}^n V_i \varphi_i(t), \quad (4.170)$$

где $m_x(t)$, $\varphi_i(t)$ — комплексные неслучайные функции, а V_i — некоррелированные комплексные случайные величины с нулевыми математическими ожиданиями.

Каноническое разложение корреляционной функции комплексной случайной функции выражается в виде

$$K_x(t, t') = \sum_{i=1}^n \varphi_i(t) \overline{\varphi_i(t')} D_i, \quad (4.171)$$

где D_i — дисперсия величины V_i , равная

$$D_i = M[|V_i|^2]. \quad (4.172)$$

При $t = t'$ из (4.171) получаем выражение для дисперсии комплексной случайной функции, заданной каноническим разложением

$$D_x(t) = \sum_{i=1}^n |\varphi_i(t)|^2 D_i. \quad (4.173)$$

Линейные преобразования случайных функций, заданных каноническим разложением

Пусть некоторая динамическая система, описываемая линейным оператором L , преобразует случайную функцию $X(t)$ (воздействие) в случайную функцию $Y(t)$ (реакцию)

$$Y(t) = L\{X(t)\}. \quad (4.174)$$

Задана случайная функция $X(t)$ каноническим разложением

$$X(t) = m_x(t) + \sum_{i=1}^n V_i \varphi_i(t). \quad (4.175)$$

Реакция системы на $X(t)$

$$Y(t) = L\{X(t)\} = L\{m_x(t)\} + \sum_{i=1}^n V_i L\{\varphi_i(t)\}. \quad (4.176)$$

Формула (4.176) является каноническим разложением случайной функции $Y(t)$ с математическим ожиданием

$$m_y(t) = L\{m_x(t)\} \quad (4.177)$$

и координатными функциями

$$\psi_i(t) = L\{\varphi_i(t)\}. \quad (4.178)$$

При линейном преобразовании канонического разложения случайной функции $X(t)$ получаем каноническое разложение случайной функции $Y(t)$. При этом математическое ожидание и координатные функции подвергаются тому же линейному преобразованию.

Если реакция на выходе линейной системы получается в виде канонического разложения

$$Y(t) = m_y(t) + \sum_{i=1}^n V_i \psi_i(t), \quad (4.179)$$

то легко можно найти ее корреляционную функцию и дисперсию

$$K_y(t, t') = \sum_{i=1}^n \psi_i(t) \psi_i(t') D_i; \quad (4.180)$$

$$D_y(t) = \sum_{i=1}^n [\psi_i(t)^2] D_i. \quad (4.181)$$

Определение координатных функций. Представим центрированную случайную функцию в виде суммы

$$\hat{X}(t) = \sum_{i=1}^n V_i \varphi_i(t), \quad (4.182)$$

где $\varphi_i(t)$ — неизвестные координатные функции, которые нам необходимо определить. Умножим (4.182) на V_j и применим операцию математического ожидания

$$M[\hat{X}(t) V_j] = \sum_{i=1}^n M[V_i V_j] \varphi_i(t). \quad (4.183)$$

Все V_i — произвольные некоррелированные случайные величины с нулевыми математическими ожиданиями и дисперсиями D_i . Поэтому в правой части уравнения (4.183) единственное слагаемое (при $i \neq j$) не равно нулю. Это слагаемое равно $D_i \varphi_i(t)$. Таким образом,

$$M[\dot{X}(t) V_i] = D_i \varphi_i(t), \quad (4.184)$$

откуда

$$\varphi_i(t) = \frac{1}{D_i} M[\dot{X}(t) V_i]. \quad (4.185)$$

Выражение (4.185) определяет координатные функции при данном выборе случайных коэффициентов V_i . Приведем доказательство того, что координатные функции $\varphi_i(t)$, определенные из (4.185), дают наилучшее приближение случайной функции $\dot{X}(t)$ любым данным числом членов ряда (4.182) при данном выборе коэффициентов V_i .

Представим $\dot{X}(t)$ в виде

$$\dot{X}(t) = \sum_{i=1}^n V_i \psi_i(t) + R_n(t), \quad (4.186)$$

где $\psi_i(t)$ — произвольные функции. Нам необходимо доказать, что $M[R_n(t)^2]$ минимально при любом t , если координатные функции определены из (4.185). Запишем

$$\begin{aligned} M[R_n(t)^2] &= M\left[\left\{\dot{X}(t) - \sum_{i=1}^n V_i \psi_i(t)\right\}^2\right] = M[\dot{X}(t)^2] - \\ &- 2 \sum_{i=1}^n \psi_i(t) M[\dot{X}(t) V_i] + \sum_{i=1}^n \sum_{j=1}^n M[V_i V_j] \psi_i(t) \psi_j(t). \end{aligned} \quad (4.187)$$

Учитывая (4.184), получаем

$$\begin{aligned} M[R_n(t)^2] &= M[\dot{X}(t)^2] - 2 \sum_{i=1}^n \psi_i(t) D_i \varphi_i(t) + \sum_{i=1}^n D_i \psi_i(t)^2 = \\ &= M[\dot{X}(t)^2] + \sum_{i=1}^n D_i \{\psi_i(t)^2 - 2\psi_i(t) \varphi_i(t)\}. \end{aligned} \quad (4.188)$$

Выражение в фигурных скобках можно представить в виде $\psi_i(t)^2 - 2\psi_i(t) \varphi_i(t) = \{\psi_i(t) - \varphi_i(t)\}^2 - \varphi_i(t)^2$. (4.189)

Тогда формулу (4.188) можно переписать так:

$$M[R_n(t)^2] = M[\dot{X}(t)^2] - \sum_{i=1}^n D_i \varphi_i(t)^2 + \sum_{i=1}^n D_i \{\psi_i(t) - \varphi_i(t)\}^2. \quad (4.190)$$

Из последнего равенства очевидно, что для любого t при любом заданном n математическое ожидание квадрата остаточного члена в (4.186) будет минимальным, если положить $\psi_i(t) \equiv \varphi_i(t)$, ($i = 1, 2, \dots, n$). Другими словами, формула (4.186) с отброшенным остаточным членом при любом n дает наилучшее приближение случайной функции $\dot{X}(t)$, если в качестве координатных функций взять функции $\varphi_i(t)$, вычисленные из (4.185).

Координатные функции, полученные в соответствии с (4.185), назовем *оптимальными*.

Из уравнения (4.190) при $\psi_i(t) \equiv \varphi_i(t)$ получим формулу для математического ожидания квадрата остаточного члена канонического разложения по оптимальным координатным функциям

$$M[R_n(t)^2] = D_x(t) - \sum_{i=1}^n D_i \varphi_i(t)^2. \quad (4.191)$$

Изложенный метод канонического представления случайных функций очень удобен с точки зрения различных преобразований, главным образом линейных. При помощи этого метода линейные операции над случайными функциями (дифференцирование, интегрирование, решение дифференциальных уравнений и т. д.) можно свести к соответствующим операциям над известными неслучайными функциями, т. е. пользоваться обычным аппаратом математического анализа.

Контрольные вопросы и задания

1. Что такое элементарная случайная функция?
2. Определите характеристики элементарной случайной функции.
3. Дайте определение каноническому разложению случайной функции.
4. Найдите характеристики случайной функции, представленной каноническим разложением.
5. Как определяются координатные функции канонического разложения?

§ 10. СТАЦИОНАРНЫЕ СЛУЧАЙНЫЕ ФУНКЦИИ

Важным классом случайных функций является класс стационарных случайных функций. Часто на практике встречаются процессы, протекающие во времени приблизительно одинаково. Они имеют вид непрерывных случайных колебаний вокруг некоторого среднего значения. При этом

с течением времени средняя амплитуда и характер этих колебаний существенно не изменяются. Это так называемые стационарные случайные процессы (рис. 4.35).

Примерами стационарных случайных процессов могут служить: колебания напряжения в сети, изменения показателей качества химических продуктов, случайные шумы в радиоприемнике, качка корабля и др.

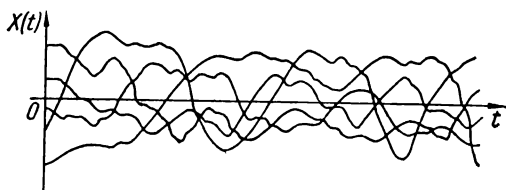


Рис. 4.35. Стационарный случайный процесс.

Стационарный процесс можно рассматривать как продолжающийся во времени неопределенно долго. За начало отсчета можно выбрать любой момент времени. При исследовании стационарного случайного процесса на любом временном участке должны быть получены одни и те же его характеристики.

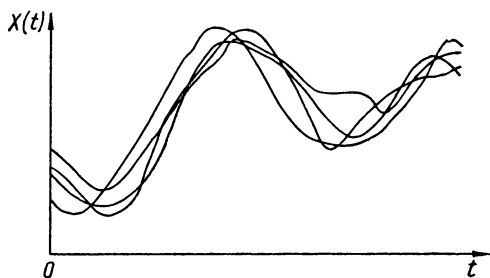


Рис. 4.36. Нестационарный случайный процесс.

С другой стороны, можно указать явно нестационарные случайные процессы. Это процессы, имеющие определенную тенденцию развития во времени. Например, затухающие колебания в системе автоматического регулирования или в электрической цепи, потребление электроэнергии в некотором районе в течение суток и т. д. На рис. 4.36 показано семейство реализаций типично нестационарного случайного процесса — изменения нагрузки энергосистемы в течение суток.

При исследовании многих объектов и систем управления можно получить точное или приближенное описание процессов с помощью стационарных случайных функций. Многие нестационарные случайные процессы существенно нестационарны не на всей протяженности. С известным приближением их можно считать стационарными на некоторых участках. Кроме того, много процессов, которые нельзя рассмотреть как стационарные, удастся выразить через неслучайные функции и стационарные случайные.

Определение стационарной случайной функции

Рассмотрим две системы случайных величин $X(t_1)$, $X(t_2)$, $X(t_3)$, ..., $X(t_n)$ и $X(t_1 + \tau)$, $X(t_2 + \tau)$, $X(t_3 + \tau)$, ..., $X(t_n + \tau)$. Эти системы соответствуют значениям случайной функции $X(t)$ в моменты t_1 , t_2 , t_3 , ..., t_n и $t_1 + \tau$, $t_2 + \tau$, $t_3 + \tau$, ..., $t_n + \tau$.

Если законы распределения системы случайных величин $X(t_1)$, $X(t_2)$, ..., $X(t_n)$ и системы случайных величин $X(t_1 + \tau)$, $X(t_2 + \tau)$, ..., $X(t_n + \tau)$ совпадают при любых значениях n и выбранных значениях t_1 , t_2 , ..., t_n и не зависят от τ , то случайная функция $X(t)$ называется стационарной в узком смысле. Вероятностные характеристики стационарной в узком смысле случайной функции не зависят от абсолютных значений аргумента, а зависят от относительных интервалов между этими значениями.

В дальнейшем, предполагая, что аргументом t случайной функции $X(t)$ является время, мы будем говорить о случайных процессах.

Стационарным в узком смысле случайным процессом называется случайный процесс $X(t)$, для которого n -мерный дифференциальный закон распределения $f(x_1, x_2, \dots, x_n; t_1, t_2, \dots, t_n)$ зависит при любом t только от интервалов времени между моментами t_1, t_2, \dots, t_n .

Так, одномерный дифференциальный закон распределения $f(x, t)$ стационарного в узком смысле процесса не будет зависеть от времени t , двумерный — $f(x_1, x_2; t_1, t_2)$ не будет зависеть отдельно от t_1 и t_2 , а только от их разности $\tau = t_2 - t_1$, n -мерный — $f(x_1, x_2, \dots, x_n; t_1, t_2, \dots, t_n)$ не будет зависеть отдельно от t_1, t_2, \dots, t_n , а только от их разностей, например,

$$\tau_1 = t_2 - t_1, \quad \tau_2 = t_3 - t_1, \quad \dots$$

При решении практических задач многомерные дифференциальные законы распределения случайных процессов

не рассматриваются. Исследования производят при помощи таких характеристик, как математическое ожидание, дисперсия и корреляционная функция. При этом используется только часть условий стационарности случайных процессов.

Случайный процесс называется стационарным в широком смысле, если его математическое ожидание не зависит от времени t , а корреляционная функция является функцией только $\tau = t_2 - t_1$:

$$m_x(t) = m_x = \text{const},$$

$$K_x(t, t + \tau) = K_x(\tau). \quad (4.192)$$

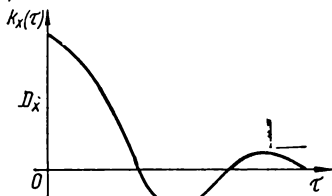


Рис. 4.37. Корреляционная функция стационарной случайной функции.

Функции, стационарные в широком смысле, часто называют просто стационарными.

Как известно, корреляционная функция любой случайной функции обладает свойством симметрии

$$K_x(t, t') = K_x(t', t). \quad (4.193)$$

Для стационарного процесса, полагая $t - t' = \tau$, получим

$$k_x(\tau) = k_x(-\tau), \quad (4.194)$$

т. е. корреляционная функция $k_x(\tau)$ — четная функция своего аргумента. Поэтому корреляционную функцию обычно определяют только для положительных значений аргумента (рис. 4.37).

Часто на практике пользуются нормированной корреляционной функцией

$$\rho_x(\tau) = \frac{k_x(\tau)}{D_x}, \quad (4.195)$$

где $D_x = k_x(0)$ — постоянная дисперсия случайного процесса. По существу функция $\rho_x(\tau)$ представляет собой коэффициент корреляции между сечениями случайной функции, разделенными интервалом τ по времени. Значение $\rho_x(\tau)$ в начальной точке (при $\tau = 0$), очевидно, равно единице

$$\rho_x(0) = 1.$$

Все эти определения справедливы и в том случае, когда мы имеем дело с функцией многих независимых переменных t_1, t_2, \dots, t_n . В соответствии с этим τ может обозначать сово-

купность разностей $t_1 - t'_1, t_2 - t'_2, \dots, t_n - t'_n$. Случайную функцию n переменных t_1, t_2, \dots, t_n называют стационарной в широком смысле, если ее математическое ожидание постоянно и корреляционная функция зависит только от разностей $t_1 - t'_1, t_2 - t'_2, \dots, t_n - t'_n$.

Для стационарных случайных функций — координат точки n -мерного пространства — разработана специальная теория, которая в физике называется теорией однородных случайных полей.

Дифференцирование стационарной случайной функции

Пусть стационарная случайная функция $X(t)$ преобразуется при помощи оператора дифференцирования в функцию $Y(t)$

$$Y(t) = \frac{dX(t)}{dt}. \quad (4.196)$$

Поскольку математическое ожидание стационарной случайной функции $X(t)$ постоянно, математическое ожидание производной тождественно равно нулю $m_Y(t) \equiv 0$.

Корреляционную функцию определим как вторую смешанную частную производную корреляционной функции $k_x(\tau)$

$$K_y(t, t') = \frac{\partial^2 k_x(\tau)}{\partial t, \partial t'}. \quad (4.197)$$

Из выражения (4.197) получаем

$$K_y(t, t') = \frac{\partial^2 k_x(\tau)}{\partial \tau^2} \frac{\partial \tau}{\partial t} \frac{\partial \tau}{\partial t'} = -k_x''(\tau). \quad (4.198)$$

Таким образом, производная стационарной случайной функции представляет собой тоже стационарную функцию, корреляционная функция которой равна взятой с обратным знаком второй производной корреляционной функции случайной функции $X(t)$.

Выражение для производной порядка s стационарной случайной функции имеет вид

$$K_{ys}(t, t') = (-1)^s k_x^{(2s)}(\tau). \quad (4.199)$$

Все производные стационарной случайной функции являются стационарными случайными функциями, причем для существования производной порядка s стационарной случайной функции необходимо и достаточно существования производной порядка $2s$ ее корреляционной функции.

Спектральное разложение стационарной случайной функции

Частотные методы анализа, основанные на использовании преобразований Фурье, широко распространены при исследованиях детерминированных процессов. Если какой-нибудь процесс можно представить в виде суммы гармонических колебаний различных частот (гармоник), то его можно характеризовать спектром процесса. Это функция, описывающая распределение амплитуд по различным частотам.

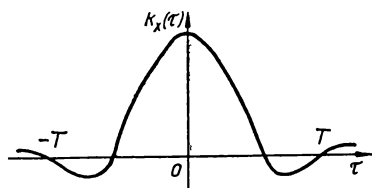


Рис. 4.38. Корреляционная функция $k_x(\tau)$.

Спектр показывает, какие гармонические составляющие преобладают в процессе, какова его внутренняя структура.

При изучении стационарных случайных функций большое значение имеет преобразование такой неслучайной функции времени, как корреляционная

функция. В зависимости от того, какие частоты и в каких соотношениях входят в состав случайной функции, ее корреляционная функция имеет тот или иной вид.

Для случайного процесса амплитуды колебаний являются величинами случайными. Спектр стационарной случайной функции характеризует распределение дисперсий по различным частотам.

Спектральное разложение стационарной случайной функции на конечном интервале времени. Рассмотрим стационарную случайную функцию $\hat{X}(t)$, которую мы наблюдаем на интервале $(0, T)$.

Корреляционная функция случайной функции $\hat{X}(t)$ задана

$$K_x(t, t') = k_x(\tau).$$

Поскольку

$$k_x(\tau) = k_x(-\tau),$$

график $k_x(\tau)$ будет иметь вид, как на рис. 4.38.

При изменении t и t' от 0 до T аргумент $(\tau) = t' - t$ изменяется от $-T$ до $+T$. Четная функция на интервале $(-T, T)$ может быть разложена в ряд Фурье по четным (косинусным) гармоникам

$$k_x(\tau) = \sum_{k=0}^{\infty} D_k \cos \omega_k \tau, \quad (4.200)$$

где

$$\omega_k = k\omega_1; \quad \omega_1 = \frac{2\pi}{2T} = \frac{\pi}{T}.$$

Коэффициенты D_k определяются по формулам

$$D_0 = \frac{1}{2T} \int_{-T}^T k_x(\tau) d\tau;$$

$$D_k = \frac{1}{T} \int_{-T}^T k_x(\tau) \cos \omega_k \tau d\tau, \quad k \neq 0. \quad (4.201)$$

Так как функции $k\tau$ и $\cos \omega_k \tau$ четные, эти формулы можно преобразовать к виду

$$D_0 = \frac{1}{T} \int_0^T k_x(\tau) d\tau;$$

$$D_k = \frac{2}{T} \int_0^T k(\tau) \cos \omega_k \tau d\tau, \quad k \neq 0. \quad (4.202)$$

В выражении для корреляционной функции $k_x(\tau)$ перейдем от аргумента τ к двум аргументам t и t' . Положим

$$\begin{aligned} \cos \omega_k \tau &= \cos \omega_k (t' - t) = \cos \omega_k t' \cos \omega_k t + \\ &+ \sin \omega_k t' \sin \omega_k t. \end{aligned} \quad (4.203)$$

Подставим это выражение в (4.200)

$$K_x(t, t') = \sum_{k=0}^{\infty} (D_k \cos \omega_k t' \cos \omega_k t + D_k \sin \omega_k t' \sin \omega_k t). \quad (4.204)$$

Это каноническое разложение корреляционной функции $K_x(t, t')$. Координатными функциями этого разложения являются косинусы и синусы частот, кратных ω_1 .

По каноническому разложению корреляционной функции можно построить каноническое разложение самой случайной функции. Координатные функции будут те же. Дисперсии будут равны коэффициентам D_k в каноническом разложении корреляционной функции.

Каноническое разложение случайной функции $\overset{\circ}{X}(t)$ будет иметь вид

$$\overset{\circ}{X}(t) = \sum_{k=0}^{\infty} (U_k \cos \omega_k t + V_k \sin \omega_k t). \quad (4.205)$$

Здесь U_k и V_k — некоррелированные случайные величины с математическими ожиданиями, равными нулю. Дисперсии

U_k и V_k одинаковы для каждой пары случайных величин с одним и тем же индексом k

$$D[U_k] = D[V_k] = D_k. \quad (4.206)$$

Эти дисперсии при различных k определяются из (4.202). Разложение такого рода называется спектральным разложением стационарной случайной функции на интервале $(0, T)$.

На представлении случайных функций в виде спектральных разложений основана спектральная теория стационарных случайных процессов.

Спектральное разложение представляет стационарную случайную функцию как совокупность периодических колебаний различных частот $\omega_1, \omega_2, \dots, \omega_k$.

Амплитуды периодических составляющих являются случайными величинами.

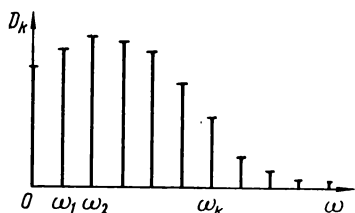


Рис. 4.39. Спектр стационарной случайной функции.

Дисперсия функции $\dot{X}(t)$ распределена определенным образом по различным частотам. Одним частотам соответствуют большие дисперсии,

другим — меньшие. Это распределение можно проиллюстрировать графически в виде спектра стационарной случайной функции. Точнее будет назвать его спектром дисперсий.

На оси абсцисс откладываются частоты $\omega_0 = 0, \omega_1, \omega_2, \dots, \omega_k$, а по оси ординат соответствующие дисперсии (рис. 4.39).

Следует заметить, что дисперсия стационарной случайной функции равна сумме дисперсий всех гармоник ее спектрального разложения. Геометрически это означает, что сумма всех ординат спектра (рис. 4.39) равна дисперсии случайной функции.

Спектральное разложение стационарной случайной функции на бесконечном интервале времени. Спектральное разложение стационарной случайной функции $\dot{X}(t)$ на конечном интервале времени $(0, T)$ дало нам спектр дисперсий в виде отдельных дискретных значений, разделенных равными промежутками. Это так называемый линейчатый спектр.

Чем больше интервал времени, в котором рассматривается случайная функция, тем полнее наши сведения о ней.

Перейдем к пределу при $T \rightarrow \infty$ и посмотрим, какой характер будет иметь при этом спектр случайной функции. При $T \rightarrow \infty$ $\omega_1 = \frac{2\pi}{2t} \rightarrow 0$. Таким образом, расстояние между частотами ω_k , на которых строится спектр, будет неограниченно уменьшаться. При этом дискретный спектр приближается к непрерывному. В этом непрерывном спектре каждому сколь угодно малому интервалу частот $\Delta\omega$ будет соответствовать элементарная дисперсия ΔD (ω).

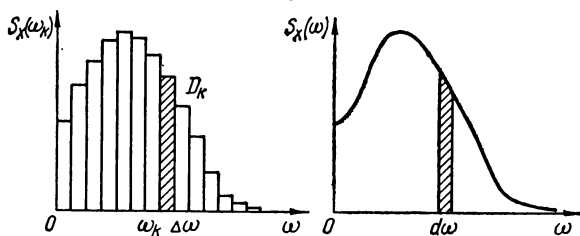


Рис. 4.40. Спектральная плотность.

Обозначим расстояние между соседними частотами $\Delta\omega$. На каждом отрезке $\Delta\omega$, как на основании, построим прямоугольник с помощью D_k (рис. 4.40). Получим ступенчатую диаграмму.

Высота на участке $\Delta\omega$, прилежащем к точке ω_k , равна

$$S_x(\omega_k) = \frac{D_k}{\Delta\omega}. \quad (4.207)$$

Это средняя плотность дисперсии на участке $\Delta\omega$. Суммарная площадь всей диаграммы равна дисперсии случайной функции.

Пусть $T \rightarrow \infty$, при этом $\Delta\omega \rightarrow 0$ и ступенчатая кривая будет неограниченно приближаться к плавной кривой $S_x(\omega)$ (рис. 4.40). Это график плотности распределения дисперсий по частотам непрерывного спектра. Функция $S_x(\omega)$ называется спектральной плотностью дисперсии или спектральной плотностью стационарной случайной функции $\dot{X}(t)$.

Площадь, ограниченная кривой $S_x(\omega)$, равна дисперсии случайной функции

$$D_x = \int_0^{\infty} S_x(\omega) d\omega. \quad (4.208)$$

Эта формула есть разложение дисперсии D_x на сумму элементарных слагаемых $S_x(\omega) d\omega$. Каждое из этих слагаемых

представляет собой дисперсию, приходящуюся на элементарный участок $d\omega$, прилежащий к точке ω .

Введенная новая характеристика стационарного случайного процесса — спектральная плотность — описывает частотный состав стационарного процесса. Эта характеристика не является самостоятельной. Спектральную плотность $S(\omega)$ можно выразить через корреляционную функцию, подобно тому, как через корреляционную функцию выражают ординаты дискретного спектра

$$k_x(\tau) = \int_0^{\infty} S_x(\omega) \cos \omega \tau d\omega; \quad (4.209)$$

$$S_x(\omega) = \frac{2}{\pi} \int_0^{\infty} k_x(\tau) \cos \omega \tau d\tau. \quad (4.210)$$

В математике выражение такого типа называется интегралом Фурье. Это обобщение разложения в ряд Фурье для случая непериодической функции, рассматриваемой на бесконечном интервале. Формулы (4.209) и (4.210) называются преобразованиями Фурье. Это частный вид преобразования Фурье — так называемое «косинус-преобразование Фурье».

Часто при решении практических задач вместо спектральной плотности $S_x(\omega)$ пользуются нормированной спектральной плотностью

$$s_x(\omega) = \frac{S_x(\omega)}{D_x}, \quad (4.211)$$

где D_x — дисперсия случайной функции.

Нормированная корреляционная функция $\rho_x(\tau)$ и нормированная спектральная плотность $s_x(\omega)$ связаны теми же преобразованиями Фурье

$$\rho_x(\tau) = \int_0^{\infty} s(\omega) \cos \omega \tau d\omega; \quad (4.212)$$

$$s_x(\omega) = \frac{2}{\pi} \int_0^{\infty} \rho_x(\tau) \cos \omega \tau d\tau.$$

Если положить $\tau = 0$ и учесть, что $\rho_x(0) = 1$, получим

$$\int_0^{\infty} s_x(\omega) d\omega = 1, \quad (4.213)$$

т. е. полная площадь, ограниченная графиком нормированной спектральной плотности, равна единице.

Свойство эргодичности стационарных случайных процессов

Пусть требуется оценить характеристики некоторого стационарного случайного процесса $X(t)$. Ранее были рассмотрены способы получения математического ожидания и корреляционной функции из опыта. Для этого необходимо знать определенное число реализаций случайного процесса $X(t)$.

В результате обработки реализаций можно найти оценки для математического ожидания $\tilde{m}_x(t)$ и корреляционной функции $\tilde{K}_x(t_1, t_2)$. Поскольку число наблюдений ограничено, функция $m_x(t)$ не будет строго постоянной. Ее придется осреднить и заменить некоторым постоянным \tilde{m}_x . Точно так же, осредняя значения $\tilde{K}_x(t_1, t_2)$ для разных $\tau = t_2 - t_1$, получим корреляционную функцию $k_x(\tau)$.

Это достаточно сложный и громоздкий метод обработки. Он состоит из двух этапов:

- 1) приближенное определение характеристик случайной функции;
- 2) приближенная оценка этих характеристик.

При обработке наблюдений над стационарной случайной функцией возникает вопрос, необходимо ли располагать несколькими реализациями? Так как случайный процесс является стационарным и протекает во времени однородно, можно предположить, что одна единственная реализация достаточной продолжительности может полностью обеспечить определение характеристик случайной функции.

Оказывается, что такая возможность существует не для всех случайных процессов. Не всегда даже достаточно длительная реализация оказывается эквивалентной множеству отдельных реализаций.

Рассмотрим два стационарных случайных процесса (рис. 4.41, а и б). Каждая реализация случайного процесса

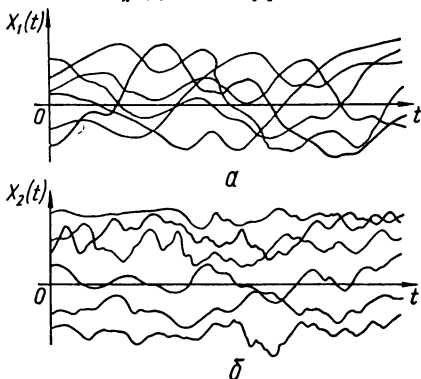


Рис. 4.41. Стационарные случайные процессы.

$X_1(t)$ характеризуется средним значением, вокруг которого происходят колебания, и средним размахом колебаний. Выберем произвольно одну из реализаций $X_1(t)$ и продолжим мысленно опыт, в результате которого она получена, в интервале времени T .

При достаточно большом T эта одна реализация сможет дать достаточно хорошее представление о свойствах случайного процесса в целом. Осредняя, например, значения этой реализации по оси времени, получаем приближенное значение математического ожидания случайной функции; осредняя квадраты отклонений от этого среднего, находим приближенное значение дисперсии и т. д.

Каждая отдельная реализация такого случайного процесса может характеризовать всю совокупность возможных реализаций. В этом случае говорят, что случайный процесс обладает эргодическим свойством.

Среднее значение для каждой реализации процесса $X_2(t)$ свое и может существенно отличаться от математического ожидания, построенного как среднее множества реализаций. О такой случайной функции говорят, что она не обладает эргодическим свойством.

Если случайный процесс $X(t)$ обладает эргодическим свойством, то для него среднее по времени приближенно равно среднему по множеству наблюдений. Это утверждение справедливо и для $X^2(t)$, $X(t)$, $X(t + \tau)$ и т. д. Таким образом, все характеристики случайного процесса можно приближенно определять по одной достаточно длинной реализации.

Частным случаем неэргодической стационарной функции может служить процесс, описываемый выражением

$$Z(t) = X(t) + Y, \quad (4.214)$$

где $X(t)$ — эргодическая стационарная случайная функция с характеристиками m_x , $k_x\tau$; Y — случайная величина с характеристиками m_y и D_y . Будем считать, что корреляция между $X(t)$ и Y отсутствует.

Характеристики случайной функции $Z(t)$

$$m_z = m_x + m_y; \quad (4.215)$$

$$k_z(\tau) = k_x(\tau) + D_y. \quad (4.216)$$

Очевидно, что функция $Z(t)$ стационарна. Но также очевидно, что она не обладает эргодическим свойством. Каждая ее реализация имеет то или иное среднее во времени в зависимости от значения, которое принимает случайная

величина Y . О наличии эргодического свойства стационарной случайной функции можно судить по виду ее корреляционной функции. Например, в рассмотренном случае корреляционная функция $k_z(t)$ отличается от корреляционной функции $k_x(\tau)$ на постоянное слагаемое D_y (рис. 4.42).

При $t \rightarrow \infty$ корреляционная функция эргодического стационарного случайного процесса $X(t)$ стремится к нулю. А корреляционная функция процесса $Z(t)$, не обладающего эргодическим свойством, стремится к постоянному значению D_y . При обработке результатов практических наблюдений может оказаться, что, начиная с некоторого значения, дальнейшее увеличение τ не приводит к убыванию корреляционной функции. Этот факт обычно свидетельствует о наличии в исследуемой случайной функции слагаемого типа обычной случайной величины, а значит и о том, что процесс не является эргодическим.

Если же при $t \rightarrow \infty$ корреляционная функция стремится к нулю, то в большинстве случаев это указывает на эргодичность процесса.

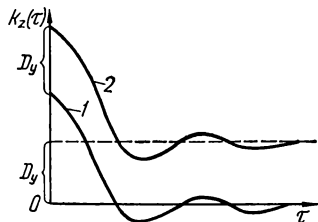


Рис. 4.42. Корреляционные функции:

1 — эргодического стационарного процесса; 2 — неэргодического стационарного процесса.

§ 11. НЕСТАЦИОНАРНЫЕ СЛУЧАЙНЫЕ ФУНКЦИИ

Строго говоря, практически все случайные функции в той или иной степени изменяют свои характеристики с изменением аргумента. Это *нестационарные функции или нестационарные процессы*, если независимой переменной является время. Оперировать с нестационарными случайными процессами гораздо сложнее, чем со стационарными. Поэтому большое практическое распространение получили методы для квазистационарных или нестационарных функций, приводимых к стационарным путем различных допущений и промежуточных преобразований. Рассмотрим некоторые примеры нестационарных случайных функций.

1. **Элементарная случайная функция.** Это простейший пример нестационарной случайной функции

$$X(t) = Z\varphi(t). \quad (4.217)$$

В выражении (4.217) Z — случайная величина с математическим ожиданием, равным нулю; $\varphi(t)$ — неслучайная функция времени.

Математическое ожидание элементарной случайной функции равно нулю. Так как $M[Z] = 0$, то

$$m_x(t) = M[X(t)] = M[Z\varphi(t)] = \varphi(t) M[Z] = 0. \quad (4.218)$$

Дисперсия $D_x(t)$ элементарной случайной функции $X(t)$ равна

$$D_x(t) = M[X^2(t)] = M[Z^2\varphi^2(t)] = \varphi^2(t) D_z. \quad (4.219)$$

Корреляционная функция $K_x(t, t')$ элементарной случайной функции $X(t)$ равна произведению дисперсии D_z случайной величины Z на значение неслучайной функции $\varphi(t)$ при $t = t$ и на ее значение при $t = t'$.

В соответствии с определением

$$K_x(t, t') = M[\hat{X}(t) \hat{X}(t')] = M[Z^2\varphi(t)\varphi(t')],$$

откуда

$$K_x(t, t') = D_z\varphi(t)\varphi(t'). \quad (4.220)$$

Согласно выражениям (4.218), (4.219), (4.220) функция $X(t)$ является нестационарной, если $\varphi(t)$ зависит от времени.

2. Сумма m элементарных случайных функций. Рассмотрим сумму

$$X(t) = \sum_{i=1}^m Z_i\varphi_i(t), \quad (4.221)$$

где Z_1, Z_2, \dots, Z_m — система случайных величин, математическое ожидание которых равно нулю; $\varphi_1(t), \varphi_2(t), \dots, \varphi_m(t)$ — неслучайные функции времени.

Математическое ожидание $m_x(t)$ функции $X(t)$ равно нулю. Корреляционную функцию и дисперсию можно выразить в виде

$$K_x(t, t') = \sum_{i=1}^m \sum_{k=1}^m K_{z_i z_k} \varphi_i(t) \varphi_k(t'); \quad (4.222)$$

$$D_x(t) = K_x(t, t) = \sum_{i=1}^m \sum_{k=1}^m K_{z_i z_k} \varphi_i(t) \varphi_k(t). \quad (4.223)$$

В выражении (4.222) $K_{z_i z_k}$ — корреляционные моменты случайных величин $Z_i Z_k$ (при $i \neq k$). Для $i = k$ $K_{z_i z_k}$ — дисперсии.

Если случайные величины Z_1, Z_2, \dots, Z_m некоррелированы, то $K_{z_i z_k} = 0$ при $i \neq k$. Тогда

$$K_x(t, t') = \sum_{i=1}^{\infty} K_{z_i z_i} \varphi_i(t) \varphi_i(t'); \quad (4.224)$$

$$D_x(t) = \sum_{i=1}^m K_{z_i z_i} \varphi_i^2(t). \quad (4.225)$$

3. Случайная функция, равная сумме неслучайной и стационарной случайной функции. Пусть случайная функция определяется выражением

$$X(t) = \psi(t) + Z(t),$$

где $\psi(t)$ — неслучайная функция, а $Z(t)$ — стационарная случайная. Если $\psi(t)$ не является величиной постоянной, то математическое ожидание случайной функции $X(t)$ зависит от времени

$$m_x(t) = \psi(t) + m_z. \quad (4.226)$$

Корреляционная функция случайной функции $X(t)$ зависит лишь от разности значений аргументов t и t' , так как она равна корреляционной функции стационарной случайной функции $Z(t)$

$$K_x(t, t') = k_x(\tau), \quad \tau = t' - t. \quad (4.227)$$

4. Случайная функция, равная произведению неслучайной функции на стационарную случайную. Выражение для такой функции будет иметь вид

$$X(t) = \varphi(t) Z(t), \quad (4.228)$$

где $\varphi(t)$ — неслучайная функция, $Z(t)$ — стационарная случайная. Если $\varphi(t)$ — непостоянная величина и математическое ожидание $m_z(t)$ стационарной случайной функции $Z(t)$ не равно нулю, то математическое ожидание m_x зависит от времени

$$m_x(t) = \varphi(t) m_z. \quad (4.229)$$

Корреляционная функция $K_x(t, t')$ случайной функции $X(t)$ определяется в соответствии с (4.87) через корреляционную функцию $k_z \tau$

$$K_x(t, t') = \varphi(t) \varphi(t') k_z(\tau), \quad \tau = t' - t. \quad (4.230)$$

Дисперсия $D_x(t)$ случайной функции $X(t)$ равна

$$D_x(t) = K_x(t, t) = \varphi^2(t) k_z(0) = \varphi^2(t) D_z, \quad (4.231)$$

где D_z — дисперсия случайной функции $Z(t)$.

Очевидно, что если $\varphi(t)$ не является постоянной величиной, то случайная функция $X(t)$ — нестационарна.

5. Случайная функция, равная сумме неслучайной функции и произведения неслучайной функции на стационарную случайную. Нестационарная случайная функция $X(t)$ определяется выражением

$$X(t) = \psi(t) + \varphi(t) Z(t), \quad (4.232)$$

где $\psi(t)$ — неслучайные функции времени; $Z(t)$ — стационарная случайная функция с математическим ожиданием m_z , корреляционной функцией $k_z(\tau)$ и дисперсией D_z .

Математическое ожидание $m_x(t)$ нестационарной случайной функции $X(t)$ будет

$$m_x(t) = M[\psi(t) + \varphi(t) Z(t)] = \psi(t) + \varphi(t) m_z. \quad (4.233)$$

Корреляционная функция

$$K_x(t, t') = \varphi(t) \varphi(t') k_z(\tau), \quad (4.234)$$

где $\tau = t' - t$. Дисперсия случайной функции $X(t)$

$$D_x(t) = K_x(t, t') = \varphi^2(t) D_z. \quad (4.235)$$

6. Более общий случай будет, когда $X(t)$ представляет собой линейную комбинацию m функций $Z_1(t), Z_2(t), \dots, Z_m(t)$. В этом случае рассматривается нестационарная случайная функция

$$X(t) = \psi(t) + \sum_{i=1}^m \varphi_i(t) Z_i(t), \quad (4.236)$$

где $\psi(t)$ и $\varphi_i(t)$ — неслучайные функции времени.

Математическое ожидание функции $X(t)$ определяется как

$$m_x(t) = \psi(t) + \sum_{i=1}^m \varphi_i(t) m_{z_i}, \quad (4.237)$$

где m_{z_i} — математическое ожидание случайной функции $Z_i(t)$.

Корреляционная функция

$$K_x(t, t') = \sum_{i=1}^m \sum_{k=1}^m \varphi_i(t) \varphi_k(t') k_{z_i z_k}(\tau), \quad (4.238)$$

где

$$k_{z_i z_k}(\tau) = M[\{Z_i(t) - m_{z_i}\} \{Z_k(t') - m_{z_k}\}], \quad i \neq k \quad (4.239)$$

— взаимная корреляционная функция случайных функций $Z_i(t)$ и $Z_k(t)$. Дисперсия нестационарной случайной

функции $X(t)$

$$D_x(t) = \sum_{i=1}^m \sum_{k=1}^m \varphi_i(t) \varphi_k(t) D_{z_i z_k}, \quad (4.240)$$

где

$$D_{z_i z_k} = M \{ [Z_i(t) - m_z] [Z_k(t) - m_z] \}, \quad i \neq k \quad (4.241)$$

— взаимная дисперсия случайных функций $Z_i(t)$ и $Z_k(t)$. $k_{z_i z_k}(\tau)$ и $D_{z_i z_k}$ — соответственно корреляционная функция и дисперсия случайной функции $Z_i(t)$.

Если случайные функции $Z_1(t)$, $Z_2(t)$, ..., $Z_m(t)$ некоррелированы, то

$$k_{z_i z_k}(\tau) = 0; \quad i \neq k; \quad (4.242)$$

$$D_{z_i z_k} = 0; \quad i \neq k.$$

Тогда

$$K_x(t, t') = \sum_{i=1}^m \varphi_i(t) \varphi_i(t') k_{z_i z_i}(\tau) \quad (4.243)$$

и

$$D_x(t) = \sum_{i=1}^m \varphi_i^2(t) D_{z_i}. \quad (4.244)$$

7. Случайная функция, приводимая путем замены независимой переменной t к сумме неслучайной функции и произведения неслучайной функции на стационарную случайную функцию. Пусть случайная функция $X(t)$ представлена суммой математического ожидания $m_x(t)$ и случайной функции $\dot{X}(t)$ с нулевым математическим ожиданием

$$X(t) = m_x(t) + \dot{X}(t), \quad (4.245)$$

и пусть независимая переменная t путем изменения масштаба и размерности преобразуется в новую независимую переменную \bar{t} . Найдем условие, при котором после такого преобразования случайная функция равна сумме неслучайной функции $\psi(\bar{t})$ и стационарной случайной функции $Z(\bar{t})$ с нулевым математическим ожиданием. Допустим, преобразование t в \bar{t} имеет вид

$$\bar{t} = \chi(t). \quad (4.246)$$

Преобразованием независимой переменной t функция $X(t)$ приводится к виду

$$X(t) = \psi(\bar{t}) + Z(\bar{t}). \quad (4.247)$$

Определим сначала, как найти функцию $\psi(\bar{t})$. Очевидно, что

$$m_x(t) = M[X(t)] = \psi(\bar{t}). \quad (4.248)$$

Если определено преобразование (4.246), т. е. известна функция $\chi(t)$, то при условии

$$t = \chi_{-1}(\bar{t}). \quad (4.249)$$

Функция $\psi(\bar{t})$ определяется как

$$\psi(\bar{t}) = m_x[\chi_{-1}(\bar{t})]. \quad (4.250)$$

Пусть, например,

$$\bar{t} = \chi(t) = t^2,$$

а

$$m_x(t) = at^2 e^{-\alpha t}.$$

Тогда

$$\psi(\bar{t}) = a\bar{t}e^{-\alpha\sqrt{\bar{t}}}.$$

Теперь найдем условие, при котором случайная функция представлена в виде (4.247). Корреляционная функция $K_x(t, t')$ случайной функции $X(t)$

$$\begin{aligned} K_x(t, t') &= M[\{X(t) - m_x(t)\} \{X(t') - m_x(t')\}] = \\ &= M[\{\psi(\bar{t}) - m_x(t) + Z(\bar{t})\} \{\psi(\bar{t}') - m_x(t') + Z(\bar{t}')\}], \end{aligned}$$

где

$$\bar{t}' = \chi(t), \quad \bar{t}' = \chi(t').$$

Так как

$$m_x(t) - \psi(\bar{t}) = 0;$$

$$m_x(t') - \psi(\bar{t}') = 0,$$

то

$$K_x(t, t') = M[Z(\bar{t})Z(\bar{t}')].$$

Условие приводимости случайной функции $X(t)$ к

$$X(t) = \psi(\bar{t}) + Z(\bar{t})$$

получаем в виде

$$K_x(t, t') = K_z(\bar{t}' - \bar{t}) = K_z[\chi(t') - \chi(t)]. \quad (4.251)$$

Условие (4.251) можно преобразовать к виду более удобному для проверки приводимости функции к стационарной.

Продифференцируем (4.251)

$$\begin{aligned} & \frac{\partial K_x(t, t')}{\partial t} dt + \frac{\partial K_x(t, t')}{\partial t'} dt' = \\ & = \frac{\partial K_z[\chi(t') - \chi(t)]}{\partial \chi} [\chi'(t') dt' - \chi'(t) dt]. \end{aligned} \quad (4.252)$$

В выражении (4.252) $\frac{\partial K_z[\chi(t') - \chi(t)]}{\partial \chi}$ представляет собой производную $K[\chi(t') - \chi(t)]$ по аргументу $\chi(t') - \chi(t)$. Так как t и t' — независимые переменные, уравнение (4.252) можно разделить на два

$$\begin{aligned} \frac{\partial K_x(t, t')}{\partial t'} &= \frac{\partial K_z[\chi(t) - \chi(t')]}{\partial \chi} \chi'(t'); \\ \frac{\partial K_x(t, t')}{\partial t} &= - \frac{\partial K_z[\chi(t) - \chi(t')]}{\partial \chi} \chi'(t). \end{aligned} \quad (4.253)$$

Из (4.253) получим

$$\frac{\frac{\partial K_x(t, t')}{\partial t}}{\frac{\partial K_x(t, t')}{\partial t'}} = - \frac{\chi'(t)}{\chi'(t')}. \quad (4.254)$$

Если равенство (4.254) выполняется, то отношение производных от корреляционной функции $K_x(t, t')$ по аргументам t и t' должно быть равно взятому с обратным знаком отношению $\chi'(t)$ и $\chi'(t')$.

Частным необходимым условием приведения нестационарной функции $X(t)$ к виду (4.247) является постоянство дисперсии $D_x(t)$

$$D_x(t) = K_x(t, t') = K_z[\chi(t) - \chi(t)] = K_z(0) = \text{const}. \quad (4.255)$$

Пусть, например, нестационарная случайная функция $X(t)$ с дисперсией D_x определяется корреляционной функцией

$$K_x(t, t') = D_x e^{-\alpha/t'^2 - t'^2}, \quad (4.256)$$

где $\alpha > 0$. Положим

$$\bar{t} = \chi(t) = t^2.$$

Тогда корреляционная функция $K_x(t, t')$ преобразуется к виду

$$K_x(t, t') = D_x e^{-\alpha/\bar{t}' - \bar{t}'}, \quad (4.257)$$

т. е. центрированная составляющая получилась стационарной.

Определим $\chi(t)$ из уравнения (4.254). Учитывая, что для любой функции $x(t)$

$$\frac{\partial |x(t)|}{\partial t} = \begin{cases} x'(t), & x(t) > 0; \\ -x'(t), & x(t) < 0, \end{cases}$$

имеем

$$\frac{\partial |x(t)|}{\partial t} = x'(t) \operatorname{sign} x(t).$$

А значит

$$\frac{\frac{\partial K_x(t, t')}{\partial t}}{\frac{\partial K_x(t, t')}{\partial t'}} = -\frac{t}{t'} = -\frac{\chi(t)}{\chi(t')}. \quad (4.258)$$

Таким образом,

$$\bar{t} = \chi(t) = t^2$$

с точностью до постоянной величины.

Понятия и методы теории вероятностей и теории случайных функций представляют собой мощный математический аппарат. Он возник в результате обобщения большого количества наблюдений и экспериментов над реальными объектами и явлениями. Воплощенный в строгую математическую форму, этот аппарат позволяет решать сложные практические задачи в области обработки экспериментальных данных, определения закономерностей, существующих в природе, предсказания их изменений, управления.

Контрольные вопросы и задания

1. Дайте определение стационарного процесса в узком и в широком смысле.
2. Опишите автокорреляционную функцию стационарного случайного процесса. Какой характер имеет ее график?
3. Расскажите о спектральном разложении стационарной случайной функции на конечном интервале.
4. Что представляет собой спектр стационарной случайной функции при разложении на бесконечном интервале?
5. Как связаны между собой спектральная плотность и корреляционная функция?
6. Расскажите о свойстве эргодичности стационарных случайных функций.
7. Чем отличаются корреляционные функции эргодической и неэргодической стационарных случайных функций?
8. Приведите основные типы нестационарных случайных функций.

§ 12. КОНЕЧНЫЕ СЛУЧАЙНЫЕ ПРОЦЕССЫ. ВЕРоятНОСТНЫЕ ПОСЛЕДОВАТЕЛЬНОСТИ

Классическая теория вероятностей часто рассматривает задачи следующего вида: «Есть две урны; первая из них содержит два черных и один белый шар, а вторая — один черный и два белых шара. Выбирают наудачу одну урну,

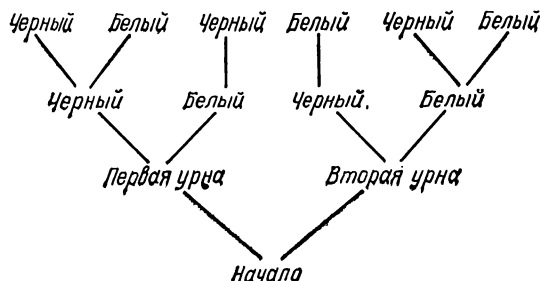


Рис. 4.43. Дерево логических возможностей.

и из нее вынимают последовательно два шара. Какова вероятность того, что первый окажется черным, а второй белым?» Рассмотрим все логические возможности. Последовательность исходов включает три шага. Первый шаг — выбор урны, второй — извлечение первого шара, третий — извле-

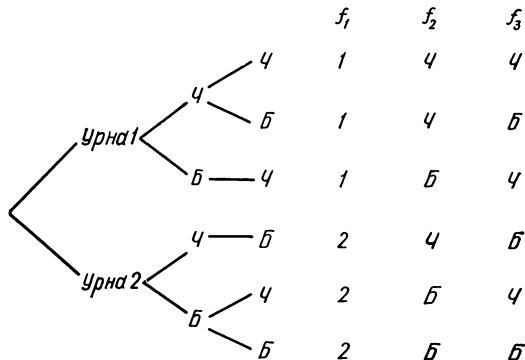


Рис. 4.44. Функции исхода на дереве логических возможностей.

чение второго шара. Весьма удобно для анализа логических возможностей вычерчивать так называемое «дерево» логических исходов (рис. 4.43).

При рассмотрении серии из n опытов множество возможных последовательностей исходов можно трактовать, как

пространство логических возможностей. Пусть f_i определяет исход i -го эксперимента. Функции f_1, f_2, \dots, f_n , определенные на множестве u всевозможных путей на дереве исходов, назовем функциями исходов.

Совокупность n функций исходов описывает пространство логических возможностей.

Для нашего примера пространство логических возможностей включает 6 возможностей (рис. 4.44); $f_1 = 1$ или 2

в зависимости от того, какая из урн была выбрана; f_2 и f_3 определяют цвет вынутых 1-го и 2-го шаров.

Рассмотрим еще один пример. Подбрасывают монету и игральную кость. Эта операция представляет собой последовательность двух экспериментов. Области возможных значений функций исходов f_1 и f_2 соответственно являются $R_1 =$ (лицевая, обратная сторона) и $R_2 = (1, 2, 3, 4, 5, 6)$. В пространстве логических возможностей каждая из 12 точек однозначно описывается значениями функций исходов. На рис. 4.45 изображено дерево для описываемого примера.

Рис. 4.45. Дерево логических возможностей для двух последовательных испытаний.

Каждая точка пространства логических возможностей u описывается определенной комбинацией значений функций f . Обратное утверждение, что любая комбинация значений этих функций описывает некоторую точку из u , не всегда справедливо. Так, в первом примере с шарами комбинация $f_1 = 1, f_2 = Б, f_3 = Б$ не может встретиться.

Логически независимыми будем называть функции f_1, f_2, \dots, f_n , определенные на заданном множестве u , если любая комбинация значений этих функций возможна. Если функции исходов описывают все пространство u и являются логически независимыми, то говорят, что они образуют базис пространства u .

Каждая конкретная последовательность возможных исходов соответствует определенной ветви дерева. Если исход каждого отдельного опыта зависит от неконтролируемых

факторов, несет элемент случайности, то последовательность называется случайной (стохастической, вероятностной). Будем рассматривать конечные случайные последовательности. Это означает, что число опытов конечно и каждый опыт имеет конечное число возможных исходов. Кроме того, предполагается, что если известны исходы всех опытов, предшествовавших данному, то для этого последнего опыта можно определить как все возможные исходы, так и вероятности каждого из этих исходов. Задача заключается в предсказании хода всей последовательности в целом. Например, в случае многократного бросания монеты нас могут интересовать вероятности истинности высказываний типа: «Более чем две трети всех бросаний приводит к выпадению лицевой стороны» или «Число выпадений оборотной и лицевой сторон одно и то же» и т. д. Для ответа на подобные вопросы необходимо всей последовательности в целом приписать некоторую вероятностную меру.

Пусть u — множество всех ветвей некоторого дерева. Весовой функцией путей для этого дерева называется весовая функция (вероятностная мера), определенная на множестве u . Веса ветвей дерева задаются так, что сумма весов всех ветвей, выходящих из любой точки ветвления, равна 1. В любом дереве ветви, выходящие из точки ветвления в j -м ряду, описывают возможные исходы j -го опыта, если исходы первых $j - 1$ опытов уже известны.

Введем систему обозначений для весов ветвей. Пусть $abc \dots st$ — одна из возможных последовательностей исходов первых j опытов при условии, что исходы первых $j - 1$ опытов описываются последовательностью $abc \dots s$. Вес исхода t j -го опыта обозначим

$$P_{abc \dots s, t}.$$

Вес исхода первого опыта обозначается P_a . В примере с урнами и шарами $P_1 = \frac{1}{2}$, $P_{1,4} = \frac{2}{3}$ и $P_{1,44} = \frac{1}{2}$.

Для того чтобы определить вероятности высказываний, связанных со всей последовательностью опытов, нужно построить вероятностную меру на множестве путей данного дерева. При этом нужно исходить из весов отдельных ветвей и меру выбирать так, чтобы эти веса были равны соответствующим условным вероятностям.

Рассмотрим последовательность из трех опытов. Обозначим черех x произвольный путь на соответствующем

дереве логических возможностей. Функции исхода описывают пространство логических возможностей так, что x — единственный элемент, содержащийся во множестве истинности высказывания типа

$$[f_1(x) = a] \wedge [f_2(x) = b] \wedge [f_3(x) = c].$$

Поэтому вес $w(x)$ элемента x равен

$$P[(f_1 = a) \wedge (f_2 = b) \wedge (f_3 = c)].$$

Эту вероятность можно переписать в виде

$$P[f_1 = a] \cdot P_{f_1 = a}[f_2 = b] \cdot P_{f_1 = a, f_2 = b}[f_3 = c].$$

Из определения весов ветвей очевидно, что

$$P[f_1 = a] = P_a;$$

$$P_{f_1 = a}[f_2 = b] = P_{ab};$$

$$P[(f_1 = a) \wedge (f_2 = b) \wedge (f_3 = c)] = P_{ab,c}.$$

Для справедливости всех этих равенств следует положить

$$w(x) = P_a P_{a,b} P_{ab,c}.$$

Таким образом, каждому пути приписывается вес, равный произведению весов ветвей, составляющих этот путь. Если веса путей и веса ветвей связаны таким образом, то веса ветвей всегда равны соответствующим условным вероятностям.

Три типа конечных случайных последовательностей

Предъявляя определенные требования к функциям исхода, можно выделить различные типы случайных последовательностей. Рассмотрим три наиболее важных типа.

1. *Последовательностью с независимыми значениями* называется конечная случайная последовательность с функциями исхода $f_1, f_2, \dots, f_n, \dots$, если для любого n и произвольных исходов t, s, r справедливо равенство

$$P(f_{n-1} = s) \wedge (f_{n-2} = r) \wedge \dots (f_1 = a) [f_n = t] = P[f_n = t].$$

Иначе говоря, в последовательности с независимыми значениями вероятность любого исхода n -го опыта (выбранного из множества возможных исходов) не зависит от предшествующих опытов. Последовательности с независимыми значениями

ми обладают важным свойством. Для последовательности с n функциями исхода

$$P[(f_1 = a) \wedge (f_2 = b) \wedge \dots \wedge (f_n = t)] = \\ = P[f_1 = a] \cdot P[f_2 = b] \cdot \dots \cdot P[f_n = t],$$

т. е. функции исхода f_1, f_2, \dots, f_n последовательности с независимыми значениями вероятностно независимы. Примером случайной последовательности с независимыми значениями может служить последовательность бросаний монет и игральной кости.

2. *Последовательностью независимых испытаний* называется конечная случайная последовательность с независимыми значениями, причем

$$P[f_n = a] = P[f_m = a]$$

для любых n, m и a .

Это, очевидно, частный случай последовательности с независимыми значениями. Из определения ясно, что для дерева такой последовательности все пучки ветвей должны быть эквивалентными, т. е. состоять из ветвей, отвечающих одним и тем же исходам, причем одинаковым исходам соответствуют одинаковые вероятности.

В качестве примера можно рассмотреть последовательность подбрасываний «неправильной» монеты, для которой вероятность выпадения оборотной стороны при каждом бросании равна $\frac{2}{3}$. Для любого числа бросаний можно построить дерево. Случай $n = 3$ показан на рис. 4.48. Все три функции исхода определены в области $R =$ (лицевая, оборотная сторона) и для каждой из трех функций исхода

$$P[f_i = 0] = \frac{2}{3}; \quad P[f_i = 1] = \frac{1}{3}.$$

3. *Марковской цепью* называется конечная случайная последовательность с функциями исхода f_0, f_1, \dots, f_n , если ее исходное состояние f_0 фиксировано,

$$P(f_{n-1} = s) \wedge (f_{n-2} = r) \wedge \dots \wedge (f_1 = a) [f_n = t] = \\ = P_{f_{n-1}} = s [f_n = t]$$

и

$$P_{f_{n-1}} = s [f_n = t] = P_{f_{m-1}} = s [f_m = t]$$

для всех $m \geq 1, n \geq 2$ и любой последовательности исходов a, b, \dots, s, t .

Иначе говоря, исход данного опыта зависит только от исхода предыдущего опыта, и, кроме того, характер этой зависимости одинаков для всех этапов последовательности опытов.

Если две точки ветвления дерева марковской последовательности характеризуются одинаковыми исходами, то, независимо от того, принадлежат эти точки одному или различным ветвям дерева, выходящие из этих точек пучки ветвей должны быть эквивалентны.

Учитывая эти особые свойства марковских цепей, им можно дать новое определение. Пусть $\{s_1, s_2, \dots, s_r\}$ — мно-

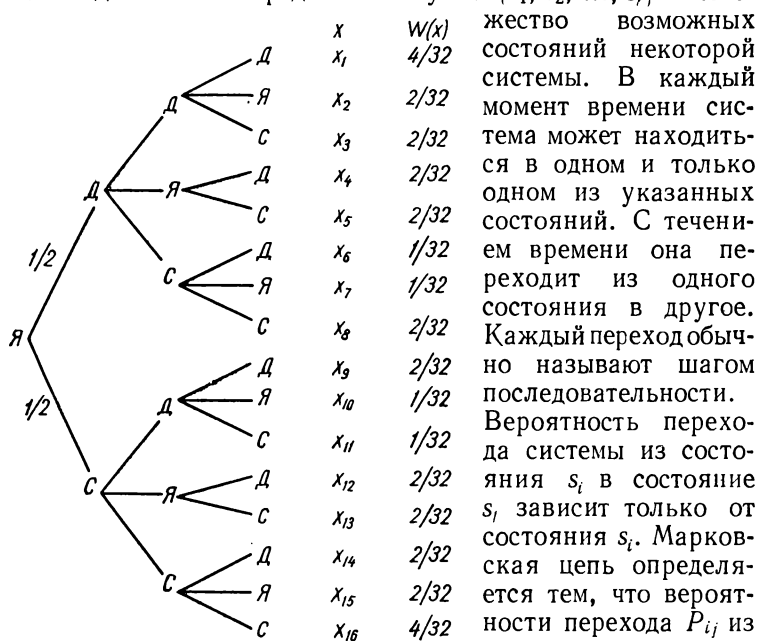


Рис. 4.46. Дерево логических возможностей марковской цепи.

жество возможных состояний некоторой системы. В каждый момент времени система может находиться в одном и только одном из указанных состояний. С течением времени она переходит из одного состояния в другое. Каждый переход обычно называют шагом последовательности. Вероятность перехода системы из состояния s_i в состояние s_j зависит только от состояния s_i . Марковская цепь определяется тем, что вероятности перехода P_{ij} из состояния s_i в состояние s_j определяются для всех упорядочен-

ных пар состояний. Кроме того, должно быть задано исходное состояние, в котором, как предполагается, система находится в начальный момент времени. По этим данным для любой конечной марковской цепи можно построить дерево логических возможностей и вероятностную меру на нем.

Рассмотрим пример построения марковской цепи. Известно, что на Земле 0, никогда не бывает двух ясных дней подряд. Если сегодня ясно, то завтра с одинаковой вероятностью будет дождь или снег. Если сегодня дождь (или снег), то с вероятностью $1/2$ погода не изменится. Если все же изменится, то в половине случаев снег заменяется дождем

или наоборот, и лишь в половине случаев на следующий день будет ясная погода.

Сегодня на Земле 0₃ ясный день. Построим марковскую цепь, используя всю имеющуюся в нашем распоряжении информацию. Различные виды погоды D , $Я$, C принимаем в качестве состояний цепи. Подсчитаем вероятности перехода из одного состояния в другое. Эти вероятности удобно свести в квадратную таблицу

$$\begin{array}{c} \begin{array}{ccc} D & Я & C \end{array} \\ \begin{array}{c} D \\ Я \\ C \end{array} \begin{pmatrix} 1/2 & 1/4 & 1/4 \\ 1/2 & 0 & 1/2 \\ 1/4 & 1/4 & 1/2 \end{pmatrix} \end{array} .$$

В первой строке записаны вероятности различной погоды после дождя. Во второй строке — вероятности различной погоды после ясного дня. В третьей — после снега. Построим дерево логических возможностей на три последовательных дня и определим на нем вероятностную меру (рис. 4.46).

Это дерево позволяет предсказать возможность дождя в каждый из трех следующих дней. Получаем

$$P[f_1 = D] = \frac{1}{2},$$

$$P[f_2 = D] = \frac{5}{8},$$

$$P[f_3 = D] = \frac{13}{32}.$$

Марковские цепи

В общем случае рассматривается последовательность опытов. Исходом каждого опыта является один из конечного числа возможных исходов a_1, a_2, \dots, a_r , причем в каждом опыте вероятность исхода a_j либо вовсе не зависит от исходов предшествующих экспериментов, либо зависит от исхода единственного эксперимента, непосредственно предшествующего данному.

Зависимость эта задается числами P_{ij} , представляющими вероятность исхода a_j заданного эксперимента при условии, что предшествующий эксперимент имел исход a_i . Исходы a_1, a_2, \dots, a_r называются состояниями, а числа P_{ij} — вероятностями перехода. Вероятности перехода представ-

ляются двумя способами. Первый способ заключается в том, что вероятности перехода записываются в виде квадратной таблицы, называемой матрицей. Например, для марковской цепи с тремя состояниями a_1 , a_2 и a_3 матрица имеет вид

$$P = \begin{pmatrix} P_{11} & P_{12} & P_{13} \\ P_{21} & P_{22} & P_{23} \\ P_{31} & P_{32} & P_{33} \end{pmatrix}.$$

Второй способ представления вероятностей перехода состоит в построении диаграммы перехода.

Например, для матрицы перехода

$$P = \begin{matrix} & \begin{matrix} a_1 & a_2 & a_3 \end{matrix} \\ \begin{matrix} a_1 \\ a_2 \\ a_3 \end{matrix} & \begin{pmatrix} 0 & 1 & 0 \\ 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{3} & 0 & \frac{2}{3} \end{pmatrix} \end{matrix}.$$

диаграмма перехода имеет вид, как на рис. 4.47. Нули в матрице означают невозможность соответствующего перехода.

В любой матрице перехода сумма элементов каждой строки равна единице. Действительно, элементы i -й строки представляют собой вероятности всех возможных исходов в последовательности, находящейся в состоянии a_i .

Предположим, последовательность начинается из состояния i . Определим вероятность того, что через n шагов процесс перейдет в состояние j . Эту вероятность можно обозначить P_{ij}^n .

В общем случае нас интересует эта вероятность для всех возможных начальных состояний i и всех возможных конечных состояний j .

Числа P_{ij}^n также удобно представлять матрицей

$$P^{(n)} = \begin{pmatrix} P_{11}^{(n)} & P_{12}^{(n)} & P_{13}^{(n)} \\ P_{21}^{(n)} & P_{22}^{(n)} & P_{23}^{(n)} \\ P_{31}^{(n)} & P_{32}^{(n)} & P_{33}^{(n)} \end{pmatrix}.$$

Для примера определим вероятности различных возможных состояний через три шага в марковской цепи, представленной диаграммой (рис. 4.47). Пусть последовательность начинается из состояния a_1 . Построим дерево логических возможностей и вероятностную меру на нем (рис. 4.48).

Вероятность $P_{13}^{(3)}$, например, есть сумма всех весов, приписанных тем путям дерева исходов, которые оканчиваются состоянием a_3

$$P_{13}^{(3)} = 1 \cdot \frac{1}{2} \cdot \frac{1}{2} + 1 \cdot \frac{1}{2} \cdot \frac{2}{3} = 7/12.$$

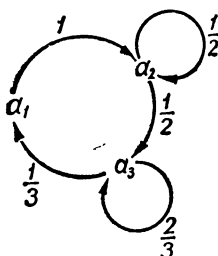


Рис. 4.47. Диаграмма перехода.

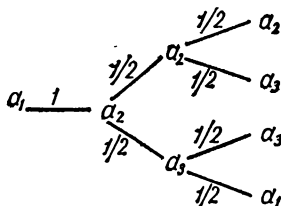


Рис. 4.48. Вероятностная мера на дереве логических возможностей.

Аналогично,

$$P_{12}^{(3)} = 1 \cdot 1/2 \cdot 1/2 = 1/4 \text{ и } P_{11}^{(3)} = 1 \cdot \frac{1}{2} \cdot \frac{1}{3} = \frac{1}{6}.$$

Если построить меру дерева в предположении, что начальным состоянием является a_2 , получим $P_{21}^{(3)}$; $P_{22}^{(3)}$; $P_{23}^{(3)}$.

Точно также определяются $P_{31}^{(3)}$; $P_{32}^{(3)}$; $P_{33}^{(3)}$.

Окончательно матрица имеет вид

$$P^{(3)} = \begin{matrix} & \begin{matrix} a_1 & a_2 & a_3 \end{matrix} \\ \begin{matrix} a_1 \\ a_2 \\ a_3 \end{matrix} & \begin{pmatrix} \frac{1}{6} & \frac{1}{4} & \frac{7}{12} \\ \frac{7}{36} & \frac{7}{24} & \frac{37}{72} \\ \frac{4}{27} & \frac{7}{18} & \frac{25}{54} \end{pmatrix} \end{matrix}.$$

Контрольные вопросы и задания

1. Дайте определение случайной последовательности.
2. Как можно истолковать пространство логических возможностей?
3. Что такое дерево исходов?
4. Как задаются веса путей на дереве исходов?
5. Расскажите о последовательностях с независимыми исходами. Приведите пример.
6. Что называется последовательностью независимых испытаний?
7. Дайте определение цепи Маркова. Приведите пример.

Глава 5

СТАТИСТИЧЕСКИЙ АНАЛИЗ

§ 1. ОСНОВНЫЕ ЗАДАЧИ

Все рассмотренные законы теории вероятностей — это не беспредметные абстракции, лишенные физического смысла. Они представляют собой математическое выражение реальных закономерностей природы.

Математические объекты, которые рассматривались ранее: случайные величины, случайные функции, законы распределения и другие — все они отражают существующие в природе явления и процессы. Все полученные характеристики прямо или косвенно опираются на результаты экспериментов, на реальные события. Математический аппарат теории вероятностей позволяет теоретическим путем определять вероятности одних событий через вероятности других, законы распределения и характеристики одних случайных величин или функций через законы распределения и характеристики других. Это значительно экономит время и средства, которые нужны для эксперимента, хотя и не исключает необходимости эксперимента. Любое исследование всегда базируется на опытных данных. Важно уметь правильно оперировать с экспериментальными данными, уметь собирать и обрабатывать их таким образом, чтобы как можно полнее осмыслить и охарактеризовать исследуемые процессы и явления.

Методы сбора, описания и анализа экспериментальных данных составляют предмет специальной науки — математической статистики.

Все задачи математической статистики связаны с обработкой наблюдений над массовыми случайными явлениями. В зависимости от характера практического вопроса и от объема экспериментального материала эти задачи могут быть различными. Рассмотрим наиболее характерные из них.

Определение закона распределения по статистическим данным

При решении практических задач результаты наблюдений обычно представляют собой множества элементов или единиц какой-либо природы. Эти множества называются статистическими совокупностями. Количество элементов в той или иной статистической совокупности называется объемом совокупности.

Реально мы всегда имеем дело с ограниченным количеством экспериментальных данных. Поэтому результаты обработки всегда характеризуются большим или меньшим элементом случайности. При увеличении числа наблюдений результаты исследований становятся более точными. Появляется возможность обнаружить определенные закономерности. Вступает в силу закон больших чисел. Поэтому к методике обработки экспериментальных данных предъявляются определенные требования.

По возможности должны быть сохранены типичные, характерные черты наблюдаемого явления. Должно быть отброшено все несущественное, второстепенное. В связи с этим возникает задача сглаживания или выравнивания статистических данных. Их надо представить в компактном виде с помощью простых аналитических зависимостей.

Проверка правдоподобия гипотез

Эта задача тесно связана с предыдущей. Допустим, принята гипотеза о том, что случайная величина распределена по закону Пуассона. Требуется подтвердить справедливость этой гипотезы или опровергнуть ее. Другой пример. В опытах наблюдается тенденция к зависимости между двумя случайными величинами. Действительно ли существует объективная зависимость между ними? Или эта тенденция объясняется случайными причинами, недостаточным объемом материала? Для решения этих вопросов в математической статистике разработаны специальные приемы.

Определение неизвестных параметров распределения

Иногда объем исходных данных настолько незначителен, что задача определения законов распределения не ставится. В других случаях характер закона распределения качественно известен еще до опыта на основании теоретических

соображений или по опыту предыдущих исследований. Тогда задача может быть поставлена более узко. Необходимо определить лишь некоторые наиболее характерные параметры (характеристики) распределения случайной величины или системы случайных величин.

Если число опытов невелико, определить точные значения характеристик нельзя. Можно ставить вопрос лишь о нахождении более или менее приближенных «оценок» или «подходящих» значений характеристик. Оценки необходимо определить так, чтобы при массовом применении они приводили в среднем к меньшим ошибкам, чем всякие другие.

Если получено подтверждение о связи между некоторыми величинами, ставятся задачи определения формы и тесноты этой связи. Эти задачи решает теория корреляции.

§ 2. РЯДЫ РАСПРЕДЕЛЕНИЯ И ИХ ХАРАКТЕРИСТИКИ

Каждый элемент статистической совокупности характеризуется некоторыми свойствами — признаками. Возьмем, например, результаты измерений напряжения в осветительной сети, производимых каждый час в течение суток, в: 220, 222, 220, 220, 222, 220, 218, 218, 220, 220, 222, 222, 220, 218, 220, 222, 216, 216, 218, 218, 214, 210, 218.

Каждое отдельное значение признака называется *вариантом* (220 в, 218 в, 216 в, ...). Поэтим признакам элементы совокупности варьируют или, как говорят, обнаруживают дисперсию.

Число, показывающее, сколько раз встречается данный вариант в совокупности, называют абсолютной частотой.

Если расположить отдельные значения признака (варианты) в возрастающем или убывающем порядке, то получится ряд распределения признака или вариационный ряд. *Дисперсией* (вариацией, колеблемостью, изменчивостью) мы будем называть способность признаков изменяться под влиянием большого количества причин.

Построение рядов распределения

Предположим, что регистрирующее устройство контроля каждый час печатает на специальном бланке величину напряжения в сети. Эта запись может иметь, например, такой вид, как на табл. 5.1. Чтобы разобраться в подобном

материале и сделать какие-то выводы, эти данные необходимо систематизировать.

Таблица 5.2 представляет собой ряд распределения в дискретной форме. Здесь отдельные значения признака отличаются друг от друга на конечную величину, т. е. даны в виде дискретных значений.

Таблица 5.1

Часы	<i>и</i>	Часы	<i>и</i>	Часы	<i>и</i>	Часы	<i>и</i>
1	220	7	220	13	222	19	216
2	222	8	218	14	220	20	218
3	220	9	218	15	218	21	218
4	220	10	220	16	220	22	214
5	220	11	220	17	220	23	216
6	222	12	222	18	216	00	218

Если значения признака могут отличаться одно от другого на сколь угодно малую величину, имеет место непрерывная вариация.

В рассматриваемом примере напряжение является непрерывно изменяющимся фактором. Дискретное представление

Таблица 5.2

Варианты	Абсолютная частота, <i>n</i>	Накопленная частота	Относительная частота <i>v</i> , %
214	1	1	4,16
216	3	4	12,48
218	6	10	24,96
220	10	20	41,76
222	4	24	16,64
Итого	24		100%

Таблица 5.3

Интервал	Абсолютная частота, <i>n</i>	Накопленная частота	Относительная частота <i>v</i> , %
213—215	1	1	4,16
215—217	3	4	12,48
217—219	6	10	24,96
219—221	10	20	41,76
221—223	4	24	16,64
Итого	24		100%

обусловлено тем, что измеряемая величина регистрируется в отдельные моменты времени. Поэтому каждый вариант можно рассматривать как центр интервала, в котором заключается действительное значение измеряемого признака. Тогда частоты будут относиться уже не к отдельным значениям признака, а к интервалам, называемым обычно

интервалами округления. Получаем ряд распределения в интервальной форме (табл. 5.3). Признаки могут занимать любое промежуточное значение внутри интервала. Это непрерывные признаки.

Существуют ряды распределения принципиально дискретные. Например, распределение приборов по числу отказов за некоторый период.

Иногда целесообразно представлять такие ряды в интервальной форме. Но это представление носит условный характер.

Кроме абсолютных частот, распределение можно характеризовать относительными частотами, которые вычисляются по формуле

$$v = \frac{n}{N} 100\%,$$

где N — объем статистической совокупности.

При точных измерениях статистическая совокупность может содержать большое количество различных значений признака. В этих случаях необходимо производить укрупненную группировку, объединять близкие значения признака. Число интервалов не следует брать особенно большим, так как в каждом интервале окажется мало наблюдений и закономерность не сможет четко проявиться. С другой стороны, при малом числе интервалов можно не обнаружить подробностей распределения.

Для выбора оптимальной величины интервала можно воспользоваться формулой

$$k \approx \frac{x_{\max} - x_{\min}}{1 + 3,2 \lg N}, \quad (5.1)$$

где x_{\max} — наибольшее значение;

x_{\min} — наименьшее значение признака в совокупности.

Рассмотрим пример. Взята партия из 200 резисторов с номинальным сопротивлением 2000 *ом*. В результате измерения сопротивления каждого образца с точностью до 5 *ом* получена таблица значений (табл. 5.4). Построим ряд распределения для исследуемой статистической совокупности.

Для этого выделим сначала из совокупности наибольшее и наименьшее значение признака.

Как видно из таблицы, $x_{\max} = 2200$ *ом*, $x_{\min} = 1780$ *ом*, диапазон изменения (вариации) признака равен 420. Определим оптимальную длину интервала группировки

$$K = \frac{420}{1 + 3,2 \cdot 2,301} \approx 50.$$

Таблица 5.4

2200	2085	2130	2085	2135	2145	2140	2150	2140	2170
2080	2120	2090	2095	2100	2100	2110	2120	2115	2090
2020	2070	2030	2060	2045	2000	2050	2035	2070	1930
2095	2115	2080	2120	2100	2080	2080	2095	2115	2095
1890	2025	1920	2035	1880	2040	1885	2055	1920	1780
2020	1885	2030	1905	2035	1920	2065	1915	2070	1880
2115	1830	1890	1845	1890	1840	1880	1855	1895	1860
2085	2025	2020	2040	2005	2060	2015	2055	2065	2020
2195	1950	2015	1945	2010	1970	1990	1960	2010	1935
2155	2065	2000	2040	2000	2045	2000	2070	2005	2030
2090	1980	2045	1995	2060	1995	2020	2000	2025	2005
1930	2020	1940	1980	1935	2005	1950	1985	1940	1965
1980	1990	1980	2000	1990	1985	2000	1990	2020	1945
1930	2035	1970	2060	1945	2040	1940	2030	1935	2020
2020	1945	2045	1965	2070	1955	2055	1965	2070	1970
1930	2005	1960	1980	1970	1990	1940	2020	1970	2010
1980	1995	2015	2020	2000	2015	1995	2010	2000	1985
1930	2000	1965	2015	1965	1990	1955	1990	1970	2020
2005	2010	2020	1980	1995	2020	2020	2005	2000	1995
2010	1940	2000	1930	1980	1935	1985	1960	2015	1930

Таблица 5.5

Сопrotивление в ом (интервалы)	Количество резисторов
1775—1825	III
1825—1875	IIII
1875—1925	III III III
1925—1975	III III III III III III III
1975—2025	III III III III III III III III III III III
2025—2075	III III III III III III III II
2075—2125	III III III III
2125—2175	III II
2175—2225	II

При определении границ интервалов следует начинать ряд, отступая примерно на $\frac{1}{2}$ интервала до наименьшего и на $\frac{1}{2}$ интервала сверх наибольшего значения признака. Границы интервалов должны согласоваться с характером округления вариантов в процессе измерения. Для этого проще всего совместить границы интервалов группировки с границами интервалов округления. В интервале группировки должно быть целое число интервалов округления (в нашем случае интервал округления равен пяти). Строим рабочую таблицу для подсчета частот признаков в каждом

интервале. Данные из табл. 5.4 переносим в табл. 5.5, представляя против каждого интервала черточки (единицы счета). После накопления четырех черточек их перечеркивают, чтобы вести счет по пять. На основании рабочей таблицы получаем ряд распределения (табл. 5.6).

Анализируя ряды распределения, можно заметить определенную закономерность в изменении признака. Один из вариантов, примерно в центре ряда, встречается наиболее

Таблица 5.6

Сопротивление в ом (интервалы)	Количество резисторов (частоты)	Накопленные частоты	Центр интервала
1775 — 1825	3	3	1800
1825 — 1875	4	7	1850
1875 — 1925	14	21	1900
1925 — 1975	38	59	1950
1975 — 2025	65	124	2000
2025 — 2075	42	166	2050
2075 — 2125	25	191	2100
2125 — 2175	7	198	2150
2175 — 2225	2	200	2200
Итого	200		

часто. Например, 220 в в первом примере или 2000 ом во втором. По мере удаления от этого варианта в сторону увеличения или уменьшения частоты постепенно уменьшаются. Иногда нарастание частот в одной части ряда происходит быстрее, чем убывание их после максимальной частоты, или наоборот. В этих случаях говорят, что ряд обладает соответственно правосторонней (положительной) или левосторонней (отрицательной) асимметрией. Это явление в большей или меньшей степени проявляется во многих рядах распределения. В связи с этим иногда целесообразно группировать ряд с неодинаковыми интервалами, так как необходимо более подробно дифференцировать одну ее часть и менее подробно другую.

Графическое изображение рядов распределения

Достаточно наглядное представление дает изображение рядов распределения в графической форме. Графики дискретных и интервальных рядов имеют свои специфические особенности. В качестве примера графического представле-

ния дискретного ряда построим ряд распределения напряжения в сети (см. табл. 5.2).

По оси абсцисс откладываем варианты (напряжения), по оси ординат — абсолютные частоты. Из каждой точки оси абсцисс, изображающей вариант, восстанавливаем перпендикуляр, длина которого выражает соответствующую частоту. Вершины этих перпендикуляров соединяем прямыми линиями (рис. 5.1). Получается фигура, которую называют полигоном (многоугольником) распределения.

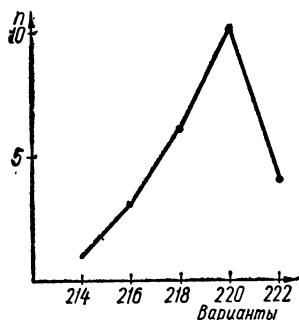


Рис. 5.1. Полигон распределения для дискретного ряда.

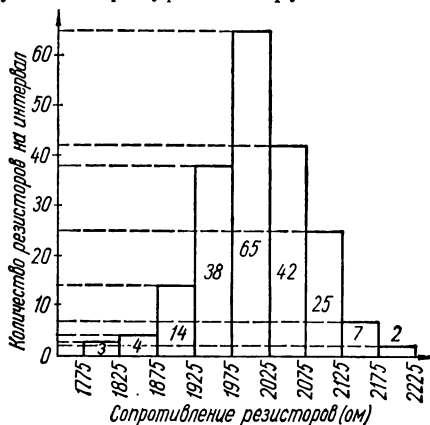


Рис. 5.2. Гистограмма распределения для интервального ряда.

Статистический смысл элементов полигона: точки на оси абсцисс — варианты, ординаты — частоты.

Для построения графика интервального ряда распределения возьмем второй пример — распределение сопротивлений в партии из 200 штук.

Пользуясь данными табл. 5.6, откладываем по оси абсцисс интервалы значений сопротивления резисторов, а по оси ординат в масштабе частоты — перпендикуляры — высоты прямоугольников. Поскольку основания всех прямоугольников равны, то их площади пропорциональны высотам. Площадь каждого прямоугольника равна частоте признака для данного интервала. Полученный график (рис. 5.2) называется гистограммой распределения.

Статистический смысл элементов гистограммы: точки на оси абсцисс — варианты; площади прямоугольников — частоты; каждая ордината представляет собой частоту, приходящуюся на единицу измерения признака. Эту характеристику называют плотностью частоты.

Приведенное выше построение гистограммы распределения основывалось на предположении, что плотность частоты неизменна внутри каждого интервала. Более справедливо было бы считать, что плотность частоты от интервала к интервалу меняется равномерно. На этом предположении основано построение интервальных рядов распределения в виде полигонов.

По оси абсцисс откладываем интервалы значений признака и отмечаем центры интервалов (рис. 5.3). Из центра каждого интервала восставляем ординату, пропорциональ-

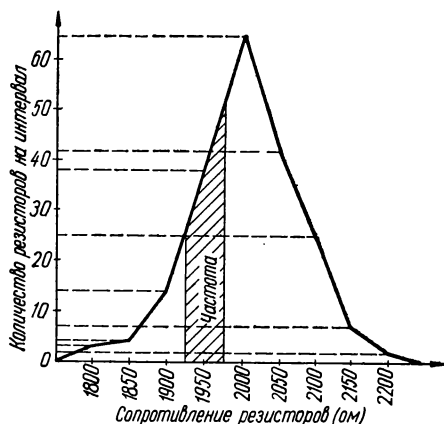


Рис. 5.3. Полигон распределения для интервального ряда.

ную плотности частоты. Соединив вершины ординат прямыми линиями, получаем графическое изображение интервального ряда в виде многоугольника. Статистический смысл элементов его: точки на оси абсцисс — варианты; ординаты — плотности частоты. Частоту можно определить как площадь многоугольника, ограниченного сверху отрезком полигона,

снизу — осью абсцисс, а слева и справа — ординатами, соответствующими границам интересующего нас интервала.

Полученный график интервального ряда распределения носит название полигона с ненагруженными ординатами в отличие от графика дискретного ряда (см. рис. 5.1), который называют полигоном с нагруженными ординатами.

При увеличении числа наблюдений график — гистограмма распределения вариантов статистической совокупности будет приближаться к кривой распределения исследуемой случайной величины. Очевидно, предельное положение графика распределения вариантов статистической совокупности представляет собой кривую плотности распределения.

По данным ряда распределения можно приближенно построить и статистическую функцию распределения. В статистике ее называют кумулятивной кривой, или кумулятой.

При построении кумулятивной кривой для дискретных распределений по оси абсцисс откладывают варианты признака. Из соответствующих точек восставляют ординаты, равные накопленным частотам признака (см. табл. 5.6). Соединив вершины ординат, получают кумулятивную кривую.

Если кумюлята строится для интервального ряда распределения, то нижней границе первого интервала соответствует нулевая частота. Верхней границе последнего интервала соответствует накопленная частота, равная объему статистической совокупности. На рис. 5.4 показана кумулятивная кривая, построенная для примера табл. 5.6.

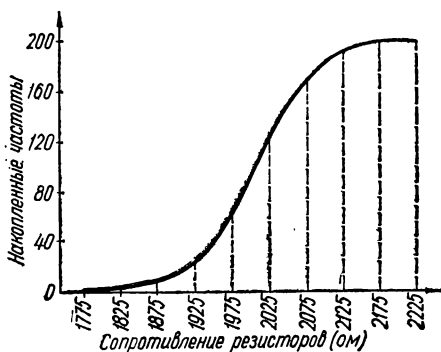


Рис. 5.4. Кумулятивная кривая.

Числовые характеристики рядов распределения

Для случайных величин были рассмотрены числовые характеристики: математическое ожидание, дисперсия, начальные и центральные моменты различных порядков. Для распределений статистических совокупностей тоже можно ввести подобные характеристики. Аналогично характеристике положения — математическому ожиданию случайной величины — для статистического распределения вводится среднее арифметическое наблюдаемых значений

$$\bar{x} = \frac{\sum x}{N}. \quad (5.2)$$

Если отдельные варианты в ряде распределения повторяются по несколько раз, нужно учесть частоту каждого варианта при вычислении средней арифметической

$$\bar{x} = \frac{\sum xm}{\sum m}, \quad (5.3)$$

где m — частоты вариантов x . Эта формула дает значение взвешенной средней арифметической, и числа m называют весами вариантов.

Аналогично можно определить и другие характеристики статистического распределения. Если в выражении для дисперсии случайной величины

$$D[X] = M[(X - m_x)^2]$$

заменить математическое ожидание средним арифметическим, получим статистическую дисперсию

$$D_x^* = \frac{\sum (x - \bar{x})^2}{N}. \quad (5.4)$$

Иногда в статистике для характеристики колеблемости пользуются величиной

$$\theta = \frac{\sum |x - \bar{x}|}{N},$$

которая называется средним абсолютным отклонением.

Чаще всего рассеяние характеризуют средним квадратическим отклонением

$$\sigma = \sqrt{D_x^*} = \sqrt{\frac{\sum (x - \bar{x})^2}{N}}.$$

Для рядов распределения среднее квадратическое отклонение определяется как взвешенный показатель

$$\sigma = \sqrt{\frac{\sum (x - \bar{x})^2 m}{\sum m}}.$$

Некоторые свойства статистических параметров

Можно вычислять \bar{x} и σ значительно проще, если использовать некоторые элементарные свойства этих статистических параметров. Сформулируем их.

С в о й с т в о 1. Если все частоты (веса) умножить на одно и то же число, то \bar{x} и σ не изменятся. Это свойство очевидно, так как если в формулах

$$\bar{x} = \frac{\sum xm}{\sum m}; \quad \sigma = \sqrt{\frac{\sum (x - \bar{x})^2 m}{\sum m}}$$

все веса m умножать на постоянное число k , то это число вынесется на знак суммы в числителе и знаменателе и сократится. Из этого свойства следует, между прочим, что, вычисляя \bar{x} и σ , мы можем воспользоваться относительными частотами вместо абсолютных. Результат будет тот же.

Свойство 2. Если все варианты признака умножить на одно и то же число, то \bar{x} умножится на это число, а σ умножится на модуль этого числа.

Для доказательства рассмотрим новые варианты $z = kx$, где k — произвольное число (положительное или отрицательное). Получим новую среднюю арифметическую

$$\bar{z} = \frac{\Sigma z}{n} = \frac{\Sigma kx}{n} = k \frac{\Sigma x}{n} = k\bar{x}$$

и новое среднее квадратическое отклонение

$$\begin{aligned}\sigma^2 &= \sqrt{\frac{\Sigma (z - \bar{z})^2}{n}} = \sqrt{\frac{\Sigma (kx - k\bar{x})^2}{n}} = \sqrt{\frac{k^2 \Sigma (x - \bar{x})^2}{n}} = \\ &= |k| \sqrt{\frac{\Sigma (x - \bar{x})^2}{n}} = |k| \sigma.\end{aligned}$$

Свойство 3. Если по всем вариантам прибавить одно и то же число a , то \bar{x} увеличится на a , а σ не изменится. Действительно, при $z = (x + a)$ имеем

$$\bar{z} = \frac{\Sigma (x + a)}{n} = \frac{\Sigma x}{n} + \frac{\Sigma a}{n},$$

но

$$\frac{\Sigma x}{n} = \bar{x}, \quad \frac{\Sigma a}{n} = \frac{na}{n} = a,$$

следовательно, $\bar{z} = \bar{x} + a$.

Далее получаем

$$\begin{aligned}\sigma_z &= \sqrt{\frac{\Sigma (z - \bar{z})^2}{n}} = \sqrt{\frac{\Sigma (x + a - \bar{x} - a)^2}{n}} = \\ &= \sqrt{\frac{\Sigma (x - \bar{x})^2}{n}} = \sigma.\end{aligned}$$

Свойство 4. Дисперсия равна среднему квадрату минус квадрат средней

$$\sigma^2 = \frac{\Sigma x^2}{n} - (\bar{x})^2 \quad (5.5)$$

и для взвешенных показателей

$$\sigma^2 = \frac{\Sigma x^2 m}{\Sigma m} - (\bar{x})^2. \quad (5.6)$$

Для формулы (5.5) имеем

$$\sigma^2 = \frac{\Sigma (x - \bar{x})^2}{n} = \frac{\Sigma x^2}{n} - \frac{\Sigma 2x\bar{x}}{n} + \frac{\Sigma (\bar{x})^2}{n}.$$

Вынесем $2\bar{x}$ за знак суммы. Учитывая, что $\Sigma (x)^2 = n (\bar{x})^2$, получим

$$\begin{aligned}\sigma^2 &= \frac{\Sigma x^2}{n} - 2\bar{x} \frac{\Sigma x}{n} + \frac{n (\bar{x})^2}{n} = \frac{\Sigma x^2}{n} - 2 (\bar{x})^2 + (\bar{x})^2 = \\ &= \frac{\Sigma x^2}{n} - (\bar{x})^2.\end{aligned}$$

Формула (5.6) доказывается аналогично. Пользуясь свойством (4), можно заменить расчет квадратов отклонений от средней вычислением квадратов самих вариантов. Это зачастую значительно проще.

Основные свойства средней арифметической

Свойство 1 (нулевое). Сумма отклонений от средней арифметической равна нулю. Запишем

$$\Sigma (x - \bar{x}) = \Sigma x - \Sigma \bar{x},$$

но

$$\Sigma x = n\bar{x}, \text{ так как } \bar{x} = \frac{\Sigma x}{n}, \text{ и}$$

$$\Sigma \bar{x} = n\bar{x}.$$

Следовательно,

$$\Sigma (x - \bar{x}) = n\bar{x} - n\bar{x}.$$

Свойство 2 (минимальное). Сумма квадратов отклонений от средней арифметической меньше, чем сумма квадратов отклонений от любого другого числа.

Обозначим через f сумму квадратов отклонений вариантов от числа c и найдем значение c , для которого функция f минимальна

$$f = \Sigma (x - c)^2 = \min. \quad (5.7)$$

Необходимое условие минимума заключается в том, чтобы

$$\frac{df}{dc} = 0.$$

Дифференцируя, имеем

$$-2\Sigma (x - c) = 0, \text{ или } \Sigma (x - c) = 0,$$

$$\Sigma x = \Sigma c, \quad \Sigma x = nc,$$

откуда

$$c = \frac{\sum x}{n} = \bar{x}.$$

Условие (5.7) называется критерием наименьших квадратов, который широко применяется в практике исследований и управления.

Заметим, что, строго говоря, кроме условия

$$\frac{df}{dc} = 0$$

для удовлетворения критерия наименьших квадратов необходимо иметь положительный знак второй производной

$$\frac{d^2f}{dc^2} > 0.$$

Это условие, очевидно, удовлетворяется, так как

$$\frac{d^2f}{dc^2} = \frac{d}{dc} [-2\sum (x - c) = 2\sum] = 2n > 0.$$

Смысл критерия наименьших квадратов заключается в следующем. Отыскиваем число c , которое ближе всего к заданной совокупности значений x . Близость c к x измеряется квадратом отклонения $(x - c)^2$. А поскольку есть целый ряд значений x , то естественно потребовать, чтобы сумма квадратов отклонений была наименьшей.

Оказывается, что число, которое ближе всего (в смысле критерия наименьших квадратов) к заданной совокупности значений, является средней арифметической этих значений.

Минимальное свойство характеризует связь между средней арифметической и средним квадратическим отклонением. Среднее квадратическое отклонение, вычисленное от средней арифметической, оказывается меньше, чем вычисленное от какой-либо другой величины (например медианы или моды)

$$\sqrt{\frac{\sum (x - \bar{x})^2}{n}} < \sqrt{\frac{\sum (x - c)^2}{n}},$$

если $c \neq \bar{x}$.

Наонец, заметим, что начальные и центральные моменты любых порядков для статистического распределения вычисляются по формулам

$$\begin{aligned} \alpha_s^* &= \frac{\sum x^s}{N}; \\ \mu_s^* &= \frac{\sum (x - \bar{x})^s}{N}. \end{aligned} \quad (5.8)$$

При увеличении объема экспериментальных данных все статистические характеристики, очевидно, будут сходиться по вероятности к соответствующим математическим характеристикам.

Любая статистическая совокупность содержит элемент случайности. Только при весьма большом количестве наблюдений эта случайность сглаживается, и существующая закономерность проявляется достаточно отчетливо. Но увеличение числа опытов не всегда возможно и поэтому приходится применять косвенные методы для определения кривой распределения. Задача подбора теоретической кривой распределения носит название задачи выравнивания рядов распределения.

Выравнивание рядов распределения

Необходимо подобрать теоретическую кривую распределения так, чтобы она наилучшим образом, в определенном смысле, отражала закономерности исследуемого статистического распределения.

Неопределенность задачи выравнивания рядов распределения заключается в требовании «наилучшего» приближения. Решение зависит от выбора критерия. Чаще всего таким критерием приближения служит критерий наименьших квадратов. При этом наилучшим приближением к эмпирической зависимости в данном классе функций является такое приближение, при котором сумма квадратов отклонений эмпирической линии от теоретической минимальна. Более подробно познакомимся с применением этого критерия в следующем разделе, посвященном задачам статистического измерения связи.

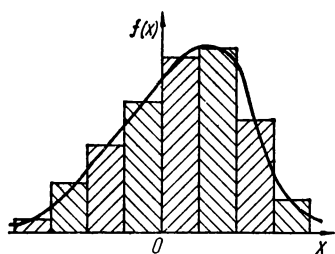


Рис. 5.5. Сглаживание рядов распределения.

Вопрос согласованности теоретического и эмпирического распределений — это вопрос проверки правдоподобия гипотез. В самом деле, как бы удачно ни была выбрана теоретическая кривая (рис. 5.5), как бы точно ни были рассчитаны ее параметры между нею и статистическим распределением, неизбежны расхождения. Требуется ответить на вопрос: эти расхождения объясняются элементом слу-

чайности, недостаточным числом наблюдений или же это существенные отклонения из-за плохого выбора теоретической кривой? Решение этого вопроса дает применение критериев согласия.

Закон распределения случайной величины X можно представить функцией распределения $\Phi(x)$, плотностью распределения $f(x)$ или совокупностью вероятностей P_i . Наиболее общей формой представления является функция распределения $\Phi(x)$, поэтому будем рассматривать некоторую гипотезу H , которая состоит в том, что величина X распределена по закону $\Phi(x)$.

Пусть расхождение теоретического и статистического распределения характеризуется некоторой величиной U . Этот показатель можно выбрать в виде суммы квадратов отклонений теоретических вероятностей от соответствующих частот, в виде суммы тех же квадратов с некоторыми весовыми коэффициентами или же максимального отклонения статистической функции распределения от теоретической и т. д.

При таком выборе показатель U , очевидно, представляет собой случайную величину. Закон распределения этой случайной величины зависит от закона распределения X и от числа испытаний N . Если гипотеза H соответствует действительности, то распределение U определяется функцией $\Phi(x)$ и числом N .

Пусть в результате ряда испытаний показатель отклонения U принял некоторое значение u и пусть закон распределения величины U известен. Требуется определить, является ли расхождение случайным или же это расхождение слишком велико и объясняется наличием существенной разницы между статистическим распределением и выбранной теоретической кривой. Предположим, что гипотеза H верна, т. е. u объясняется случайными причинами. С учетом этого предположения вычислим вероятность того, что за счет случайных причин, связанных с недостаточным объемом статистической совокупности, показатель отклонения U окажется не меньше, чем наблюдаемое при испытании значение u .

Если вероятность $P = (U \geq u)$ мала, гипотезу H можно считать маловероятной. Значительная вероятность события $U \geq u$ свидетельствует о правдоподобии H .

Показатель отклонения U можно выбрать так, чтобы закон распределения этой величины практически не зависел от $\Phi(x)$ при достаточно большом n . Рассмотрим один

из наиболее распространенных критериев согласия — критерий χ^2 Пирсона.

Пусть результаты N независимых испытаний представлены рядом распределения, состоящим из k интервалов (табл. 5.7). Пусть из теоретических соображений выбран закон распределения $\Phi(x)$ случайной величины X . Зная теоретический закон распределения, можно определить вероятности попадания случайной величины в каждый интервал

$$P_1, P_2, \dots, P_k.$$

Расхождение между теоретическим и статистическим распределениями будем характеризовать суммой квадратов

Таблица 5.7

Варианты	x_1, x_2	x_2, x_3	x_k, x_{k+1}
Частоты	n_1	n_2	n_k

отклонений $(v_i - p_i)$, взятых с некоторыми коэффициентами c_i

$$U = \sum_{i=1}^k c_i (v_i - p_i)^2, \quad (5.9)$$

где v — статистическая вероятность или относительная частота, равная n/N .

Весовые коэффициенты необходимы, потому что отклонения, относящиеся к различным интервалам, нельзя считать равноправными. Веса c_i берут обратно пропорциональными вероятностям интервалов. Причем коэффициент пропорциональности обусловлен выражением

$$c_i = \frac{N}{p_i}. \quad (5.10)$$

При таком выборе, как показал Пирсон, закон распределения показателя U практически не зависит от функции распределения $\Phi(x)$ и от числа опытов N , а зависит только от числа интервалов k . Этот закон при увеличении N приближается к так называемому распределению χ^2 . Показатель отклонения U при таком выборе коэффициентов c_i

обозначают χ^2 и определяют по формуле

$$\chi^2 = n \sum_{i=1}^k \frac{(N_i - p_i)^2}{p_i}. \quad (5.10)$$

С учетом $v_i = \frac{n_i}{N}$, где n_i абсолютная частота i -го интервала, получаем

$$U = \chi^2 = \sum_{i=1}^k \frac{(n_i - N p_i)^2}{N p_i}. \quad (5.11)$$

Распределением χ^2 с r «степенями свободы» называется распределение суммы квадратов r независимых случайных величин, каждая из которых подчинена нормальному закону с нулевым математическим ожиданием и единичной дисперсией. Плотность распределения χ^2 имеет вид

$$f_r(u) = \begin{cases} \frac{1}{2^{\frac{r}{2}} \Gamma\left(\frac{r}{2}\right)} u^{\frac{r}{2}-1} \cdot e^{-\frac{u}{2}} & \text{при } u > 0 \\ 0 & \text{при } u < 0, \end{cases}$$

где $\Gamma(\alpha) = \int_0^{\infty} t^{\alpha-1} e^{-t} dt$ — известная гамма-функция.

Параметр r называется числом «степеней свободы» распределения. Количество «степеней свободы» r равно числу интервалов k минус число независимых условий, наложенных на частоты v_i . Можно, например, наложить условие

$$\sum_{i=1}^k v_i = 1, \quad \cdot$$

это условие необходимо выполнять во всех случаях.

Если теоретическая кривая подбирается из условия совпадения средних значений, накладывается условие

$$\sum_{i=1}^k \bar{x}_i v_i = m \bar{x}.$$

Если, кроме того, потребовать совпадения дисперсий теоретического и статистического распределения, то необходимо выполнить условие

$$\sum_{i=1}^k (\bar{x}_i - \bar{x})^2 v_i = D_x,$$

и т. д.

Существуют специальные таблицы распределения χ^2 , с помощью которых для каждого значения χ^2 и числа «степеней свободы» r можно найти вероятность P того, что величина, распределенная по закону χ^2 , превзойдет это значение.

Насколько мала должна быть вероятность P для того, чтобы опровергнуть гипотезу, вопрос неопределенный. Практически установлено, что если $P < 0,1$, целесообразно проверить эксперимент и в случае повторного появления существенных отклонений искать более подходящий закон

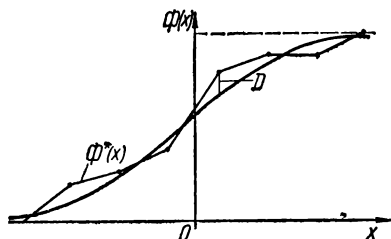


Рис. 5.6. Сглаживание статистической функции распределения.

распределения. При пользовании критерием χ^2 необходимо как достаточно большое количество наблюдений N , так и существенное количество наблюдений внутри каждого интервала (не менее $5 \div 10$). Если в отдельных интервалах мало наблюдений ($1 \div 3$), рекомендуется некоторые интервалы объединять.

Критерием согласия для оценки степени совпадения теоретического распределения со статистическим является критерий А. Н. Колмогорова. Показателем отклонения служит максимальное значение модуля разности между статистической функцией распределения $\Phi^*(x)$ и аппроксимирующей теоретической (рис. 5.6)

$$U = D = \max |\Phi^*(x) - \Phi(x)|. \quad (5.12)$$

Этот показатель отклонения очень просто вычисляется и имеет простой закон распределения. А. Н. Колмогоров показал, что независимо от $\Phi(x)$ при неограниченном возрастании числа независимых наблюдений n вероятность события

$$D\sqrt{n} \geq \lambda$$

стремится к пределу

$$P(\lambda) = 1 - \sum_{k=-\infty}^{\infty} (-1)^k e^{-2k^2 \lambda^2}. \quad (5.13)$$

Значения $P(\lambda)$ приведены в таблице 5.8.

Если $P(\lambda)$ оказывается малой, гипотезу следует отвергнуть; при значительных $P(\lambda)$ можно считать, что она не противоречит опытным данным.

Критерий А. Н. Колмогорова значительно проще критерия χ^2 , но его можно применить только в случае, когда $\Phi(x)$ полностью известна заранее. Причем известен не только вид функции, но и все входящие в нее параметры. Это на практике встречается редко. Обычно из теоретических соображений известен только вид $\Phi(x)$. Параметры же определяются по конкретному числовому материалу.

Таблица 5.8

λ	$P\lambda$	λ	$P\lambda$	λ	$P\lambda$
0,0	1,000	0,7	0,711	1,4	0,040
0,1	1,000	0,8	0,544	1,5	0,022
0,2	1,000	0,9	0,393	1,6	0,012
0,3	1,000	1,0	0,270	1,7	0,006
0,4	0,997	1,1	0,178	1,8	0,003
0,5	0,964	1,2	0,112	1,9	0,002
0,6	0,864	1,3	0,068	2,0	0,001

В заключение следует обратить внимание на тот факт, что применение критериев согласия (в том числе и рассмотренных) лишь в некоторых случаях позволяет опровергнуть выбранную гипотезу. Если же вероятность P велика, то это еще не доказывает справедливость гипотезы H . В этом случае можно говорить лишь о том, что принятая гипотеза не противоречит опытным данным.

Рассмотренные задачи статистики описываются в рамках теории вероятностей и связаны с понятием случайной величины. Если же мы имеем дело с несколькими случайными величинами, необходимо дать характеристики не только каждой в отдельности, но и охарактеризовать их взаимосвязь. В статистическом анализе вопросы измерения связи между различными показателями реальных явлений рассматриваются в теории корреляции.

Контрольные вопросы и задания

1. Сформулируйте основные задачи статистического анализа.
2. Что такое ряд распределения? Какая разница между дискретным и интервальным рядом?
3. Покажите, как графически представляются ряды распределения.
4. Перечислите и охарактеризуйте основные статистические параметры.
5. Какие важные свойства статистических параметров вы знаете?
6. Сформулируйте критерий наименьших квадратов.
7. Что представляет собой задача выравнивания (сглаживания) рядов распределения?

§ 3. СТАТИСТИЧЕСКОЕ ИЗМЕРЕНИЕ СВЯЗИ

В любой отрасли науки и техники приходится изучать зависимость между различными показателями. Причем качественное проявление этих зависимостей стремятся выразить в количественной форме. В математическом анализе, классической физике рассматриваются методы количественного измерения связи — функциональной зависимости.

Как определяется функциональная зависимость? В основе лежит идея однозначного соответствия между величинами. Каждому значению одной переменной — аргумента — строго соответствует значение другой переменной — функции.

Таблица 5.9

$I(a)$	$U(e)$
0,1	50,0
0,2	100,0
0,3	150,0
0,4	200,0
0,5	250,0
0,6	300,0

Таблица 5.10

$I(a)$	$U(e)$				
0,1	49,9	49,8	49,9	50,1	50,0
0,2	100,1	100,1	99,9	99,9	99,8
0,3	150,2	150,0	150,0	151,0	150,1
0,4	200,1	199,8	200,2	200,0	199,9
0,5	250,0	250,2	250,0	250,1	249,8
0,6	300,2	300,0	300,1	299,8	300,0

Однако каждый экспериментатор знает, что это понятие является лишь абстракцией. Как бы точно ни производился эксперимент, как бы строго ни закреплялись условия опыта и побочные факторы, неизбежен разброс результатов. И нельзя прийти к однозначным выводам об интересующей нас зависимости.

Известный в электротехнике закон Ома связывает функциональной зависимостью падение напряжения на активном сопротивлении и силу тока, протекающего по этому сопротивлению

$$U = IR.$$

В соответствии с этим законом функция $U = f(I)$ при постоянной величине R , например при $R = 500 \text{ ом}$, в табличном виде может быть представлена таблицей 5.9. Но если выполнить ряд измерений падения напряжения на этом сопротивлении, задавая указанные значения силы тока, то практически при каждом опыте будут получаться несколько отличные результаты. Наблюдается колеблемость функции $U = f(I)$, например, как в таблице 5.10. Каковы же причины колеблемости функции? Дело в том, что наблюдаемая величина падения напряжения U на самом деле функция не одного только аргумента — силы тока.

Она зависит и от многих других факторов (аргументов): качества сопротивления, изменений температуры, стабильности источников питания, погрешностей измерительных приборов и т. д. Многие условия изменяются от опыта к опыту, колеблются по своим собственным законам. Эти колебания часто случайны, никак не связаны с интересующей нас зависимостью. Тем самым они обуславливают случайную колеблемость интересующей нас функции. Если бы удалось закрепить все посторонние факторы, полностью повторять в последующем опыте условия предыдущего, можно было бы получить строгую зависимость. Но это практически невозможно. Часть посторонних влияний с трудом поддается контролю, а часть вообще ускользает от внимания.

Тем не менее, если колеблемость невелика, точность приемлема (дальние знаки после запятой), можно с известным приближением считать зависимость функциональной.

Во многих областях научных и технических исследований и особенно в практических исследованиях на производстве, в экономике, в социологии подчас невозможно устранить влияние посторонних факторов. Во-первых, многие из них часто неизвестны. Во-вторых, затруднен, а иногда и вовсе невозможен эксперимент (например, в экономических и социологических исследованиях). Приходится ограничиваться наблюдениями явлений в их естественных условиях.

Так возникают две основные задачи измерения связи:

1. Определить на основе большого количества данных, как изменялась бы функция при изменении одного из своих аргументов, если бы другие ее аргументы не изменялись. Причем задача должна решаться на материале, где прочие аргументы на самом деле изменяются и своей изменчивостью искажают интересующую нас зависимость.

2. Определить степень искажающего влияния прочих факторов на интересующую нас зависимость. Другими словами, нужно определить силу, с которой данная зависимость проявляется среди многообразия нарушающих ее воздействий.

Корреляционная зависимость

Задачи статистического измерения связи всегда решаются при заданном числе учитываемых признаков. Остальные признаки, значения которых для каждого элемента совокупности неизвестны, — неучитываемые.

Рассмотрим пример, на который мы будем опираться в дальнейшем при решении основных задач измерения связи.

Сложная радиоэлектронная аппаратура после сборки и наладки на предприятии проходит стадию испытательной работы или тренировки. В процессе тренировки устра-

Таблица 5.11

$\begin{matrix} X \\ \backslash \\ Y \end{matrix}$	50—70	70—80	90—110	110—130	130—150	Итого
130—150		1	4	6	15	26
110—130		5	20	13	4	42
90—110	2	23	49	7	4	85
70—90	2	16	16	2	2	38
50—70	6	3				9
Итого	10	48	89	28	25	200

няются различные скрытые и ранее не обнаруженные дефекты, заменяются детали низкого качества. От тренировки существенно зависит эксплуатационная надежность приборов.

В качестве показателя надежности можно воспользоваться временем безотказной работы прибора, причем гарантийное время примем за 100 %

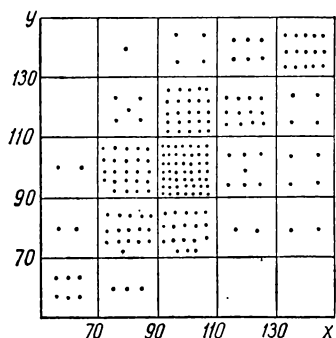


Рис. 5.7. Поле корреляции.

надежности: Нас интересует зависимость надежности (в % к гарантийному сроку) от тренированности аппаратуры (в % к номинальному сроку, установленному техническими условиями).

Интересующую нас функцию — надежность — обозначим Y , а аргумент — тренированность — X .

Для исследования этой зависимости воспользуемся данными о партии из 200 приборов. Для отдельных приборов этой партии значения Y менялись в пределах 50—150 и значения X также менялись в пределах 50—150. Эти данные представлены в виде таблицы 5.11. В табл. 5.11 каждому фиксированному значению аргумента X соответствует не одно, а несколько значений функции Y . Материал мож-

но представить графически, как показано на рис. 5.7. Результаты каждого наблюдения отмечаются точкой в системе координат.

Глядя на представленный материал, можно сразу же сказать, что между Y и X существует зависимость. Однако то не функциональная зависимость и ей надо дать определение.

Считается, что Y корреляционно зависит от X , если:

1) каждому значению аргумента X соответствует ряд распределения функции Y и

2) с изменением X эти ряды закономерно изменяют свое положение.

Представление исследуемой зависимости в виде табл. 5.11 называется корреляционной таблицей. Соответственно построение на рис. 5.7 дает нам поле корреляции.

Если с изменением X ряды не изменяют своего положения или изменяют его случайно, то Y корреляционно не зависит от X .

§ 4. ИССЛЕДОВАНИЕ ФОРМЫ СВЯЗИ. ЭМПИРИЧЕСКАЯ ЛИНИЯ РЕГРЕССИИ

Мы установили, что корреляционная зависимость характеризуется закономерным смещением рядов распределения функции при изменении значений аргумента. Смещение может быть более или менее быстрым. Положение рядов может изменяться в сторону увеличения или в сторону уменьшения. При исследовании корреляционной зависимости необходимо определить, в какую сторону и с какой скоростью смещаются ряды распределения функции на тех или иных участках изменения аргумента. Для этого нужна точная оценка положения рядов распределения на оси y . Такой оценкой являются средние показатели.

Рассчитаем для приведенного ранее примера (см. табл. 5.11) средние показатели для всех рядов распределения y , соответствующих заданным значениям x . В теории корреляции вычисления обычно производятся по формулам средних арифметических.

Для первого ряда распределения (при $x = 60$):

y	n	ny
60	6	360
80	2	160
100	2	200
Итого	10	720

Для второго ряда распределения (при $x = 80$):

y	n	n_j
60	3	180
80	16	1280
100	23	2300
120	5	600
140	1	140
Итого	48	4500

$$\bar{y}_2 = \frac{\sum ny}{\sum n} = \frac{4500}{48} \approx 93,8.$$

Удобнее рассчитывать средние \bar{y}_i одновременно для всех рядов. Расчеты располагают в общей таблице (табл. 5.12). При этом целесообразно пользоваться упрощенным способом вычислений, который заключается в следующем.

Таблица 5.12

y'	$x \backslash y$	60	80	100	120	140	Итого
2	140		2 1	8 4	12 6	30 15	26
1	120		5 5	20 20	13 13	4 4	42
0	100	0 2	0 23	0 49	0 7	0 4	85
-1	80	-2 2	-16 16	-16 16	-2 2	-2 2	38
-2	60	-12 6	-6 3				9
Итого	n_i	10	48	89	28	25	200
	$\sum y'_i n$	-14	-15	12	23	32	
	\bar{y}'_i	-1,4	-0,31	0,13	0,82	1,28	
	\bar{y}_i	72	93,8	102,6	116,4	125,6	

Варианты y заменяем упрощенными вариантами y' в соответствии с преобразованием

$$y' = \frac{y - C_y}{i_y},$$

где C_y — новое начало отсчета; i_y — интервал группировки по Y .

Выберем $C_y = 100$; $i_y = 20$. Упрощенные варианты y' заносим в левый столбец табл. 5.12.

Частоту в каждой клетке корреляционной таблицы умножаем на соответствующее значение y' и произведение записываем в правом верхнем углу клетки. Итоги этих произведений по столбцам $\Sigma y' n$ записываем в нижней строке таблицы. Каждый итог затем делим на число наблюдений в столбце n_i . Полученные упрощенные средние \bar{y}_i записываем строкой ниже. Искомые средние \bar{y} определяем из выражения

$$\bar{y}_i = i_y \bar{y}'_i + C_y,$$

или в нашем случае

$$\bar{y}_i = 20 \bar{y}'_i + 100.$$

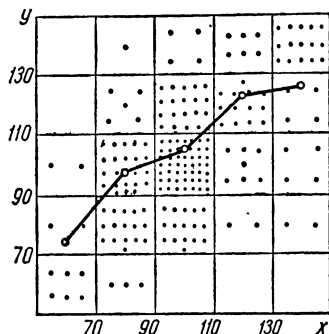


Рис. 5.8. Эмпирическая линия регрессии.

Переносим результаты расчета на поле корреляции (рис. 5.8). Из середины интервалов аргумента восстанавливаем ординаты, соответствующие значениям \bar{y}_i . Вершины ординат соединяем прямыми. Полученная ломаная линия называется эмпирической линией регрессии y по x . Она характеризует смещение рядов распределения y с увеличением x , т. е. показывает, как в среднем изменяется y с увеличением x .

На рис. 5.8 отчетливо видна тенденция к росту надежности аппаратуры в связи с увеличением времени тренировки. Линия регрессии имеет зигзаги, носящие случайный характер. Эмпирическую линию регрессии необходимо рассматривать в поле корреляции на фоне отдельных точек, которые она осредняет. Очевидно, с большим «доверием» следует отнестись к тем участкам линии регрессии, которые проходят в «густых» местах корреляционного поля. Менее достоверны те части линии, которым соответствуют «разреженные» места поля (малое количество наблюдений).

В соответствии с законом больших чисел можно утверждать, что при увеличении числа наблюдений эмпирическая линия регрессии будет все точнее отражать исследуемую закономерность.

Предельное положение, к которому стремится эмпирическая линия регрессии при неограниченном увеличении числа наблюдений, называют предельной теоретической линией регрессии.

Учитывая изложенное, можно уточнить данное ранее определение корреляционной зависимости.

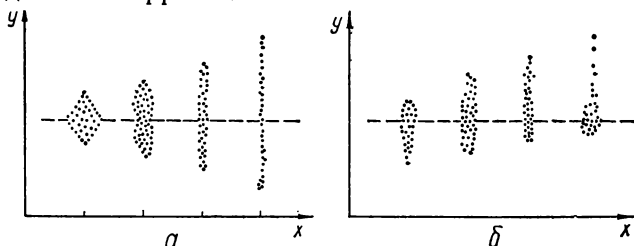


Рис. 5.9. Статистическая зависимость:

а — по дисперсии; б — по асимметрии.

Переменная y находится в корреляционной зависимости от x , если:

- 1) каждому значению аргумента x соответствует ряд распределения функции y и
- 2) предельная теоретическая линия регрессии y по x не параллельна оси абсцисс.

Более общим является понятие статистической зависимости. Переменная y считается статистически зависимой от x , если:

- 1) каждому значению аргумента x соответствует ряд распределения функции y и
- 2) с изменением x закономерно изменяются статистические характеристики этих рядов. Другими словами, при статистической зависимости может изменяться не только положение рядов распределения, но и другие параметры рядов, например, рассеяние, степень асимметрии и т. п.

Таким образом, корреляционная зависимость — это частный случай зависимости статистической. Любая корреляционная зависимость является статистической, но не всякая статистическая зависимость корреляционная. На рис. 5.9 представлены примеры статистических зависимостей, которые не являются в то же время зависимостями корреляционными.

В первом примере (рис. 5.9, а) с изменением x изменяется рассеяние рядов распределения y . Во втором примере (рис. 5.9, б) с изменением x изменяется асимметрия рядов распределения y . В том и другом случае линия регрессии y по x параллельна оси абсцисс, следовательно, переменная y корреляционно не зависит от x .

Если с изменением значений аргумента x ряды распределения y не изменяются или их характеристики изменяются случайным образом, то переменная y статистически не зависит от x .

Иногда характеристика положения ряда с помощью средней арифметической недостаточна. Тогда рассчитывают другие характеристики положения (моды, медианы). В подобных случаях всегда следует указывать, применительно к какой форме средних строится линия регрессии. Если же нет дополнительных указаний на форму средних, считается, что используются средние арифметические.

Ранее были определены две основные задачи теории корреляции: определение формы связи и определение тесноты связи. При определенных условиях линия регрессии представляет собой решение первой из этих задач. Требуется установить, какова была бы зависимость между функцией и одним из ее аргументов, если бы прочие аргументы этой функции не изменялись. При этом для решения задачи мы располагаем материалом, где прочие аргументы на самом деле изменяются — варьируют. Понятие «прочие аргументы» довольно неопределенное. Для того чтобы поставленная задача имела смысл, необходимо это понятие уточнить, поскольку различное определение «прочих аргументов» может привести к различным выражениям исследуемой зависимости. Рассмотрим пример.

При массовом изготовлении тороидальных сердечников (рис. 5.10) размеры сердечника варьируют.

Между весом сердечника и другими варьирующими признаками существует зависимость

$$g = 2\pi^2 R r^2 \gamma (1 + \epsilon), \quad (5.14)$$

где g — вес сердечника; r — радиус сечения; R — радиус осевой окружности тора; γ — удельный вес материала; $(1 + \epsilon)$ — коэффициент искажения формы (включающий и ошибки измерения).

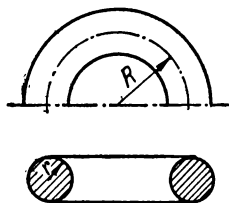


Рис. 5.10. Тороидальный сердечник.

Из формулы (5.14) выходит, что вес сердечника g изменялся бы пропорционально r^2 , если бы «прочие аргументы» — R , γ , $(1 + \epsilon)$ не изменялись.

Но вместо (5.14) для определения веса сердечника можно воспользоваться выражением

$$g = \frac{1}{2} S r \gamma (1 + \epsilon), \quad (5.15)$$

где S — величина боковой поверхности, равная $4\pi^2 R r$. Из формулы (5.15) можно заключить, что вес g изменялся бы пропорционально r , если бы «прочие аргументы» — S , γ , $(1 + \epsilon)$ — не изменялись.

Получены два существенно различных ответа на поставленный вопрос. Причина этого в том, что не была учтена зависимость между основным и «прочим» аргументами. Очевидно, если мы будем закреплять аргументы, зависящие от основного, получится иное решение задачи по сравнению с тем, когда закрепляются аргументы, не зависящие от основного.

Таким образом, «прочие» факторы или аргументы, от которых зависит наша функция, можно разделить на два класса:

- 1) побочные факторы, зависящие от основного X ;
- 2) побочные факторы, не зависящие от основного.

Скажем, в нашем примере надежность рассматривается как функция, а тренированность — как основной аргумент. Очевидно, что надежность зависит еще от качества монтажа, транспортировки и механических воздействий, режима эксплуатации, надежности отдельных элементов прибора при тех или иных отклонениях питания, нагрузки и т. д. Все эти факторы (аргументы) отнесены к «прочим». Среди них есть аргументы, не зависящие от основного, например, режим эксплуатации, механические повреждения. Есть и зависящие от основного аргумента, например, надежность отдельных элементов при различных нагрузках или изменении питания.

Очевидно, и в данном случае при закреплении аргументов, зависящих от основного, получится решение, отличное от того, которое мы получили бы, закрепляя аргументы, не зависящие от основного.

При формулировке основной задачи всегда необходимо оговорить, как зависят «прочие» аргументы от аргумента, влияние которого на функцию мы собираемся проследить.

Схема А—М—Н. Введем следующие допущения:

а) будем считать, что влияние прочих аргументов на функцию сочетается с влиянием x по законам сложения и умножения

$$y = u = v\varphi(x) + w\psi(x) + \dots, \quad (5.16)$$

где $\varphi(x)$, $\psi(x)$ — функции x (x^2 , $\sin x$, $\ln x$...); u , v , w — прочие аргументы y , каждый из которых может оказаться результатом действия многих факторов;

б) «прочие» аргументы будем считать корреляционно не зависящими от основного аргумента x .

При выполнении условий а и б считают, что «прочие» аргументы действуют аддитивно (сложением), мультипликативно (умножением) и независимо от аргумента x , короче говоря, по схеме А—М—Н.

Теперь сформулируем задачу определения формы связи следующим образом.

Определить, какой была бы зависимость между функцией y и одним из ее аргументов x , если бы прочие аргументы функции, действующие по схеме А—М—Н, не изменялись. Причем определить эту зависимость на материале, где прочие аргументы на самом деле изменяются, а их значения нам неизвестны. Решение поставленной задачи дается предельной теоретической линией регрессии y по x .

Рассмотренный метод определения регрессии y по x позволяет освободиться от влияния всех аргументов функции y , действующих совместно с аргументом x по схеме А—М—Н. При этом нет необходимости знать конкретные значения этих аргументов и их материальную природу.

В том случае, когда аргументы u , v , w , ... корреляционно связаны с x или, в общем случае, не действуют по схеме А—М—Н, метод регрессии не позволяет освободить зависимость между y и x от влияния колеблемости этих аргументов. Приходится применять более сильные статистические методы, например метод множественной корреляции, который требует дополнительно учитывать все признаки, находящиеся с x в корреляционной связи. В некоторых задачах можно получить хорошее решение при помощи метода функциональной корреляции, который будет рассмотрен далее.

Теоретическая линия регрессии

В соответствии с определением предельной теоретической линии регрессии для ее нахождения необходимо увеличивать число наблюдений (одновременно сокращая

интервалы группировки) до тех пор, пока не выявится с достаточной точностью закономерность, лежащая в основе изучаемого процесса. Но на практике всегда приходится иметь дело с ограниченным и иногда небольшим числом наблюдений. Чтобы в этих условиях обнаружить закономерность, проявляющуюся отчетливо лишь в массе наблюдений, приходится пользоваться косвенными приемами анализа.

При этом получается не действительная предельная теоретическая линия регрессии, а лишь приближенная ее оценка, которую называют просто «теоретической линией регрессии». Термин «предельная» относится к действительной линии, к которой стремится эмпирическая при увеличении числа наблюдений.

Процесс нахождения теоретической линии регрессии заключается в обоснованном выборе аппроксимирующей кривой и расчете параметров ее уравнения.

Предельная теоретическая линия регрессии представляет собой плавную кривую. Кривая выражается математическим уравнением того или иного вида. В таком же виде дается ее оценка, т. е. теоретическая линия регрессии. Процесс ее нахождения называют выравниванием эмпирической линии регрессии.

Чаще всего для этой цели используют кривые, уравнения которых выражаются многочленами целых положительных степеней

$$\begin{aligned}\bar{y}_x &= a + bx; \\ \bar{y}_x &= a + bx + c \cdot x^2; \\ \bar{y}_x &= a + bx + cx^2 + dx^3\end{aligned}$$

и т. д.

Известно, что любую сколь угодно сложную непрерывную функцию на заданном отрезке изменения аргумента можно с необходимой точностью аппроксимировать многочленом n -го порядка, при достаточно большом n и правильно выбранных параметрах. Семейство многочленов настолько разнообразно, что может показаться, будто нет необходимости использовать для выравнивания другие формулы. Но это не так.

Как бы точно многочлен ни воспроизводил зависимость в заданном интервале изменения аргумента, это не свидетельствует о том, что вне этого интервала данный многочлен будет соответствовать действительному ходу изменения функции.

На рис. 5.11 ломаной линией показана эмпирическая линия регрессии. На участке T ее можно с любой точностью аппроксимировать многочленом вида $y = a + bx + cx^2 + dx^3 + ex^4$ при соответствующем выборе коэффициентов. Но это вовсе не означает, что найденное выражение описывает характер изменения за пределами интервала T .

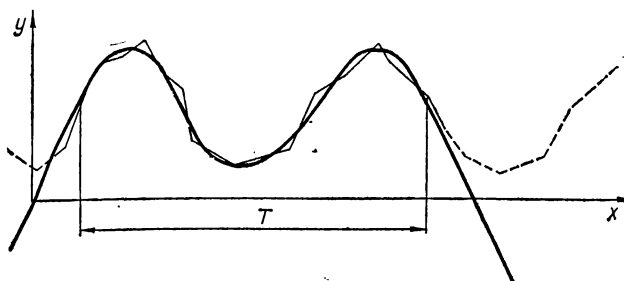


Рис. 5.11. Вываживание на заданном участке.

Таким образом, при помощи многочлена высокой степени можно воспроизвести значения произвольной функции внутри некоторого промежутка (приближенное интерполирование), но, в общем случае, нельзя оценить течение процесса вне заданного промежутка (задача приближенного экстраполирования или предсказания процесса).

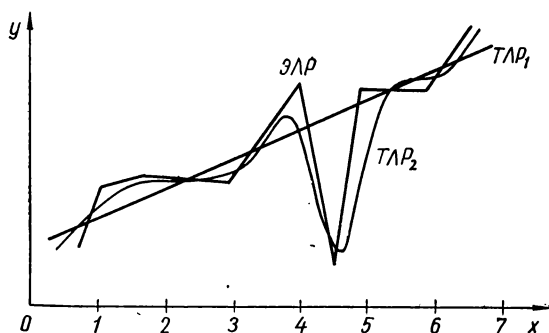


Рис. 5.12. Выбор теоретической линии регрессии.

Мы увидим далее, что способ расчета параметров (коэффициентов) искомой аппроксимирующей функции основан на требовании максимальной близости ее к эмпирической линии регрессии. Чем выше степень аппроксимирующего многочлена, чем больше коэффициентов содержит его уравнение, тем лучшее приближение при этом достигается.

Но близость аппроксимирующей кривой к эмпирической линии регрессии не всегда означает близость ее к предельной теоретической линии регрессии. Увеличивая порядок многочлена и число коэффициентов, мы начинаем воспроизводить не только закономерные изменения в ходе эмпирической линии, но и случайные ее зигзаги (рис. 5.12).

Точка эмпирической линии регрессии, соответствующая значению X между 4 и 5, явно выпадает из общей закономерности. Это можно объяснить, например, малым количеством наблюдений в этой области значений X .

Увеличение порядка и числа параметров выравнивающей кривой хотя и способствует лучшему приближению к эмпирической линии, одновременно может привести к удалению от действительной теоретической. Этим налагается предел увеличения точности выравнивания при помощи повышения степени многочленов. И отсюда возникает задача нахождения в каждом отдельном случае кривой, которая при меньшем числе параметров лучше передает закономерный ход линии регрессии по сравнению с параболой того или иного порядка.

Выбор и обоснование типа кривой регрессии

Для выбора и обоснования типа кривой регрессии нет универсального метода.

Существует несколько основных путей решения этой задачи, причем на практике каждый из них используется не изолировано, а в сочетании с другими.

Эмпирический подход основан на законе больших чисел.

О типе теоретической кривой можно судить по виду эмпирической линии регрессии. Удовлетворительные результаты можно получить лишь при большом количестве наблюдений, так как в этом случае график воспроизводит характерные особенности теоретической линии.

Однако при малом числе наблюдений этот путь приводит к неясным результатам. Резкие зигзаги эмпирической линии регрессии очень затрудняют выявление закономерности.

Ясно, что если существующая зависимость достаточно сложная, то ее очень трудно уловить при эмпирическом подходе, когда число наблюдений невелико. Поэтому, кроме суждения о зависимости по виду эмпирической линии регрессии, желательно привлечь дополнительные соображения.

Теоретический подход. Если судить о зависимости только по виду эмпирической линии регрессии, то остается в

стороне вопрос о реальной физической природе исследуемых переменных. Но ведь изучаемые y и x не абстрактные символы. Они описывают конкретные физические явления и процессы. Поэтому естественно использовать сведения, известные из теории изучаемых процессов. Если исследуется зависимость количества частиц в потоке, пронизывающем определенную площадку, и характеристик ускоряющего поля, то существенную помощь при определении типа зависимости окажут теоретические положения физики. Например, при исследовании зависимости безотказной работы аппаратуры от тренированности следует опираться на теорию надежности и т. д.

В любом случае необходимо использовать данные той конкретной области науки и техники, на базе которой возникает задача измерения связи. При этом возможны различные пути использования теории.

1. *Опыт предыдущих исследований.* Очень часто необходимо проводить исследования, аналогичные в какой-то степени проведенным ранее. Нас может интересовать задача определения связи между признаками при каких-то новых условиях. Похожие исследования уже были проведены и может быть даже описаны в литературе. Поэтому есть на что опереться при решении конкретной задачи.

Однако этот путь, очевидно, неприемлем для новых исследований.

2. *Эксперимент.* Это первичный и почти всегда основной путь исследования существующих закономерностей. Сильная сторона его заключается в том, что во многих случаях удается закреплять или сокращать колеблемость ряда факторов. А это, в свою очередь, приводит к уменьшению случайных отклонений эмпирической линии регрессии. Так, например, если определять зависимость между падением напряжения в некотором электрическом аппарате (с эквивалентным сопротивлением R) и силой тока при значительных колебаниях питания или температуры окружающей среды, то зависимость будет «размыта». Но при стабилизации питания, при поддержании постоянной температуры, при закреплении других факторов линейная зависимость проявилась бы достаточно отчетливо.

Тем не менее у экспериментального подхода есть и своя слабая сторона. Дело в том, что только при закреплении побочных факторов, действующих по схеме $A-M-N$, исследование может привести к теоретически верному результату. Если же закреплять факторы, корреляционно

связанные с основным аргументом, то эмпирическая линия регрессии хотя и получается более закономерной, но систематически отличается от опыта к опыту. Но ведь на практике, особенно при исследовании зависимостей впервые, трудно бывает установить, какие из факторов связаны, а какие нет. Для этого необходимы отдельные исследования. Более того, часто даже не все факторы, влияющие на изменение функции, известны.

Кроме того, иногда экспериментирование затруднено, а в ряде задач невозможно, например, при экономических или социологических исследованиях. Хотя именно там, где экспериментирование затруднено, применение статистических методов измерения связи особенно необходимо.

3. *Логический анализ.* Каким бы путем мы не исследовали тип зависимости между переменными, всегда приходится логически анализировать исходные данные и получаемые результаты. Логический анализ зависимости может проявляться в самых различных формах. Это могут быть и рассуждения, основанные на простом здравом смысле, и составление уравнений, связывающих функцию с ее аргументами. Простой здравый смысл подсказывает нам, что при увеличении надежности элементов время безотказной работы аппаратуры при прочих равных условиях должно возрастать.

Логический анализ, в основном, должен помочь уяснить общий характер зависимости между функцией y и ее аргументом x при «прочих равных» условиях, действующих по схеме $A—M—N$. Логический анализ должен помочь нам сделать выводы о наличии или отсутствии корреляции между различными факторами. При этом часто нужно анализировать материальную природу этих факторов, высказывать соображения о ходе процесса при их изменениях.

Все изложенные приемы и методы помогают в выборе и обосновании типа уравнения регрессии. Но для того, чтобы форму связи охарактеризовать количественно, необходимо определить численные значения параметров (коэффициенты) в уравнении регрессии.

Расчет параметров уравнения регрессии

Прямолинейная зависимость. Данные не сгруппированы. Методику определения параметров уравнения регрессии разберем на примере. Мы располагаем сведениями о времени заводской тренировки и времени безотказной

работы в условиях эксплуатации для десяти приборов. Эти данные представлены в табл. 5.13. Требуется оценить зависимость между этими характеристиками. Задача решается в несколько этапов.

1. Устанавливаем, какой показатель — функция, а какой — аргумент. В рассматриваемом примере нас интересует зависимость времени безотказной работы от времени тренировки приборов. Первый показатель — функцию — обозначим y , второй — аргумент — x .

Таблица 5.13

Условный номер прибора	Время тренировки (в % к номинальному)	Время безотказной работы (в % к установленному)
1	70	78
2	120	135
3	140	138
4	90	108
5	150	142
6	100	110
7	130	138
8	60	74
9	110	98
10	80	72

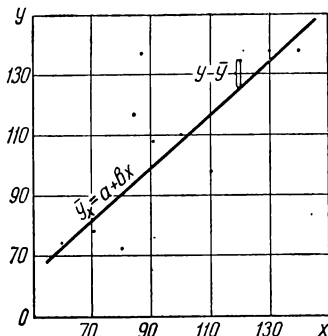


Рис. 5.13. Поле корреляции для партии из 10 приборов.

2. Систематизируем данные, строим поле корреляции (рис. 5.13). Составлять корреляционную таблицу нет необходимости, так как объем исходных данных незначителен. По расположению точек в поле корреляции можно судить о наличии прямой зависимости между x и y .

3. При большом количестве наблюдений необходимо считать эмпирическую линию регрессии при помощи корреляционной таблицы. В нашем примере этот этап опускается из-за малочисленности экспериментальных данных.

4. Строим теоретическую линию регрессии. По расположению точек на поле корреляции можно заключить, что зависимость y от x близка к линейной. Выразим ее уравнением

$$\bar{y}_x = a + bx,$$

где a и b — неизвестные параметры.

Это уравнение при различных значениях a и b дает нам множество прямых на плоскости. Из них надо выбрать ту, которая наилучшим образом соответствует эксперимен-

тальным данным. Такая прямая должна быть ближе к заданным точкам корреляционного поля, чем любая другая. Расстояние от точки до прямой будем измерять отрезком ординаты, соединяющей точку с прямой

$$\bar{y} - y_x,$$

где y — истинная ордината точки, а \bar{y}_x — ордината соответствующей точки линии. Очевидно, что отклонение точки от прямой может быть положительным и отрицательным. Но поскольку знаки отклонений нас не интересуют, мы будем рассматривать квадраты этих отклонений $(y - \bar{y}_x)^2$. Потребуем, чтобы сумма квадратов отклонений фактических ординат от ординат, вычисленных по уравнению прямой, была минимальной

$$\Sigma (y - \bar{y}_x)^2 = \min. \quad (5.17)$$

Выражение 5.17 называют критерием наименьших квадратов. Метод определения параметров аппроксимирующей кривой, основанной на этом критерии, известен под названием метода наименьших квадратов.

Будем искать параметры прямой линии. Запишем из (5.17)

$$f = \Sigma (y - a - bx)^2 = \min. \quad (5.18)$$

Для того чтобы (5.18) удовлетворялось, необходимо выполнить условия

$$\frac{\partial f}{\partial a} = 0, \quad \frac{\partial f}{\partial b} = 0. \quad (5.19)$$

Условия (5.19) для положительной квадратической функции от параметров a , b являются не только необходимыми, но и достаточными условиями минимума. Из (5.18) и (5.19) получаем

$$\frac{\partial f}{\partial a} = -2\Sigma (y - a - bx) = 0$$

или

$$\Sigma (y - a - bx) = 0,$$

что дает

$$\Sigma y = na + b\Sigma x,$$

где n — число точек.

Далее

$$\frac{\partial f}{\partial b} = -2\Sigma (y - a - bx)x = 0,$$

или

$$\Sigma (y - a - bx)x = 0,$$

что дает

$$\Sigma xy = a\Sigma x + b\Sigma x^2.$$

В результате получена система двух уравнений первой степени с двумя неизвестными a и b

$$\left. \begin{aligned} \Sigma y &= na + b\Sigma x, \\ \Sigma xy &= a\Sigma x + b\Sigma x^2. \end{aligned} \right\} \quad (5.20)$$

Эту систему называют системой нормальных уравнений по способу наименьших квадратов для определения пара-

Таблица 5.14

x	y	x^2	xy	\bar{y}_x
60	74	3600	4440	73,285
70	78	4900	5460	81,955
80	72	6400	5760	90,625
90	108	8100	9720	99,295
100	110	10 000	11 000	107,965
110	98	12 100	10 780	116,635
120	135	14 400	16 200	125,305
130	138	16 900	17 940	133,975
140	138	19 600	19 320	142,640
150	142	22 500	21 300	151,310
$\Sigma = 1050$	1093	118 500	121 920	

метров. Расчеты можно производить при помощи табл. 5.14. Подставим полученные величины Σx , Σx^2 , Σy , Σxy в (5.20) и решим систему относительно a и b

$$\begin{array}{rcl} 1093 & = & 10a + 1050b \quad | : 10 \\ 121\,920 & = & 1050a + 118\,500b \quad | : 1050 \\ \hline 109,3 & = & a + 105b \\ 116,114 & = & a + 112,857b \\ 6,814 & = & 7,857b \end{array}$$

$$b = \frac{6,814}{7,857} = 0,867,$$

$$109,3 = a + 105 \cdot 0,867,$$

$$a = 18,265.$$

Уравнение теоретической линии регрессии y по x теперь принимает вид

$$\bar{y}_x = 18,265 + 0,867x. \quad (5.21)$$

Для различных значений x из уравнения (5.21) определяем соответствующие значения \bar{y}_x и по этим данным строим прямую регрессии в поле корреляции (см. рис. 5.13).

Прямолинейная зависимость. Данные сгруппированы. В рассматриваемом примере мы располагаем небольшим количеством исходных данных ($n = 10$). Полученные данные во многом случайны и могут не повториться для другой партии приборов. В соответствии с законом больших чисел более надежные результаты можно получить при массовых наблюдениях. При этом будем иметь значительный объем исходных данных, которые нужно предварительно сгруппировать и свести в корреляционную таблицу. На основе корреляционной таблицы производится расчет теоретической линии регрессии.

Рассмотрим пример, для которого мы уже рассчитывали эмпирическую линию регрессии. Речь снова идет о зависимости времени безотказной работы аппаратуры (в % к установленному) от времени заводской тренировки (в % к номинальному).

В первом приближении будем считать теоретическую линию регрессии прямой линией с уравнением

$$\bar{y}_x = a + bx.$$

В предыдущем параграфе показано, что a и b находят из системы нормальных уравнений

$$\left. \begin{aligned} \Sigma y &= na + b\Sigma x, \\ \Sigma xy &= a\Sigma x + b\Sigma x^2. \end{aligned} \right\}$$

В корреляционной таблице варианты x , y встречаются с определенными частотами, поэтому суммы в нормальных уравнениях удобнее вычислять во взвешенном виде. Будем обозначать их соответственно Σx , Σy , Σx^2 , Σxy .

Система нормальных уравнений принимает вид

$$\left. \begin{aligned} \Sigma y &= na + b\Sigma x, \\ \Sigma xy &= a\Sigma x + b\Sigma x^2. \end{aligned} \right\} \quad (5.22)$$

Вычисление производим с помощью корреляционной таблицы (табл. 5.15).

Проводим замену переменных

$$x' = \frac{x - C_x}{i_x}, \quad y' = \frac{y - C_y}{i_y},$$

где C_x и C_y — новые начала отсчета; i_x и i_y — интервалы по x и по y . Примем $C_x = 100$, $C_y = 100$, $i_x = 20$, $i_y = 20$.

Таблица 5.15

	x'	-2	-1	0	1	2	Итого l	ly'
y'	$\begin{array}{c} x \\ y \end{array}$	60	80	100	120	140		
2	140		1 2	4 8	6 12	15 30	26	52
1	120		5 5	20 20	13 13	4 4	42	42
0	100	2 0	23 0	49 0	7 0	4 0	85	0
-1	80	2 -2	16 -16	16 -16	2 -2	2 -2	38	-38
-2	60	6 -12	3 -6				9	-18
	Итого n	10	48	89	28	25	200	38
	nx'	-20	-48	0	28	50	10	—
	nx'^2	40	48	0	28	100	216	—
	$\Sigma ly'$	-14	-15	12	23	32	38	—
	$x' \Sigma ly'$	28	15	0	23	64	130	—

Параметры теоретической линии регрессии отыскиваем из системы нормальных уравнений

$$\begin{aligned} \dot{\Sigma} y' &= na' + b' \dot{\Sigma} x', \\ \dot{\Sigma} x' y' &= a' \dot{\Sigma} x' + b' \dot{\Sigma} x'^2. \end{aligned} \quad (5.23)$$

Подставив числовые значения, получим

$$\begin{cases} 38 = 200a' + 10b', \\ 130 = 10a' + 216b'. \end{cases}$$

Решив систему, определим значения a' и b' :

$$a' = -0,160; \quad b' = 0,594$$

запишем

$$\bar{y}_{x'} = -0,16 + 0,594x'.$$

Перейдем от параметров a' и b' к параметрам a и b . Для этого в уравнение $\bar{y}_{x'} = a' + b'x'$ подставим

$$x' = \frac{x - C_x}{i_x}; \quad \bar{y}_{x'} = \frac{\bar{y}_x - C_y}{i_y};$$

$$\frac{\bar{y}_x - C_y}{i_y} = a' + b' \frac{x - C_x}{i_x}.$$

Выполнив необходимые преобразования, получаем

$$\bar{y}_x = \left[C_y + i_y a' - b' \frac{i_y}{i_x} C_x \right] + b' \frac{i_y}{i_x} x,$$

где

$$a = C_y + i_y a' - b' \frac{i_y}{i_x} C_x;$$

$$b = b' \frac{i_y}{i_x}.$$

После подстановки числовых значений из табл. 5.15

$$b = 0,594 \frac{20}{20} = 0,594,$$

$$a = 100 + 20 \cdot (-0,16) - 0,594 \cdot 100 = 37,4.$$

Уравнение теоретической линии регрессии будет иметь вид

$$\bar{y}_x = 37,4 + 0,594x.$$

По этому уравнению на поле корреляции строим прямую регрессии (рис. 5.14). С помощью найденной зависимости можно оценивать увеличение срока безотказной работы исследуемых приборов при увеличении времени их тренировки.

В нашем случае увеличение времени тренировки на 10% приводит к увеличению времени безотказной работы в среднем на 5,94%.

Нелинейная зависимость. Попробуем несколько глубже проанализировать разобранный пример, привлекая сведения из теории надежности и логически анализируя физическую природу изучаемой зависимости. То, что с увеличением тренированности эксплуатационная надежность аппаратуры возрастает, очевидно. Но рост этот не может быть равномерным при изменении значений аргумента.

Вначале время безотказной работы с увеличением тренированности возрастает довольно быстро. Но с некоторого момента начинает сказываться старение деталей и соединений, и в связи с этим понижается надежность. Можно даже сказать, что существует некоторое предельное значение, к которому приближается надежность при достаточно больших значениях аргумента.

Исходя из всего этого более правильным было бы предположить наличие нелинейной зависимости между изучаемыми переменными.

При нелинейной зависимости аппроксимацию теоретической линии регрессии также можно осуществить с помощью метода наименьших квадратов.

Предположим, что зависимость выражается квадратичной параболой вида

$$\bar{y}_x = a + bx + cx^2. \quad (5.24)$$

Будем определять параметры a , b , c , пользуясь критерием наименьших квадратов

$$f = \Sigma (y - \bar{y}_x)^2 = \Sigma (y - a - bx - cx^2)^2 = \min. \quad (5.25)$$

Условия обращения f в \min записывают в виде

$$\frac{\partial f}{\partial a} = 0; \quad \frac{\partial f}{\partial b} = 0; \quad \frac{\partial f}{\partial c} = 0 \quad (5.26)$$

при

$$\frac{\partial^2 f}{\partial a^2} > 0; \quad \frac{\partial^2 f}{\partial b^2} > 0; \quad \frac{\partial^2 f}{\partial c^2} > 0.$$

Из (5.25) и (5.26) получаем систему нормальных уравнений

$$\left. \begin{aligned} \Sigma y &= na + b\Sigma x + c\Sigma x^2, \\ \Sigma xy &= a\Sigma x + b\Sigma x^2 + c\Sigma x^3, \\ \Sigma x^2y &= a\Sigma x^2 + b\Sigma x^3 + c\Sigma x^4. \end{aligned} \right\} \quad (5.27)$$

При большом количестве исходных данных суммы в уравнениях вычисляют в виде взвешенных. Тогда систему

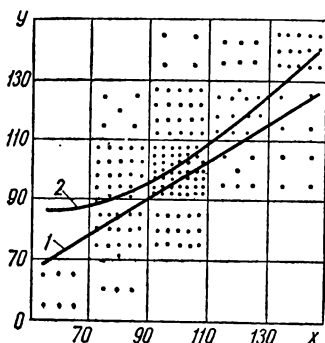


Рис. 5.14. Поле корреляции для партии из 200 приборов.

(5.27) можно записать так:

$$\left. \begin{aligned} \dot{\Sigma}y &= na + b\dot{\Sigma}x + c\dot{\Sigma}x^2, \\ \dot{\Sigma}xy &= a\dot{\Sigma}x + b\dot{\Sigma}x^2 + c\dot{\Sigma}x^3, \\ \dot{\Sigma}x^2y &= a\dot{\Sigma}x^2 + b\dot{\Sigma}x^3 + c\dot{\Sigma}x^4. \end{aligned} \right\}$$

Таблица 5.16

	x'	-2	-1	0	1	2	Итог l	ly'
y'	$\begin{array}{c} x \\ y \end{array}$	60	80	100	120	140		
2	140		1 2	4 8	6 12	15 30	26	52
1	120		5 5	20 20	13 13	4 4	42	42
0	100	2 0	23 0	49 0	7 0	4 0	85	0
-1	80	2 -2	16 -16	16 -16	2 -2	2 -2	38	-38
-2	60	6 -12	3 -6				9	-18
	Итог n	10	48	89	28	25	200	38
	nx'	-20	-48	0	28	50	$\dot{\Sigma}x'=10$	
	nx'^2	40	48	0	28	100	$\dot{\Sigma}x'^2=216$	
	nx'^3	-80	-48	0	28	200	$\dot{\Sigma}x'^3=100$	
	nx'^4	160	48	0	28	400	$\dot{\Sigma}x'^4=636$	
	$\Sigma_l vy'$	-14	-15	12	23	32	$\dot{\Sigma}y'=38$	
	$x^2 \Sigma_l vy'$	28	15	0	23	64	$\dot{\Sigma}x'y'=130$	
	$x^2 \Sigma_l vy'$	-56	-15	0	23	128	$\dot{\Sigma}x'^2y'=128$	

Расчеты производят при помощи табл. 5.16. Заменим переменные

$$x' = \frac{x - C_x}{i_x}; \quad y' = \frac{y - C_y}{i_y},$$

где $C_x = 100$; $C_y = 100$; $i_x = 20$; $i_y = 20$ и будем искать зависимость

$$\bar{y}_{x'} = a' + b'x + c'x^2. \quad (5.48)$$

Неизвестные параметры a' , b' , c' определяем из системы нормальных уравнений

$$\left. \begin{aligned} \dot{\Sigma} y' &= na' + b' \dot{\Sigma} x' + c' \dot{\Sigma} x'^2, \\ \dot{\Sigma} x' y' &= a' \dot{\Sigma} x' + b' \dot{\Sigma} x'^2 + c' \dot{\Sigma} x'^3, \\ \dot{\Sigma} x'^2 y' &= a' \dot{\Sigma} x'^2 + b' \dot{\Sigma} x'^3 + c' \dot{\Sigma} x'^4. \end{aligned} \right\} \quad (5.29)$$

Подставив числовые данные из расчетной таблицы, получаем

$$\begin{aligned} 38 &= 200a' + 10b' + 216c', \\ 130 &= 10a' + 216b' + 100c', \\ 128 &= 216a' + 100b' + 636c'. \end{aligned}$$

Решив систему, определим значения a' , b' , c' :

$$a' = 0,061; \quad b' = 0,557; \quad c' = 0,093.$$

Теперь уравнение (5.28) можно записать в виде

$$\bar{y}_x' = 0,061 + 0,557x + 0,093x^2.$$

После замены переменных

$$x' = \frac{x - C_x}{i_x}; \quad y' = \frac{y - C_y}{i_y}$$

получим

$$\frac{\bar{y}_x - C_y}{i_y} = a' + b' \frac{x - C_x}{i_x} + c' \frac{(x - C_x)^2}{i_x^2}. \quad (5.30)$$

После промежуточных преобразований (5.30) принимает вид

$$\begin{aligned} \bar{y}_x = & \left[C_y + i_y a' - b' \frac{i_y}{i_x} C_x + C' \frac{i_y}{i_x^2} C_x^2 \right] + \\ & + \left[b' \frac{i_y}{i_x} - 2C' \frac{i_y}{i_x^2} C_x \right] x + C' \frac{i_y}{i_x^2} x^2, \end{aligned} \quad (5.31)$$

где

$$C_y + i_y a' - b' \frac{i_y}{i_x} C_x + C' \frac{i_y}{i_x^2} C_x^2 = a;$$

$$b' \frac{i_y}{i_x} - 2C' \frac{i_y}{i_x^2} C_x = b;$$

$$C' \frac{i_y}{i_x^2} = c.$$

Подставляя числовые значения, записываем уравнение найденной теоретической линии регрессии в окончательном виде

$$\bar{y}_x = 92,02 - 0,373x + 0,00465x^2. \quad (5.32)$$

Придавая различные значения аргументу x , определяем соответствующие значения \bar{y}_x и строим линию регрессии в поле корреляции (см. рис. 5.14). Оказывается, что в той области изменения аргумента, где исследовалась зависимость, наблюдается ускоренный рост функции. Следовательно, наши данные не могут дать сведений о предельном уровне надежности. По-видимому, имея достаточно большое количество наблюдений при больших значениях аргумента, можно было бы решить эту задачу. Теоретическую линию регрессии пришлось бы искать в виде более сложной кривой, и аналитическое выражение оказалось бы более сложным. Это привело бы к большему объему вычислений. Но принципиально задача разрешима с помощью того же метода наименьших квадратов.

Когда уравнение регрессии представляет собой многочлен заданной степени k

$$\bar{y}_x = a + a_1x + a_2x^2 + \dots + a_kx^k \quad (5.33)$$

или в общем случае, когда \bar{y}_x предполагается линейной функцией параметров a, b, c, \dots

$$\bar{y}_x = a + bp(x) + cq(x), \quad (5.34)$$

коэффициенты вычисляют на основании требования

$$\Sigma (y - \bar{y}_x)^2 = \min,$$

приводящего к системе нормальных уравнений

$$\left. \begin{aligned} \Sigma y &= na + b\Sigma p(x) + c\Sigma q(x) + \dots, \\ \Sigma yp(x) &= a\Sigma p(x) + b\Sigma p^2(x) + c\Sigma q(x)p(x) + \dots, \\ \Sigma yq(x) &= a\Sigma q(x) + b\Sigma p(x)q(x) + c\Sigma q^2(x) + \dots. \end{aligned} \right\} \quad (5.35)$$

При составлении системы нормальных уравнений можно обойтись без дифференцирования функции

$$f = \Sigma (y - \bar{y}_x)^2.$$

Для этого можно воспользоваться следующим общим правилом. Уравнение искомой линии регрессии

$$\bar{y}_x = a + bp(x) + cq(x) \quad (5.36)$$

нужно последовательно умножить на коэффициенты при параметрах. После этого в обеих частях уравнения берется сумма.

Например, коэффициент при параметре a равен 1; после суммирования получаем первое нормальное уравнение (5.36). Коэффициент при b равен $p(x)$; после умножения (5.36) на $p(x)$ и суммирования получаем второе нормальное уравнение и т. д.

Таблица 5.17

Выпуск продукции (тыс. шт.)	Средняя себестоимость единицы	Число предприятий
1	16,50	6
1—2	13,75	6
2—3	13,31	8
3—4	12,50	7
4—5	13,52	4
5—6	12,75	4
6—7	12,30	3
7—8	12,83	2
8—9	12,28	2
9—10	12,34	2
		44

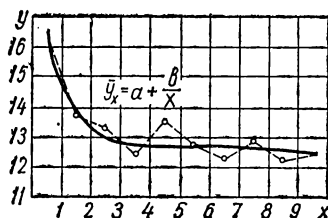


Рис. 5.15. Корреляционная зависимость.

Приведем в качестве примера расчет параметров уравнения регрессии себестоимости продукции по величине выпуска.

В табл. 5.17 приведены данные для 44 предприятий. На рис. 5.15 изображена эмпирическая линия регрессии, построенная по данным табл. 5.18. По виду этой линии можно предположить наличие между y и x гиперболической зависимости вида: $\bar{y}_x = a + \frac{b}{x}$.

Запишем систему нормальных уравнений со взвешенными суммами

$$\left. \begin{aligned} \sum n \bar{y}_i &= na + b \sum \frac{n}{x}, \\ \sum n \frac{y_i}{x} &= a \sum \frac{n}{x} + b \sum \frac{n}{x^2}. \end{aligned} \right\} \quad (5.37)$$

Для расчетов составляем табл. 5.18. Подставляя числовые значения, получаем

$$\left. \begin{aligned} 592,36 &= 44a + 23,991b, \\ 356,47 &= 23,991a + 29,005b. \end{aligned} \right\}$$

Коэффициенты равны

$$a = 12,318; \quad b = 2,101,$$

и уравнение линии регрессии получаем в виде

$$\bar{y}_x = 12,318 + \frac{2,101}{x}.$$

На рис. 5.15 нанесена теоретическая линия регрессии, соответствующая полученному уравнению.

Таблица 5.18

x	\bar{y}_l	n	$\frac{n}{x}$	$\frac{n}{x^2}$	\overline{ny}_l	$\frac{\overline{ny}_l}{x}$
0,5	15,50	6	12,000	24,000	99,00	198,00
1,5	13,75	6	4,000	2,667	82,50	55,00
2,5	13,31	8	3,200	1,280	106,48	42,59
3,5	12,50	7	2,000	0,571	87,50	25,00
4,5	13,52	4	0,889	0,198	54,08	12,02
5,5	12,75	4	0,727	0,132	51,00	9,27
6,5	12,30	3	0,462	0,071	36,90	5,68
7,5	12,83	2	0,267	0,036	25,66	3,42
8,5	12,28	2	0,236	0,028	24,56	2,89
9,5	12,34	2	0,211	0,022	24,68	2,60
		44	23,991	29,005	592,36	356,47

В тех случаях, когда зависимость между \bar{y}_x и параметрами a , b , c ... нелинейна, для расчета параметров приходится пользоваться специальными методами: последовательными приближениями, заменой переменных.

Функциональная корреляция

Корреляционная зависимость может оказаться более простой, если рассматривать не сами переменные, а некоторые функции (логарифмы, обратные величины и т. д.).

При этом можно значительно облегчить исследование корреляционной зависимости, если произвести предварительную замену переменных.

Допустим, что на основе предварительного материального анализа между переменными y и x предполагается зависимость по схеме

$$y = \sqrt{u + vt^x}, \quad (5.38)$$

где u и v — неучитываемые аргументы.

Предполагается, что они корреляционно от x не зависят. Обозначим

$$\left. \begin{aligned} z &= y^2, \\ t &= t^x, \end{aligned} \right\} \quad (5.39)$$

тогда (5.38) можно записать

$$z = u + vt. \quad (5.40)$$

Следовательно, уравнение теоретической линии регрессии z по t должно иметь вид

$$\bar{z}_i = \bar{u} + \bar{v}t, \quad (5.41)$$

где \bar{u} , \bar{v} — неизвестные параметры, оцениваемые по опытным данным; \bar{u} и \bar{v} определяются из условия

$$\Sigma (z - \bar{z}_i)^2 = \min. \quad (5.42)$$

Система нормальных уравнений

$$\left. \begin{aligned} \Sigma z &= n\bar{u} + \bar{v}\Sigma t \\ \Sigma zt &= \bar{u}\Sigma t + \bar{v}\Sigma t^2 \end{aligned} \right\}. \quad (5.43)$$

После нахождения \bar{u} и \bar{v} восстановим прежнюю систему переменных. Формула эмпирической линии регрессии z по t

$$\bar{z}_i = \frac{\Sigma t^z}{n_i},$$

где i — номер значения t (или соответствующего x); n_i — частота.

В прежней системе координат эта формула запишется в виде

$$\hat{y}_i = \sqrt{\frac{\Sigma t y^2}{n_i}}, \quad (5.44)$$

где суммирование под знаком радикала производится по всем y^2 , соответствующим определенному значению x_i ; \hat{y}_i — значение y , квадрат которого равен \bar{z}_i .

Уравнение теоретической линии регрессии z по t в прежней системе координат

$$\hat{y}_x = \sqrt{\bar{u} + \bar{v}e^x}, \quad (5.45)$$

где $x = \ln t$; \hat{y}_x — значение y , квадрат которого равен \bar{z}_i ; \bar{u} , \bar{v} — найдены из (5.43).

Замена переменных дала возможность привести сложную схему (5.38) к более простой (5.41), для которой рассчитать уравнение регрессии значительно проще. Этот пример показывает, что в ряде случаев влияние изменений неучитываемых факторов на функцию y можно исключить, введя функциональное преобразование зависимой переменной

$$z = \Phi(y) \quad (5.46)$$

с последующим расчетом линии регрессии z по x .

При этом схему действия факторов можно представить в виде

$$y = \Phi^{-1}[u + v\varphi(x) + w\psi(x) + \dots], \quad (5.47)$$

где u, v, w — аргументы, корреляционно не зависящие от x ; Φ^{-1} — функциональная зависимость, обратная Φ .

Схема (5.47) называется схемой Φ —А—М—Н, а Φ — ее определяющей функцией. Применяя к (5.47) преобразование (5.46), получаем

$$z = u + v\varphi(x) + w\psi(x) + \dots \quad (5.48)$$

Уравнение предельной теоретической линии регрессии z по x должно иметь вид

$$\bar{z}_x = \bar{u} + \bar{v}\varphi(x) + \bar{w}\psi(x) + \dots \quad (5.49)$$

Принципиально возможно и дальнейшее упрощение уравнения. Например, функциональное преобразование $t = \varphi(x)$ приводит уравнение

$$z = u + v\varphi(x) \quad (5.50)$$

к линейному

$$z = u + v(t). \quad (5.51)$$

Для проверки правильности уравнения (5.49) и оценки факторов $\bar{u}, \bar{v}, \bar{w}, \dots$ по опытным данным рассчитываем эмпирическую и теоретическую линии регрессии z по x (или z по t). Ординаты эмпирической линии находим из уравнения

$$\bar{z}_i = \frac{\sum_i z}{n_i}, \quad (5.52)$$

где суммирование \sum_i ведется для значений z , соответствующих данному значению x ; n_i — частота данного x . Параметры $\bar{u}, \bar{v}, \bar{w}, \dots$ оцениваются из требования наименьших квадратов

$$\sum (z - \bar{z}_x)^2 = \min. \quad (5.53)$$

Затем при помощи обратных преобразований

$$\hat{y}_i = \Phi^{-1}(\bar{z}_i)$$

и

$$\hat{y}_x = \Phi^{-1}(\bar{z}_x)$$

получаем формулу эмпирической линии функциональной регрессии y по x

$$\hat{y}_i = \Phi^{-1} \left[\frac{\sum_i \Phi(y)}{n_i} \right] \quad (5.54)$$

и уравнение теоретической линии функциональной регрессии y по x

$$\hat{y}_x = \Phi^{-1} [\bar{u} + \bar{v}\varphi(x) + \bar{w}\Psi(x) + \dots]. \quad (5.55)$$

Функциональные средние

Переменная x дает при помощи (5.54) эмпирическую линию функциональной регрессии y по x . Для постоянного значения x формула (5.54) дает функциональную среднюю значений y , соответствующих заданному значению x . Различные определяющие функции Φ дают различные функциональные средние.

Рассмотрим наиболее интересные разновидности функциональных средних на числовом примере. Заданы следующие значения y при $x = \text{const}$: 5, 2, 1, 4, 4.

1. Пусть схема действия факторов выражается в виде

$$y = u + v\varphi(x) + w\Psi(x), \quad (5.56)$$

где u, v, w — корреляционно не зависят от x .

Очевидно, для этой схемы нет необходимости в функциональном преобразовании y

$$\Phi(y) = y.$$

Тогда по формуле (5.54) получаем

$$\hat{y} = \frac{y_1 + y_2 + \dots + y_n}{n},$$

а это ничто иное как средняя арифметическая $\hat{y} = \bar{y}$. По нашим данным

$$\hat{y} = \frac{5 + 2 + 1 + 4 + 4}{5} = 3,2.$$

2. Теперь рассмотрим схему действия факторов в виде

$$y = \frac{1}{u + v\varphi(x) + w\Psi(x)}. \quad (5.57)$$

Определяющую функцию схемы можно выразить $\Phi(y) = \frac{1}{y}$. Тогда по формуле (5.54)

$$\hat{y} = \frac{h}{\frac{1}{y_1} + \frac{1}{y_2} + \dots + \frac{1}{y_n}}, \quad (5.58)$$

или по нашим данным

$$\hat{y} = \frac{5}{\frac{1}{5} + \frac{1}{2} + \frac{1}{1} + \frac{1}{4} + \frac{1}{4}} = 2,27.$$

Функциональная средняя вида (5.58) называется средней гармонической.

3. Следующая схема вида

$$y = a^{u+v\varphi(x)+w\Psi(x)},$$

u, v, w — корреляционно не зависят от x .

Имеем $\Phi(y) = \lg y$ и, следовательно,

$$\hat{y} = a^{\frac{\lg y_1 + \lg y_2 + \dots + \lg y_n}{n}},$$

или

$$\hat{y} = \sqrt[n]{y_1 y_2 \dots y_n}. \quad (5.59)$$

Числовые значения дают

$$\hat{y} = \sqrt[5]{5 \cdot 2 \cdot 1 \cdot 4 \cdot 4} = 2,76.$$

Выражение (5.59) определяет среднюю геометрическую.

4. Для схемы вида

$$y = \sqrt[k]{u + v\varphi(x) + w\Psi(x)} \quad (5.60)$$

имеем определяющую функцию $\Phi(y) = y^k$.

Функциональную среднюю получаем в виде

$$\hat{y} = \sqrt[k]{\frac{y_1^k + y_2^k + \dots + y_n^k}{n}}. \quad (5.61)$$

Это выражение для степенной средней порядка k . В частности,

при $k = 1$ — имеем среднюю арифметическую,
 при $k = 2$ — среднюю квадратическую,
 при $k = 3$ — среднюю кубическую и т. д.

Различные функциональные средние, вычисленные по одним и тем же числовым данным, различаются по величине. Расхождения определяются числовыми данными. Но порядок возрастания значений различных функциональных средних во многих случаях не зависит от числовых данных, а определяется видом функции Φ . Это свойство сравнимости, или мажорантности, средних. А. Я. Боярский установил достаточные условия сравнимости функциональных средних — признак мажорантности средних. При положительных значениях y должен быть следующий порядок: средняя гармоническая \leq средней геометрической \leq средней арифметической \leq средней квадратической \leq средней кубической для одних и тех же числовых данных.

В каждом конкретном случае определяется та средняя, которая характеризует данный ряд распределения в зависимости от его особенностей. Например, средняя гармоническая вычисляется тогда, когда средняя предназначается для расчета сумм слагаемых, обратно пропорциональных величине данного признака, т. е. когда нужно складывать не сами варианты, а величины, обратные им.

Рассмотрим пример. В таблице 5.19 приведены данные о работе 22 рабочих в течение 6 ч. Определим количество изготовленных ими деталей. Поскольку все рабочие работали по 6 ч, можно рассматривать количество рабочих как величину, определяющую общие затраты времени. Взвешивание заключается в делении количества рабочих (m) в каждой группе на затраты времени на изготовление одной детали (x). Таким образом, в этом примере мы имеем дело со средней гармонической

$$\bar{x} = \frac{\sum n}{\sum \frac{n}{x}} = \frac{22}{1,7} \approx 12,94.$$

При использовании средней гармонической удобно пользоваться таблицами обратных чисел.

Таблица 5.19

Количество времени на изготовление одной детали в минутах, x	Количество рабочих, n	$\frac{n}{x}$
10	2	0,20
12	9	0,75
14	7	0,50
16	4	0,25
Итого	22	1,70

Средняя геометрическая используется главным образом при изучении динамики, например, средних коэффициентов и темпов роста каких-либо показателей.

Средняя квадратическая используется только в тех случаях, когда варианты представляют собой отклонения фактических величин от средней арифметической или от заданной нормы.

Контрольные вопросы и задания

1. Сформулируйте основные задачи измерения связи.
2. Покажите, как по опытным данным строится поле корреляции и корреляционная таблица.
3. Как строится эмпирическая линия регрессии? Как определить корреляционную зависимость по виду эмпирической линии регрессии?
4. Какое соотношение между корреляционной и статистической зависимостью?
5. Что такое предельная теоретическая линия регрессии?
6. Каким образом выбирается форма теоретической линии регрессии?
7. Покажите на примере, как рассчитывать параметры уравнения регрессии.
8. Что такое функциональная корреляция? Назовите основные типы функциональных средних.

§ 5. ИССЛЕДОВАНИЕ ТЕСНОТЫ СВЯЗИ

В предыдущем разделе мы рассмотрели первую основную задачу теории корреляции — исследование формы связи. Для строгой функциональной зависимости эта задача — единственная. При корреляционной зависимости решение этой задачи не дает нам полной характеристики зависимости функций и аргументов. Посмотрим на рис. 5.16. На обоих корреляционных полях линии регрессии расположены одинаково, зависимость описывается одним и тем же уравнением регрессии. Но точки поля (б) значительно менее рассеяны относительно линии регрессии по сравнению с точками поля (а).

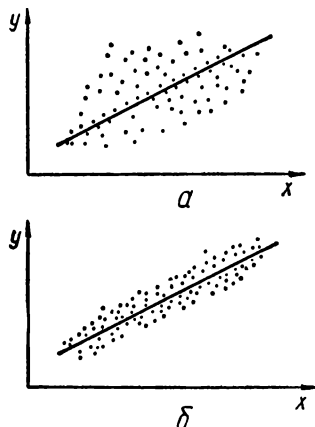


Рис. 5.16. К задаче измерения тесноты связи.

Известно, что отклонения точек поля от линии регрессии объясняются влиянием неучитываемых факторов. В случае (б) влияние основного аргумента x

в меньшей степени осложняется влиянием побочных факторов, чем в случае (а), и связь между y и x более тесная.

Измерение тесноты корреляционной зависимости — вторая основная задача теории корреляции. При большой тесноте связи, зная аргумент, можно с большой точностью предсказывать значение функции. Если же теснота связи мала, влияние основного аргумента x на функцию обнаруживается лишь в среднем по уравнению регрессии.

Изменчивость функции под влиянием изменчивости различных аргументов характеризуется дисперсией. Возвращаясь к нашему примеру зависимости времени безотказ-

Таблица 5.20

$\begin{matrix} X \\ \backslash \\ Y \end{matrix}$	50—70	70—90	90—110	110—130	130—150	Итого
130—150		1	4	6	15	26
110—130		5	20	13	4	42
90—110	2	23	49	7	4	85
70—90	2	16	16	2	2	38
50—70	6	3				9
Итого	10	48	89	28	25	200

ной работы приборов от тренированности. Колеблемость функции объясняется многими факторами: условиями эксплуатации, транспортировки, надежностью и качеством отдельных элементов, узлов и т. д.

Предположим, что нам удалось измерить колеблемость функции под влиянием какого-либо одного фактора, остальные в это время были закреплены. Тогда мы по сути выяснили степень влияния этого фактора на функцию. Эта цель может быть достигнута, если удастся разложить полную дисперсию функции y на две части. Одна из них измеряет влияние интересующего нас аргумента x , другая — влияние всех остальных, действующих независимо от x . Проведем такое разложение на нашем примере. Нас интересует аргумент x — время тренировки приборов (табл. 5.20).

Эмпирическое корреляционное отношение

Вся статистическая совокупность делится на ряд частных совокупностей (столбцы), соответствующих определенным фиксированным значениям аргумента x . Иными

слsвами, каждому постоянному значению аргумента x ($x = 60, 80, \dots$) соответствует ряд распределения y . Всю статистическую совокупность, которую часто называют генеральной, можно охарактеризовать средней арифметической

$$\bar{y} = \frac{\sum y}{n}$$

и дисперсией

$$\sigma^2 = \frac{\sum (y - \bar{y})^2}{n}.$$

Для каждой частной совокупности аналогичные показатели равны

$$\bar{y}_i = \frac{\sum_i y}{n_i} \quad \text{и} \quad \sigma_i^2 = \frac{1}{n_i} \sum_i (y - \bar{y}_i)^2,$$

где i — номер столбца.

В соответствии с основными свойствами статистических параметров можно записать

$$\bar{y} = \frac{\sum \bar{y}_i n_i}{n}, \quad (5.62)$$

$$\sigma^2 = \bar{\sigma}_i^2 + \bar{\delta}_i^2, \quad (5.63)$$

где $\bar{\sigma}_i^2 = \frac{\sum \sigma_i^2 h_i}{n}$ — средняя из частных дисперсий; отклонение частной средней от генеральной $\delta_i = \bar{y}_i - \bar{y}$; $\bar{\delta}_i^2 = \frac{\sum \delta_i^2 h_i}{n}$ — дисперсия частных средних.

Проанализируем выражение (5.63). В каждой частной совокупности значение аргумента x неизменно и поэтому дисперсия $\bar{\sigma}_i^2$ не зависит от x . Другими словами, первый член правой части формулы (5.63) характеризует колеблемость y в зависимости от прочих факторов, не зависящих от x . Но поскольку полная дисперсия функции σ^2 обусловлена также и изменчивостью x , остается предположить, что эта изменчивость характеризуется вторым компонентом формулы (5.63), а именно $\bar{\delta}_i^2$. На рис. 5.17 показана геометрическая интерпретация формулы (5.63). Линия генеральной средней $\bar{y} = c$ параллельна оси абсцисс. Общая дисперсия

$$\sigma^2 = \frac{\sum (y - \bar{y})^2}{n}$$

представляет собой средний квадрат отклонений точек корреляционного поля от линии $\bar{y} = c$. Значение

$$\bar{\sigma}_i^2 = \frac{\sum (y - \bar{y}_i)^2}{n}$$

дает средний квадрат отклонений точек корреляционного поля от эмпирической линии регрессии. А дисперсия частных средних

$$\bar{\delta}_i^2 = \frac{\sum (\bar{y}_i - \bar{y})^2}{n}$$

представляет собой средний квадрат отклонений эмпирической линии регрессии от линии $\bar{y} = c$.

Рассмотрим показатель, равный отношению дисперсии частных средних к генеральной дисперсии

$$\eta^2 = \frac{\bar{\delta}_i^2}{\sigma^2}. \quad (5.64)$$

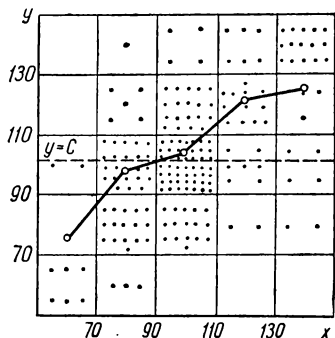


Рис. 5.17. Генеральная средняя.

Величина η показывает, какая часть полной колеблемости y обусловлена изменчивостью аргумента x .

Корень квадратный из величины η^2 в математической статистике называют эмпирическим корреляционным отношением η по x . Показатель η характеризует тесноту связи y с x . Очевидно, что значение η всегда удовлетворяет неравенству

$$0 \leq \eta \leq 1.$$

При $\eta = 0$ $\bar{\delta}_i^2 = 0$ и, следовательно, частные средние \bar{y}_i не изменяются. В этом случае \bar{y} корреляционно не зависит от x , так как линия регрессии параллельна оси абсцисс и совпадает с линией $\bar{y} = c$. Здесь, конечно, не следует забывать о законе больших чисел. При малом количестве наблюдений может получиться $\eta = 0$ и при наличии некоторой корреляции. С другой стороны, при отсутствии корреляции малое число наблюдений может дать $\eta > 0$.

Если $\eta = 1$, то $\bar{\sigma}_i^2 = 0$, и тогда все точки корреляционного поля лежат на эмпирической линии регрессии. В подобных случаях связь между y и x — точная или функциональная. Но это опять-таки справедливо только для бесконечного числа наблюдений, так как при малом числе

наблюдений равенство $\eta = 1$ может случайно получиться и при неточной зависимости.

В общем случае, когда $0 < \eta < 1$, говорят о более или менее тесной корреляционной зависимости функции от интересующего нас аргумента. В математической статистике разработан ряд приемов для удобства расчета корреляционных отношений. При неограниченно большом числе наблюдений формула (5.64) достаточно полно решает задачу о тесноте связи.

При ограниченном количестве экспериментальных данных, а это часто встречается в практических задачах, наблюдаются не только неизбежные случайные погрешности величин

$$\sigma^2, \bar{\sigma}_i^2, \bar{\delta}_i^2,$$

но и систематические преувеличения $\bar{\delta}_i^2$ и преуменьшения $\bar{\sigma}_i^2$. Частные средние \bar{y}_i изменяются не только под влиянием аргумента x , но случайно колеблются также под воздействием других, не зависящих от x , факторов. Эта случайная вариация \bar{y}_i преувеличивает $\bar{\delta}_i^2$ и в то же время преуменьшает $\bar{\sigma}_i^2$, так как в последнем влияние побочных факторов недоучитывается. Если это учесть, то станет ясно, что эмпирическое корреляционное отношение систематически дает большую тесноту связи между y и x , причем тем большую, чем меньше количество наблюдений и чем больше столбцов в корреляционной таблице. В каждом из этих столбцов оказывается меньше наблюдений с увеличением числа столбцов в корреляционной таблице, а это ведет к более значительным случайным колебаниям \bar{y}_i . Для устранения погрешностей вместо эмпирической линии регрессии при расчете показателей $\bar{\sigma}_i^2$ и $\bar{\delta}_i^2$ вводят в рассмотрение теоретическую линию регрессии.

Теоретическое корреляционное отношение

Пусть \bar{y}_x — частная средняя, которая получена из уравнения для теоретической линии регрессии $\bar{y}_x = f_x$. Заменим этой средней эмпирическую среднюю \bar{y}_i в выражениях для $\sigma_i^2, \bar{\sigma}_i^2, \bar{\delta}_i^2$. Тогда получим

$$\sigma_{ix}^2 = \frac{\sum_i (y - \bar{y}_x)^2}{n_i},$$

$$\overline{\sigma}_{i\tau}^2 = \frac{\sum \sigma_{i\tau}^2 n_i}{n} = \frac{\sum (y - \bar{y}_x)^2}{n},$$

$$\bar{\delta}_{i\tau}^2 = \frac{\sum \delta_{i\tau}^2 n_i}{n} = \frac{\sum (\bar{y}_x - \bar{y})^2}{n}.$$

Графически $\overline{\sigma}_{i\tau}^2$ выражает средний квадрат отклонений точек корреляционного поля от теоретической линии регрессии, $\bar{\delta}_{i\tau}^2$ — средний квадрат отклонений точек теоретической линии регрессии от линии генеральной средней $\bar{y} = c$.

Вместо уравнения (5.63) воспользуемся уравнением

$$\sigma^2 = \overline{\sigma}_{i\tau}^2 - \bar{\delta}_{i\tau}^2. \quad (5.65)$$

Вместо показателя η можно теперь рассматривать показатель

$$\eta_\tau = \sqrt{\frac{\bar{\delta}_{i\tau}^2}{\sigma^2}},$$

который называется теоретическим корреляционным отношением y по x . Очевидно, что

$$0 \leq \eta \leq 1.$$

При $\eta_\tau = 0$, $\bar{\delta}_{i\tau}^2 = 0$, и теоретическая линия регрессии параллельна оси абсцисс. Для достаточно большого числа наблюдений это свидетельствует об отсутствии корреляции между y и x .

Если $\eta_\tau = 1$, то $\overline{\sigma}_{i\tau}^2 = 0$, и все точки поля лежат на теоретической линии регрессии. При большом количестве наблюдений это может свидетельствовать о наличии функциональной зависимости между y и x .

По сравнению с уравнением (5.63) дисперсия перераспределяется между двумя составляющими. В самом деле, исходя из минимального свойства средней арифметической, в каждом столбце корреляционной таблицы имеем

$$\sigma_i^2 \leq \sigma_{i\tau}^2,$$

а значит, и

$$\overline{\sigma}_i^2 \leq \overline{\sigma}_{i\tau}^2.$$

Следовательно,

$$\bar{\delta}_i^2 \geq \bar{\delta}_{i\tau}^2.$$

Таким образом,

$$\eta_\tau \leq \eta,$$

т. е. теоретическое корреляционное отношение не больше эмпирического.

Теоретическое корреляционное отношение в определенной степени исправляет систематическую ошибку, свойственную эмпирическому корреляционному отношению. Однако полностью эту систематическую ошибку устранить нельзя. Дело в том, что теоретическая линия регрессии, по которой рассчитывались $\bar{\sigma}_{it}^2$ и $\bar{\delta}_{it}^2$, не является предельной теоретической линией регрессии. Она получена косвенным путем на основе ограниченного количества наблюдений. На теоретическую линию регрессии также влияют случайные факторы, хотя и меньше, чем на эмпирическую. Это связано с тем, что теоретическая линия регрессии, рассчитанная по критерию наименьших квадратов, ближе к точкам реального корреляционного поля, чем предельная теоретическая линия регрессии, так как теоретическая линия регрессии рассчитана именно на основании этих точек. Поэтому η_T дает нам преувеличенное значение тесноты связи между y и x , хотя это преувеличение в общем случае меньше, чем при использовании эмпирической линии регрессии.

Все приведенные рассуждения справедливы, когда вид зависимости выбран правильно, а отклонения \bar{y}_x от предельных $Y(x)$ — случайны. В противном случае (например, когда выбранный вид зависимости слишком упрощен) теоретическая линия регрессии, вычисленная по критерию наименьших квадратов, может оказаться более удаленной от точек корреляционного поля, чем предельная (выбор прямой вместо параболы и т. д.). Это может привести к систематическому занижению тесноты связи при использовании теоретического корреляционного отношения η_T . Занижение оказывается тем значительнее, чем сильнее упрощает выбранный тип уравнения регрессии действительную зависимость.

Коэффициент корреляции

В простейшем случае теоретическая линия регрессии выражается уравнением прямой линии

$$\bar{y}_x = a + bx.$$

В этом уравнении угловой коэффициент b называют обычно коэффициентом регрессии y по x . Он показывает, на сколько единиц в среднем изменяется y , когда x увеличи-

вается на одну единицу. Может показаться, что коэффициент регрессии является достаточно удобной характеристикой тесноты зависимости y от x .

Но дело в том, что b — величина размерная, и числовое значение этого коэффициента зависит от выбора единицы измерения по x и по y . Обычно рассматриваются зависимости между признаками различной природы. Единицы измерения этих признаков могут быть выбраны из различных соображений. Например, в уравнении регрессии массы человеческого тела по росту при измерении роста в сантиметрах коэффициент b получается в 100 раз меньший, чем при измерении роста в метрах. А при измерении массы в граммах вместо килограммов коэффициент регрессии b окажется в 1000 раз большим. При произвольном выборе единиц измерения признаков коэффициент регрессии b не может быть универсальным показателем тесноты связи. Если бы мы имели в своем распоряжении такую систему единиц измерения, используя которую можно было бы сравнивать данные по различным признакам, то произвольный выбор единиц измерения не влиял бы на коэффициент регрессии. В качестве такой системы используется система измерения в «сигмах» (стандартизованная). За единицу измерения признака в этой системе принимается его среднее квадратическое отклонение σ , начало отсчета при этом обычно переносится в точку, соответствующую средней арифметической.

Коэффициент регрессии в стандартизованной системе единиц (стандартизованный коэффициент регрессии) показывает, на сколько сигм изменяется в среднем y , когда x увеличивается на одну сигму.

Стандартизованный коэффициент регрессии называется коэффициентом корреляции между y и x и обозначается r_{yx} , или просто r .

Из определения b и r получаем

$$r = b \frac{\sigma_x}{\sigma_y}. \quad (5.66)$$

Легко показать, что при прямолинейной регрессии коэффициент корреляции действительно является показателем тесноты связи.

Рассмотрим формулу теоретического корреляционного отношения

$$\eta^2 = \frac{\bar{\delta}_{i\tau}^2}{\sigma^2} = \frac{\sum (\bar{y}_x - \bar{y})'}{n\sigma^2}.$$

Заменим \bar{y}_x и \bar{y} их выражениями

$$\begin{aligned}\bar{y}_x &= a + bx, \\ \bar{y} &= a + b\bar{x}.\end{aligned}$$

Получаем

$$\eta_r^2 = \frac{\frac{1}{n} \sum b^2 (x - \bar{x})^2}{\sigma^2} = \frac{b^2 \frac{1}{n} \sum (x - \bar{x})^2}{\sigma^2} = b^2 \frac{\sigma_x^2}{\sigma_y^2}$$

σ^2 в знаменателе приписываем индекс y , чтобы отличить (от σ_x^2 в числителе).

Сравнивая последнее выражение с формулой (5.66), получаем

$$\eta_r^2 = r^2, \quad (5.67)$$

или

$$\eta_r = |r|, \quad (5.68)$$

т. е. коэффициент корреляции при регрессии равен по абсолютной величине теоретическому корреляционному отношению.

Из (5.68) ясно, что модуль коэффициента корреляции, так же, как и η_r , может изменяться от 0 до 1.

Но в отличие от η_r коэффициент корреляции характеризуется также знаком. Знак при r совпадает со знаком при коэффициенте регрессии b

$$1 \leq r \leq 1.$$

При $r = -1$ и $r = 1$ все точки корреляционного поля лежат на линии регрессии и существует строгая пропорциональность между y и x . Знак «+» при r характеризует положительную корреляцию (\bar{y}_x растет с увеличением x). Знак «-» характеризует отрицательную корреляцию (\bar{y}_x уменьшается с ростом x).

Модификации формулы коэффициента корреляции

Кроме основной формулы (5.66) на практике часто удобно пользоваться некоторыми ее видоизменениями.

Из системы нормальных уравнений

$$\left. \begin{aligned}\sum y &= na + b\sum x, \\ \sum xy &= a\sum x + b\sum x^2\end{aligned}\right\}$$

делением на n получаем

$$\left. \begin{aligned} \bar{y} &= a + b\bar{x}, \\ \bar{y}\bar{x} &= a\bar{x} + b \frac{\Sigma x^2}{n}. \end{aligned} \right\} \quad (5.69)$$

Определим из системы (5.69) коэффициент регрессии

$$b = \frac{\left| \frac{1}{x} \frac{\bar{y}}{y\bar{x}} \right|}{\left| \frac{1}{x} \frac{\Sigma x^2}{n} \right|} = \frac{\frac{\bar{y}\bar{x} - \bar{y}\bar{x}}{\Sigma x^2 - (\bar{x})^2}}{\frac{\bar{y}\bar{x} - \bar{y}\bar{x}}{\sigma_x^2}}. \quad (5.70)$$

Подставляя (5.70) в (5.71), имеем

$$r = \frac{\bar{y}\bar{x} - \bar{y}\bar{x}}{\sigma_x \sigma_y}. \quad (5.71)$$

Из выражения (5.71) легко получить еще одну модификацию основной формулы

$$r = \frac{\frac{1}{n} \Sigma (x - \bar{x})(y - \bar{y})}{\sigma_x \sigma_y}. \quad (5.72)$$

Действительно, числители (5.72) и (5.71) равны

$$\begin{aligned} \frac{1}{n} \Sigma (x - \bar{x})(y - \bar{y}) &= \frac{1}{n} \Sigma xy - \frac{1}{n} \Sigma \bar{x}y - \\ &- \frac{1}{n} \Sigma x\bar{y} + \frac{1}{n} \Sigma \bar{x}\bar{y} = \bar{x}\bar{y} - \bar{x} \frac{1}{n} \Sigma y - \bar{y} \frac{1}{n} \Sigma x + \\ &+ \frac{1}{n} \Sigma \bar{x}\bar{y} = \bar{x}\bar{y} - \bar{x}\bar{y} - \bar{y}\bar{x} + \bar{x}\bar{y} = \bar{x}\bar{y} - \bar{x}\bar{y}. \end{aligned}$$

Все приведенные формулы представляют определенный интерес, особенно с точки зрения теоретических исследований. Практически же удобнее вычислять коэффициент корреляции, пользуясь выражением

$$r = \frac{n \Sigma xy - \Sigma x \Sigma y}{\sqrt{n \Sigma x^2 - (\Sigma x)^2} \sqrt{n \Sigma y^2 - (\Sigma y)^2}}, \quad (5.73)$$

которое получается при умножении (5.71) на n^2 .

В качестве примера рассчитаем коэффициент корреляции для случая прямой регрессии времени безотказной работы приборов по времени тренировки (табл. 5.20). Расчет удобно вести, записывая промежуточные данные в

таблицу (табл. 5.21). Подставляя данные в формулу (5.73), получаем

$$r = \frac{10 \cdot 121\,920 - 1050 \cdot 1093}{\sqrt{10 \cdot 118 \cdot 500 - 1050^2} \sqrt{10 \cdot 126\,549 - 1093^2}} = \\ = \frac{71\,550}{\sqrt{72\,500} \sqrt{70\,841}} = \frac{71\,550}{71\,665,2} = 0,948.$$

Коэффициент корреляции между надежностью аппаратуры и ее тренированностью оказался весьма большим. Но этот результат, разумеется, нельзя считать надежным,

Таблица 5.21

x	y	x^2	y^2	xy
60	74	3600	5476	4440
70	78	4900	6064	5460
80	72	6400	5184	5760
90	108	8100	11 664	9720
100	110	10 000	12 100	11 000
110	98	12 100	9604	10 780
120	135	14 400	18 225	16 200
130	138	16 900	19 044	17 940
140	138	19 600	19 044	19 320
150	142	22 500	20 164	21 300
$\Sigma = 1050$	1093	118 500	126 549	121 920

поскольку число наблюдений в нашем примере невелико ($n = 10$).

При большом числе наблюдений, когда материал группируют и составляют корреляционную таблицу, коэффициент корреляции вычисляют по формуле

$$r_{y'x'} = \frac{n\sum x'y' - \sum x'\sum y'}{\sqrt{n\sum x'^2 - (\sum x')^2} \sqrt{n\sum y'^2 - (\sum y')^2}}, \quad (5.74)$$

$$x' = \frac{x - C_x}{i_x}; \quad y' = \frac{y - C_y}{i_y}, \quad (5.75)$$

где C_x, C_y — новые начала отсчета по x и y , а i_x, i_y — интервалы группировки по x и y соответственно. Эти преобразования не отражаются на величине коэффициента корреляции

$$r_{xy} = r_{y'x'}.$$

Действительно, поскольку из (5.75)

$$x = xi_x + c_x, \quad y = y'i_y + c_y,$$

Таблица 5.22

	x'	-2	-1	0	1	2	Итого	iy'	iy'^2
y'	x	60	80	100	120	140			
2	140		1 2	4 8	6 12	15 30	26	52	104
1	120		5 5	20 20	13 13	4 4	42	42	42
0	100	2 0	23 0	49 0	7 0	4 0	85	0	0
-1	80	2 -2	16 -16	16 -16	2 -2	2 -2	38	-38	38
-2	60	6 -12	3 -6				9	-18	36
Итого n		10	48	89	28	25	200	38	220
nx'		-20	-48	0	28	50	10		
nx'^2		40	48	0	28	100	216		
$\Sigma iy'y'$		14	-15	12	23	32	38		
$x'\Sigma iy'y'$		28	15	0	23	64	130		

имеем

$$\bar{x} = \bar{x}'i_x + c_x, \quad \bar{y} = \bar{y}'i_y + c_y;$$

$$\sigma_x = i_x\sigma'_x, \quad \sigma_y = i_y\sigma'_y;$$

$$\frac{1}{n} \Sigma (x - \bar{x})(y - \bar{y}) = i_x i_y \frac{1}{n} \Sigma (x' - \bar{x}')(y' - \bar{y}').$$

Следовательно,

$$\begin{aligned}
 r_{yx} &= \frac{\frac{1}{n} \Sigma (x - \bar{x})(y - \bar{y})}{\sigma_x \sigma_y} = \frac{i_x i_y \frac{1}{n} \Sigma (x' - \bar{x}')(y' - \bar{y}')}{i_x i_y \sigma'_x \sigma'_y} = \\
 &= \frac{\frac{1}{n} \Sigma (x' - \bar{x}')(y' - \bar{y}')}{\sigma'_x \sigma'_y} = r_{y'x'}.
 \end{aligned}$$

Для примера рассчитаем коэффициент корреляции между временем безотказной работы приборов и временем тренировки по имеющимся у нас данным о партии в 200 приборов. Для расчета составляем таблицу (табл. 5.22). Коэффициент корреляции получим, подставляя данные таблицы в выражение (5.74)

$$r = \frac{n\sum x'y' - \sum x' \sum y'}{\sqrt{n\sum x'^2 - (\sum x')^2} \sqrt{n\sum y'^2 - (\sum y')^2}} =$$

$$= \frac{200 \cdot 130 - 10 \cdot 38}{\sqrt{200 \cdot 216 - (10)^2} \sqrt{200 \cdot 220 - (38)^2}} = \frac{25\,620}{42828,5} = 0,598.$$

Коэффициент корреляции при нелинейной зависимости

Как уже говорилось, мерой тесноты связи переменных y и x служит корреляционное отношение. Если η или $\eta_T = 1$, то зависимость функциональная. Чем меньше значение η , тем дальше зависимость y и x от функциональной.

Иногда требуется решить более простой вопрос: определить, в какой мере соблюдается строгая пропорциональность в изменении переменных y и x . Для этого вычисляют коэффициент корреляции r . Если $r = \pm 1$, говорят, что налицо строгая пропорциональность в изменении y и x . Чем ближе к нулю значение r , тем дальше характер изменения y и x от пропорционального. Поэтому коэффициент корреляции часто называют мерой пропорционального изменения или мерой так называемой спрямленной зависимости переменных. Этот смысл коэффициент корреляции сохраняет независимо от того, прямолинейна или криволинейна теоретическая линия регрессии y по x .

При прямолинейной зависимости коэффициент корреляции является одновременно и показателем тесноты корреляционной связи. При нелинейной зависимости величина коэффициента корреляции может оказаться значительно меньше теоретического корреляционного отношения, рассчитанного в предположении нелинейной зависимости. Может оказаться даже, что при спрямлении криволинейных зависимостей коэффициент корреляции будет близким к нулю, в то время как корреляционное отношение близко к единице (рис. 5.18). Если теоретическая линия регрессии монотонно возрастает или убывает и не слишком уклоняется

от прямой, то коэффициент корреляции можно с успехом использовать для предварительной ориентировки в вопросе о тесноте связи.

Вычисление коэффициента корреляции при нелинейной зависимости позволяет:

- 1) охарактеризовать степень приближения исследуемой корреляционной зависимости к линейной функциональной;
- 2) ориентировочно определить тесноту корреляционной зависимости.

Второй пункт имеет большое практическое значение, так как коэффициент корреляции вычислить намного проще, чем η или η_T . В большинстве практических задач измерение тесноты связи обычно начинают с нахождения коэффициента корреляции, иногда даже до определения линии регрессии.

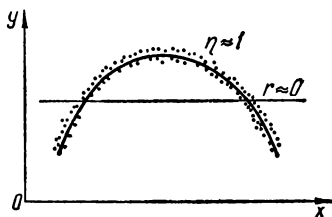


Рис. 5.18. Спрямленная нелинейная регрессия.

Определение прямой регрессии по основным статистическим параметрам

Для ряда распределения с одним варьирующим признаком в качестве основных статистических показателей мы рассматриваем \bar{x} и σ .

Для двух переменных x и y таблица основных показателей имеет вид

\bar{x}	\bar{y}
σ_x	σ_y
r_{yx}	

Эти показатели рассчитывают в первую очередь наряду с эмпирической линией регрессии. Затем уже строится теоретическая линия регрессии и более точно определяется теснота связи с помощью корреляционных отношений.

Если в качестве теоретической линии регрессии выбрана прямая вида

$$\bar{y}_x = a + bx,$$

то ее параметры легко найти с помощью основных статистических показателей. Для этого используются формулы

$$b = r \frac{\sigma_y}{\sigma_x}, \quad (5.76)$$

$$a = \bar{y} - b\bar{x}. \quad (5.77)$$

Выражение (5.76) получено из общей формулы для коэффициента корреляции (5.66), а (5.77) — из первого нормального уравнения $\Sigma y = na + b\Sigma x$ путем деления его на n .

Параметры a и b , вычисленные из (5.76) и (5.77), дают уравнение регрессии, в точности совпадающее с уравнением, полученным методом наименьших квадратов. Предлагаем убедиться в этом самостоятельно на одном из рассмотренных ранее примеров.

Сопряженные показатели корреляции

До сих пор мы четко определяли одну из переменных как функцию y , а другую — как аргумент x . При любом конкретном исследовании этот выбор зависит от конечной цели.

Если изменение одной переменной обусловлено изменением другой, то естественно первую назвать функцией, аргументом которой является вторая переменная.

Но очень часто приходится рассматривать пары коррелированных величин, когда обе можно рассматривать и как причины, и как следствия друг друга. Известно, например, что увеличение объема продукции способствует снижению себестоимости (из-за сокращения так называемых постоянных или слабо меняющихся расходов). С другой стороны, снижение себестоимости — это существенный фактор увеличения объема продукции. В этом случае корреляция между объемом продукции и уровнем себестоимости имеет двустороннюю причинную направленность.

Встречаются задачи, в которых ни одну из коррелированных величин нельзя рассматривать как причину изменения другой. Корреляция обусловлена общими причинами, которые влияют на обе переменные. Например, показатели надежности транзисторов различных типов из года в год возрастают и обнаруживают достаточно тесную связь. Но очевидно, что между ними нет причинной зависимости. Корреляция возникает под влиянием общих факторов, влияющих на повышение качества (улучшение сырья, совершенствование технологии, повышение квалификации рабочих и т. д.).

Интересен класс задач, в которых следствие принимается в качестве аргумента, а причина — в качестве функции. Например, выпуск продукции и количество потребленного сырья на предприятиях связаны прямой корреляционной зависимостью. В данном случае расходуемое сырье — это действительная причина возникновения продукции.

Но с экономической точки зрения не количество сырья обуславливает объем продукции, а наоборот, производственная программа определяет потребность в сырье. В подобных случаях для оценки потребности в исходных материалах

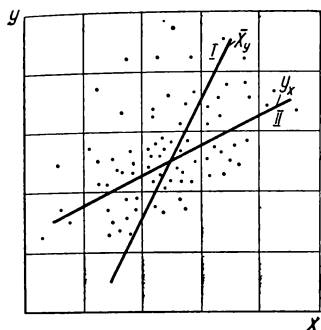


Рис. 5.19. Взаимно-сопряженные прямые регрессии.

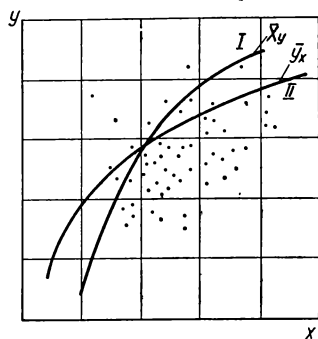


Рис. 5.20. Взаимно-сопряженные линии регрессии при нелинейной зависимости.

возникает задача измерения связи, когда объем продукции принимается в качестве аргумента, а количество исходных материалов — в качестве функции.

Вообще говоря, во многих задачах статистического измерения связи коррелированные переменные равноправны с точки зрения выбора их в качестве функции или аргумента. Вот почему, кроме определения статистических параметров и уравнений регрессии, характеризующих форму и тесноту связи y по x , представляет интерес вычисление аналогичных показателей, характеризующих форму и тесноту связи x по y . Такие показатели называют взаимно-сопряженными. Методика расчета взаимно-сопряженных показателей совершенно аналогична той, которая применяется для расчета основных показателей. Не решая числовые примеры, перейдем к характеристике некоторых особенностей статистических показателей связи, присущих корреляционной зависимости.

На рис. 5.19 и 5.20 показаны взаимно-сопряженные линии регрессии в предположении линейной (рис. 5.19) и

параболической (рис. 5.20) зависимости переменных y и x на одном и том же поле корреляции. Линии \bar{y}_x определены как характеристики корреляционной зависимости y от x . При линейной зависимости

$$\bar{y}_x = a + bx,$$

и в случае параболы

$$\bar{y}_x = a + bx + cx^2.$$

Здесь переменная x рассматривалась как аргумент, а переменная y — как функция.

Линии \bar{x}_y построены как характеристики зависимости x от y . Для линейной зависимости

$$\bar{x}_y = a + by,$$

для параболы

$$\bar{x}_y = a + by + cy^2.$$

Здесь переменная y рассматривалась как аргумент, а x — как функция.

Сразу же заметим, что наличие различающихся сопряженных линий регрессии является особенностью корреляционной зависимости. Для функциональной зависимости безразлично, относительно какой переменной решено уравнение, — графическое представление от этого не изменяется. Сопряженные уравнения при этом получают одно из другого тождественным преобразованием.

Для корреляционной зависимости сопряженные уравнения нельзя получить одно из другого тождественным преобразованием, они необратимы. Как видно из рисунков, y изменяется по отношению к x значительно быстрее на линии II (x по y) по сравнению с линией I (y по x). Это говорит о том, что признак изменяется медленнее, когда он выступает в корреляционной функции на правах функции по сравнению с изменением его на правах аргумента. Описанное явление носит название закона регрессии.

Теперь проанализируем сопряженные показатели тесноты связи. Совершенно очевидно, что коэффициент корреляции r не зависит от того, какая переменная, x или y , принята в качестве аргумента, а какая — в качестве функции. Все формулы, введенные для коэффициента корреляции, симметричны относительно переменных x и y

$$r_{yx} = r_{xy} = r.$$

Из равенства сопряженных коэффициентов корреляции ясно, что кроме формулы

$$r = b \frac{\sigma_x}{\sigma_y} \quad (5.78)$$

можно пользоваться выражением

$$r = b_1 \frac{\sigma_y}{\sigma_x}, \quad (5.79)$$

где b_1 — коэффициент регрессии x по y .

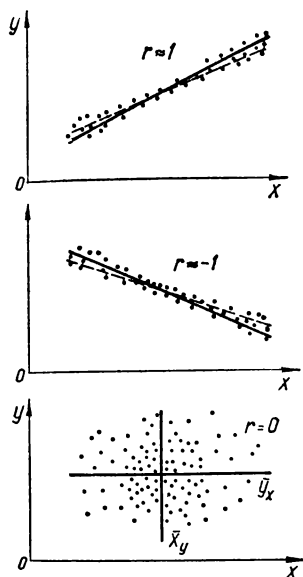


Рис. 5.21. Связь взаимно-сопряженных линий регрессии с коэффициентом корреляции.

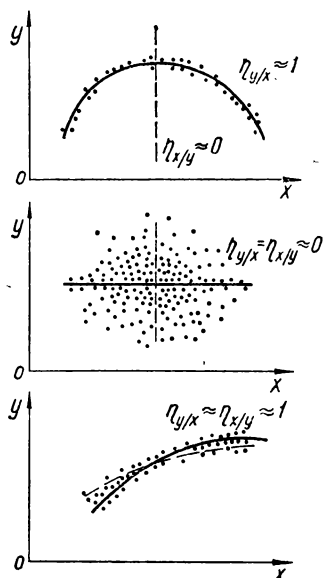


Рис. 5.22. Взаимно-сопряженные корреляционные отношения

Неравенство $|r| \leq 1$ показывает, что среднее изменение функции меньше изменения аргумента, если обе переменные выражены в стандартизованной системе единиц (сигмах). Это другая форма выражения закона регрессии.

Сопоставляя формулы (5.78) и (5.79), получаем

$$r^2 = b b_1, \quad (5.80)$$

т. е. произведение сопряженных коэффициентов регрессии равно квадрату коэффициента корреляции.

При $r = \pm 1$, т. е. при строгой пропорциональности изменения x и y , из формулы (5.80) имеем

$$b = \frac{1}{b_1},$$

что подтверждает совпадение в этом случае прямых регрессии.

Чем меньше абсолютная величина коэффициента корреляции, тем больше расхождение между взаимно-сопряженными прямыми регрессии. В пределе, при $r = 0$, прямые регрессии взаимно перпендикулярны, так как из выражений (5.78) и (5.79) следует, что b и b_1 здесь равны нулю (рис. 5.21).

Заметим также, что знаки b и b_1 совпадают со знаком коэффициента корреляции и между собой: они либо оба положительны, либо оба отрицательны.

В отличие от коэффициента корреляции корреляционные отношения y по x и x по y (теоретические и эмпирические) при криволинейной зависимости имеют различные значения. Чаще всего $\eta_{y/x}$ и $\eta_{x/y}$ близки друг другу, но бывают случаи, когда расхождение значительно. Не исключена также возможность, скажем, $\eta_{y/x} \approx 1$, а $\eta_{x/y} \approx 0$ (рис. 5.22). При $\eta_{y/x} = \eta_{x/y} = 0$ сопряженные линии регрессии представляют собой взаимно перпендикулярные прямые. При $\eta_{y/x} = \eta_{x/y} = 1$ обе линии регрессии совпадают.

Контрольные вопросы и задания

1. Что характеризует собой эмпирическое корреляционное отношение?
2. В чем характерная особенность теоретического корреляционного отношения?
3. Почему корреляционное отношение дает завышенную оценку по сравнению с фактической теснотой связи?
4. Дайте геометрическое представление дисперсий на поле корреляции.
5. Докажите, что при прямолинейной зависимости корреляционное отношение равно коэффициенту корреляции.
6. Что характеризует коэффициент корреляции при нелинейной зависимости?
7. Как определить уравнение линии регрессии по основным статистическим параметрам?
8. Охарактеризуйте сопряженные показатели корреляции.

Корреляция многих переменных.

Уравнение множественной регрессии

Зависимость функции от двух или нескольких аргументов исследуют при помощи множественно-корреляционного анализа. В простейшем случае прямолинейной зависимости уравнение регрессии p переменных имеет вид

$$\bar{x}_{1,2,3,\dots,p} = b_1 + b_2 x_2 + b_3 x_3 + \dots + b_p x_p, \quad (5.81)$$

где $\bar{x}_{1,2,3\dots p}$ — среднее значение функции x_1 при заданных значениях аргументов x_2, x_3, \dots, x_p . Значение функции x_1 (выходной параметр) называют результативным признаком, а параметры x_2, x_3, \dots, x_p — факториальными признаками. Рассмотрим вначале линейную зависимость результативного признака (x_1) от двух факториальных (x_2, x_3). Уравнение регрессии можно записать в виде

$$\bar{x}_{1,2,3} = b_1 + b_2x_2 + b_3x_3. \quad (5.82)$$

Параметры этого уравнения находят из решения системы нормальных уравнений по способу наименьших квадратов

$$\left. \begin{aligned} n \cdot b_1 + b_2 \Sigma x_2 + b_3 \Sigma x_3 &= \Sigma x_1, \\ b_1 \Sigma x_2 + b_2 \Sigma x_2^2 + b_3 \Sigma x_2 x_3 &= \Sigma x_1 x_2, \\ b_1 \Sigma x_3 + b_2 \Sigma x_2 x_3 + b_3 \Sigma x_3^2 &= \Sigma x_1 x_3, \end{aligned} \right\} \quad (5.83)$$

где n — число одновременных наблюдений по трем признакам;

$\Sigma x_1, \Sigma x_2, \Sigma x_3$ — суммы соответствующих значений по этим признакам.

Для расчетов удобно составить таблицу промежуточных данных $\Sigma x_1, \Sigma x_2, \Sigma x_3, \Sigma x_1 x_2$ и т. д.

С возрастанием числа факториальных признаков возрастает и число членов уравнения регрессии. Для трех факториальных признаков x_2, x_3, x_4 имеем

$$\bar{x}_{1,2,3,4} = b_1 + b_2x_2 + b_3x_3 + b_4x_4. \quad (5.84)$$

Параметры уравнения (5.84) находят из системы четырех нормальных уравнений

$$\left. \begin{aligned} nb_1 + b_2 \Sigma x_2 + b_3 \Sigma x_3 + b_4 \Sigma x_4 &= \Sigma x_1, \\ b_1 \Sigma x_2 + b_2 \Sigma x_2^2 + b_3 \Sigma x_2 x_3 + b_4 \Sigma x_2 x_4 &= \Sigma x_1 x_2, \\ b_1 \Sigma x_3 + b_2 \Sigma x_2 x_3 + b_3 \Sigma x_3^2 + b_4 \Sigma x_3 x_4 &= \Sigma x_1 x_3, \\ b_1 \Sigma x_4 + b_2 \Sigma x_2 x_4 + b_3 \Sigma x_3 x_4 + b_4 \Sigma x_4^2 &= \Sigma x_1 x_4. \end{aligned} \right\} \quad (5.85)$$

Рассмотрим расчет уравнения множественной регрессии в общем случае.

Прежде всего все переменные зависимости между ними выражаем в стандартизованном масштабе с помощью формулы перехода

$$t_x = \frac{x - \bar{x}}{\sigma_x}, \quad (5.86)$$

где x — значение признака в натуральном масштабе;
 t_x — соответствующее значение в стандартизованном масштабе.

Из выражения (5.86) видно, что

$$\bar{t} = 0; \sigma_t = 1,$$

т. е. среднее значение признака равно нулю, а среднеквадратическое отклонение — единице. Это существенно упрощает формулы основных показателей линейной корреляции. Коэффициент корреляции между стандартизованными переменными x и y выражается формулой

$$r = \frac{1}{n} \sum t_x t_y. \quad (5.87)$$

Очевидно, что этот коэффициент равен коэффициенту корреляции между переменными, выраженными в натуральном масштабе. В уравнении прямой регрессии в стандартизованном масштабе

$$\bar{t}_{y,x} = r_{xy} t_x \quad (5.88)$$

отсутствует свободный член, а коэффициентом регрессии является r_{xy} .

Множественную прямолинейную регрессию можно представить уравнением в стандартизованном масштабе

$$\bar{t}_{1,23\dots p} = \beta_2 t_2 + \beta_3 t_3 + \dots + \beta_p t_p, \quad (5.89)$$

где t_2, t_3, \dots, t_p — стандартизованные значения переменных x_2, x_3, \dots, x_p ;

$\bar{t}_{1,23\dots p}$ — среднее значение стандартизованного резуль-
 тативного признака x_1 , соответствующее заданным значениям переменных x_2, x_3, \dots, x_p ;

$\beta_2, \beta_3, \dots, \beta_p$ — стандартизованные коэффициенты множественной регрессии.

Коэффициенты $\beta_2, \beta_3, \dots, \beta_p$ находят из условия

$$\sum [t_1 - \bar{t}_{1,23\dots p}]^2 = \min. \quad (5.90)$$

Если взять частные производные по $\beta_2, \beta_3, \dots, \beta_p$ функции

$$\begin{aligned} f &= \frac{1}{n} \sum [t_1 - \bar{t}_{1,23\dots p}]^2 = \\ &= \frac{1}{n} \sum [t_1 - \beta_2 t_2 - \beta_3 t_3 - \dots - \beta_p t_p] \end{aligned}$$

ального признака) на функцию (результативный признак). Для этого требуется вычислить соответствующие парные коэффициенты корреляции.

Коэффициенты $\beta_2, \beta_3, \dots, \beta_p$ показывают, на какую часть сигмы изменяется среднее значение результативного признака, если соответствующий факториальный признак увеличился на сигму, а прочие факториальные признаки остались без изменения. Другими словами, эти коэффициенты характеризуют скорость изменения среднего значения функции по каждому аргументу при постоянных значениях прочих. Так как все переменные выражены в сравнимых единицах измерения (сигмах), то коэффициенты $\beta_2, \beta_3, \dots, \beta_p$ характеризуют степень влияния изменения каждого факториального признака на результативный.

Уравнение чистой регрессии

Как изменится среднее значение результативного признака с изменением ряда факториальных признаков x_2, x_3, \dots, x_k , если другие факториальные признаки $x_{k+1}, x_{k+2}, \dots, x_p$ закрепить на среднем уровне?

Ответ на этот вопрос дает уравнение чистой регрессии.

Уравнение чистой регрессии получаем подстановкой в уравнение множественной регрессии значений $\bar{x}_{k+1}, \bar{x}_{k+2}, \dots, \bar{x}_p$ вместо $x_{k+1}, x_{k+2}, \dots, x_p$. Эта операция влияет лишь на свободный член уравнения множественной линейной регрессии. Для уравнения чистой линейной регрессии свободный член рассчитывается по формуле

$$b_{1(k+1, k+2, \dots, p)} = b_1 + b_{k+1} \bar{x}_{k+1} + b_{k+2} \bar{x}_{k+2} + \dots + b_p \bar{x}_p. \quad (5.93)$$

Тогда само уравнение имеет вид

$$\bar{x}_{1,2,3 \dots k(k+1, k+2, \dots, p)} = b_{1(k+1, k+2, \dots, p)} + b_2 x_2 + b_3 x_3 + \dots + b_k x_k, \quad (5.94)$$

где b_1, b_2, \dots, b_k — коэффициенты, непосредственно взятые из уравнения множественной регрессии.

Для упрощения уравнения чистой регрессии можно записать в стандартизованном масштабе. Для этого нужно просто опустить члены с закрепленными переменными в уравнении множественной регрессии

$$t_{1,2,3 \dots k(k+1, k+2, \dots, p)} = \beta_2 t_2 + \beta_3 t_3 + \dots + \beta_k t_k. \quad (5.95)$$

Вообще, уравнение чистой регрессии целесообразно рассматривать при незначительной тесноте связи между исключаемыми и остающимися признаками. При существенно тесной связи между всеми признаками нужно вычислять так называемые уравнения частных регрессий. В уравнениях частных регрессий исключаемые переменные закрепляются не на средних, а на других уровнях. Эти уровни следует выбирать, исходя из интересующих нас участков изменения основных (остающихся) признаков.

Коэффициент множественной корреляции

Для измерения тесноты связи, как и при двух переменных, можно воспользоваться корреляционным отношением. Теоретическое корреляционное отношение при множественной корреляции выражается как

$$\eta_{1,23\dots p}^2 T = \frac{\bar{\delta}_{1,23\dots p}^2 T}{\sigma_1^2}, \quad (5.96)$$

где

$$\bar{\delta}_{1,23\dots p}^2 T = \frac{1}{n} \sum [x_{1,23\dots p} - \bar{x}]^2;$$

$$\sigma_1^2 = \frac{1}{n} \sum (x_1 - \bar{x}_1)^2.$$

При линейной корреляции выражение (5.96) называется коэффициентом множественной корреляции $R_{1,23\dots p}$.

Коэффициент множественной корреляции можно представить в виде

$$R_{1,23\dots p} = \sqrt{\beta_2 r_{12} + \beta_3 r_{13} + \dots + \beta_p r_{1p}}. \quad (5.97)$$

Приведем вывод этого выражения. Если выразить переменные в стандартизованном масштабе, получим

$$R_{1,23\dots p}^2 = \frac{1}{n} \sum [\bar{t}_{1,23\dots p}]^2, \quad (5.98)$$

где

$$\bar{t}_{1,23\dots p} = \beta_2 t_2 + \beta_3 t_3 + \dots + \beta_p t_p.$$

Перепишем (5.98) в виде

$$\begin{aligned} R_{1,23\dots p}^2 &= \frac{1}{n} \sum \bar{t}_{1,23\dots p} (\beta_2 t_2 + \beta_3 t_3 + \dots + \beta_p t_p) = \\ &= \beta_2 \frac{1}{n} \sum \bar{t}_{1,23\dots p} t_2 + \beta_3 \frac{1}{n} \sum \bar{t}_{1,23\dots p} t_3 + \dots \\ &\quad \dots + \beta_p \frac{1}{n} \sum \bar{t}_{1,23\dots p} t_p. \end{aligned} \quad (5.99)$$

Из условия наименьших квадратов

$$\Sigma |t_1 - \bar{t}_{1,23 \dots p}|^2 = \min \quad (5.100)$$

получаем систему нормальных уравнений

$$\left. \begin{aligned} \Sigma [t_1 - \bar{t}_{1,23\dots p}] t_2 &= 0, \\ \Sigma [t_1 - \bar{t}_{1,23\dots p}] t_3 &= 0, \\ \Sigma [t_1 - \bar{t}_{1,23\dots p}] t_p &= 0. \end{aligned} \right\} \quad (5.101)$$

Из (5.101) имеем

$$\begin{aligned}\frac{1}{n} \Sigma \bar{t}_{1,23\dots p} t_2 &= \frac{1}{n} \Sigma t_1 t_2 = r_{12}, \\ \frac{1}{n} \Sigma \bar{t}_{1,23\dots,p} &= \frac{1}{n} \Sigma t_1 t_3 = r_{13}, \\ . &. \\ \frac{1}{n} \Sigma \bar{t}_{1,23\dots p} t_p &= \frac{1}{n} \Sigma t_1 t_p = r_{1p}.\end{aligned}\tag{5.102}$$

Подставляя (5.102) в (5.99), получаем

$$R_{1,23\dots p}^2 = \beta_2 r_{12} + \beta_3 r_{13} + \dots + \beta_p r_{1p},$$

откуда вытекает (5.97).

Если обозначить

$$\Delta' = \begin{vmatrix} r_{12} & r_{13} & \dots & r_{1p} & 0 \\ 1 & r_{23} & \dots & r_{2p} & r_{21} \\ r_{32} & 1 & \dots & r_{3p} & r_{31} \\ \dots & \dots & \dots & \dots & \dots \\ r_{p2} & r_{p3} & \dots & 1 & r_{p1} \end{vmatrix},$$

то коэффициент множественной корреляции можно рассчитать по формуле

$$R_{1,23\dots p} = \sqrt{\frac{\Delta'}{\Lambda}}, \quad (5.103)$$

где Δ — определитель системы нормальных уравнений по способу наименьших квадратов для множественной корреляции.

Эмпирические меры тесноты связи

Тесноту связи между различными переменными или факторами можно охарактеризовать с помощью эмпирических показателей. Эти показатели были введены различными ис-

следователями, занимавшимися задачами статистического анализа. Рассмотрим некоторые из эмпирических мер тесноты связи.

Коэффициент ассоциации A применяется для характеристики связи двух качественных признаков, представленных только двумя группами. Чтобы вычислить коэффициент A , необходимо построить четырехклеточную таблицу корреляции, которая характеризует связь между двумя явлениями. Каждое из явлений, в свою очередь, должно быть альтернативным (табл. 5.23).

Таблица 5.23

$y \backslash x$	+	-	Всего
+	a	b	$a+b$
-	c	d	$c+d$
Всего	$a+c$	$b+d$	

Таблица 5.24

$y \backslash x$	Без нарушений	С нарушениями	Всего
Удовлетворительное	68	5	73
Неудовлетворительное	11	16	27
Всего	79	21	100

Рассмотрим в качестве примера зависимость качества нефтепродукта y от качества регулирования технологического процесса (x). Выделим две группы по качеству нефтепродукта: удовлетворительное, неудовлетворительное и две группы режимов — с нарушением технологии и без нарушений. Взятые в разное время 100 проб нефтепродукта распределяются в корреляционной таблице (табл. 5.24).

Коэффициент ассоциации вычисляется по формуле

$$A = \frac{ad + bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}. \quad (5.104)$$

Подставив конкретные значения из табл. 5.24, получим:

$$A = \frac{68 \cdot 16 - 5 \cdot 11}{\sqrt{73 \cdot 27 \cdot 79 \cdot 21}} = 0,515.$$

Вычисленное значение коэффициента ассоциации свидетельствует о достаточно тесной прямой связи между качеством продукта и соблюдением технологических норм производства.

При обратной зависимости между признаками получаем $ad < bc$, и коэффициент ассоциации оказывается отрицательным.

Коэффициент взаимной сопряженности характеризует тесноту связи для качественных признаков, представленных более чем двумя группами. Для расчета этих коэффициентов применяют формулу Пирсона

$$C = \sqrt{\frac{\varphi^2}{1 + \varphi}} \quad (5.105)$$

или формулу Чупрова

$$K = \sqrt{\frac{\varphi^2}{(k_1 - 1)(k_2 - 1)}} \quad (5.106)$$

Таблица 5.25

$\begin{matrix} x \\ y \end{matrix}$	Больше номинального	Номинальное	Меньше номинального	
Больше гарантийного	$\begin{matrix} h_1 & h_1^2 \\ h_1^2 & \\ h_1 + h_4 + h_7 = a_1 \end{matrix}$	$\begin{matrix} h_2 & h_2^2 \\ h_2^2 & \\ h_2 + h_5 + h_8 = a_2 \end{matrix}$	$\begin{matrix} h_3 & h_3^2 \\ h_3^2 & \\ h_3 + h_6 + h_9 = a_3 \end{matrix}$	$\begin{matrix} h_1 + h_2 + h_3 = b_1 \\ a_1 + a_2 + a_3 = c_1 \end{matrix}$
Гарантийное	$\begin{matrix} h_4 & h_4^2 \\ h_4^2 & \\ h_1 + h_4 + h_7 = a_4 \end{matrix}$	$\begin{matrix} h_5 & h_5^2 \\ h_5^2 & \\ h_2 + h_5 + h_8 = a_5 \end{matrix}$	$\begin{matrix} h_6 & h_6^2 \\ h_6^2 & \\ h_3 + h_6 + h_9 = a_6 \end{matrix}$	$\begin{matrix} h_4 + h_5 + h_6 = b_2 \\ a_4 + a_5 + a_6 = c_2 \end{matrix}$
Меньше гарантийного	$\begin{matrix} h_7 & h_7^2 \\ h_7^2 & \\ h_1 + h_4 + h_7 = a_7 \end{matrix}$	$\begin{matrix} h_8 & h_8^2 \\ h_8^2 & \\ h_2 + h_5 + h_8 = a_8 \end{matrix}$	$\begin{matrix} h_9 & h_9^2 \\ h_9^2 & \\ h_3 + h_6 + h_9 = a_9 \end{matrix}$	$\begin{matrix} h_7 + h_8 + h_9 = b_3 \\ a_7 + a_8 + a_9 = c_3 \end{matrix}$
	$h_1 + h_4 + h_7$	$h_2 + h_5 + h_8$	$h_3 + h_6 + h_9$	$N; c_1 + c_2 + c_3$

Величина φ^2 в выражениях (5.105) и (5.106) называется показателем взаимной сопряженности. Он определяется суммой отношений квадратов частот каждой клетки корреляционной таблицы к произведению итоговых частот соответствующего столбца и строки. При вычитании из этой суммы единицы получается φ^2 . В формуле (5.106) k_1 — число групп по столбцам таблицы, k_2 — число групп по строкам.

Рассмотрим пример расчета коэффициента взаимной сопряженности между временем безотказной работы приборов y и временем их тренировки x . В каждой клетке таблицы 5.25 записаны частоты h_i , их квадраты и частные от

деления квадратов частот на сумму частот по столбцу. В итоговом столбце записаны суммы частот и отношения для вычисления показателей взаимной сопряженности. По определению

$$\varphi^2 = C_1 + C_2 + C_3 - 1. \quad (5.107)$$

Подставим конкретные значения, полученные при изучении партии из 150 приборов ($N = 150$). Данные сведены в таблицу 5.26 и расчет произведен с точностью до двух зна-

Таблица 5.26

$x \backslash y$	Больше номинального	Номинальное	Меньше номинального	Итого
Больше гарантийного	15 225 9,00	20 400 4,71	5 25 0,63	40 $\frac{14,34}{40} = 0,36$
Гарантийное	8 64 2,56	50 2500 29,41	20 400 0,00	78 $\frac{41,97}{78} = 0,54$
Меньше гарантийного	2 4 0,16	15 225 2,65	15 225 5,63	32 $\frac{8,44}{32} = 0,26$
Итого	25	85	40	150; 1,16

ков после запятой. Из формулы (5.107) получаем

$$\varphi^2 = 1,16 - 1,00 = 0,16.$$

По формуле Пирсона вычисляем

$$C = \sqrt{\frac{0,16}{1,16}} = 0,37,$$

что характеризует достаточно тесную связь между временем безотказной работы и временем тренировки.

Формула Чупрова дает несколько иной результат

$$K = \sqrt{\frac{0,16}{2 \cdot 2}} = 0,20.$$

Коэффициент взаимной сопряженности Чупрова более гибкий, он учитывает число групп k_1 и k_2 , образуемых по каждому признаку. Поэтому результат 0,20 более точный

по сравнению с коэффициентом взаимной сопряженности, вычисленной по формуле Пирсона.

Ранговый коэффициент корреляции позволяет определить тесноту связи между взаимосвязанными признаками в количественном выражении. Для вычисления рангового коэффициента корреляции необходимо записать все значения факториального признака в возрастающем (или убывающем) порядке (ранжировать). Соответственно запи-

Таблица 5.27

20,0	1
21,0	2
21,5	4
21,5	4
21,5	4
22,0	6
22,5	8,5
22,5	8,5
22,5	8,5
22,5	8,5
23,0	12
23,0	12
23,0	13
23,5	14
24,0	15
26,0	16

Таблица 5.28

8,0	21,5	1	4	—3	9
9,0	21,0	2,5	2	+0,5	0,25
9,0	22,5	2,5	8,5	—6	36
10,0	22,5	4	8,5	—4,5	20,25
10,5	21,5	5	4	+1	1
11,0	22,0	6	6	0	0
11,5	22,5	7,5	8,5	—1	1
11,5	22,5	7,5	8,5	—1	1
12,0	23,0	9	12	—3	9
13,0	20,0	10	1	+9	81
13,5	23,0	11,5	12	—0,5	0,25
13,5	26,0	11,5	16	—4,5	20,25
14,5	23,5	13,5	14	—0,5	0,25
14,5	21,5	13,5	4	+9,5	90,25
15,0	24,0	15	15	0	0
15,0	23,0	12	12	+4	16
					285,50

сываются значения результативного признака. Определяем ранг по обоим признакам, т. е. номер каждого признака в ранжированных рядах. В качестве примера рассмотрим зависимость между двумя признаками x и y . По обоим признакам определяются ранги R_x и R_y . Для одинаковых значений ранг определяется как частное от деления суммы их рангов на число этих одинаковых значений. Например, ряд y и соответствующие ранги приведены в таблице 5.27. В свободной таблице 5.28 записаны ранги R_x и R_y , разности рангов $R_x - R_y$ и квадраты этих разностей.

Значение рангового коэффициента корреляции определяется выражением

$$\rho = 1 \frac{\sum d^2}{n - (n^2 - 1)} \cdot \quad (5.108)$$

Если между признаками функциональная прямая зависимость, то $\Sigma d^2 = 0$ и $\rho = 1$, если зависимость обратная, то $\rho < 0$.

Для нашего примера

$$\rho = 1 - \frac{6 \cdot 6 \cdot 285,5}{16 \cdot 255} = 0,58,$$

что свидетельствует о достаточно тесной связи между исследуемыми переменными.

Таблица 5.29

x	\bar{x}	y	\bar{y}	$\text{sign } x - \bar{x}$	$\text{sign } y - \bar{y}$	Совпадение и несовпа- дение
8	12,0	21,5	22,5	—	—	<i>a</i>
9		21,0		—	—	<i>a</i>
9		22,5		—	0	
10		22,5		—	0	
10,5		21,5		—	—	<i>a</i>
11,0		22,0		—	—	<i>a</i>
11,5		22,5		—	0	
11,5		22,5		—	0	
12,0		23,0		0	+	
13,0		20,0		+	—	<i>b</i>
13,5		23,0		+	+	<i>a</i>
13,5		26,0		+	+	<i>a</i>
14,5		23,5		+	+	<i>a</i>
14,5		21,5		+	—	<i>b</i>
15,0		24,0		+	+	<i>a</i>
15,5		23,0		+	+	<i>a</i>
192		360,0				

Коэффициент Фехнера вычисляется на основе первых степеней отклонений всех значений взаимосвязанных признаков от среднего значения каждого признака. После определения отклонений от среднего сравниваются знаки отклонений по первому и второму признакам. Совпадения знаков отклонений обозначим через *a*, а несовпадения — через *b*.

Величина

$$l = \frac{\Sigma a - \Sigma b}{\Sigma a + \Sigma b} \quad (5.109)$$

позволяет судить о степени тесноты связи между признаками.

Рассчитаем коэффициент Фехнера для предыдущего примера (табл. 5.29)

$$\Sigma a = 9; \quad \Sigma b = 2;$$

$$i = \frac{9-2}{9+2} = 0,636.$$

Коэффициент Фехнера весьма близок к значению рангового коэффициента корреляции и свидетельствует о тесной связи между признаками.

Контрольные вопросы и задания

1. Что такое уравнение множественной регрессии?
2. Как применять метод наименьших квадратов для расчета параметров регрессии при множественно-корреляционной зависимости?
3. Покажите, каким образом можно определить уравнение чистой регрессии.
4. Выведите формулу для коэффициента множественной корреляции.
5. Перечислите и охарактеризуйте основные эмпирические показатели тесноты связи.

§ 6. ЭЛЕМЕНТЫ ТЕОРИИ ОШИБОК

Ошибки измерения. При различных измерениях на производстве и в научных исследованиях ошибки неизбежны. Каждое конкретное измерение дает лишь приближенное значение действительной измеряемой величины. Ошибкой измерения обычно называют разность между результатом измерения и истинным значением параметра

$$e = x - X.$$

В теории ошибок рассматриваются два вида ошибок измерения:

- 1) систематические ошибки, которые при данных условиях измерения имеют вполне определенное значение. К ним относятся, например, ошибки измерительной аппаратуры;
- 2) случайные, которые появляются в результате взаимодействия большого числа незначительных в отдельности факторов. Случайные ошибки в каждом отдельном случае имеют различные значения.

В статистическом анализе рассматриваются методы предсказания возникновения систематических ошибок и сведения их к минимуму или полной ликвидации. Что касается случайных ошибок, то при большом количестве измерений крупные ошибки встречаются реже мелких. Кроме того, число положительных ошибок приблизительно совпадает

с числом отрицательных и поэтому сумма всех ошибок равна нулю.

Обычно малыми ошибками пренебрегают. Обращают внимание лишь на наибольшую возможную ошибку, чтобы обезопасить результаты измерений от влияния случайных неточностей.

Существуют случайные ошибки, которыми нельзя пренебречь, так как они достаточно велики по абсолютной величине. С другой стороны, эти ошибки подчинены определенному закону, позволяющему установить зависимость между величиной ошибки и вероятностью ее появления.

Средняя ошибка сводного результата измерения Если в качестве действительного значения измеряемой величины принять среднюю арифметическую из всех результатов n измерений, точность одного измерения можно оценить при помощи средней арифметической из абсолютных значений ошибок

$$\bar{\varepsilon} = \frac{\sum |x - \bar{x}|}{n}, \quad (5.110)$$

где n — число измерений; x — численное значение отдельных измерений; \bar{x} — средняя арифметическая результатов измерений.

Мерой точности соответствия принятой средней арифметической \bar{x} истинному значению измеряемой величины x служит средняя ошибка сводного результата измерения

$$\bar{\varepsilon}_x = \frac{\bar{\varepsilon}}{\sqrt{n}}. \quad (5.111)$$

Пусть, например, произведено 10-кратное измерение размера детали (в мм). Результаты измерений в возрастающем порядке: 138; 139; 140; 141; 141; 142; 142; 143; 144; 145.

Вычислим среднюю арифметическую абсолютных значений ошибок. Средняя арифметическая результатов измерений

$$\bar{x} = \frac{138 + 139 + \dots + 145}{10} = 141,5 \text{ мм.}$$

Ошибки измерения:

$$138 - 141,5 = -3,5;$$

$$139 - 141,5 = -2,5;$$

$$140 - 141,5 = -1,5;$$

$$141 - 141,5 = -0,5;$$

$$141 - 141,5 = -0,5;$$

$$142 - 141,5 = +0,5;$$

$$142 - 141,5 = +0,5;$$

$$143 - 141,5 = +1,5;$$

$$144 - 141,5 = +2,5;$$

$$145 - 141,5 = +3,5.$$

Средняя арифметическая из абсолютных значений ошибок

$$\bar{e} = \frac{3,5 + 2,5 + \dots + 2,5 + 3,5}{10} = 1,7.$$

Вычислим теперь среднюю ошибку сводного результата измерения

$$\bar{e}_x = \frac{\bar{e}}{\sqrt{n}} = \frac{1,7}{\sqrt{10}} \approx 0,54 \text{ мм.}$$

Таким образом, точность соответствия величины 141,5 истинному размеру характеризуется средней ошибкой, равной 0,54 мм.

Средняя квадратическая ошибка Примем в качестве меры точности одного измерения среднюю квадратическую ошибок измерений, т. е.

$$\sigma_x = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}}. \quad (5.112)$$

В этом случае средняя квадратическая ошибка найденной средней арифметической ошибок измерения равна

$$\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}}. \quad (5.113)$$

Зависимость между средней квадратической ошибкой и средней ошибкой сводного результата измерения можно записать в виде

$$\frac{\sigma_{\bar{x}}}{\bar{e}_x} \approx 1,25. \quad (5.114)$$

Рассчитаем среднюю квадратическую ошибку по данным предыдущего примера. Мера точности одного измерения

$$\begin{aligned} \sigma_x &= \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}} = \\ &= \sqrt{\frac{(-3,5)^2 + (-2,5)^2 + \dots + (2,5)^2 + (3,5)^2}{9}} = 2,15. \end{aligned}$$

Средняя квадратическая ошибка найденной средней арифметической, равной 141,5 мм,

$$\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}} = \frac{2,15}{\sqrt{10}} \approx 0,68.$$

Сопоставление средней квадратической ошибки со средней ошибкой свободного результата дает

$$\frac{\sigma_{\bar{x}}}{\varepsilon_{\bar{x}}} = \frac{0,68}{0,54} \approx 1,26.$$

Вероятная ошибка. Мерой точности одного измерения может служить вероятная ошибка

$$\delta_x = 0,6745 \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}} \approx \frac{2}{3} \sigma_x. \quad (5.115)$$

При этом вероятная ошибка сводного результата измерений определяется как

$$\delta_{\bar{x}} \approx \frac{2}{3} \sigma_{\bar{x}}. \quad (5.116)$$

По данным предыдущего примера вероятная ошибка сводного результата измерений

$$\delta_{\bar{x}} = \frac{2}{3} \cdot 0,68 = 0,46.$$

Наиболее вероятные границы сводных результатов. Примем в качестве действительного значения измеряемой величины среднюю арифметическую всех измерений. В теории ошибок отклонения результатов измерений (x) от средней (\bar{x}) называют «кажущимися ошибками». С их помощью можно оценить точность соответствия средней арифметической неизвестному истинному значению измеряемой величины (X).

С этой целью обычно используют удвоенную или утроенную среднеквадратичную ошибку сводного результата измерений или его вероятную ошибку. Получаем

$$X = \bar{x} \pm 3\sigma_{\bar{x}}, \quad (5.117)$$

или

$$X = \bar{x} \pm 3\delta_{\bar{x}}, \quad (5.118)$$

Если ошибки подчинены нормальному закону распределения, то найденные из (5.117) или (5.118) границы известной истинной величины соблюдаются с большой вероятностью (соответственно 0,997 и 0,954).

Для рассмотренного примера границы истинного значения размера

$$X = \bar{x} \pm 3\sigma_{\bar{x}} = 141,5 \pm 3 \cdot 0,68 = 141,5 \pm 2,04.$$

Контрольные вопросы и задания

1. Какие два вида ошибок рассматриваются при измерениях и расчетах?
2. Что такое средняя ошибка сводного результата измерения?
3. Как связаны между собой средняя квадратическая ошибка и средняя ошибка сводного результата измерения?
4. Что такое вероятная ошибка? Как определить вероятную ошибку сводного результата измерений?
5. Каким образом определяются наиболее вероятные границы сводных результатов?

Глава 6

ТЕОРИЯ СПЕКТРОВ

Спектральный способ описания различных явлений получил широкое признание. Спектральный язык сегодня стал всеобщим для тех, кто имеет дело с техническим применением различного рода колебаний. Спектральными методами можно описывать не только явления, но и многочисленные свойства различных инженерных устройств.

Эволюция спектральных представлений выглядит следующим образом.

В результате интегрирования функции времени в бесконечных пределах получаются *спектры*, зависящие только от частоты. В дальнейшем учет реальных условий различных экспериментов заставляет ввести понятие *текущего спектра*, который определяется также преобразованием Фурье, но с переменным верхним пределом интегрирования, в качестве которого фигурирует текущее время. Таким образом, появляется спектральная функция, зависящая не только от частоты, но и от времени. Это некоторое промежуточное понятие, сближающее частотные и временные представления. Далее процесс сближения этих двух представлений продолжается: вводится понятие *мгновенного спектра*, которое близко к понятию мгновенной частоты. На этой стадии можно говорить о *синусоиде с переменной частотой*.

§ 1. ГАРМОНИЧЕСКИЙ АНАЛИЗ

На практике часто приходится иметь дело с *периодическими явлениями*, т. е. воспроизводимыми в прежнем виде через определенный промежуток времени T , называемый периодом. Различные величины, связанные с рассматриваемым периодическим явлением, по истечении периода T возвращаются к своим прежним значениям и представля-

ют, следовательно, периодические функции от времени t , характеризующиеся равенством

$$x(t + T) = x(t). \quad (6.1)$$

При этом предполагается, что $x(t)$ — функция, определенная на всей числовой прямой. Далее: если T есть период функции, то nT (где n — любое целое число) есть тоже период рассматриваемой функции.

Таким образом, функция, имеющая период отличный от нуля, называется периодической.

Отметим, что (6.1) выражает основное свойство периодической функции, состоящее в том, что ход явления периодически повторяется и что периодичность эта существует вечно, т. е. для всех времен от $-\infty$ до $+\infty$. Из этого сразу можно заключить, что периодических явлений в строгом смысле определения (6.1) в действительности нет и быть не может. Периодическая же функция есть математическая абстракция, полезность которой станет очевидной немного ниже. Основные свойства периодических функций следующие:

1. Если функция $x(t)$ имеет период T , то функция $y(t) = x(at)$ имеет период $\frac{T}{a}$. В самом деле,

$$y\left(t + \frac{T}{a}\right) = x\left[a\left(t + \frac{T}{a}\right)\right] = x(at + T) = x(at) = y(t).$$

2. Если $x(t)$ имеет период T , то интеграл этой функции, взятый в пределах, отличающихся на T , не зависит от выбора нижнего предела интегрирования, т. е.

$$\int_b^{b+T} x(t) dt = \int_0^T x(t) dt.$$

Причем это равенство справедливо при всяком b . Действительно, пусть, например, $0 < b < T$, тогда

$$\int_b^{b+T} = \int_b^T + \int_T^{b+T} = \int_b^T + \int_0^b = \int_0^T,$$

если учесть, что вследствие периодичности $\int_T^{b+T} = \int_0^b$. Наглядная иллюстрация второго свойства периодических функций дана на рис. 6.1.

Простейшей из периодических функций (если не счи-

татъ постоянной) является синусоидальной вида: $A \sin \times \times (\omega t + \alpha)$, где ω есть частота, связанная с периодом T соотношением

$$\omega = \frac{2\pi}{T}. \quad (6.2)$$

Из подобных простейших периодических функций можно составить и более сложные. Очевидно, составляющие синусоидальные величины должны быть разных частот, так как сложение синусоидальных величин одной и той же частоты приводит опять к синусоидальной величине той же частоты, хотя и другой амплитуды. Если же сложить несколько величин вида

$$\begin{aligned} x_1 &= A_1 \sin \omega t; \\ x_2 &= A_2 \sin 2\omega t; \\ x_3 &= A_3 \sin 4\omega t; \\ &\dots \dots \dots \end{aligned} \quad (6.3)$$

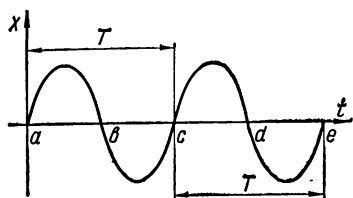


Рис. 6.1. Графическая интерпретация второго свойства периодической функции.

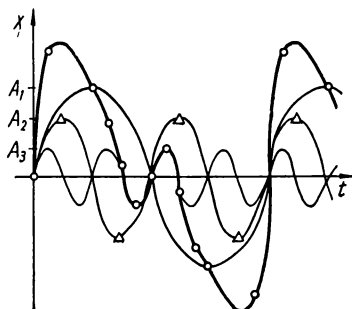


Рис. 6.2. Сложение периодических функций

то получим в результате периодическую функцию, существенно отличающуюся от каждой из величин вида (6.3) с периодом T . Это наглядно показано на рис. 6.2. Естественно, что при сложении величин из бесконечного ряда, составленного по типу (6.3), эффект отличия результирующей периодической функции от синусоиды проявится еще больше.

Однако на практике может возникнуть обратная задача: представить определенную периодическую функцию $x(t)$ с периодом T в виде суммы конечного или хотя бы бесконечного множества синусоидальных величин вида (6.3).

Можно доказать, что если периодическая функция ограничена, кусочно-непрерывна и имеет на протяжении периода конечное число экстремумов, то обратная задача разрешима. Другими словами, при выполнении указанных усло-

вий функция $x(t)$ может быть разложена в тригонометрический ряд

$$x(t) = A_0 + A_1 \sin(\omega t + \alpha_1) + A_2 \sin(2\omega t + \alpha_2) + \\ + A_3 \sin(3\omega t + \alpha_3) + \dots = A_0 + \sum_{k=1}^{\infty} A_k \sin(k\omega t + \alpha_k), \quad (6.4)$$

где

$$A_0, A_1, \dots, A_k, \dots; \alpha_2, \alpha_3, \dots, \alpha_k, \dots$$

— постоянные, имеющие особое значение для каждой функции, а частота ω определяется формулой (6.2).

Геометрически это означает, что любую периодическую функцию можно представить как результат наложения ряда синусоид.

Отдельные синусоидальные величины, входящие в состав разложения (6.4), называют *гармоническими составляющими (гармониками)* функции $x(t)$. Сам же процесс разложения периодической функции на гармоники носит название *гармонического анализа*.

Если в качестве независимой переменной выбрать τ при условии, что

$$\tau = \omega t = \frac{2\pi t}{T},$$

то получим

$$f(\tau) = x\left(\frac{\tau}{\omega}\right).$$

Эта функция тоже периодическая, но уже с периодом 2π . Тогда разложение (6.4) примет вид

$$f(\tau) = A_0 + A_1 \sin(\tau + \alpha_1) + A_2 \sin(2\tau + \alpha_2) + \\ + A_3 \sin(3\tau + \alpha_3) + \dots = A_0 + \sum_{k=1}^{\infty} A_k \sin(k\tau + \alpha_k). \quad (6.5)$$

Развернув члены ряда (6.5) по формуле для синуса суммы и положив

$$A_0 = a_0; \quad A_k \sin \alpha_k = a_k; \quad A_k \cos \alpha_k = b_k; \\ (k = 1, 2, 3, \dots),$$

придем к окончательной форме тригонометрического разложения функции $f(\tau)$

$$f(\tau) = a_0 + (a_1 \cos \tau + b_1 \sin \tau) + (a_2 \cos 2\tau + b_2 \sin 2\tau) + \\ + (a_3 \cos 3\tau + b_3 \sin 3\tau) + \dots = \\ = a_0 + \sum_{k=1}^{\infty} (a_k \cos k\tau + b_k \sin k\tau), \quad (6.6)$$

или в более развернутом виде

$$f(\tau) = a_0 + \sum_{k=1}^{\infty} \left(a_k \cos 2\pi k \frac{t}{T} + b_k \sin 2\pi k \frac{t}{T} \right). \quad (6.7)$$

Выражения (6.6) и (6.7) представляют собой ряд Фурье, записанный в вещественной форме.

§ 2. МЕТОД ЭЙЛЕРА — ФУРЬЕ ДЛЯ ОПРЕДЕЛЕНИЯ КОЭФФИЦИЕНТОВ РЯДА ФУРЬЕ

Для того чтобы разложение заданной функции в ряд Фурье было возможно, необходимо знать коэффициенты $a_0, a_1 b_1, \dots, a_k b_k \dots$ в выражениях (6.6) и (6.7).

Один из методов определения этих коэффициентов был указан Эйлером и независимо от него Фурье.

Предполагается, что заданная функция $f(\tau)$ интегрируема в промежутке $[-\pi; \pi]$, в собственном или в несобственном смысле. Кроме того, в последнем случае предполагается, что функция абсолютно интегрируема.

Допустим, что разложение $f(\tau)$ в ряд Фурье существует и представлено выражением (6.6). Проинтегрируем это выражение от $-\pi$ до π .

В результате будем иметь

$$\int_{-\pi}^{\pi} f(\tau) d\tau = 2\pi a_0 + \sum_{k=1}^{\infty} \left[a_k \int_{-\pi}^{\pi} \cos k\tau d\tau + b_k \int_{-\pi}^{\pi} \sin k\tau d\tau \right], \quad (6.8)$$

но очевидно, что

$$\begin{aligned} \int_{-\pi}^{\pi} \cos k\tau d\tau &= \left. \frac{\sin k\tau}{k} \right|_{-\pi}^{\pi} = 0, \\ \int_{-\pi}^{\pi} \sin k\tau d\tau &= - \left. \frac{\cos k\tau}{k} \right|_{-\pi}^{\pi} = 0. \end{aligned} \quad (6.9)$$

Следовательно, все члены под знаком суммы в выражении (6.8) будут нулями. Тогда легко найти, что

$$a_0 = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(\tau) d\tau. \quad (6.10)$$

Теперь установим аналитическое выражение для вычисления коэффициента a_k .

Для этого умножим обе части равенства (6.6) на $\cos k\tau$ и проинтегрируем полученные выражения для обеих частей

в том же промежутке $[-\pi, \pi]$:

$$\begin{aligned} \int_{-\pi}^{\pi} f(\tau) \cos k\tau d\tau &= a_0 \int_{-\pi}^{\pi} \cos k\tau d\tau + \\ &+ \int_{-\pi}^{\pi} \left\{ \sum_{k=1}^{\infty} (a_k \cos k\tau + b_k \sin k\tau) \right\} [\cos k\tau] d\tau = \\ &= \left(a_1 \int_{-\pi}^{\pi} \cos \tau \cos k\tau d\tau + b_1 \int_{-\pi}^{\pi} \sin \tau \cos k\tau d\tau \right) + \\ &+ \left(a_2 \int_{-\pi}^{\pi} \cos 2\tau \cos k\tau d\tau + b_2 \int_{-\pi}^{\pi} \sin 2\tau \cos k\tau d\tau \right) + \dots \\ &\dots + \left(a_k \int_{-\pi}^{\pi} \cos^2 k\tau d\tau + b_k \int_{-\pi}^{\pi} \sin k\tau \cos k\tau d\tau \right) + \dots, \quad (6.11) \end{aligned}$$

так как

$$\begin{aligned} \int_{-\pi}^{\pi} \sin k'\tau \cos k\tau d\tau &= \frac{1}{2} \int_{-\pi}^{\pi} [\sin(k' + k)\tau + \\ &+ \sin(k' - k)\tau] d\tau = 0; \end{aligned} \quad (6.12)$$

$$\begin{aligned} \int_{-\pi}^{\pi} \cos k'\tau \cos k\tau d\tau &= \frac{1}{2} \int_{-\pi}^{\pi} [\cos(k' + k)\tau + \\ &+ \cos(k' - k)\tau] d\tau = 0; \quad (k' \neq k) \end{aligned} \quad (6.13)$$

и

$$\int_{-\pi}^{\pi} \cos^2 k\tau d\tau = \int_{-\pi}^{\pi} \frac{1 + \cos 2k\tau}{2} d\tau = \pi. \quad (6.14)$$

Таким образом, под знаком суммы в выражении (6.11) обращаются в нуль все интегралы, кроме того, в котором $k' = k$.

Следовательно,

$$a_k = \frac{1}{\pi} \int_{-\pi}^{\pi} f(\tau) \cos k\tau d\tau, \quad (k = 1, 2, 3, \dots). \quad (6.15)$$

Аналогично для выражения коэффициента b_k умножим разложение (6.6) на $\sin k\tau$ и затем, интегрируя полученное

выражение почленно, определим коэффициент при синусе

$$b_k = \frac{1}{\pi} \int_{-\pi}^{\pi} f(\tau) \sin k\tau d\tau, \quad (k = 1, 2, 3, \dots). \quad (6.16)$$

Замечательное свойство ряда Фурье заключается в том, что если взять конечное число членов ряда, т. е. аппроксимировать периодическую функцию полиномом, представив ее в виде

$$f(\tau) \approx a_0 + \sum_{k=1}^N (a_k \cos k\tau + b_k \sin k\tau),$$

то для любого N получается наименьшее квадратичное отклонение от точного значения $f(\tau)$, если коэффициенты полинома определены по формулам (6.10), (6.15), (6.16). С увеличением числа N приближение, разумеется, улучшается, и в пределе, при $N \rightarrow \infty$ приближенное равенство переходит в точное.

§ 3. РЯД ФУРЬЕ В КОМПЛЕКСНОЙ ФОРМЕ

Ряд Фурье можно записать в комплексной форме.

Рассмотрим снова произвольную функцию $f(\tau)$ с периодом 2π , абсолютно интегрируемую в любом конечном промежутке, и связанный с ней ряд Фурье

$$f(\tau) = \frac{a_0}{2} + \sum_{k=1}^{\infty} (a_k \cos k\tau + b_k \sin k\tau). \quad (6.17)$$

По формулам Эйлера

$$\begin{aligned} \cos k\tau &= \frac{e^{jk\tau} + e^{-jk\tau}}{2}; \\ \sin k\tau &= \frac{e^{jk\tau} - e^{-jk\tau}}{2j} = j \frac{-e^{jk\tau} + e^{-jk\tau}}{2}. \end{aligned}$$

Подстановка этих выражений в (6.17) дает

$$f(\tau) \approx \frac{a_0}{2} + \sum_{k=1}^{\infty} \left(\frac{a_k - jb_k}{2} e^{jk\tau} + \frac{a_k + jb_k}{2} e^{-jk\tau} \right). \quad (6.18)$$

Полагаем

$$\begin{aligned} c_0 &= \frac{a_0}{2}; \quad c_k = \frac{a_k - jb_k}{2}; \quad c_{-k} = \frac{a_k + jb_k}{2}; \\ &(k = 1, 2, 3, \dots). \end{aligned}$$

Тогда l -я частная сумма ряда (6.18), а следовательно, и ряда (6.17) может быть записана в виде

$$S_k(\tau) = c_0 + \sum_{k=1}^l (c_k e^{jk\tau} + c_{-k} e^{-jk\tau}) = \sum_{k=-l}^l c_k e^{jk\tau}.$$

Поэтому комплексная запись ряда Фурье будет выглядеть так:

$$f(\tau) \approx \sum_{k=-\infty}^{\infty} c_k e^{jk\tau}. \quad (6.19)$$

Коэффициенты для записи ряда Фурье в комплексной форме определяются из следующих выражений:

$$c_0 = \frac{a_0}{2} = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(\tau) d\tau; \quad (6.20)$$

$$\begin{aligned} c_k &= \frac{a_k - jb_k}{2} = \frac{1}{2\pi} \left[\int_{-\pi}^{\pi} f(\tau) \cos k\tau d\tau - j \int_{-\pi}^{\pi} f(\tau) \sin k\tau d\tau \right] = \\ &= \frac{1}{2\pi} \int_{-\pi}^{\pi} f(\tau) (\cos k\tau - j \sin k\tau) d\tau = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(\tau) e^{-jk\tau} d\tau, \\ &\quad (k = 1, 2, 3, \dots); \end{aligned} \quad (6.21)$$

$$\begin{aligned} c_{-k} &= \frac{a_k + jb_k}{2} = \frac{1}{2\pi} \left[\int_{-\pi}^{\pi} f(\tau) \cos k\tau d\tau + j \int_{-\pi}^{\pi} f(\tau) \sin k\tau d\tau \right] = \\ &= \frac{1}{2\pi} \int_{-\pi}^{\pi} f(\tau) (\cos k\tau + j \sin k\tau) d\tau = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(\tau) e^{jk\tau} d\tau, \\ &\quad (k = 1, 2, 3, \dots). \end{aligned} \quad (6.22)$$

Формулы (6.20), (6.21), (6.22), очевидно, можно объединить в одну

$$\begin{aligned} c_k &= \frac{1}{2\pi} \int_{-\pi}^{\pi} f(\tau) e^{-jk\tau} d\tau, \\ &\quad (k = 0, \pm 1, \pm 2, \pm 3, \pm \dots). \end{aligned} \quad (6.23)$$

§ 4. РЯД ФУРЬЕ И НАИМЕНЬШАЯ СРЕДНЕКВАДРАТИЧНАЯ ОШИБКА

Пусть $f(\tau)$ — всюду непрерывная периодическая функция с периодом 2π . Обозначим через $T_n(\tau)$ любой тригонометрический многочлен вида

$$T_n(\tau) = \frac{a_0}{2} + (a_1 \cos \tau + b_1 \sin \tau) + \dots \\ \dots + (a_n \cos n\tau + b_n \sin n\tau).$$

Если $f(\tau)$ заменить $T_n(\tau)$ то, как указывалось выше, мы можем совершить некоторую ошибку. Оценим ее. Для этого введем интеграл

$$I = \int_0^{2\pi} [f(\tau) - T_n(\tau)]^2 d\tau,$$

называемый *средней квадратичной погрешностью*. Задача должна быть решена так, чтобы $T_n(\tau)$ был наилучшим приближением к $f(\tau)$ с точки зрения средней квадратичной погрешности. Для этого интеграл I должен быть наименьшим по величине, что можно обеспечить, подобрав соответствующим образом коэффициенты a_0, a_1, \dots, a_n , и b_0, b_1, \dots, b_n .

Минимум I можно найти обычными методами для функций многих переменных

$$\frac{\partial I}{\partial a_k} = 0; \quad \frac{\partial I}{\partial b_k} = 0,$$

где

$$k = 0, 1, 2, \dots, n.$$

Тогда

$$\frac{\partial I}{\partial a_k} = \int_0^{2\pi} 2 [f(\tau) - T_n(\tau)] \cos k\tau d\tau, \\ \frac{\partial I}{\partial b_k} = \int_0^{2\pi} 2 [f(\tau) - T_n(\tau)] \sin k\tau d\tau.$$

Вычисляя, получим

$$\frac{\partial I}{\partial a_k} = 2 \int_0^{2\pi} f(\alpha) \cos k\alpha d\alpha - 2 \int_0^{2\pi} T_n(\tau) \cos k\alpha d\alpha = \\ = 2 \int_0^{2\pi} f(\alpha) \cos k\alpha d\alpha - 2\pi a_k.$$

Аналогично

$$\frac{\partial I}{\partial b_k} = 2 \int_0^{2\pi} f(\alpha) \sin k\alpha d\alpha - 2\pi b_k.$$

Откуда, учитывая условия для минимума интеграла I ,

$$\left. \begin{aligned} a_k &= \frac{1}{\pi} \int_0^{2\pi} f(\alpha) \cos k\alpha d\alpha, \\ b_k &= \frac{1}{\pi} \int_0^{2\pi} f(\alpha) \sin k\alpha d\alpha. \end{aligned} \right\}$$

Полученные значения для a_k и b_k есть ничто иное как коэффициенты ряда Фурье.

Следовательно, из всех тригонометрических многочленов $T_n(\tau)$ только тот делает среднюю квадратичную погрешность минимальной, у которого коэффициенты являются коэффициентами Фурье.

§ 5. ИНТЕГРАЛ ФУРЬЕ

Как видно из предыдущего, ряд Фурье дает разложение периодической функции по тригонометрическим. Можно обобщить это разложение и для случая непериодической функции.

Приведем, вначале без доказательства, некоторые вспомогательные факты

$$1) \int_0^{\infty} \frac{\sin \tau}{\tau} d\tau = \pi;$$

2) если $f(\tau)$ имеет на каждом конечном интервале не более конечного числа точек разрыва и абсолютно интегрируема на $(-\infty, +\infty)$, то

$$\int_{-\infty}^{\infty} f(\tau) \sin a\tau d\tau \rightarrow 0.$$

$$a \rightarrow +\infty.$$

Это утверждение известно под названием леммы Римана для бесконечного промежутка. Доказательство обоих утверждений можно найти в специальной литературе.

Выведем *достаточное условие представимости функции интегралом Фурье*. Пусть $f(\tau)$ — функция, имеющая на каждом конечном сегменте не более конечного числа точек разрыва и абсолютно интегрируемая на $(-\infty, +\infty)$, что

означает следующее: несобственный интеграл $\int_{-\infty}^{\infty} |f(\tau)| d\tau$ есть конечная величина, т. е. *интеграл сходится*.

Из предыдущего изложения известно, что в каждой точке дифференцируемости τ_0 функции $f(\tau)$, заданной на конечном промежутке $[l; -l]$, при $l > |\tau_0|$ будет справедливо

$$f(\tau_0) = \frac{a_0}{2} + \sum_{k=1}^{\infty} \left(a_k \cos \frac{k\pi\tau_0}{l} + b_k \sin \frac{k\pi\tau_0}{l} \right),$$

где коэффициенты определяются приведенными выше формулами.

Следовательно,

$$\begin{aligned} f(\tau_0) &= \frac{1}{2l} \int_{-l}^l f(\tau) d\tau + \sum_{k=1}^{\infty} \left(\frac{1}{l} \int_{-l}^l f(\tau) \cos \frac{k\pi\tau}{l} d\tau \cos \frac{k\pi\tau_0}{l} + \right. \\ &\quad \left. + \frac{1}{l} \int_{-l}^l f(\tau) \sin \frac{k\pi\tau}{l} d\tau \sin \frac{k\pi\tau_0}{l} \right) = \\ &= \frac{1}{2l} \int_{-l}^l f(\tau) d\tau + \sum_{k=1}^{\infty} \frac{1}{l} \int_{-l}^l f(\tau) \cos \frac{k\pi(\tau - \tau_0)}{l} d\tau. \end{aligned}$$

Далее, полагая

$$\omega_k = \frac{k\pi}{l}, \quad \left(\text{тогда } \Delta\omega_k = \frac{\pi}{l}; \quad \frac{1}{l} = \frac{1}{\pi} \Delta\omega_k \right),$$

получим

$$f(\tau_0) = \frac{1}{2l} \int_{-l}^l f(\tau) d\tau + \frac{1}{\pi} \sum \Delta\omega_k \int_{-l}^l f(\tau) \cos \omega_k(\tau - \tau_0) d\tau.$$

При $l \rightarrow +\infty$ очевидно, что

$$\frac{1}{2l} \int_{-l}^l f(\tau) d\tau \rightarrow 0, \quad \text{если предположить, что интеграл}$$

$$\int_{-\infty}^{\infty} f(\tau) d\tau \text{ сходится.}$$

Кроме того, естественно предположить

$$\sum_{k=1}^{\infty} \Delta\omega_k \int_{-l}^l f(\tau) \cos \omega_k(\tau - \tau_0) d\tau \rightarrow \int_0^{\infty} d\omega \int_{-\infty}^{\infty} f(\tau) \cos \omega(\tau - \tau_0) d\tau.$$

Если это действительно имеет место, то выражение, полученное для $f(\tau_0)$, дает в пределе

$$f(\tau_0) = \frac{1}{\pi} \int_0^\infty d\omega \int_{-\infty}^\infty f(\tau) \cos \omega(\tau - \tau_0) d\tau. \quad (6.24)$$

Докажем, что (6.24) действительно справедливо.

Предположим, что

$$I(a) = \frac{1}{\pi} \int_0^a d\omega \int_{-\infty}^\infty f(\tau) \cos \omega(\tau - \tau_0) d\tau.$$

Так как $|f(\tau) \cos \omega(\tau - \tau_0)| \leq |f(\tau)|$ и $\int_{-\infty}^\infty |f(\tau)| d\tau$ сходится, можно изменить последовательность интегрирования, в результате чего получим

$$\begin{aligned} I(a) &= \frac{1}{\pi} \int_{-\infty}^\infty f(\tau) d\tau \int_0^a \cos \omega(\tau - \tau_0) d\omega = \\ &= \frac{1}{\pi} \int_{-\infty}^\infty f(\tau) \frac{\sin a(\tau - \tau_0)}{\tau - \tau_0} d\tau = \frac{1}{\pi} \int_{-\infty}^\infty f(\tau_0 + \tau) \frac{\sin a\tau}{\tau} d\tau. \end{aligned}$$

Здесь интеграл получен из предыдущего заменой τ на $\tau_0 + \tau$.

Далее, если $f(\tau_0) = 0$, то $\frac{f(\tau_0 + \tau)}{\tau} \rightarrow f'(\tau_0)$ при $\tau \rightarrow 0$ и, следовательно, функция $y(\tau) = \frac{f(\tau_0 + \tau)}{\tau}$ после необходимого дополнения в точке $\tau = 0$ становится непрерывной в окрестности нуля и будет абсолютно интегрируема на $(-\infty, +\infty)$, так как при достаточно малом ε она непрерывна на интервале $(-\varepsilon, \varepsilon)$, а вне его

$$|y(\tau)| \leq \frac{|f(\tau_0 + \tau)|}{\varepsilon}.$$

Тогда в силу леммы Римана для бесконечного промежутка имеем

$$I(a) = \frac{1}{\pi} \int_{-\infty}^\infty y(\tau) \sin a\tau d\tau \rightarrow 0.$$

В общем случае (когда значение $f(\tau_0)$ может быть любым) предположим, что

$$z(\tau) = \begin{cases} f(\tau_0) & \text{при } \tau_0 - 1 \leq \tau \leq \tau_0 + 1, \\ 0 & \text{для всех других;} \end{cases}$$

$$\bar{f}(\tau) = f(\tau) - z(\tau).$$

Тогда

$$I(a) \frac{1}{\pi} \int_{-\infty}^{\infty} f(\tau_0 + \tau) \frac{\sin a\tau}{\tau} d\tau + \frac{1}{\pi} \int_{-\infty}^{\infty} z(\tau_0 + \tau) \frac{\sin a\tau}{\tau} d\tau.$$

Первое слагаемое правой части стремится к нулю при $a \rightarrow +\infty$, так как $\bar{f}(\tau_0) = 0$. Второе слагаемое правой части равно

$$\begin{aligned} \frac{f(\tau_0)}{\pi} \int_{-1}^1 \frac{\sin a\tau}{\tau} d\tau &= \frac{f(\tau_0)}{\pi} \int_{-a}^a \frac{\sin \tau}{\tau} d\tau \rightarrow \\ &\rightarrow \frac{f(\tau_0)}{\pi} \int_{-\infty}^{\infty} \frac{\sin \tau}{\tau} d\tau = f(\tau_0) \end{aligned}$$

при $a \rightarrow +\infty$ и замене $a\tau$ на τ . Таким образом, в точках дифференцируемости τ_0 функции $f(\tau)$ имеем

$$f(\tau_0) = \lim I(a) = \frac{1}{\pi} \int_0^{\infty} d\omega \int_{-\infty}^{\infty} f(\tau) \cos \omega(\tau - \tau_0) d\tau.$$

Заменяв τ_0 на τ и τ на t , доказанное можно сформулировать следующим образом.

Если $f(\tau)$ имеет на каждом конечном интервале не более конечного числа точек разрыва и абсолютно интегрируема на $(-\infty, +\infty)$, то в каждой точке τ , в которой $f(\tau)$ дифференцируема, имеем

$$f(\tau) = \frac{1}{\pi} \int_0^{\infty} d\omega \int_{-\infty}^{\infty} f(t) \cos \omega(t - \tau) dt. \quad (6.25)$$

Выражение, стоящее в правой части равенства (6.25), называется двойным интегралом Фурье данной функции. Учитывая, что

$$\cos \omega(t - \tau) = \cos \omega t \cos \omega \tau + \sin \omega t \sin \omega \tau,$$

после внесения множителя $\frac{1}{\pi}$ внутренний интеграл в выражении (6.25) можно преобразовать следующим образом

$$\begin{aligned} \frac{1}{\pi} \int_{-\infty}^{\infty} f(t) \cos \omega(t - \tau) d\tau &= \frac{1}{\pi} \int_{-\infty}^{\infty} f(t) \cos \omega t dt \cdot \cos \omega \tau + \\ &+ \frac{1}{\pi} \int_{-\infty}^{\infty} f(t) \sin \omega t dt \cdot \sin \omega \tau = a(\omega) \cos \omega \tau + b(\omega) \sin \omega \tau, \end{aligned}$$

где

$$\left. \begin{aligned} a(\omega) &= \frac{1}{\pi} \int_{-\infty}^{\infty} f(t) \cos \omega t dt, \quad (\omega \geq 0); \\ b(\omega) &= \frac{1}{\pi} \int_{-\infty}^{\infty} f(t) \sin \omega t dt, \quad (\omega \geq 0). \end{aligned} \right\} \quad (6.26)$$

Тогда выражение (6.25) примет вид

$$f(\tau) = \int_0^{\infty} [a(\omega) \cos \omega \tau + b(\omega) \sin \omega \tau] d\omega. \quad (6.27)$$

Выражение, стоящее в правой части формулы (6.27), называется интегралом Фурье для функции $f(\tau)$.

§ 6. ПРИЗНАКИ СХОДИМОСТИ ИНТЕГРАЛА ФУРЬЕ

При описании функций с помощью интеграла Фурье на практике часто ставится вопрос о сходимости интеграла. Решению этого вопроса помогают ряд признаков.

Признак Дини. Предположим, что значение интеграла, определяемое правой частью выражения (6.25), равно постоянному числу A_0 . Введем функцию

$$\varphi(t) = f(\tau_0 + t) + f(\tau_0 - t) - 2S_0.$$

Предположим далее, что функция $f(\tau)$ в точке τ_0 либо непрерывна и $S_0 = f(\tau_0)$, либо имеет в этой точке с обеих сторон разрывы лишь первого рода и тогда

$$S_0 = \frac{f(\tau_0 + 0) + f(\tau_0 - 0)}{2}.$$

При этих предположениях признак Дини о сходимости интеграла Фурье формулируется следующим образом:

Интеграл Фурье функции $f(\tau)$ в точке τ_0 сходится и имеет $S_0 = \text{const}$, если при некотором $h > 0$ сходится интеграл

$$\int_0^h \frac{|\varphi(t)|}{t} dt.$$

Признак Дирихле — Жордана формулируется следующим образом:

Интеграл Фурье функции $f(\tau)$ в точке τ_0 сходится и имеет значение S_0 , если в некотором промежутке $[\tau_0 - n; \tau_0 + n]$ с центром в этой точке функция $f(\tau)$ имеет ограниченное изменение.

В основе наших рассуждений до сих пор лежало предположение, что функция $f(\tau)$ *абсолютно интегрируема* во всем бесконечном промежутке от $-\infty$ до $+\infty$. Затем уже, налагая дополнительно различные условия на поведение функции в непосредственной окрестности интересующей нас точки τ_0 , получим достаточные признаки представимости функции в этой точке интегралом Фурье.

Если сохраним лишь допущение, что функция $f(\tau)$ абсолютно интегрируема в каждом конечном промежутке, и далее, условие на бесконечности заменим следующим: для $|\tau| \geq H$ функция $f(\tau)$ монотонна, точнее она монотонна для $\tau \geq H$ и для $\tau \leq -H$ в отдельности и притом

$$\begin{aligned} \lim_{\tau \rightarrow \pm \infty} f(\tau) &= 0, \\ \tau &\rightarrow \pm \infty, \end{aligned} \quad (6.28)$$

то признаки Дини и Дирихле — Жордана останутся в силе и при новых предположениях относительно функции $f(\tau)$.

Из всего сказанного, в частности, вытекает условие применимости формулы Фурье к практическим задачам:

если функция $f(\tau)$ имеет ограниченное применение во всем бесконечном промежутке $[-\infty; +\infty]$ и, сверх того, выполняется предельное равенство (6.28), то в каждой точке τ_0 интеграл Фурье сходится и имеет значение S_0 .

§ 7. КОМПЛЕКСНАЯ ФОРМА ИНТЕГРАЛА ФУРЬЕ

Преобразуем подынтегральное выражение в формуле (6.27), используя формулы Эйлера:

$$a(\omega) \cos \omega \tau + b(\omega) \sin \omega \tau = a(\omega) \frac{e^{j\omega\tau} + e^{-j\omega\tau}}{2} +$$

$$\begin{aligned}
& + b(\omega) \frac{e^{j\omega\tau} - e^{-j\omega\tau}}{2j} = a(\omega) \frac{e^{j\omega\tau} + e^{-j\omega\tau}}{2} - \\
& - j b(\omega) \frac{e^{j\omega\tau} - e^{-j\omega\tau}}{2} = \frac{a(\omega) - j b(\omega)}{2} e^{j\omega\tau} + \\
& + \frac{a(\omega) + j b(\omega)}{2} e^{-j\omega\tau} = c(\omega) e^{j\omega\tau} + c(-\omega) e^{-j\omega\tau}.
\end{aligned}$$

Здесь положено

$$c(\omega) = \frac{a(\omega) - j b(\omega)}{2}; \quad c(-\omega) = \frac{a(\omega) + j b(\omega)}{2}.$$

Следовательно

$$\begin{aligned}
& \int_0^{\lambda} [a(\omega) \cos \omega\tau + b(\omega) \sin \omega\tau] d\omega = \\
& = \int_0^{\lambda} [c(\omega) e^{j\omega\tau} + c(-\omega) e^{-j\omega\tau}] d\omega = \\
& = \int_0^{\lambda} c(\omega) e^{j\omega\tau} d\omega + \int_0^{\lambda} c(-\omega) e^{-j\omega\tau} d\omega = \\
& = \int_0^{\lambda} c(\omega) e^{j\omega\tau} d\omega + \int_{-\lambda}^0 c(\omega) e^{j\omega\tau} d\omega = \int_{-\lambda}^{\lambda} c(\omega) e^{j\omega\tau} d\omega.
\end{aligned}$$

Следует учитывать, что после замены $-\omega$ на ω интеграл

$$\int_0^{\lambda} c(-\omega) e^{-j\omega\tau} d\omega \text{ переходит в } \int_{-\lambda}^0 c(\omega) e^{j\omega\tau} d\omega.$$

Доопределим $c(\omega)$

$$\begin{aligned}
c(\omega) &= \frac{a(\omega) - j b(\omega)}{2} = \\
&= \frac{1}{2} \left(\frac{1}{\pi} \int_{-\infty}^{\infty} f(t) \cos \omega t dt - j \frac{1}{\pi} \int_{-\infty}^{\infty} f(t) \sin \omega t dt \right) = \\
&= \frac{1}{2\pi} \int_{-\infty}^{\infty} f(t) (\cos \omega t - j \sin \omega t) dt = \frac{1}{2\pi} \int_{-\infty}^{\infty} f(t) e^{-j\omega t} dt.
\end{aligned}$$

($\omega \geq 0$)

Если учесть, что $c(-\omega) = \bar{c}(\bar{\omega})$, то становится очевидным, что эти формулы верны и для $\omega < 0$.

Из формулы (6.27) теперь можно получить

$$\begin{aligned} f(\tau) &= \lim_{\lambda \rightarrow \infty} \int_0^{\lambda} [a(\omega) \cos \omega \tau + b(\omega) \sin \omega \tau] d\omega = \\ &= \lim_{\lambda \rightarrow \infty} \int_{-\lambda}^{\lambda} c(\omega) e^{j\omega \tau} d\omega = \int_{-\infty}^{\infty} c(\omega) e^{j\omega \tau} d\omega, \end{aligned}$$

если принять, что

$$\int_{-\infty}^{\infty} = \lim_{\lambda \rightarrow \infty} \int_{-\lambda}^{\lambda}.$$

Таким образом, в точках дифференцируемости функции $f(\tau)$ имеем

$$f(\tau) = \int_{-\infty}^{\infty} c(\omega) e^{j\omega \tau} d\omega, \quad (6.29)$$

где

$$c(\omega) = \frac{1}{2\pi} \int_{-\infty}^{\infty} f(t) e^{-j\omega t} dt. \quad (6.30)$$

Выражение (6.29) является *представлением интеграла Фурье в комплексной форме*.

Далее, если в формулу (6.29) вместо $c(\omega)$ подставим его значение, то после внесения $e^{j\omega \tau}$ под знак внутреннего интеграла получим

$$f(\tau) = \frac{1}{2\pi} \int_{-\infty}^{\infty} d\omega \int_{-\infty}^{\infty} f(t) e^{j\omega(\tau-t)} dt. \quad (6.31)$$

Выражение (6.31) называется *двойным интегралом Фурье в комплексной форме*.

§ 8. ПРЕОБРАЗОВАНИЕ ФУРЬЕ

Перепишем формулу (6.31) так:

$$\begin{aligned} f(\tau) &= \frac{1}{2\pi} \int_{-\infty}^{\infty} d\omega \int_{-\infty}^{\infty} f(t) e^{j\omega(\tau-t)} dt = \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} f(t) e^{-j\omega t} dt \right] e^{j\omega \tau} d\omega. \end{aligned} \quad (6.32)$$

Далее полагаем

$$F(\omega) = \int_{-\infty}^{\infty} f(t) e^{-j\omega t} dt. \quad (6.33)$$

Тогда вместо (6.32) получим

$$f(\tau) = \frac{1}{2\pi} \int_{-\infty}^{\infty} F(\omega) e^{j\omega\tau} d\omega. \quad (6.34)$$

Формулы (6.33) и (6.34) представляют собой пару преобразований Фурье, связывающих между собой две функции, вещественную функцию времени $f(\tau)$ и комплексную функцию частоты $F(\omega)$. Действительно, формулу (6.34) можно записать в вещественной форме. Введя обозначение

$$F(\omega) = A(\omega) + jB(\omega),$$

получим

$$f(\tau) = \frac{1}{\pi} \int_0^{\infty} [A(\omega) \cos \omega\tau - B(\omega) \sin \omega\tau] d\omega.$$

При этом учитывается, что A — четная функция, а B — нечетная.

Другая форма записи формулы (6.34) в вещественной форме получится, если выражение (6.34) представить как

$$f(\tau) = \frac{1}{2\pi} \int_0^{\infty} [F(\omega) e^{j\omega\tau} + F(-\omega) e^{-j\omega\tau}] d\omega.$$

Здесь в квадратных скобках стоит сумма сопряженных комплексов, которая равна удвоенной вещественной части,

$$f(\tau) = \frac{1}{\pi} \operatorname{Re} \int_0^{\infty} F(\omega) e^{j\omega\tau} d\omega.$$

Выражение (6.33) для вычисления комплексной функции $F(\omega)$ называется *спектральной функцией для $f(\tau)$* . Понятие спектральной функции имеет важное значение в электро- и радиотехнике и их приложениях. Формулы (6.33) и (6.34) являются основными в теории спектров. Формула (6.34) представляет собой интеграл Фурье в комплексной форме, смысл этой формулы состоит в том, что функция $f(\tau)$ представлена суммой синусоидальных составляющих. А так как функция $f(\tau)$ предполагалась непериодической, то естественно, что она может быть представлена только суммой бесконечно большого числа бесконечно малых колебаний, бесконечно близких по частоте.

Комплексная амплитуда такого колебания бесконечно мала и равна

$$dA = \frac{1}{\pi} F(\omega) d\omega. \quad (6.35)$$

Частотный интервал между двумя соседними колебаниями также бесконечно мал и равен $d\omega$.

Таким образом, физически становится ощутимо различие между понятиями о ряде Фурье и интеграле Фурье.

Если ряд Фурье представляет периодическую функцию суммой бесконечного числа амплитуд, но с частотами, имеющими определенные дискретные значения, то интеграл Фурье представляет непериодическую функцию суммой амплитуд с непрерывной последовательностью частот. Другими словами, можно сказать, что в непериодической функции имеются все частоты.

Следовательно, ряд Фурье представляет периодическую функцию как сумму периодических составляющих, тогда как интеграл Фурье представляет непериодическую функцию суммой периодических составляющих. То есть, в случае представления функции интегралом Фурье сумма не обладает существенным свойством своих слагаемых, что следует обязательно учитывать при практических приложениях интеграла Фурье.

Контрольные вопросы и задания

1. Какая функция называется периодической?
2. Какими основными свойствами характеризуются периодические функции?
3. Каким условием должна удовлетворять функция для возможности определения коэффициентов ряда Фурье по методу Эйлера — Фурье?
4. Запишите выражения для ряда и интеграла Фурье в вещественной и в комплексной форме.
5. Сформулируйте признаки сходимости интеграла Фурье.
6. В чем заключается физическое различие понятий ряда и интеграла Фурье?

§ 9. СПЕКТР АМПЛИТУД И СПЕКТР ФАЗ

Перепишем формулу (6.7) в несколько иной удобной для дальнейших рассуждений форме.

В § 3 мы имели

$$f(\tau) \approx \frac{a_0}{2} + \sum_{k=1}^{\infty} \left(a_k \cos 2\pi k \frac{t}{T} + b_k \sin 2\pi k \frac{t}{T} \right).$$

Учитывая, что

$$a_k = c_k \cos \varphi_k; \quad b_k = c_k \sin \varphi_k,$$

и, кроме того,

$$c_k = \sqrt{a_k^2 + b_k^2}; \quad \operatorname{tg} \varphi_k = \frac{b_k}{a_k}; \quad c_0 = \frac{a_0}{2},$$

можно записать ряд Фурье в вещественной форме в виде

$$f(\tau) = c_0 + \sum_{k=1}^{\infty} c_k \cos \left(2\pi k \frac{t}{T} - \varphi_k \right). \quad (6.38)$$

Далее полагая, что основная частота (частота первой гармоники) $\omega_0 = \frac{2\pi}{T}$, перепишем формулу (6.38) таким образом:

$$f(\tau) = c_0 + \sum_{k=1}^{\infty} c_k \cos (k\omega_1 t - \varphi_k). \quad (6.39)$$

Здесь c_0 — постоянная составляющая; c_k — амплитуда k -ой гармоники; $k\omega_1 = \omega_k$ — угловая частота k -ой гармоники; φ_k — начальная фаза k -ой гармоники.

Поскольку угловая частота ω_1 основной гармоники связана с периодом функции $f(\tau)$ соотношением $\omega_1 = \frac{2\pi}{T}$, т. е.

частота первой гармоники $f_1 = \frac{1}{T}$, а частоты любой гармоники кратны частоте основной гармоники, то, следовательно, все они определяются простыми кратными соотношениями. Судя по выражению (6.39) любая периодическая функция $f(\tau)$ вполне может быть определена совокупностью величин c_k , φ_k и ω_k . Совокупность величин c_k называется спектром амплитуд. Совокупность величин φ_k называется спектром фаз.

Рассмотрим отдельно вопрос о спектральном представлении функции $f(\tau)$ для случаев, когда эта функция периодическая, почти периодическая и непериодическая.

1. $f(\tau)$ — периодическая функция.

Как было указано в предыдущих параграфах, при выполнении известных условий функция $f(\tau)$ может быть полностью определена выражением (6.39). Следовательно, любой периодический сигнал можно рассматривать как результат наложения друг на друга бесконечного количества гармоник, а также постоянной составляющей.

Понятие спектров удобно интерпретировать графически. Для этого, как показано на рис. 6.3, периодический сигнал представляется в виде ряда отдельных спектральных линий в соответствующих координатах. Длины линий пропорциональны амплитудам (спектр амплитуд) или фазам (спектр фаз) соответствующих гармоник. При этом расстояние между соседними линиями постоянно при равномерности шкалы частот.

Исходя из этих соображений спектр периодического сигнала $f(\tau)$ называют дискретным (или линейчатым) гармоническим спектром.

В зависимости от конкретного вида сигнала $f(\tau)$ в его разложении могут отсутствовать некоторые гармоники, общее число гармоник может оказаться конечным. Однако всегда гармоничность спектра сохраняется, так как всегда частоты имеющихся гармоник находятся в простых кратных соотношениях. Относительно отсутствующих гармоник нужно иметь в виду, что амплитудам их приписаны нулевые значения.

Непрерывная кривая, соединяющая концы линий спектра, которая показана на рис. 6.3, называется огибающей спектра амплитуд (или спектра фаз).

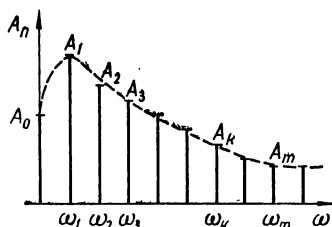


Рис. 6.3. Огибающая спектра.

Очевидно, огибающая спектра — это геометрическое место концов линий спектра амплитуд (спектра фаз). Для получения уравнения огибающей спектра амплитуд (спектра фаз), необходимо из выражения (6.39) определить зависимость амплитуды C_k не от дискретного номера гармоники k , а от текущего значения частоты. Имеем

$$f(\tau) \approx \sum_{-\infty}^{\infty} C_k e^{jk\tau},$$

причем функция $f(\tau)$ задана на отрезке времени $t_1 \leq \tau \leq t_2$ и повторяется с периодом $T = t_2 - t_1$. Тогда

$$C_k = \int_{t_1}^{t_2} f(\tau) e^{-jk\tau} d\tau = \int_{t_1}^{t_2} f(\omega t) e^{-jk\omega t} dt = \{|C_k|(\omega)\} e^{-j\varphi(\omega)}.$$

Здесь $\{|C_k|(\omega)\}$ — зависимость модуля амплитуды C_k от частоты, ω соответствует огибающей спектра амплитуд, а зависимость $\varphi(\omega)$ представляет собой огибающую спектра фаз.

2. $f(\tau)$ — почти периодическая функция.

В результате сложения синусоид с некратными частотами (например $\sin \omega + \sin \sqrt{2} \omega$) получится заведомо непериодическая функция.

Одно из основных свойств таких функций в том, что для нее можно определить лишь приближенный период — почти период. Такие сигналы в технике встречаются довольно часто. В математике такие функции получили название почти периодических функций. Спектр таких функций также дискретен и представляется в виде суммы гармонических составляющих с произвольными частотами.

Таким образом, дискретные, или линейчатые спектры могут принадлежать как периодическим, так и непериодическим функциям. В первом случае линейчатый спектр обязательно гармонический.

Для практических целей большое значение имеет частный случай почти периодических функций вида

$$f(\tau) = \sum C_k \cos[(\omega_0 + k\omega_1)t - \varphi_k]. \quad (6.40)$$

Здесь k может принимать как положительное, так и отрицательное значение. Линии спектра такого разложения эквидистантны, поэтому такой спектр обычно называют квазигармоническим. Примером является спектр периодически модулированных колебаний, где ω_0 — несущая частота. Таким образом, функция, обладающая дискретным спектром из произвольно расположенных по шкале частот спектральных линий называется почти периодической функцией.

3. $f(\tau)$ — непериодическая функция.

Как было указано в § 8, формула, отвечающая обратному преобразованию Фурье, имеет вид

$$f(\tau) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} F(\omega) e^{i\omega\tau} d\omega.$$

Этой формулой по аналогии со случаем, когда $f(\tau)$ — периодическая функция, можно представить непериодическую функцию $f(\tau)$ в виде суммы бесконечно большого числа гармоник с бесконечно малыми комплексными амплитудами

$$dC e^{i\omega\tau} = \frac{1}{\pi} F(\omega) e^{i\omega\tau} d\omega. \quad (6.41)$$

Отсюда следует:

$$F(\omega) = \pi \frac{dC}{d\omega}. \quad (6.42)$$

Таким образом, $F(\omega)$ не амплитуда, а так называемая спектральная плотность. Часто эту величину именуют спектральной характеристикой или комплексным спектром непериодической функции, а модуль $|F(\omega)|$ — спектром. Посколь-

ку спектральная характеристика — величина комплексная, ее можно представить в виде

$$F(j\omega) = a(\omega) + jb(\omega) = F(\omega) e^{-j\psi(\omega)}.$$

Пользуясь прямым преобразованием Фурье (6.33), можем определить

$$a(\omega) = \int_{-\infty}^{\infty} f(\tau) \cos \omega \tau d\tau,$$

$$b(\omega) = \int_{-\infty}^{\infty} f(\tau) \sin \omega \tau d\tau.$$

Следовательно, модуль и фаза спектральной характеристики соответственно равны

$$|F(\omega)| = \sqrt{[a(\omega)]^2 + [b(\omega)]^2},$$

$$\varphi(\omega) = \arctg \frac{b(\omega)}{a(\omega)}.$$

Спектр непериодической функции, таким образом, полностью определяется модулем и фазой спектральной характеристики, зависящими от частоты. Обычно зависимости модуля $F(\omega)$ и фазы $\varphi(\omega)$ спектральной характеристики от частоты называют спектром амплитуд и спектром фаз непериодического сигнала.

Очевидно, что непериодический сигнал можно представить периодической функцией времени с бесконечно большим периодом. Из выражения для частоты основной гармоники периодического сигнала $f_1 = \frac{1}{T}$ следует, что при возрастании периода T разность между частотами соседних гармоник уменьшается, стремясь в пределе к нулю. Следовательно, в пределе (а это именно то, что нас интересует с точки зрения непериодической функции) спектр становится сплошным. Таким образом, спектр непериодических функций содержит колебания всех частот, но огибающая спектра при этом остается неизменной.

§ 10. ОСНОВНЫЕ ТЕОРЕМЫ О СПЕКТРАХ

Теоремы о спектрах выражают основные свойства преобразования Фурье. Из рассмотрения прямого и обратного преобразования Фурье (6.33), (6.34) можно сделать некоторые общие заключения о характере спектральной характеристики $F(\omega)$, заданной непериодической функцией $f(\tau)$, и,

наоборот, о функции $f(\tau)$ — по заданной ее спектральной характеристике $F(\omega)$.

1. Теорема сложения спектров (принцип суперпозиции). Допустим, что задана непериодическая функция

$$f(\tau) = \sum_k f_k(\tau).$$

Тогда, воспользовавшись (6.33), можем записать

$$\begin{aligned} F(\omega) &= \int_{-\infty}^{\infty} f(\tau) e^{-j\omega\tau} d\tau = \int_{-\infty}^{\infty} \sum_k f_k(\tau) e^{-j\omega\tau} d\tau = \\ &= \sum_k \int_{-\infty}^{\infty} f_k(\tau) e^{-j\omega\tau} d\tau = \sum_k F_k(\omega). \end{aligned}$$

Другими словами, спектр суммы функций равен сумме спектров слагаемых, т. е. преобразование Фурье является линейным, а следовательно для него справедлив принцип суперпозиции.

2. Теорема об изменении масштаба. Пусть задана непериодическая функция

$$f_k(\tau) = f(k\tau).$$

Тогда, используя формулу прямого преобразования Фурье (6.33), получим

$$\begin{aligned} F_k(\omega) &= \int_{-\infty}^{\infty} f_k(\tau) e^{-j\omega\tau} d\tau = \\ &= \int_{-\infty}^{\infty} f(k\tau) e^{-j\omega\tau} d\tau = \frac{1}{k} \int_{-\infty}^{\infty} f(\tau_1) e^{-j\frac{\omega}{k}\tau_1} d\tau_1, \end{aligned}$$

иначе

$$F_k(\omega) = \frac{1}{k} F\left(\frac{\omega}{k}\right),$$

т. е. при изменении масштаба времени в k раз масштаб частот для спектра меняется в $\frac{1}{k}$ раз. Отсюда ясно, что единственный способ сжатия спектра (сокращения ширины) без изменения характера его состоит в том, чтобы явления, описываемые функцией $f(\tau)$, растянуть во времени. Наоборот, для расширения спектра нужно ускорить ход явления, описываемого этой функцией.

3. Теорема запаздывания. Пусть заданная функция, описывающая ход непериодического явления, имеет вид

$$\hat{f}_{\tau}(\tau) = f(\tau - \hat{\tau}),$$

где $\hat{\tau}$ — запаздывание во времени.

Используя прямое преобразование Фурье, можем записать

$$F_{\hat{\tau}}(\omega) = \int_{-\infty}^{\infty} f_{\hat{\tau}}(\tau) e^{-j\omega\tau} d\tau = \int_{-\infty}^{\infty} f(\tau - \hat{\tau}) e^{-j\omega\tau} d\tau.$$

Далее произведем замену переменных: $t = \tau - \hat{\tau}$. В результате получим

$$F_{\hat{\tau}}(\omega) = e^{-j\omega\hat{\tau}} \int_{-\infty}^{\infty} f(t) e^{-j\omega t} dt,$$

• или иначе

$$F_{\hat{\tau}}(\omega) = e^{-j\omega\hat{\tau}} F(\omega).$$

Таким образом, если рассматриваемое непериодическое явление описывается функцией $f(\tau)$, а необходимо получить спектр функции $f(\tau - \hat{\tau})$, то для этого надо умножить спектр $F(\omega)$ процесса $f(\tau)$ на $e^{-j\omega\hat{\tau}}$. Такое преобразование изменит только спектральную плотность фаз. Спектральная плотность амплитуд останется при этом неизменной.

4. Теорема о спектре производной и интеграла. Пусть задана $f(\tau)$, описывающая непериодическое явление; требуется найти спектр функции

$$f'(\tau) = \frac{d[f(\tau)]}{d\tau}.$$

Подставив это значение в выражение для прямого преобразования Фурье и проинтегрировав полученное выражение по частям, имеем

$$F_{(1)}(\omega) = \int_{-\infty}^{\infty} f'(\tau) e^{-j\omega\tau} d\tau = f(\tau) e^{-j\omega\tau} \int_{-\infty}^{\infty} + j\omega \int_{-\infty}^{\infty} f(\tau) e^{-j\omega\tau} d\tau.$$

Далее, если

$$\lim_{\tau \rightarrow \pm\infty} f(\tau) = 0,$$

то

$$F_{(1)}(\omega) = j\omega F(\omega).$$

Используя индуктивный метод доказательства теоремы для k -й производной, получим

$$F_{(k)}(\omega) = \int_{-\infty}^{\infty} \frac{d^k f(\tau)}{d\tau^k} e^{-j\omega\tau} d\tau = (j\omega)^k F(\omega).$$

Если для всех производных до $(n - 1)$ -го порядка включительно выполняется условие

$$\lim_{\tau \rightarrow \pm \infty} \frac{d^k f(\tau)}{d\tau^k} = 0,$$

$$F_n(\omega) = (j\omega)^n F(\omega).$$

Применяя тот же прием, легко показать, что спектр интеграла от функции $f(\tau)$, взятого в пределах от $-\infty$ до τ ,

$$F_{(-1)}(\omega) = \frac{1}{j\omega} F(\omega)$$

при

$$\int_{-\infty}^{\infty} f(\tau) d\tau = 0.$$

5. Теорема Рейли (теорема энергий). Перепишем выражение для обратного преобразования Фурье с несколько измененной индексацией

$$f_1(\tau) = \frac{1}{2\pi} \int_{-\infty}^{\infty} F_1(\omega) e^{j\omega\tau} d\omega.$$

Умножим обе части полученного выражения на $f_2(\tau)$ и проинтегрируем полученное выражение в пределах $\pm\infty$. В результате получим

$$\begin{aligned} \int_{-\infty}^{\infty} f_1(\tau) f_2(\tau) d\tau &= \frac{1}{2\pi} \int_{-\infty}^{\infty} f_2(\tau) \int_{-\infty}^{\infty} F_1(\omega) e^{j\omega\tau} d\omega d\tau \\ &= \int_{-\infty}^{\infty} F_1(\omega) d\omega \int_{-\infty}^{\infty} f_2(\tau) e^{j\omega\tau} d\tau. \end{aligned}$$

Следовательно, учитывая выражение для прямого преобразования Фурье (6.33), окончательно получим

$$\int_{-\infty}^{\infty} f_1(\tau) f_2(\tau) d\tau = \frac{1}{2\pi} \int_{-\infty}^{\infty} F_1(\omega) F_2(-\omega) d\omega. \quad (6.43)$$

Учитывая, что $F_1(\omega)$ и $F_2(-\omega)$ есть сопряженные комплексные числа и используя тригонометрическую форму представления комплексного числа в вещественной форме, в общем случае при $f_1(\tau) \neq f_2(\tau)$ будем иметь

$$\int_{-\infty}^{\infty} f_1(\tau) f_2(\tau) d\tau = \frac{1}{\pi} \int_0^{\infty} \Phi_1(\omega) \Phi_2(\omega) \cos(\varphi_1 - \varphi_2) d\omega.$$

Для частного случая, когда

$$f_1(\tau) = f_2(\tau),$$

$$\int_{-\infty}^{\infty} f^2(\tau) d\tau = \frac{1}{\pi} \int_0^{\infty} \Phi^2(\omega) d\omega. \quad (6.44)$$

Соотношение (6.44) известно под названием теоремы Рейли. По физическому смыслу функция (6.44) представляет собой спектральную плотность энергии. Кроме того, выражение (6.44) означает, что энергию некоторого процесса можно вычислять двояким способом: либо интегрируя квадрат функции времени (мгновенная мощность), либо интегрируя квадрат амплитудного спектра.

6. Теорема о транспозиции (переносе) спектра. Пусть задана функция $f(\tau)$ с известным спектром $F(\omega)$. Требуется определить, какой функции будет соответствовать спектр при смещении частоты на Ω . По формуле (6.33) для прямого преобразования Фурье имеем

$$F(\omega + \Omega) = \int_{-\infty}^{\infty} f(\tau) e^{-j(\omega + \Omega)\tau} d\tau,$$

откуда можно определить искомую функцию

$$f_{\Omega}(\tau) = e^{-j\Omega\tau} f(\tau). \quad (6.45)$$

7. Теорема о спектре произведения двух функций. Пусть имеем две функции $f_1(\tau)$ и $\psi_2(\tau) = f_2(\tau) e^{-j\omega\tau}$. Перепишем формулу (6.43) в новых обозначениях

$$\int_{-\infty}^{\infty} f_1(\tau) \psi_2(\tau) d\tau = \frac{1}{2\pi} \int_{-\infty}^{\infty} F_1(\Omega) \Phi_2(-\Omega) d\Omega.$$

Если у функции $f_2(\tau)$ был спектр F_2 , то на основании теоремы о транспозиции спектра

$$\Phi_2(\Omega) = F(\omega - \Omega).$$

Следовательно,

$$\int_{-\infty}^{\infty} f_1(\tau) \psi_2(\tau) d\tau = \frac{1}{2\pi} \int_{-\infty}^{\infty} F_1(\Omega) F_2(\omega - \Omega) d\Omega. \quad (6.46)$$

Таким образом, если F_1 и F_2 соответственно — спектры двух функций $f_1(\tau)$ и $\psi_2(\tau)$, а F — спектр произведения этих функций, то

$$F(\omega) = \frac{1}{2\pi} \int_{-\infty}^{\infty} F_1(\Omega) F_2(\omega - \Omega) d\Omega. \quad (6.47)$$

Другими словами, спектр произведения двух функций равен произведению спектров этих функций.

8. Теорема о спектре свертки. Сверткой (или складкой) двух функций называется интеграл вида

$$f(\tau) = \int_{-\infty}^{\infty} f_1(t) f_2(\tau - t) dt. \quad (6.48)$$

Используя формулу (6.33) для прямого преобразования Фурье, вычислим спектр такой функции:

$$\begin{aligned} S(\omega) &= \int_{-\infty}^{\infty} f(\tau) e^{-j\omega\tau} d\tau = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_1(t) f_2(\tau - t) e^{-j\omega\tau} dt d\tau = \\ &= \int_{-\infty}^{\infty} e^{-j\omega\tau} d\tau \int_{-\infty}^{\infty} f_1(t) f_2(\tau - t) dt = \\ &= \int_{-\infty}^{\infty} f_1(t) dt \int_{-\infty}^{\infty} f_2(\tau - t) e^{-j\omega\tau} d\tau = \\ &= \int_{-\infty}^{\infty} f_1(t) e^{-j\omega t} dt \int_{-\infty}^{\infty} e^{-j\omega\mu} f_2(\mu) d\mu. \end{aligned} \quad (6.49)$$

(Здесь была сделана замена переменной $\mu = \tau - t$). Таким образом, видно, что спектр функции $f(\tau)$ будет иметь вид

$$S(\omega) = S_1(\omega) S_2(\omega). \quad (6.50)$$

Спектр свертки равен произведению спектров свертываемых функций.

§ 11. ТЕКУЩИЙ И МГНОВЕННЫЙ СПЕКТРЫ

Мы определили периодическую функцию следующим соотношением:

$$f(\tau + T) = f(\tau). \quad (6.51)$$

Но выше неоднократно отмечалось, что такое определение есть матическая абстракция, поскольку не может существовать реального физического процесса, отвечающего выражению (6.51).

Для вычисления спектра по формуле (6.33) для прямого преобразования Фурье необходимо интегрирование по времени в бесконечных пределах

$$F(\omega) = \int_{-\infty}^{\infty} f(\tau) e^{-j\omega\tau} d\tau.$$

Поскольку функция $f(\tau)$ есть отображение некоторого реального физического процесса, информация о его ходе известна лишь до настоящего текущего момента.

Здесь имеется в виду, что ход процесса от настоящего момента до $+\infty$ нельзя предсказать на основании одних рассуждений. Информацию о ходе процесса от $-\infty$ до текущего момента можно считать в определенном приближении нам известной. Тогда формулу (6.33) для вычисления спектра можно преобразовать к виду

$$F_{\tau}(\omega) = \int_{-\infty}^{\tau} f(\tau) e^{-j\omega\tau} d\tau. \quad (6.52)$$

Величина $F_{\tau}(\omega)$, определенная выражением (6.52), является функцией не только частоты, но и времени. При более строгом подходе к вопросу о нижнем пределе интегрирования в выражении (6.52) мы должны учитывать, что фактически информация о ходе данного физического процесса, описываемого функцией $f(\tau)$, известна нам лишь начиная с некоторого момента t_0 , условно принимаемого нами за начало отсчета времени. На основании всего этого, выражение (6.52) можно переписать в другой форме

$$F_{\tau}(\omega) = \int_0^t f(\tau) e^{-j\omega\tau} d\tau. \quad (6.53)$$

Выражения (6.52) и (6.53) называются текущим спектром, связывающим спектральные представления функций с временными.

Текущий спектр выражает со спектральной точки зрения развитие процесса во времени. Спектр короткого отрезка процесса за небольшое время от его начала — однороден, так как короткий отрезок любого процесса есть по сути дела просто короткий импульс. Если с течением времени происходит периодическое повторение некоторого цикла явления, то на текущем спектре начинают формироваться максимумы на основной частоте и ее гармониках. Здесь мерилom периодичности является число периодов: если число периодов повторения отрезка любого процесса много больше единицы, то можно считать действительный циклический процесс периодическим. С течением времени максимумы становятся все более острыми, амплитуда их растет, а значение спектральной плотности между максимумами убывает и в пределе при длительности периодического процесса $t \rightarrow \infty$ сплошной текущий спектр вырождается в линейчатый

текущий спектр периодического, в строгом смысле, сигнала. Таким образом, основным в приведенных выше рассуждениях является допущение, что периодический процесс — это тот предел, к которому может стремиться с течением времени реальный повторяющийся процесс.

Дальнейшим развитием связи между частотным и временным представлением явлений является введение понятия мгновенного спектра. Если первоначальное определение спектра (6.33) дает функцию частоты, где зависимость от времени выпадает, так как спектр отражает процесс в целом:

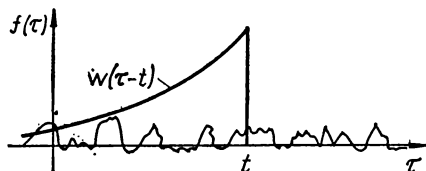


Рис. 6.4. Пример весовой функции при определении мгновенного спектра.

определение текущего спектра отражает предысторию процесса вплоть до текущего момента, то понятие о мгновенном спектре дает представление о спектре, изменяющемся во времени и отражающем

свойства процесса в данный момент.

Приведем некоторые примеры различных определений мгновенного спектра.

Простейшее определение мгновенного спектра можно дать в виде

$$F_T(\omega; \tau) = \int_{\tau-T}^{\tau} f(t) e^{-i\omega t} dt. \quad (6.54)$$

Здесь осуществляется «скользящее» интегрирование. Дело в том, что при таком определении мгновенного спектра интервал интегрирования имеет постоянную длину, но перемещается по оси времени. Кроме того, расположение этого интервала интегрирования неизменно относительно текущего времени.

Наиболее общим является описание мгновенного спектра с помощью введения в подынтегральное выражение (6.33) скользящей весовой функции, вид которой определяется конкретными требованиями. Так, например, с введением весовой функции вида $w(t - \tau)$ определение мгновенного спектра преобразуется к виду

$$F_w(\omega; \tau) = \int_{-\infty}^{\infty} w(t - \tau) f(t) e^{-i\omega t} dt. \quad (6.55)$$

Весовая функция вида $w(t - \tau) = e^{\alpha\tau}(\tau - t)$ учитывает предысторию процесса с весом, экспоненциально убываю-

щим по мере удаления от настоящего момента. Эта функция представляет собой практическое требование, определяемое результатом спектрального анализа при помощи некоторых реальных фильтров (рис. 6.4). Выражение для мгновенного спектра примет вид

$$F_{\omega}(\omega; \tau) = \int_{-\infty}^{\infty} e^{i\omega\tau} (\tau - t) f(t) e^{-i\omega t} dt. \quad (6.56)$$

Таким образом, конкретный вид выражения для мгновенного спектра определяется выбором целесообразной, с точки зрения необходимости, скользящей весовой функции.

Контрольные вопросы

1. Как определяется спектр амплитуд и спектр фаз?
2. Какой сигнал характеризуется дискретным (или линейчатым) спектром?
3. Чем отличается спектр периодической функции от спектра непериодической?
4. Чем определяется спектр непериодической функции?
5. Справедлив ли принцип суперпозиции для преобразования Фурье?
6. В чем заключается способ сжатия спектра?
7. Каким образом можно вычислять энергию какого-либо процесса?
8. Что выражают со спектральной точки зрения понятия текущего и мгновенного спектров?

§ 12. МОДУЛЯЦИЯ.

СПЕКТРЫ МОДУЛИРОВАННЫХ КОЛЕБАНИЙ

Процесс модуляции определяется тем, что некоторый параметр переносчика информации о передаваемом сигнале изменяется во времени в соответствии с законом изменения передаваемого сигнала. Аналитическое представление гармонического характера изменения некоторого параметра представляется в виде

$$f(\tau) = A \cos(\omega\tau + \theta) = A \cos \varphi. \quad (6.57)$$

Если A — амплитуда, ω — частота и θ — начальная фаза, которые могут быть либо постоянными, либо медленно меняющимися величинами, а полная фаза гармонического колебания, описываемого выражением (6.57), определяется параметром $\varphi = \omega\tau + \theta$, то выражение (6.57) представляет собой немодулированное колебание. Теперь, если один из параметров — A или φ — изменять по закону сигнала, передающего информацию, то колебание $f(\tau)$ становится модулированным. Изменение параметра A немодулированного

колебания по закону изменения сигнала, несущего информацию, называется *амплитудной модуляцией* (АМ).

Изменение полной фазы, как правило высокочастотного гармонического колебания, составляемого выражением (6.57), по закону изменения сигнала, несущего информацию, называется *угловой модуляцией*. При воздействии на параметр ω угловая модуляция называется *частотной модуляцией* (ЧМ). При воздействии на параметр θ угловая модуляция носит название *фазовой модуляции* (ФМ).

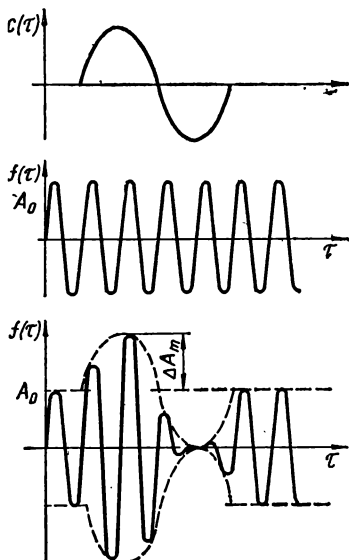


Рис. 6.5. Амплитудно-модулированные колебания.

Рассмотрим прежде вопрос о спектре амплитудно-модулированных колебаний. Пусть сигнал, несущий полезную информацию, изменяется по закону $c(\tau)$ (рис. 6.5, а). При условии, что высокочастотная несущая имеет вид, показанный на рис. 6.5, б, амплитудно-модулированное колебание будет иметь вид, показанный на рис. 6.5, в.

Аналитически такой случай амплитудной модуляции можно представить выражением

$$f(\tau) = [A_0 + kc(\tau)] \times \cos(\omega_0\tau + \theta_0) = A(\tau) \times \cos(\omega_0\tau + \theta_0). \quad (6.58)$$

Здесь A_0 — амплитуда несущего колебания при отсутствии модуляции; ω_0 — несущая частота; θ_0 — начальная фаза; k — коэффициент пропорциональности; $A(\tau)$ — огибающая модулированного колебания.

Поскольку $c(\tau)$ — гармоническое колебание, то уравнение огибающей можно представить в виде

$$A(\tau) = A_0 + \Delta A_m \cos(\Omega\tau + \gamma), \quad (6.59)$$

где ΔA_m — изменение амплитуды огибающей высокочастотного колебания, Ω — частота модулирующего сигнала, γ — начальная фаза модулирующего сигнала. Отношение $\frac{\Delta A_m}{A_0} = M$ называется коэффициентом глубины модуляции.

Для модуляции без искажений должно выполняться условие $M \leq 1$. В случае рассмотренной тональной модуляции, т. е. модуляции одним тоном, когда модулирующий сигнал является гармоническим, аналитическое выражение для модулированного сигнала имеет вид

$$f(\tau) = A_0 [1 + M \cos(\Omega\tau + \gamma)] \cos(\omega_0\tau + \theta_0), \quad (6.60)$$

или в общем случае

$$f(\tau) = A_0 [1 + Mc(\tau)] \cos(\omega_0\tau + \theta_0).$$

Причем, следует иметь в виду, что амплитудно-модулированное колебание изменяется от $A_{\min} = A_0 (1 - M)$ до $A_{\max} = A_0 (1 + M)$.

Рассмотрим вопрос о спектре амплитудно-модулированного колебания в зависимости от вида модулирующего воздействия.

1. Спектр модулированного колебания в случае тональной модуляции. Перепишем выражение (6.60) в виде

$$f(\tau) = A_0 \cos(\omega_0\tau + \theta_0) + A_0 M \cos(\Omega\tau + \gamma) \cos(\omega_0\tau + \theta_0).$$

Далее, заменив произведение косинусов суммой косинусов, получим

$$f(\tau) = A_0 \cos(\omega_0\tau + \theta_0) + \frac{A_0 M}{2} \cos[(\omega_0 + \Omega)\tau + \theta_0 + \gamma] + \frac{A_0 M}{2} \cos[(\omega_0 - \Omega)\tau + \theta_0 - \gamma]. \quad (6.61)$$

Как видно из выражения (6.61), первое слагаемое есть исходное немодулированное колебание с несущей частотой ω_0 . Второе и третье слагаемые появились в результате амплитудной модуляции.

Частота $(\omega_0 + \Omega)$ называется верхней боковой частотой, а частота $(\omega_0 - \Omega)$ — нижней боковой частотой. Амплитуды этих двух колебаний одинаковы и равны $\frac{A_0 M}{2}$, а фазы симметричны относительно фазы несущего колебания. Таким образом, тональное амплитудно-модулированное колебание имеет спектр амплитуд, который графически изображен на рис. 6.6. Этот спектр дискретный, так как содержит составляющие трех частот. Ширина полосы частот, занимаемая таким спектром, равна 2Ω .

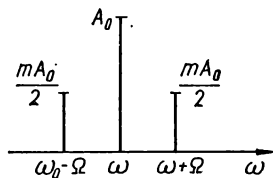


Рис. 6.6. Тональное АМ колебание.

Следует иметь в виду, что амплитудная модуляция — это не суммирование гармонических колебаний, а гораздо более сложный процесс преобразования спектра частот.

2. Спектр амплитудно-модулированного колебания, когда модулирующее воздействие содержит ряд гармоник, т. е.

$$c(\tau) = \sum_{i=1}^n c_i \cos(\Omega_i \tau + \gamma_i). \quad (6.62)$$

Спектр сигнала, описываемого выражением (6.62), показан на рис. 6.7, а. Используя выражение (6.62), выражение (6.61) перепишем в виде

$$f(\tau) = \left[A_0 + \sum_{i=1}^n \Delta A_{m_i} \cos(\Omega_i \tau + \gamma_i) \right] \cos(\omega_0 \tau + \theta_0). \quad (6.63)$$

Прделаем над выражением (6.63) несложное преобразование

$$\begin{aligned} f(\tau) &= A_0 \left[1 + \sum_{i=1}^n \cos(\Omega_i \tau + \gamma_i) \right] \cos(\omega_0 \tau + \theta_0) = \\ &= A_0 \cos(\omega_0 \tau + \theta_0) + A_0 \sum_{i=1}^n M_i \cos(\Omega_i \tau + \gamma_i) \cos(\omega_0 \tau + \theta_0) = \\ &= A_0 \cos(\omega_0 \tau + \theta_0) + \frac{A_0}{2} \sum_{i=1}^n M_i \cos[(\omega_0 + \Omega_i) \tau + \theta_0 + \gamma_i] + \\ &\quad + \frac{A_0}{2} \sum_{i=1}^n M_i \cos[(\omega_0 - \Omega_i) \tau + \theta_0 - \gamma_i]. \end{aligned} \quad (6.64)$$

Спектр амплитуд для выражения (6.64) изображен на рис. 6.7, б.

Таким образом, каждое гармоническое колебание, входящее в состав $c(\tau)$, обусловило появление в спектре амплитудно-модулированного колебания двух боковых частот. Следовательно, ширина полосы частот спектра в рассмотренном случае равна $2\Omega_n$, где Ω_n — максимальная частота в разложении $c(\tau)$.

3. Спектр амплитудно-модулированного колебания при условии, что модулирующее воздействие является непериодическим сигналом. Такой случай амплитудной модуляции изображен на рис. 6.8, а, б. Модулирующий сигнал $c(\tau)$ можно представить рядом Фурье в тригонометрической форме.

Тригонометрическую форму интеграла Фурье можно

получить из обратного преобразования Фурье:

$$f(\tau) = \frac{1}{2\pi} \int_{-\infty}^{\infty} F(\omega) e^{i(\omega\tau - \varphi)} d\omega = \frac{1}{2\pi} \int_{-\infty}^{\infty} F(\omega) \cos(\omega\tau - \varphi) d\omega + \\ + \frac{1}{2\pi} \int_{-\infty}^{\infty} F(\omega) \sin(\omega\tau - \varphi) d\omega,$$

так как в первом интеграле подынтегральная функция четная, а во втором — нечетная, окончательно получим

$$f(\tau) = \frac{1}{2\pi} \int_{-\infty}^{\infty} F(\omega) \cos(\omega\tau - \varphi) d\omega = \frac{1}{\pi} \int_0^{\infty} F(\omega) \times \\ \times \cos(\omega\tau - \varphi) d\omega. \quad (6.65)$$

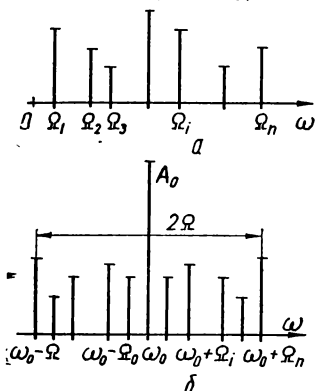


Рис. 6.7. Спектр АМ сигнала, когда модулирующее воздействие содержит ряд гармоник: a — спектр модулирующего колебания; b — спектр модулированного сигнала.

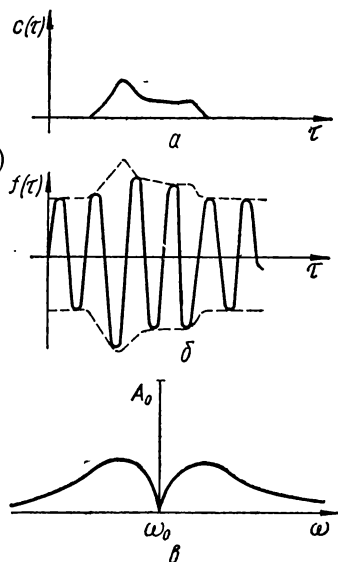


Рис. 6.8. Спектр АМ колебаний при модулировании непериодическим сигналом.

Учитывая (6.65), имеем

$$c(\tau) = \frac{1}{\pi} \int_0^{\infty} c(\Omega) \cos(\Omega\tau - \varphi) d\Omega, \quad (6.66)$$

где $c(\Omega)$ — модуль спектральной плотности сигнала.

Подставим (6.66) в (6.61) и проделаем преобразование, как в пунктах 1 и 2:

$$f(\tau) = A_0 \cos(\omega_0\tau + \theta_0) + k \left[\frac{1}{\pi} \int_0^{\infty} c(\Omega) \cos(\Omega\tau - \varphi) d\Omega \right] \times$$

$$\begin{aligned}
& \times \cos(\omega_0 \tau + \theta_0) = A_0 \cos(\omega_0 \tau + \theta_0) + \\
& + k \left[\frac{1}{\pi} \int_0^{\infty} c(\Omega) \cos(\Omega t - \varphi) \cos(\omega_0 \tau + \theta_0) d\Omega = \right. \\
& = A_0 \cos(\omega_0 \tau + \theta_0) + \frac{k}{\pi} \int_0^{\infty} \frac{c(\Omega)}{2} \cos[(\omega_0 + \Omega) \tau + \theta_0 - \varphi] d\Omega + \\
& \left. + \frac{k}{\pi} \int_0^{\infty} \frac{c(\Omega)}{2} \cos[(\omega_0 - \Omega) \tau + \theta_0 + \varphi] d\Omega. \quad (6.67)
\end{aligned}$$

Из выражения (6.67) заключаем, что при амплитудной модуляции непериодическим сигналом, обладающим сплошным спектром, по обе стороны от несущей частоты ω_0 образуются (рис. 6.8, в) две сплошные полосы боковых частот $(\omega_0 + \Omega)$ и $(\omega_0 - \Omega)$. Амплитуды колебаний этих боковых частот пропорциональны амплитудам составляющих с частотами Ω , входящими в модулирующее воздействие $c(\tau)$.

Амплитудно-модулированное колебание не является периодическим даже тогда, когда модулирующее воздействие изменяется по периодическому закону (если только несущая частота ω_0 и основная частота модуляции Ω не находятся в простом кратном соотношении).

В общем случае амплитудно-модулированное колебание, представленное рядом (6.64), относится к классу почти периодических функций, следовательно, амплитудно-модулированное колебание является почти периодическим сигналом.

Существует также понятие *балансной модуляции*, спектр которой состоит только из боковых частот. Математически это выражается тем, что амплитуда несущей частоты умножается не на $[1 + M c(\tau)]$, а просто на $c(\tau)$. Таким образом, при балансной модуляции осуществляется простое перемножение модулирующей функции и колебаний несущей частоты.

Спектр гармонических колебаний с угловой модуляцией. Учитывая введенное ранее понятие об угловой модуляции гармонических колебаний, остановимся на связи между ЧМ и ФМ. Очевидно, что в общем случае угловая частота ω определяется как производная по времени от полной фазы

$$\omega = \frac{d\varphi}{dt}. \quad (6.68)$$

Следовательно, полная фаза

$$\varphi = \int \omega dt + \theta_0. \quad (6.69)$$

Таким образом, если изменять частоту ω , то изменится фаза φ и наоборот. Поскольку при частотной модуляции мгновенное значение частоты $\omega(t)$ высокочастотного колебания $f(t)$ изменяется по закону модулирующего воздействия $c(\tau)$ (рис. 6.9), мгновенное значение частоты ω модулированного колебания можно записать

$$\omega = \omega_0 + k_{\text{чм}} c(\tau). \quad (6.70)$$

Здесь ω_0 — частота модулированного колебания; $k_{\text{чм}}$ — коэффициент, устанавливающий связь между модулирующим воздействием и изменением несущей частоты. Подставляя (6.70) в (6.69), получим выражение для полной фазы

$$\varphi = \omega_0 t + k_{\text{чм}} \int c(\tau) d\tau + \theta_0. \quad (6.71)$$

Из выражения, (6.71) видно, как изменялась фаза колебаний при частотной модуляции.

Аналитически выражение для сигнала, подвергнутого ЧМ, можно записать в виде

$$f(\tau) = A_0 \cos(\omega_0 \tau + k_{\text{чм}} \int c(\tau) d\tau + \theta_0).$$

Для введения некоторых дополнительных понятий, рассмотрим частный случай ЧМ, когда модулирующий сигнал представляет собой гармоническое колебание вида

$$c(\tau) = c_m \cos(\Omega \tau + \gamma).$$

Мгновенная частота

$$\omega = \omega_0 + k_{\text{чм}} c_m \cos(\Omega \tau + \gamma).$$

Обозначая

$$k_{\text{чм}} c_m = \omega_d,$$

получим

$$\omega = \omega_0 + \omega_d \cos(\Omega \tau + \gamma). \quad (6.72)$$

Здесь ω_d — максимальное отклонение частоты модулированного колебания от несущей частоты, называемое *девиацией*.

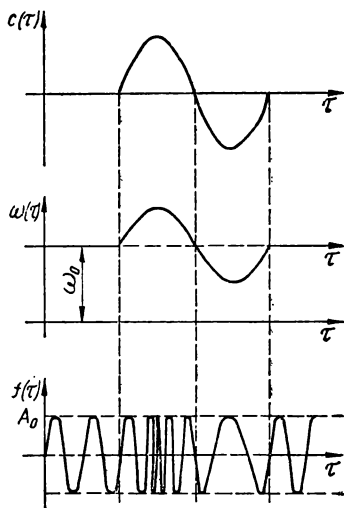


Рис. 6.9. Частотная модуляция.

цией. Подставляя (6.72) в (6.69), получим выражения для полной фазы

$$\varphi = \omega_0 \tau + \frac{\omega_d}{\Omega} \sin(\Omega t + \gamma) + \theta_0. \quad (6.73)$$

Величина $m = \frac{\omega_d}{\Omega}$ называется *индексом модуляции*. Окончательно ЧМ колебание при модуляции гармоническим сигналом запишется так:

$$f(\tau) = A_0 \cos(\omega_0 \tau + m \sin(\Omega \tau + \gamma + \theta_0)). \quad (6.74)$$

Рассмотрим подробнее вопрос о фазовой модуляции. Как уже указывалось выше, при фазовой модуляции по закону модулирующего возмущения изменяется начальная фаза θ_0 высокочастотного немодулированного колебания. Если принять, что модулирующее возмущение изменяется по закону $c(\tau)$, то изменение начальной фазы будет иметь вид

$$\theta = \theta_0 + k_{\Phi M} c(\tau).$$

Здесь $k_{\Phi M}$ — коэффициент, учитывающий связь между модулирующим сигналом и изменением фазы колебания.

Выражение для полной фазы фазово-модулированного колебания примет вид

$$\varphi = \omega_0 \tau + \theta_0 + k_{\Phi M} c(\tau).$$

Мгновенная частота по формуле (6.68)

$$\omega = \omega_0 + k_{\Phi M} \frac{dc(\tau)}{d\tau}.$$

Аналитическое выражение для ФМ колебания можно получить в следующем виде:

$$f(\tau) = A_0 \cos[\omega_0 \tau + k_{\Phi M} c(\tau) + \theta_0]. \quad (6.75)$$

В простейшем случае, когда $c(\tau)$ — гармоническое колебание вида

$$c(\tau) = c_m \cos(\Omega \tau + \gamma)$$

полная фаза ФМ колебания преобразуется к виду

$$\varphi = \omega_0 \tau + k_{\Phi M} c_m \cos(\Omega \tau + \gamma) + \theta_0.$$

Введем понятие *индекса модуляции*, представляющего собой максимальное отклонение фазы $m = k_{\Phi M} c_m$. Полная фаза ФМ колебания определяется из равенства

$$\varphi = \omega_0 \tau + m \cos(\Omega \tau + \gamma) + \theta_0,$$

а ФМ колебание аналитически записывается так:

$$f(\tau) = A_0 \cos [\omega_0 \tau + m \cos (\Omega \tau + \gamma) + \theta_0]. \quad (6.76)$$

Отсюда мгновенное значение частоты

$$\omega = \omega_0 + m\Omega \sin (\Omega \tau + \gamma) = \omega_0 + \omega_d \sin (\Omega \tau + \gamma).$$

Здесь, как и ранее, ω_d — девиация частоты.

Сравнивая выражение (6.74) для ЧМ колебания и выражение (6.76) для ФМ колебания, видим, что типы угловой модуляции отличаются друг от друга только фазой гармонической функции, определяющей изменение полной фазы высокочастотного колебания.

Заметим, что для того, чтобы определить, с какой модуляцией мы имеем дело — с частотной или с фазовой — недостаточно знать лишь характер модулированного колебания: необходимо знать и закон изменения модулирующего воздействия. Так, если аналитическое выражение для модулированного сигнала имеет вид

$$f(\tau) = A_0 \cos (\omega_0 \tau + m \sin (\Omega \tau + \gamma) + \theta_0], \quad (6.77)$$

то в случае модуляции высокочастотного сигнала в таком виде:

$$c(\tau) = c_m \cos (\Omega \tau + \gamma)$$

выражение (6.77) описывает ЧМ колебания. А в случае

$$c(\tau) = c_m \sin (\Omega \tau + \gamma)$$

выражение (6.77) описывает ФМ колебания.

Рассмотрим теперь элементы теории спектров колебаний с угловой модуляцией. Воспользуемся выражением (6.77) для вывода интересующих нас соотношений. Преобразуя косинус суммы двух углов по тригонометрической формуле, получим

$$\begin{aligned} f(\tau) &= A_0 \cos [m \sin (\Omega \tau + \gamma)] \cos (\omega_0 \tau + \theta_0) - \\ &- A_0 \sin [m \sin (\Omega \tau + \gamma)] \sin (\omega_0 \tau + \theta_0). \end{aligned} \quad (6.78)$$

Для выражения (6.78) целесообразно рассмотреть два случая. Рассмотрим сначала спектр модулированного колебания, когда индекс модуляции $m \ll 1$, тогда можно принять:

$$\left. \begin{aligned} \sin [m \sin (\Omega \tau + \gamma)] &\approx m \sin (\Omega \tau + \gamma), \\ \cos [m \sin (\Omega \tau + \gamma)] &\approx 1. \end{aligned} \right\} \quad (6.79)$$

Подставляя выражение (6.79) в (6.78), получим

$$f(\tau) \approx A_0 \cos (\omega_0 \tau + \theta_0) - A_0 m \sin (\Omega \tau + \gamma) \sin (\omega_0 \tau + \theta_0).$$

Далее, заменяя произведение синусов по тригонометрической формуле, имеем

$$f(\tau) \approx A_0 \cos(\omega_0 \tau + \theta_0) - \frac{A_0 m}{2} \cos[(\omega_0 - \Omega) \tau + (\theta_0 - \gamma)] + \\ + \frac{A_0 m}{2} \cos[(\omega_0 + \Omega) \tau + (\theta_0 + \gamma)]. \quad (6.80)$$

Таким образом, при наложении ограничений $m \ll 1$ видим, что спектр колебаний с угловой модуляцией не отличается от спектра колебаний с амплитудной модуляцией, однако фаза колебаний нижней боковой частоты при угловой модуляции сдвинута (в формуле (6.80) знак минус) по отношению к амплитудной модуляции. Спектр колебания с угловой модуляцией для этого случая изображен на рис. 6.10.

Рис. 6.10. Спектр амплитуд колебаний с угловой модуляцией при $m \ll 1$.

Рассмотрим второй, более общий случай, когда m — любая величина.

В выражении (6.78) функции $\cos[m \sin(\Omega \tau + \gamma)]$ и $\sin[m \sin(\Omega \tau + \gamma)]$ по теории бесселевых функций разложим в тригонометрические ряды вида

$$\begin{cases} \sin[m \sin(\Omega \tau + \gamma)] = 2 \sum_{n=1}^{\infty} I_{2n-1}(m) \sin[(2n-1)(\Omega \tau + \gamma)], \\ \cos[m \sin(\Omega \tau + \gamma)] = I_0(m) + 2 \sum_{n=1}^{\infty} I_{2n}(m) \cos[2n(\Omega \tau + \gamma)], \end{cases} \quad (6.81)$$

где $I_n(m)$ — бесселева функция первого порядка. Учитывая формулы (6.81), выражение (6.78) перепишем в виде

$$f(\tau) = A_0 I_0(m) \cos(\omega_0 \tau + \theta_0) - \\ - 2A_0 I_1(m) \sin(\Omega \tau + \gamma) \sin(\omega_0 \tau + \theta_0) + \\ + 2A_0 I_2(m) \cos(2\Omega \tau + 2\gamma) \cos(\omega_0 \tau + \theta_0) - \\ - 2A_0 I_3(m) \sin(3\Omega \tau + 3\gamma) \sin(\omega_0 \tau + \theta_0) + \\ + 2A_0 I_4(m) \cos(4\Omega \tau + 4\gamma) \cos(\omega_0 \tau + \theta_0) + \\ + 2A_0 I_5(m) \sin(5\Omega \tau + 5\gamma) \sin(\omega_0 \tau + \theta_0) + \dots$$

Далее, путем замены произведений косинусов и синусов по тригонометрическим формулам, получим

$$f(\tau) = A_0 I_0(m) \cos(\omega_0 \tau + \theta_0) -$$

$$\begin{aligned}
& - A_0 I_1(m) \cos [(\omega_0 - \Omega) \tau + \theta_0 - \gamma] + A_0 I_1(m) \cos [(\omega_0 + \\
& + \Omega) \tau + \theta_0 + \gamma] + A_0 I_2(m) \cos [(\omega_0 - 2\Omega) \tau + \theta_0 - 2\gamma] + \\
& + A_0 I_2(m) \cos [(\omega_0 + 2\Omega) \tau + \theta_0 + 2\gamma] - A_0 I_3(m) \cos [(\omega_0 - \\
& - 3\Omega) \tau + \theta_0 - 3\gamma] + A_0 I_3(m) \cos [(\omega_0 + 3\Omega) \tau + \theta_0 + 3\gamma] + \\
& + A_0 I_4(m) \cos [(\omega_0 - 4\Omega) \tau + \theta_0 - 4\gamma] + A_0 I_4(m) \cos [(\omega_0 + \\
& + 4\Omega) \tau + \theta_0 + 4\gamma] - A_0 I_5(m) \cos [(\omega_0 - 5\Omega) \tau + \theta_0 - 5\gamma] + \\
& + A_0 I_5(m) \cos [(\omega_0 + 5\Omega) \tau + \theta_0 + 5\gamma] + \dots
\end{aligned}$$

Следовательно, при угловой модуляции в общем случае спектр колебания состоит из бесконечного числа боковых частот, отличающихся от несущей на $\pm n\Omega$, где $n = 1 \div \infty$, а амплитуды этих боковых частот равны $A_n = I_n(m)A_0$, причем A_0 — амплитуда немодулированного колебания, а m — индекс модуляции. Вид спектра колебаний с угловой модуляцией при $m = 5$ и $\Omega = \text{const}$ показан на рис. 6.11, а. Хотя теоретически спектр колебаний с угловой модуляцией безграничен, практически он конечен, на что указывают свойства

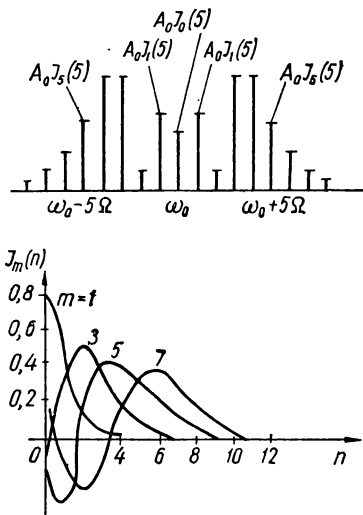


Рис. 6.11. Спектр колебаний с угловой модуляцией.

бесселевых функций. Так, на рис. 6.11, б представлены бесселевы функции для различных m в функции от n .

Из графиков видно, что функция $I_n(m)$ начинает убывать, когда $n = m$. Обычно при $n > m$ в спектре с угловой модуляцией отбрасывают все составляющие, амплитуды которых меньше $0,01 A_0$.

Очевидно, что полная ширина спектра колебаний с угловой модуляцией равна $2n\Omega$. С некоторым приближением можно считать, что в случае $m \gg 1$, $n \approx m$, и тогда $2n\Omega \approx 2m\Omega = 2\omega_d$, т. е. практическая ширина спектра полосы с некоторым приближением равна удвоенной девиации частоты.

Проследим, как проявляется различие между ЧМ и ФМ колебаниями. Для этого рассмотрим, как будут изменяться

спектры ЧМ и ФМ колебаний при изменении частоты Ω тогда, когда $m \gg 1$. Поскольку в этом случае $m \approx n$, то $n \approx m = \frac{\omega_d}{\Omega}$. Поэтому, например, при увеличении частоты модуляции Ω и постоянной амплитуде модулирующего сигнала число спектральных составляющих уменьшается, и, наоборот, с уменьшением частоты Ω число спектральных составляющих возрастает. Однако и в том и в другом случае согласно выражению

$$2n\Omega \approx 2m\Omega = 2\omega_d$$

практическая ширина спектра ЧМ колебаний остается постоянной.

Для случая ФМ колебаний, при условии $m \gg 1$, ширина спектра в соответствии с выражением

$$\omega_d = m\Omega; \quad 2n\Omega \approx 2m\Omega = 2\omega_d \\ (m = k_{\text{ФМ}} c_M)$$

равна

$$2n\Omega \approx 2m = 2k_{\text{ФМ}} c_M \Omega c_m \Omega,$$

т. е. она зависит как от амплитуды, так и от частоты модулирующего воздействия. Таким образом, из этих четырех соотношений видно, что ФМ число спектральных линий спектра, если $c_M = \text{const}$, остается неизменным. С изменением Ω при $c_M = \text{const}$ изменяется интервал между соседними гармониками, а следовательно, общая ширина спектра.

В заключение без доказательства отметим, что практическая ширина спектра колебаний с угловой модуляцией примерно в m раз больше ширины спектра АМ колебаний.

Контрольные вопросы

1. Что такое модуляция?
2. Что такое амплитудная, угловая, фазовая и частотная модуляция?
3. Что такое тональная модуляция?
4. К какому классу относится амплитудно-модулированное колебание?
5. Что такое балансная модуляция?
6. Чем отличаются друг от друга типы угловой модуляции?

§ 13. ПЕРЕНОС СПЕКТРА

Иногда необходимо такое преобразование заданной временной функции, в результате которого спектр функции сместился бы по шкале частот.

Как известно при обычной модуляции получаются две боковые полосы. Нужно получить спектр, состоящий из одной боковой полосы, требуемое преобразование которого показано на рис. 6.12.

Отметим, что здесь речь идет о смещении вещественного спектра амплитуд, чего не может дать теорема о переносе спектра (см. § 10).

Простейший технический способ осуществления необходимого смещения спектра состоит в том, что посредством балансной модуляции несущей частоты образуют двухполосный модуляционный спектр, а затем при помощи фильтров

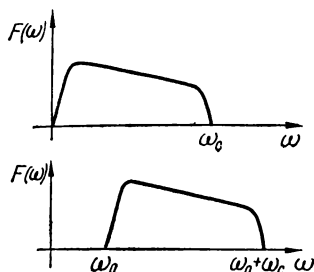


Рис. 6.12. Перенос спектра.

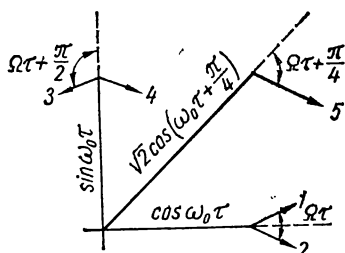


Рис. 6.13. Векторная диаграмма.

подавляют нижнюю боковую частоту. Существуют и другие способы, простейший из которых мы рассмотрим в аналитическом описании.

Формулу (6.34) для обратного преобразования Фурье в вещественной форме можно записать в виде

$$f(\tau) = \frac{1}{\pi} \operatorname{Re} \int_0^{\infty} F(\omega) e^{i\omega\tau} d\omega.$$

Задача состоит в построении функции, спектр которой $F(\omega)$, но на частоте $\omega_0 + \omega$. Для искомой функции можно записать

$$\begin{aligned} \varphi(\tau) &= \frac{1}{\pi} \operatorname{Re} \int_0^{\infty} F(\omega) e^{i(\omega_0 + \omega)\tau} d\omega = \\ &= \frac{1}{\pi} \left(\cos \omega_0 \tau \operatorname{Re} \int_0^{\infty} F(\omega) e^{i\omega\tau} d\omega - \sin \omega_0 \tau \operatorname{Im} \int_0^{\infty} F(\omega) e^{i\omega\tau} d\omega \right), \end{aligned}$$

или

$$\varphi(\tau) = f(\tau) \cos \omega_0 \tau + f^*(\tau) \sin \omega_0 \tau, \quad (6.82)$$

где

$$f^*(\tau) = -\frac{1}{\pi} \operatorname{Im} \int_0^{\infty} F(\omega) e^{j\omega\tau} d\omega. \quad (6.83)$$

Поскольку

$$F(\omega) = A(\omega) + jB(\omega),$$

то выражение (6.83) преобразуется к виду

$$\begin{aligned} f^*(\tau) &= -\frac{1}{\pi} \int_0^{\infty} (A \sin \omega\tau + B \cos \omega\tau) d\omega = \\ &= \frac{1}{\pi} \int_0^{\infty} \left[A \cos \left(\omega\tau + \frac{\pi}{2} \right) - B \sin \left(\omega\tau + \frac{\pi}{2} \right) \right] d\omega. \end{aligned} \quad (6.84)$$

Для $f(\tau)$ аналогичными преобразованиями получаем

$$f(\tau) = \frac{1}{\pi} \int_0^{\infty} (A \cos \omega\tau - B \sin \omega\tau) d\omega. \quad (6.85)$$

Сравнивая выражения (6.84) и (6.85), видим, что $f^*(\tau)$ отличается от $f(\tau)$ только тем, что ее составляющие повернуты на $\frac{\pi}{2}$ по фазе.

Графическая интерпретация сущности переноса спектра, отвечающая выражению (6.82), показана на рис. 6.13.

Здесь принимается, что

$$f(\tau) = \cos \Omega\tau;$$

$$f^*(\tau) = -\sin \Omega\tau = \cos \left(\Omega\tau + \frac{\pi}{2} \right).$$

Для простоты и наглядности принята не балансная, а обычная АМ. Следовательно, составляющая с несущей частотой не уничтожается. По горизонтали откладывают вектор несущего колебания $\cos \omega_0 t$, в результате модуляции которого получают две боковые составляющие, представленные векторами 1 и 2. Векторы расположены под углом $\pm \Omega\tau$ по отношению к основному вектору. Направления векторов 2 и 4 совпадают, а направления 1 и 3 противоположны. В результате сложения векторов несущей частоты и векторов боковых составляющих получается один вектор несущей частоты и один вектор боковой составляющей, что и требовалось получить.

§ 14. ДЕТЕКТИРОВАНИЕ.

ПРЕОБРАЗОВАНИЕ СПЕКТРОВ ПРИ ДЕТЕКТИРОВАНИИ

Термин детектирование означает обнаружение.

Детектирование — это преобразование сигнала, обратное модулированию. Цель его выделить из модулированного сигнала составляющую с частотой модуляции, закон изменения которой несет полезную информацию о передаваемом сигнале.

Соответственно трем типам модуляции различают три вида детектирования: амплитудное, частотное и фазовое. В результате детектирования получается сложное колебание, в состав которого входит интересующая нас составляющая с частотой модуляции. Поскольку математическое описание процессов детектирования и преобразования спектров при этом является очень громоздким, ограничимся рассмотрением вопроса о преобразовании спектров при детектировании на наглядных примерах.

Пусть детектированию подвергаются АМ колебания при модуляции одним тоном вида

$$f(\tau) = (1 + M \sin \Omega \tau) \sin \omega_0 \tau. \quad (6.86)$$

Спектр такого колебания, как известно, состоит из трех составляющих с частотами ω_0 , $\omega_0 + \Omega$, $\omega_0 - \Omega$. Один из методов выделения интересующего нас колебания с частотой модуляции состоит в применении к модулированному сигналу операции *линейного детектирования*, которую аналитически можно выразить следующим образом:

$$f^*(\tau) = |f(\tau)|, \quad (6.87)$$

(рис. 6.14, а). Попутно следует заметить, что название «линейное детектирование» явно неудачно, так как детектор, выполняющий эту операцию, на самом деле является нелинейным устройством.

Воспользовавшись тем, что абсолютная величина произведения равна сумме абсолютных величин сомножителей,

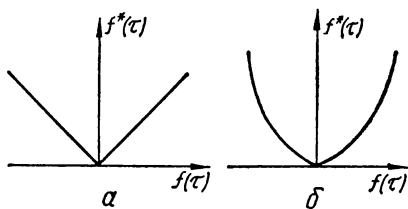


Рис. 6.14. Характеристики «вход — выход» детекторов:

а — линейного; б — квадратичного.

перепишем выражение (6.86) согласно (6.87) следующим образом: $f^*(\tau) = |f(\tau)| = (1 + M \sin \Omega \tau) |\sin \omega_0 \tau|$. (6.88)

Функцию $|\sin \omega_0 \tau|$ разложим в ряд Фурье:

$$|\sin \omega_0 \tau| = \frac{2}{\pi} \left(1 - \sum_{k=1}^{\infty} \frac{1}{4k^2 - 1} \cos 2k\omega_0 \tau \right).$$

Тогда

$$f^*(\tau) = \frac{2}{\pi} \left\{ (1 + M \sin \Omega \tau) - \sum_{k=1}^{\infty} \frac{1}{4k^2 - 1} \left[\cos 2k\omega_0 \tau + \frac{M}{2} \sin (2k\omega_0 + \Omega) \tau - \frac{M}{2} \sin (2k\omega_0 - \Omega) \tau \right] \right\}. \quad (6.89)$$

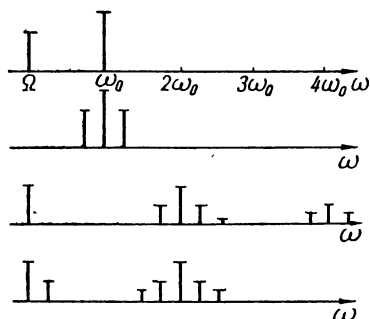


Рис. 6.15. Спектральные представления для случаев линейного и квадратичного детектирования.

В полученном выражении первое слагаемое $(1 + M \sin \Omega \tau)$ и есть та модулирующая функция, которую мы стремились выделить. Отделить составляющие с частотами $2k\omega_0$; $2k\omega_0 + \Omega$; $2k\omega_0 - \Omega$ технически не трудно.

Другой возможный вариант выделения из выражения (6.86) составляющей с частотой модулирования — преобразование этого выражения посредством

$$f^*(\tau) = [f(\tau)]^2$$

(рис. 6.14, б). Тогда выражение (6.86) преобразуется к виду

$$\begin{aligned} f^*(\tau) &= [f(\tau)]^2 = [(1 + M \sin \Omega \tau) \sin \omega_0 \tau]^2 = \\ &= (1 + M \sin \Omega \tau)^2 \sin^2 \omega_0 \tau = \\ &= \frac{1}{2} \left\{ 1 + \frac{M^2}{2} + 2M \sin \Omega \tau - \frac{M^2}{2} \cos 2\Omega \tau - \right. \\ &\quad - \left(1 + \frac{M^2}{2} \right) \cos 2\omega_0 \tau - M \sin (2\omega_0 - \Omega) \tau + \\ &\quad + M \sin (2\omega_0 + \Omega) \tau + M^2 \cos 2(\omega_0 - \Omega) \tau + \\ &\quad \left. + M^2 \cos 2(\omega_0 + \Omega) \tau \right\}. \end{aligned}$$

Следовательно, в спектре детектированного колебания содержатся составляющие с частотами

$$2\omega_0; \quad 2\omega_0 + \Omega; \quad 2\omega_0 - \Omega; \quad 2(\omega_0 + \Omega); \quad 2(\omega_0 - \Omega)$$

и, кроме того, интересующая нас составляющая с низкой частотой Ω , которая искажается составляющей с низкой частотой 2Ω .

Таким образом, данный вид детектирования рекомендуется применять только при очень малой глубине модуляции, так как отношение амплитуд второй и первой гармоники равно $\frac{M}{4}$.

Полученные спектральные представления для случаев линейного и квадратичного детектирования проиллюстрированы рис. 6.15.

§ 15. СПЕКТР СУММЫ ПЕРИОДИЧЕСКИХ ФУНКЦИЙ. СПЕКТРЫ СУММЫ И РАЗНОСТИ ДВУХ СДВИНУТЫХ ВО ВРЕМЕНИ КОЛЕБАНИЙ

Пусть задана некоторая периодическая функция $f(\tau)$. Тогда ее ряд Фурье в комплексной форме согласно выражению (6.19) запишется в виде

$$f(\tau) \approx \sum_{-\infty}^{\infty} c_k e^{jk\tau},$$

где величина c_k согласно выражению (6.23) будет

$$c_k = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(\tau) e^{-jk\tau} d\tau.$$

Если есть k функции вида $f(\tau)$, то может быть поставлена задача определить спектр суммой этих периодических функций. Для этого необходимо знать действительную амплитуду k -й гармоники спектра суммы функций.

Поскольку преобразование Фурье линейно, для него справедлив принцип суперпозиции, а следовательно, комплексная амплитуда k -й гармоники спектра суммы функций

$$c_k = \frac{1}{2\pi} \int_{-\pi}^{\pi} d\tau \sum_i f_i(\tau) e^{-jk\tau} = \sum_i c_{ik},$$

откуда для действительных амплитуд можно записать

$$c_k = 2|c_k| = 2 \left| \sum_i c_{ik} \right|.$$

Для более четкого представления о характере решения поставленной задачи рассмотрим пример.

Пусть даны два синусоидальных колебания с комплексными амплитудами

$$2c_1 = c_1 e^{-j\varphi_1}; \quad 2c_2 = c_2 e^{-j\varphi_2}.$$

Тогда

$$\begin{aligned} 2c &= 2(c_1 + c_2) = c_1 e^{-j\varphi_1} + c_2 e^{-j\varphi_2} \\ c &= |c_1 e^{-j\varphi_1} + c_2 e^{-j\varphi_2}| = |c_1 \cos \varphi_1 + c_2 \cos \varphi_2 - \\ &\quad - j(c_1 \sin \varphi_1 + c_2 \sin \varphi_2)| = \\ &= \sqrt{(c_1 \cos \varphi_1 + c_2 \cos \varphi_2)^2 + (c_1 \sin \varphi_1 + c_2 \sin \varphi_2)^2} = \\ &= \sqrt{c_1^2 + c_2^2 + 2c_1 c_2 \cos(\varphi_1 - \varphi_2)}. \end{aligned}$$

Кроме рассмотренной выше задачи, можно поставить также задачу получить спектр функции, являющейся результатом сложения двух одинаковых, но сдвинутых во времени периодических функций.

Используя выражение (6.23), для исходной функции имеем

$$c_{1k} = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(\tau) e^{-jk\tau} d\tau.$$

Для функции, сдвинутой во времени, вида $f(\tau - t)$ имеем

$$c_{2k} = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(\tau - t) e^{-jk\tau} d\tau.$$

Обозначая $(\tau - t) = \tau_1$, получим

$$c_{2k} = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(\tau_1) e^{-jk(\tau_1 + t)} d\tau_1. \quad (6.90)$$

Поскольку $\tau = \frac{2\pi t}{T}$ и $\omega = \frac{2\pi}{T}$, то преобразование интеграла (6.90) дает

$$c_{2k} = e^{-jk\omega t} c_{1k}.$$

Теперь, если сложить функции $f(\tau)$ и $f(\tau - t)$, то комплексная амплитуда k -ой гармоники спектра их суммы

$$c_k = c_{1k} + c_{2k} = c_{1k}(1 + e^{-jk\omega t}),$$

а интересующая нас действительная амплитуда k -й гармоники спектра их суммы

$$a_k = 2 |c_k| = c_{1k} |1 - e^{-jk\omega t}| = 2c_{1k} \left| \cos \frac{k\omega t}{2} \right|.$$

Таким образом, для того, чтобы получить спектр суммы двух одинаковых функций, сдвинутых во времени на t , необходимо умножить амплитуду каждой гармоники на $2 \left| \cos \frac{k\omega t}{2} \right|$.

Если составить не сумму, а разность таких двух функций, сдвинутых во времени, то проделав выкладки аналогичные выкладкам для суммы функций, будем иметь

$$c_k = c_{1k} |1 - e^{-jk\omega t}| = 2c_{1k} \left| \sin \frac{k\omega t}{2} \right|.$$

Для получения спектра разности двух одинаковых функций, сдвинутых во времени на t , необходимо умножить амплитуду каждой гармоники на $2 \left| \sin \frac{k\omega t}{2} \right|$.

§ 16. СПЕКТРЫ НЕКОТОРЫХ СИГНАЛОВ

Структура частотного спектра сигнала полностью определяется модулем и аргументом спектральной плотности $F(\omega)$. Для отыскания спектров интересующих нас сигналов будем пользоваться выражением (6.33) для прямого преобразования Фурье.

Спектр единичного скачка. Если заданная функция $f(\tau)$ определена следующим образом:

$$\left. \begin{aligned} f(\tau) &= 1, & \tau \geq 0 \\ f(\tau) &= 0, & \tau < 0 \end{aligned} \right\}, \quad (6.91)$$

то она называется *единичным скачком*. Для такой функции интеграл вида $\int_0^{\infty} |f(\tau)| d\tau \rightarrow \infty$. Следовательно, выражения (6.33) и (6.34) для прямого и обратного преобразования Фурье нельзя применить непосредственно.

Однако, такое затруднение можно обойти, если искомую спектральную плотность функции $f(\tau)$, заданную выражением (6.91), представить как предел $F(\omega)$ для функции $f(\tau) e^{-c\tau}$, где c — положительное число, стремящееся к нулю

Таким образом, учитывая все сказанное выше, выражение (6.33) можно преобразовать следующим образом:

$$F(\omega) = \lim_{c \rightarrow 0} \int_{-\infty}^{\infty} f(\tau) e^{-c\tau} e^{-j\omega\tau} d\tau = \lim_{c \rightarrow 0} \int_{-\infty}^{\infty} f(\tau) e^{-\tau(j\omega+c)} d\tau =$$

$$= \lim_{c \rightarrow 0} \left[\int_{-\infty}^0 f(\tau) e^{-\tau(j\omega+c)} d\tau + \int_0^{\infty} f(\tau) e^{-\tau(j\omega+c)} d\tau \right].$$

В этом выражении первый интеграл, стоящий в квадратных скобках, равен 0 по определению функции $f(\tau)$.

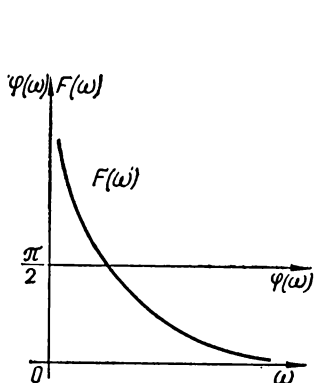


Рис. 6.16. Спектр единичного скачка.

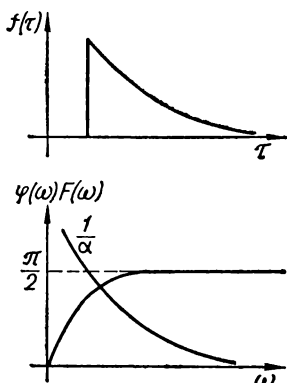


Рис. 6.17. Спектр экспоненциального импульса.

Тогда

$$F(\omega) = \lim_{c \rightarrow 0} \int_0^{\infty} f(\tau) e^{-\tau(j\omega+c)} d\tau =$$

$$= \lim_{c \rightarrow 0} \frac{1}{j\omega + c} = \frac{1}{j\omega} = \frac{1}{\omega} e^{-j\frac{\pi}{2}}.$$

Таким образом, значения модуля и аргумента спектральной плотности для единичного скачка будут

$$\left. \begin{aligned} F(\omega) &= \frac{1}{\omega}, \\ \varphi(\omega) &= \frac{\pi}{2} \end{aligned} \right\}. \quad (6.92)$$

Графически полученные результаты для $F(\omega)$ и $\varphi(\omega)$ изображены на рис. 6.16.

Спектр экспоненциального импульса. Экспоненциальный импульс (рис. 6.17, а) можно определить так:

$$\left. \begin{aligned} f(\tau) &= e^{-\alpha\tau}, & \tau \geq 0 \\ f(\tau) &= 0, & \tau < 0 \end{aligned} \right\},$$

где $\alpha > 0$.

Если в выражении (6.92) величину c заменить на α и не совершать предельного перехода, то получим следующее выражение для спектральной плотности экспоненциального импульса с единичной амплитудой

$$F(\omega) = \frac{1}{\alpha + j\omega} = \frac{1}{\sqrt{\alpha^2 + \omega^2}} e^{-j \operatorname{arctg} \frac{\omega}{\alpha}}. \quad (6.93)$$

Отсюда следует, что

$$\left. \begin{aligned} F(\omega) &= \frac{1}{\sqrt{\alpha^2 + \omega^2}} \\ \varphi(\omega) &= \operatorname{arctg} \frac{\omega}{\alpha} \end{aligned} \right\}.$$

Графически это выражение представлено на рис. 6.17, б.

Спектр прямоугольного импульса. Спектральное представление о прямоугольном импульсе легко получить, если записать прямоугольный импульс в виде разности двух скачков, сдвинутых на время $\hat{\tau}$ (рис. 6.18).

Тогда для первого скачка спектральная плотность

$$F_1(\omega) = \frac{A}{j\omega}.$$

Для второго скачка согласно теореме запаздывания спектральная плотность

$$F_2(\omega) = F_1(\omega) e^{-j\omega\hat{\tau}} = \frac{A}{j\omega} e^{-j\omega\hat{\tau}}.$$

А по теореме о спектре суммы спектральная плотность прямоугольного импульса

$$F(\omega) = F_1(\omega) + F_2(\omega) = \frac{A}{j\omega} (1 - e^{-j\omega\hat{\tau}}).$$

Учитывая, что

$$e^{-j\omega\hat{\tau}} = \cos \omega\hat{\tau} - j \sin \omega\hat{\tau},$$

определим модуль выражения для спектральной плотности прямоугольного импульса

$$F(\omega) = \frac{A}{\omega} \sqrt{(1 - \cos \omega \hat{\tau})^2 + \sin^2 \omega \hat{\tau}} =$$

$$= \left| \frac{2A}{\omega} \sin \frac{\omega \hat{\tau}}{2} \right| = A \hat{\tau} \frac{\left| \sin \frac{\omega \hat{\tau}}{2} \right|}{\frac{\omega \hat{\tau}}{2}}.$$

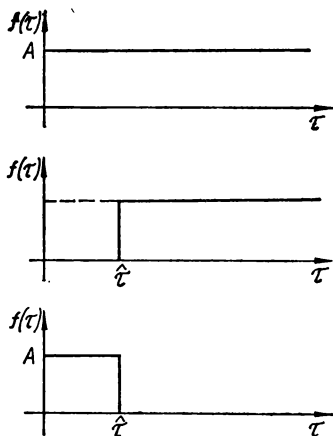


Рис. 6.18. Прямоугольный импульс.

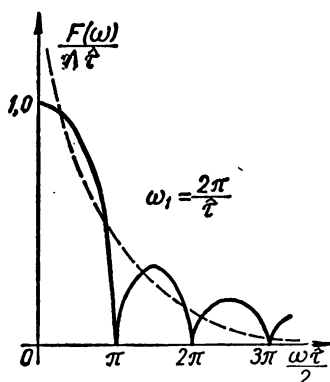


Рис. 6.19. Спектр прямоугольного импульса.

Далее, принимая во внимание, что

$$\lim_{\omega \rightarrow 0} \frac{\sin \frac{\omega \hat{\tau}}{2}}{\frac{\omega \hat{\tau}}{2}} = 1,$$

видим, что при нулевой частоте спектральная плотность прямоугольного импульса равна площади этого импульса, т. е. $F_0 = A \hat{\tau}$. Зависимость нормированного модуля спектральной плотности прямоугольного импульса $\frac{F(\omega)}{A \hat{\tau}}$ от без-

размерной переменной $\frac{\omega \hat{\tau}}{2}$ изображена на рис. 6.19 сплошной линией. Пунктирной линией на этом рисунке показан модуль спектра скачка с амплитудой A .

Появление нулей в спектре прямоугольного импульса является результатом взаимной компенсации тех гармоник, для которых сдвиг фаз равен целому числу 2π . А это справедливо для частот, отвечающих условию $\omega\tau = n2\pi$, где n — любое целое число. На частотах, отвечающих условию $\omega\tau = (2n - 1)\pi$, наоборот, вычитание спектров $F_1(\omega)$ и $F_2(\omega)$ приводит к удвоению амплитуд модуля спектральной плотности прямоугольного импульса.

При удлинении импульса расстояние между нулями $F(\omega)$ (рис. 6.19) сокращается, а начальное значение $F(\omega)$

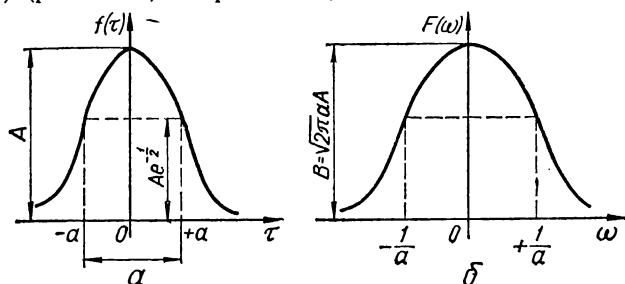


Рис. 6.20. Спектр колокольного импульса.

возрастает. В пределе при $\tau \rightarrow \infty$ спектр прямоугольного импульса вырождается в спектр скачка.

При укорочении импульса, наоборот, точка $\omega_1 = \frac{2\pi}{\tau}$, соответствующая первому нулю функции $F(\omega)$, удаляется от точки $\omega = 0$, а величина $F(\omega)$ уменьшается. И в пределе, при $\tau \rightarrow 0$, точка $\omega = \frac{2\pi}{\tau}$ удаляется в бесконечность, а спектральная плотность становится равномерной в полосе частот от 0 до ∞ .

Спектр «гауссова импульса» («колокольного» импульса). «Гауссов импульс» изображен на рис. 6.20, а, его можно определить как

$$f(\tau) = Ae^{-\frac{\tau^2}{2a^2}}.$$

Здесь постоянная a — половина длительности импульса, которая определяется на уровне $e^{-\frac{1}{2}} = \frac{1}{e^{\frac{1}{2}}} = 0,606$ от

амплитуды импульса. Таким образом, полная длительность импульса $t_{\Sigma} = 2a$.

Применяя прямое преобразование Фурье (6.33), получим

$$F(\omega) = A \int_{-\infty}^{\infty} e^{-\frac{\tau^2}{2a^2}} e^{-j\omega\tau} d\tau. \quad (6.95)$$

Для удобства вычисления интеграла дополним показатель степени в подынтегральной функции до квадрата суммы

$$\begin{aligned} -\left(\frac{\tau^2}{2a^2} + j\omega\tau\right) &= -\left[\left(\frac{\tau^2}{2a^2} + j\omega\tau + d^2\right) - d^2\right] = \\ &= -\left[\left(\frac{\tau}{\sqrt{2}a} + d\right)^2 - d^2\right], \end{aligned}$$

где « d » определяется из условия

$$j\omega\tau = 2 \frac{\tau}{\sqrt{2}a} d,$$

откуда

$$d = \frac{j\omega a}{\sqrt{2}}.$$

Следовательно, выражение (6.95) можно преобразовать к виду

$$F(\omega) = Ae^{d^2} \int_{-\infty}^{\infty} e^{-\left(\frac{\tau}{\sqrt{2}a} + d\right)^2} d\tau.$$

Переходя к новой переменной $x = \frac{\tau}{\sqrt{2}a} + d$, получим

$$F(\omega) = Ae^{d^2} \sqrt{2}a \int_{-\infty}^{\infty} e^{-x^2} dx.$$

Поскольку

$$\int_{-\infty}^{\infty} e^{-x^2} dx = \sqrt{\pi},$$

то

$$F(\omega) = A \sqrt{2\pi} a e^{-\frac{\omega^2}{2(1/a)^2}} = B e^{-\frac{\omega^2}{2b^2}}.$$

Здесь

$$B = aA \sqrt{2\pi}; \quad b = \frac{1}{a}.$$

График для спектральной плотности «гауссова импульса» изображен на рис. 6.20, б.

Таким образом, на основании сравнения выражения для «гауссова импульса» с выражением для его спектральной плотности можно заключить, что они обладают свойством симметрии, т. е. для получения одной из них по заданной другой достаточно заменить τ на ω . Тогда спектральная полоса, определяемая на уровне 0,606 от максимального значения,

$$\Delta\omega = 2b = 2 \frac{1}{a} = 2 \frac{2}{t_H} = \frac{4}{t_H}.$$

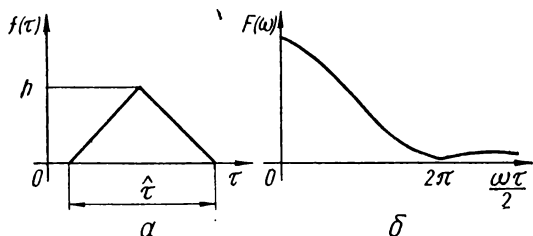


Рис. 6.21. Спектр треугольного импульса.

Спектр импульса треугольной формы. Пусть дан импульс треугольной формы с амплитудой h и основанием $\hat{\tau}$ (рис. 6.21, а). Тогда его можно определить следующим образом:

$$f(\tau) = \begin{cases} h \left(1 + \frac{2\tau}{\hat{\tau}}\right), & -\frac{\hat{\tau}}{2} < \tau < 0; \\ h \left(1 - \frac{2\tau}{\hat{\tau}}\right), & 0 < \tau < \frac{\hat{\tau}}{2}; \\ 0, & \frac{\hat{\tau}}{2} < \tau \text{ и } \tau < -\frac{\hat{\tau}}{2}. \end{cases}$$

Следовательно, спектральная плотность треугольного импульса

$$\begin{aligned} F(\omega) &= h \int_{-\frac{\hat{\tau}}{2}}^0 \left(1 + \frac{2\tau}{\hat{\tau}}\right) e^{-j\omega\tau} d\tau + h \int_0^{\frac{\hat{\tau}}{2}} \left(1 - \frac{2\tau}{\hat{\tau}}\right) e^{-j\omega\tau} d\tau = \\ &= \frac{1}{2} h \hat{\tau} \frac{1 - \cos \omega \frac{\hat{\tau}}{2}}{\frac{1}{2} \left(\omega \frac{\hat{\tau}}{2}\right)^2}; \end{aligned} \quad (6.96)$$

Графически это изображено на рис. 6.21, б. Нетрудно убедиться, что при $\omega \frac{\tau}{2} \rightarrow 0$ отношение в выражении (6.96) стремится к единице.

Спектр косинусоидального импульса. Для косинусоидального импульса с длительностью $\hat{\tau}$ и амплитудой h спектральная плотность определится из выражения

$$F(\omega) = h \int_{-\frac{\hat{\tau}}{2}}^{\frac{\hat{\tau}}{2}} e^{-j\omega\tau} \cos \pi \frac{\tau}{\hat{\tau}} d\tau = \frac{2}{\pi} h \hat{\tau} \frac{\cos \omega \frac{\hat{\tau}}{2}}{1 - \left(\frac{2}{\pi} \omega \frac{\hat{\tau}}{2} \right)^2},$$

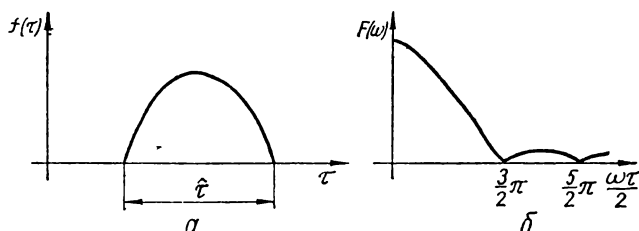


Рис. 6.22. Спектр косинусоидального импульса.

графически спектральная плотность для косинусоидального импульса показана на рис. 6.22.

Спектр отрезка синусоиды. Отрезок синусоиды, содержащий целое число периодов T , можно определить следующим образом:

$$f(\tau) = \begin{cases} 0, & \tau < -\frac{1}{2} nT, \\ \cos \omega_0 \tau, & -\frac{1}{2} nT < \tau < \frac{1}{2} nT, \\ 0, & \tau > \frac{1}{2} nT. \end{cases}$$

Тогда спектр отрезка синусоиды будет

$$\begin{aligned} F(\omega) &= \int_{-\frac{1}{2} nT}^{\frac{1}{2} nT} \cos \omega_0 \tau e^{-j\omega\tau} d\tau = \\ &= \frac{e^{-j\omega\tau}}{\omega_0^2 - \omega^2} \cdot (-j\omega \cos \omega_0 \tau + \omega_0 \sin \omega_0 \tau) \Big|_{-\frac{1}{2} nT}^{\frac{1}{2} nT}, \end{aligned}$$

а так как

$$T = \frac{2\pi}{\omega_0}, \quad \frac{1}{2} n \omega_0 T = n\pi,$$

то выражение для спектра отрезка синусоиды можно переписать в виде

$$F(\omega) = (-1)^n \frac{2\omega}{\omega_0^2 - \omega^2} \sin n\pi \frac{\omega}{\omega_0},$$

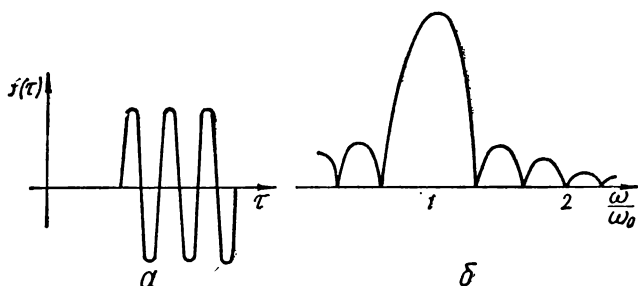


Рис. 6.23. Спектр отрезка синусоиды.

откуда модуль спектра отрезка синусоиды

$$|F(\omega)| = \frac{2}{\omega_0} \frac{\frac{\omega}{\omega_0}}{\left|1 - \left(\frac{\omega}{\omega_0}\right)^2\right|} \left|\sin n\pi \frac{\omega}{\omega_0}\right|.$$

При значении $\omega = \omega_0$ получается максимум. Раскрывая неопределенность, находим

$$|F(\omega_0)| = \frac{n\pi}{\omega_0}.$$

Графически полученные выражения для спектра отрезка синусоиды интерпретируются рис. 6.23. При увеличении отрезка синусоиды его сплошной спектр вырождается в пределе в одну спектральную линию (рис. 6.24).

Если рассмотренные импульсы за отрезок испытания могут периодически повторяться, то нетрудно установить связь между спектром одиночного импульса и спектром периодической последовательности таких же импульсов.

Так как одиночный импульс представляет собой непериодическую функцию, то естественно, что его спектр сплошной. Если же импульсы какой угодно формы будут

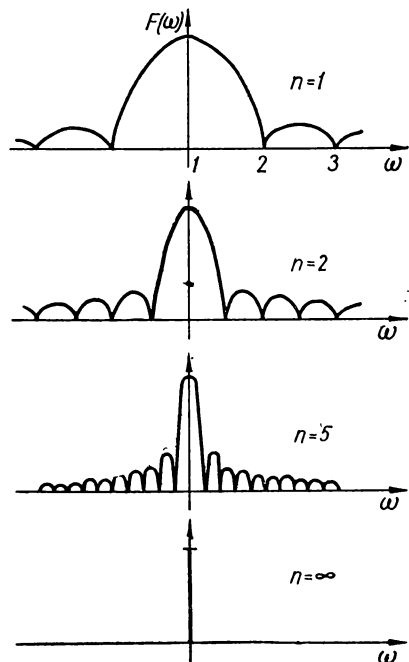
периодически повторяться, то мы получим периодическую функцию, обладающую дискретным гармоническим спектром.

Если спектр одиночного импульса определить как

$$F_0(\omega) \int_{-\infty}^{\infty} f(\tau) e^{-i\omega\tau} d\tau, \quad (6.97)$$

и импульс, описываемый функцией $f(\tau)$, повторять через промежутки времени T n раз, то получим периодическую

функцию с периодом T . Спектр такой функции можно записать следующим образом:



$$c_k = \frac{1}{T} \int_{-\frac{T}{2}}^{\frac{T}{2}} \times \\ \times e^{-i\pi k \frac{\tau}{T}} f(\tau) d\tau. \quad (6.98)$$

Сравнивая выражения (6.97) и (6.98), видим, что при значениях аргумента

$$\omega = 2\pi \frac{k}{T} = k\omega_1,$$

где ω_1 — круговая частота повторения, значения непрерывной функции $F_0(\omega)$ с точностью до постоянного множителя $\frac{1}{T}$ совпадает со значениями дискретной функции c_k .

Рис. 6.24. Вырождение спектра отрезка синусоиды.

Таким образом, с указанной точностью спектр одиночного импульса является огибающей спектра периодической последовательности таких импульсов. Или, другими словами, совокупность точек Tc_k дискретного спектра периодической последовательности импульсов аппроксимирует спектр одиночного импульса из этой последовательности.

Контрольные вопросы

1. В чем состоит сущность процесса переноса спектра?
2. Что такое детектирование? Какие различают типы детектирования?
3. Какие операции необходимо проделать для получения спектра суммы и разности двух сдвинутых во времени колебаний?
4. Чему равны значения модуля и аргумента спектральной плотности для одиночного скачка, экспоненциального, прямоугольного, треугольного, гауссова и косинусоидального импульсов?
5. Чем определяется спектр отрезка синусоиды?
6. Какая связь между спектром одиночного импульса и периодической последовательностью таких же импульсов?

§ 1. ОБЩИЕ ПОЛОЖЕНИЯ

Теория информации — одна из молодых естественно-научных дисциплин, составляющих теоретический фундамент кибернетики. Стремительно развиваясь, теория информации в настоящее время распространяется на все новые области исследований.

Применение теоретико-информационных идей уже теперь оказалось весьма плодотворным в ряде областей науки. Часто именно теория информации открывает дорогу применению математического аппарата там, где это до сих пор удавалось лишь с трудом (например, в некоторых областях биологии). В последние годы все чаще применяются методы теории информации при исследовании психических процессов, при изучении отдельных сторон социальных явлений и т. п. Определенные успехи в этом отношении уже достигнуты в лингвистике, педагогике и других науках. Информационный подход становится мощным источником эвристических методов современного знания.

Идеи, лежащие в основе теории информации, можно понять лишь из рассмотрения либо систем связи в самом широком смысле, через которые проходит информация, либо из рассмотрения статистических систем, применяющихся для хранения (накопления информации). Следует отметить, что с точки зрения математики между этими подходами нет существенной разницы.

Определяющей чертой теории информации, отделяющей ее от других ветвей статистической теории связи, является использование специфического метода измерения информации.

В статье, которая появилась в 1948 году, Шеннон дал определение разумной меры информации и применил ее для доказательства замечательных теорем, которые дали критерий оценки и сравнения различных систем связи. Однако не только инженеры связи обратили внимание на эту статью. Математики нашли в этой статье золотое дно

статистических и комбинаторных проблем. Физики заинтересовались новой интерпретацией энтропии как меры информации. Психологи нашли, что новая мера информации дает удобную количественную оценку трудности некоторых экспериментальных задач. Появились работы, описывающие применение новой теории в лингвистике, музыке, теории игр.

Многие выводы классической теории информации Шеннона получили важные приложения в задачах проектирования систем любой природы. Основой для таких приложений является лишь более широкое понимание системы связи, предложенной Шенноном.

§ 2. МОДЕЛЬ СИСТЕМЫ СВЯЗИ ШЕННОНА

Центральная задача классической теории информации состоит в том, чтобы воспроизвести в некоторой точке с заданной степенью точности сообщение, выбранное в другой точке. Предположим, что имеется множество n возможных событий, вероятности которых равны

$$p_1, p_2, \dots, p_k, \dots, p_n.$$

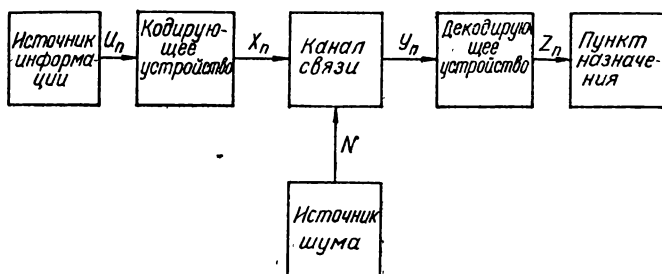


Рис. 7.1. Блок-схема информационного канала.

Считается, что система связи работает удовлетворительно, если при любом конкретном выборе сообщения из множества всех возможных сообщений ошибки нет. Это условие определяет необходимые свойства каналов, а также приемных и передающих устройств.

Обычно в теории связи операцию преобразования сообщения в физические сигналы, пригодные для передачи, называют *кодированием*, а операции восстановления сообщения по сигналу — *декодированием*.

На рис. 7.1 приведена блок-схема модели системы связи, информационного канала связи, предложенная Шенноном.

Эту модель в работах по теории информации почти всегда берут за основу исследований. Модель содержит следующие шесть частей:

1. Источник информации, который производит сообщение или последовательность сообщений.

2. Кодировущее устройство, которое преобразует сообщения, вырабатывая сигналы, пригодные для передачи по каналу связи,— передатчик.

3. Канал передачи сигналов от передатчика к приемнику.

4. Декодировущее устройство, выполняющее функции восстановления исходного сообщения по сигналу,— приемник.

5. Пункт назначения — объект, которому предназначается сообщение.

6. Источник шума.

В самом начале развития теории информации было получено несколько важных результатов, определивших дальнейшее развитие теории информации.

Наиболее важный из этих результатов известен под названием теоремы Котельникова или теоремы о выборке Найквиста (см. гл. 6). В 20-е годы Хартли установил, что количество информации, которое можно передать по каналу с шириной f за время T , определяется формулой

$$I = 2fT \log k, \quad (7.1)$$

где k — число различных значений сигнала. Таким образом, для передачи заданного объема информации необходимо иметь определенное значение произведения ширины полосы на интервал времени.

Впоследствии Габор установил «принцип неопределенности», основанный на обратном отношении между продолжительностью сигнала и эффективной шириной полосы спектра.

Для сигналов с ограниченной, но переменной шириной полосы Δf можно измерить лишь тот сигнал, который обладает продолжительностью не менее Δt , такого, что

$$\Delta f \Delta t \approx \text{const} \approx 1. \quad (7.2)$$

Габор сравнил это свойство с принципом неопределенности Гейзенберга и показал, что некоторые математические идеи квантовой теории можно применить и к изучению сигналов. Естественно, что применима не сама квантовая теория, а только некоторые ее математические методы.

Во многих практических случаях информацию приходится передавать при воздействии *шума* (*помехи*). Влияние шума на передачу информации исследовано в работах Шеннона, Фано, Котельникова и др.

Шеннон обобщил результаты Хартли о количестве информации. В своих работах Шеннон предполагал, что есть дискретный источник сообщения, генерирующий информацию в виде последовательности знаков. Последовательные знаки из источника выбираются в соответствии с определенными вероятностями, вообще говоря, зависящими от предыдущего выбора (выборов). Источник является входом дискретного передающего устройства, на вход которого поступает одна последовательность знаков, а на выходе образуется другая. При этих условиях приемник может лишь «догадываться» о том, какой именно знак послан передатчиком. С точки зрения приемника источник информации принадлежит к классу устройств, определяемых математикой как случайный. Сообщения, выбираемые из такого источника, представляют собой случайные процессы.

Для исследования систем связи с подобным источником сообщения можно применить аппарат дискретных марковских процессов, дискретных потому, что рассматриваются дискретные последовательности знаков. При конечном числе состояний кодирующего устройства его выходной сигнал в данный момент времени зависит как от текущего состояния, так и от входного символа, поступающего в этот момент. Следующее состояние представляет собой некоторую другую функцию от этих переменных. Таким образом, кодирующее устройство можно описать уравнением

$$x_n = \Phi(u_n; a_n) \text{ и } a_{n+1} = \Psi(u_n; a_n).$$

Здесь

u_n — n -ый входной символ; a_n — состояние преобразователя в момент, когда поступил сигнал u_n ; x_n — выходной символ, получаемый в момент, когда сигнал u_n поступил в преобразователь с состоянием a_n .

§ 3. ИНФОРМАЦИЯ

Поскольку в действительности информация бывает чрезвычайно разнообразной, целесообразно рассмотреть меры информации с абстрактной точки зрения.

До сих пор понятие «информация» мы интуитивно толковали как некоторую совокупность сведений, определяющих меру наших знаний о тех или иных событиях, явлениях,

фактах. Однако такое представление ничего не дает для построения количественной теории информации, которую можно было бы использовать для решения инженерных задач.

Очевидно, что всякую информацию мы получаем в результате того или иного опыта. Можно сказать, что до опыта мы не можем ответить однозначно на интересующий нас вопрос во всех случаях. Мы можем высказать лишь ряд предположений. Таким образом, до опыта есть большая или меньшая неопределенность в интересующей нас ситуации или в исходе тех или иных событий. После опыта, т. е. получения информации, можно ответить либо однозначно, либо по крайней мере количество возможных ответов уменьшится и, следовательно, уменьшится существовавшая ранее неопределенность.

Рассмотрим несколько примеров, иллюстрирующих сказанное выше. Если подброшенная вверх монета может с одинаковым успехом упасть как на лицевую сторону, так и на обратную, то говорят, что вероятность падения монеты на каждую из сторон равна $\frac{1}{2}$. Известно, что вероятность одновременного наступления двух независимых событий равна произведению вероятностей каждого из этих событий. Подсчитаем вероятность того, что при подбрасывании двух костей сумма выпавших очков составит число семь. Сумма выпавших очков равна семи в следующих взаимно исключающих друг друга случаях:

1) на одной кости выпадает единица, а на другой — шестерка;

2) на одной — двойка, на другой — пятерка;

3) на одной — тройка, на другой — четверка;

4) на одной — четверка, на другой — тройка;

5) на одной пятерка, на другой — двойка;

6) на одной — шестерка, на другой — единица.

Вероятность каждой из 6 указанных комбинаций равна

$$\frac{1}{6} \times \frac{1}{6} = \frac{1}{36}.$$

Итак, сумма выпавших очков будет равна семи в шести взаимноисключающих случаях, вероятность каждого из которых равна $\frac{1}{36}$. Следовательно, вероятность того, что при подбрасывании двух костей сумма выпавших очков окажется равной семи, равна

$$\frac{1}{36} + \frac{1}{36} + \frac{1}{36} + \frac{1}{36} + \frac{1}{36} + \frac{1}{36} = \frac{6}{36} = \frac{1}{6}.$$

Рассмотрим еще один пример. Пусть по каналу связи передаются два равновероятных сообщения в виде двух последовательностей импульсов. Последовательность, соответствующая сообщению 0, состоит из трех отрицательных импульсов, а соответствующая сообщению 1 — из трех положительных импульсов. Полярности трех последовательных выходных импульсов таковы: (+ — +). Вычисленные вероятности сообщений после приема каждого из трех последовательных импульсов сведены в таблицу 7.1.

Таблица 7.1

Номера сообщений	Последовательность на входе	Вероятности сообщений			
		Первоначальные	После +	После —	После +
0	— — —	$\frac{1}{2}$	p	$\frac{1}{2}$	p
1	+ + +	$\frac{1}{2}$	$1 - p$	$\frac{1}{2}$	$1 - p$

Символом p обозначена вероятность того, что полярность любого выходного импульса отличается от полярности входного импульса.

Вероятность выходной последовательности (+ — +) равна $p(1 - p)^2$, если на вход поступает последовательность (+ + +) и $p^2(1 - p)$, если на вход поступает последовательность (— — —).

Поскольку обе возможные входные последовательности априори равновероятны, то после приема (+ — +) вероятность того, что на входе была последовательность (+ + +), равна

$$\frac{[p(1 - p^2)]}{[p(1 - p)^2 + p^2(1 - p)]} = 1 - p.$$

Если величина $p < \frac{1}{2}$, то появление первого импульса увеличивает вероятность сообщения 1 и уменьшает вероятность сообщения 0. Появление второго импульса имеет противоположный эффект. Он снова придает вероятностям их первоначальные значения.

Появление третьего импульса снова увеличивает вероятность сообщения 1 и уменьшает вероятность сообщения 0 точно так же, как и появление первого импульса.

Таким образом, если фактически переданным является сообщение 1, то прием первого импульса увеличивает апостериорную вероятность этого сообщения, прием второго импульса уменьшает ее, прием последнего импульса снова увеличивает ее. Следовательно, эти изменения вероятностей выражают тот вклад, который вносят принятые импульсы при идентификации входного сообщения.

Таким образом, для оценки количества получаемой информации необходимо найти меру неопределенности той или иной ситуации.

Уменьшение неопределенности в результате опыта можно принять за наиболее общую меру количества получаемой информации. В этом смысле говорят, что информация обратна неопределенности. Для введения абстрактной оценки количества информации целесообразно рассмотреть определение меры информации с абстрактной точки зрения. Вместо того, чтобы говорить о сообщениях и символах, будем говорить о точках некоторого абстрактного пространства и о множествах точек, получающихся в результате задания распределения вероятностей в соответствующих пространствах.

Вначале рассмотрим дискретные пространства для абстрактного обоснования теории дискретных систем связи.

§ 4. ИЗМЕРЕНИЕ ВЗАИМНОЙ ИНФОРМАЦИИ

Рассмотрим два дискретных пространства X и Y . Обозначим через x_k некоторую точку пространства X , а через y_i некоторую точку пространства Y .

Множество X порождается заданием на пространстве X распределения вероятностей $p(x)$, приписывающего вероятность $p(x_k)$ каждой точке пространства X . Множество Y или любое другое может быть порождено подобным же образом. Так множество XY порождается заданием совместного распределения вероятностей $p(x, y)$ на произведении пространств.

Распределение вероятностей $p(x)$ и $p(y)$ выражается через $p(x, y)$ следующим образом:

$$p(x) \equiv \sum_y p(x, y),$$

$$p(y) \equiv \sum_x p(x, y),$$

где суммирование производится по всему пространству X или Y .

Условное распределение вероятностей определяется так:

$$p(y/x) \equiv \frac{p(x, y)}{p(x)},$$

$$p(x/y) \equiv \frac{p(x, y)}{p(y)}.$$

Ясно, что для произведения множеств XY

$$\sum_{XY} p(x, y) = 1.$$

Итак, пусть $p(x, y)$ — распределение вероятностей на произведении пространств XY . Требуется ввести измерения информации, содержащейся в y_i , относительно x_k . Пусть, кроме того, x_k и y_i трактуются соответственно как некоторые события на входе и выходе одного из блоков рис. 7.1. Тогда можно определить меру количества информации, переданной через этот блок.

Ранее приведенные примеры показывают, что информация относительно x_k , содержащаяся в y_i , сводится к изменению вероятности x_k от ее априорного значения $p(x_k)$ к ее апостериорному значению $p(x_k/y_i)$. Количество информации, содержащееся в событии y_i относительно появления события x_k , определяется как

$$I(x_k; y_i) \equiv \log \frac{p(x_k/y_i)}{p(x_k)}. \quad (7.4)$$

Основание логарифма, используемого в этом определении, фиксирует величину единицы измерения информации. Чаще всего используют основание 2. В этих случаях единицу информации относительно x_k получают, если вероятность x_k увеличивается в 2 раза.

Часто (так как математически это более удобно) вместо основания 2 используют основание натурального логарифма e . Соответствующая единица информации получается, если вероятность события увеличивается в e раз. Очевидно, что увеличение вероятности в 10 раз дает единицу информации, получающуюся при использовании десятичных логарифмов. Для обозначения этих единиц обычно используют наименование бит (сокращение слов *binary digit*), нат (сокращение слов *natural unit*), хартли (в честь Л. Хартли, одного из основоположников теории связи).

Наиболее распространенным наименованием единицы информации является бит, поскольку применение в качестве

основания логарифмов 2 имеет в теории информации ряд преимуществ.

Например,

$$\log_2 2 = 1,$$

на языке теории информации означает, что произошло одно из двух равновероятных событий.

§ 5. ИЗМЕРЕНИЕ КОЛИЧЕСТВА СОБСТВЕННОЙ ИНФОРМАЦИИ

Рассмотрим произведение двух множеств XU . Пусть пара x_k, y_i — точка этого множества.

Так как

$$p(x_k/y_i) \leq 1 \text{ и } p(y_i/x_k) \leq 1, \quad (7.5)$$

то взаимная информация между x_k и y_i , определенная в § 4,

$$I(x_k, y_i) = \log \frac{p(x_k/y_i)}{p(x_k)} = \log \frac{p(y_i/x_k)}{p(y_i)}. \quad (7.6)$$

Или, иначе говоря,

$$I(x_k, y_i) \leq \begin{cases} \log \frac{1}{p(x_k)} \equiv I(x_k), \\ \log \frac{1}{p(y_i)} \equiv I(y_i), \end{cases} \quad (7.7)$$

где знак равенства справедлив лишь в том случае, когда имеет место знак равенства в выражении (7.5).

О величинах

$$I(x_k) \equiv -\log p(x_k)$$

и

$$I(y_i) \equiv -\log p(y_i)$$

говорят, что они соответственно являются количеством собственной информации для x_k и y_i . Таким образом, информация в каком-либо событии измеряется логарифмом величин обратной вероятности его появления.

§ 6. СВОЙСТВА КОЛИЧЕСТВА ИНФОРМАЦИИ

Количество информации, определенное в § 4, обладает очень важным свойством симметрии по отношению к x_k и y_i .

Эту симметрию легко можно обнаружить, умножая числитель и знаменатель выражения (7.4) на $p(y_i)$

$$I(x_k, y_i) = \log \frac{p(x_k/y_i) p(y_i)}{p(x_k) p(y_i)} = \log \frac{p(x_k, y_i)}{p(x_k) p(y_i)},$$

откуда следует, что

$$I(x_k, y_i) = I(y_i, x_k).$$

Или, иначе говоря, информация, содержащаяся в y_i относительно x_k , равна информации, содержащейся в x_k относительно y_i . Именно поэтому введенная величина (7.4) и определена как мера взаимной информации между x_k и y_i .

Рассмотрим теперь произведение множеств XYZ . Пусть некоторая точка этого ансамбля появляется с вероятностью $p(x_k, y_i, z_j)$.

Взаимная информация между x_k и y_i при заданном z_j в соответствии с выражениями (7.4) и (7.6) определяется как

$$I(x_k, y_i/z_j) \equiv \log \frac{p(x_k/y_i/z_j)}{p(x_k/z_j)} = \log \frac{p(x_k, y_i/z_j)}{p(x_k/z_j) p(y_i/z_j)}. \quad (7.8)$$

Иначе говоря, *условная взаимная информация* определяется точно так же, как и в выражении (7.4), только априорные и апостериорные вероятности должны быть взяты при одном и том же условии.

Пусть далее $x_k y_i z_j$ — элемент множества XYZ и пусть

$$I(x_k, y_i z_j) \equiv \log \frac{p(x_k/y_i z_j)}{p(x_k)} \quad (7.9)$$

— взаимная информация между $y_i z_j$ и x_k . Теперь, чтобы выразить эту взаимную информацию через информацию, содержащуюся в символе y_i относительно сообщения x_k , и информацию, содержащуюся в символе z_j относительно x_k , надо умножить числитель и знаменатель выражения (7.9) на $p(x_k/y_i)$. Получим

$$\begin{aligned} I(x_k, y_i z_j) &= \log \frac{p(x_k/y_i)}{p(x_k)} + \\ &+ \log \frac{p(x_k/y_i z_j)}{p(x_k/y_i)} = I(x_k, y_i) + I(x_k, z_j/y_i). \end{aligned} \quad (7.10)$$

Иначе говоря, информация, содержащаяся в паре $y_i z_j$ относительно x_k , равна сумме информации, содержащейся в y_i относительно x_k , и информации, содержащейся в z_j относительно x_k , при условии, что значение y_i известно. Этот результат выражает *свойство аддитивности количества информации*.

Контрольные вопросы

1. Сформулируйте центральную задачу классической теории информации.
2. Что такое кодирование и декодирование?
3. В чем заключается «принцип неопределенности» Габора?
4. Дайте определение информации. В каких единицах она измеряется?
5. Перечислите основные свойства количества информации.

§ 7. ЭНТРОПИЯ

В предыдущих параграфах было показано, что собственную информацию сообщения можно трактовать как количество информации, требуемое для однозначного определения этого сообщения.

Таким образом, среднее значение информации есть то количество информации, которое в среднем должно иметься в нашем распоряжении для того, чтобы выделить любое сообщение из некоторого множества X . Сказанное можно аналитически записать следующим образом:

$$I_{\text{ср}}(X) \equiv - \sum_x p(x) \log p(x) \equiv H(X). \quad (7.11)$$

Здесь $p(x)$ — вероятность сообщения x из множества X .

Если логарифм берется при основании 2, то $H(X)$ измеряется в битах на один знак. Функция $H(X)$ имеет тот же вид, который был получен в статистической механике для термодинамической величины, известной как средняя энтропия канонической системы, если $p(x)$ рассматривается как вероятность одного из возможных состояний системы.

Аналогично в теории информации для функции $H(X)$, определяемой формулой (7.11), применяется термин «энтропия».

Можно сказать, что энтропия любой системы служит мерой «беспорядка» в ней, тогда как информация является мерой порядка в системе — таково толкование знака минус в формуле (7.11).

Энтропия множества символов равна количеству информации, которое в среднем может содержать какой-либо символ. Таким образом, энтропия является мерой эффективности использования различных символов. В этой связи важно отметить основные свойства функции $H(X)$. Свойства энтропии могут быть установлены посредством трех теорем, которые будут доказаны ниже.

Теорема 1. Энтропия $H(X)$ удовлетворяет неравенству

$$H(X) \leq \log M, \quad (7.12)$$

где M — число точек пространства X . Знак равенства есть тогда и только тогда, когда $p(x)$ равна одному и тому же значению $\frac{1}{M}$ для всех точек пространства X .

Доказательство. Прежде всего сделаем одно замечание. Пусть есть прямая $u = a - 1$ и линия, определяемая функцией $u = \ln a$, наклон которой является монотонно убывающей функцией a . Тогда, поскольку эти линии касаются лишь в точке $a = 1$, справедливо следующее выражение:

$$\ln a \leq a - 1. \quad (7.13)$$

На основании этого замечания строится доказательство теоремы.

Рассмотрим разность

$$\begin{aligned} H(X) - \log M &= \sum_x p(x) \log \frac{1}{p(x)} - \sum_x p(x) \log M = \\ &= \sum_x p(x) \log \frac{1}{Mp(x)}. \end{aligned} \quad (7.14)$$

Подставляя правую часть неравенства (7.13) в соответствующий член правой части выражения (7.14), получаем

$$H(X) - \log M \leq \sum_x \left[\frac{1}{M} - p(x) \right] \log e = 0. \quad (7.15)$$

Знак равенства справедлив тогда и только тогда, когда

$$a = \frac{1}{Mp(x)} = 1,$$

поскольку только при этом значении есть знак равенства в формуле (7.13).

Основной смысл свойства функции $H(X)$, выраженного формулой (7.12), можно сформулировать следующим образом.

Для любого заданного алфавита символов количество информации, которое в среднем может содержаться в одном символе, достигает максимума, когда все символы равновероятны. В теории информации это максимальное значение называют информационной пропускной способностью алфавита. Информационная пропускная способность алфавита измеряется логарифмом числа символов в алфавите.

Рассмотрим далее множество XY . Для него среднее значение условной собственной информации

$$I(Y/X) \equiv - \sum_{xy} p(x, y) \log p(y/x) \equiv H(Y/X). \quad (7.16)$$

Выражение (7.16) определяет величину, называемую условной энтропией Y при заданном X .

Следует отметить, что энтропия

$$H(XY) \equiv - \sum_{xy} p(x, y) \log p(x, y) \quad (7.17)$$

связана с функциями $H(X)$ и $H(Y/X)$ соотношением

$$H(XY) = H(X) + H(Y/X). \quad (7.18)$$

Это есть общее правило для определения энтропии сложного опыта. Его легко проверить, поскольку соотношение (7.18) сводится к простому осреднению выражения

$$I(xy) = I(x) + I(y/x)$$

по множеству XY . Очевидно, что в случае двух взаимно независимых символов правило сложения энтропий упрощается

$$H(XY) = H(X) + H(Y).$$

Известно, что вероятность совместного события $p(x, y) = p(x) p(y_i/x_i)$, и тогда формула (7.16) для вычисления условной энтропии будет иметь вид:

$$H(Y/X) = - \sum_{xy} p(x_i) p(y_i/x_i) \log p(y_i/x_i).$$

Теорема 2. Для заданного пространства XY условная энтропия $H(Y/X)$ удовлетворяет неравенству

$$H(Y/X) \leq H(Y), \quad (7.19)$$

в котором знак равенства есть тогда и только тогда, когда y статистически не зависит от x , т. е. когда:

$$p(y/x) = p(y).$$

Эта теорема доказывается также на основании замечания к теореме 1.

Рассмотрим разность

$$H(Y/X) - H(Y) = \sum_{xy} p(x, y) \log \frac{p(y)}{p(y/x)}. \quad (7.20)$$

Подставляя правую часть выражения (7.12) вместо логарифма в выражение (7.20), получим:

$$H(Y/X) - H(Y) \leq \sum_{xy} p(x, y) \left[\frac{p(y)}{p(y/x)} - 1 \right] \log e = 0.$$

Причем знак равенства будет тогда и только тогда, когда

$$a = \frac{p(x)}{p(y/x)} = 1.$$

Теорема 3. Для заданного пространства XYZ условные энтропии $H(Z/X)$ и $H(Z/Y)$ удовлетворяют неравенству

$$H(Z/X) \leq H(Z/Y), \quad (7.21)$$

в котором знак равенства есть тогда и только тогда, когда Z статистически не зависит от x при любом заданном y , т. е.

$$p(z/yx) = p(z/y).$$

Доказательство этой теоремы аналогично доказательству теоремы 2.

§ 8. УСЛОВНАЯ СРЕДНЯЯ ВЗАИМНАЯ ИНФОРМАЦИЯ

Кроме рассмотренных в предыдущих параграфах величин для количественной оценки информации и каналов связи, важное значение имеет понятие об *условном среднем значении взаимной информации*.

Условное среднее значение взаимной информации для пространства XY определяется следующим равенством:

$$I(X; y_i) \equiv \sum_x p(X/y_i) \log \frac{p(x/y_i)}{p(x)}. \quad (7.22)$$

Справедлива следующая **теорема** для заданного пространства: условное среднее значение взаимной информации $I(X; y_i)$ удовлетворяет неравенству

$$I(X; y_i) \geq 0, \quad (7.23)$$

в котором знак равенства есть тогда и только тогда, когда

$$p(x/y_i) = p(x), \quad (7.24)$$

т. е. когда x статистически не зависит от y .

Доказательство. Эту теорему легко доказать, используя выражение (7.13)

$$\begin{aligned} -I(X; y_i) &= \sum_x p(x/y_i) \log \frac{p(x)}{p(x/y_i)} \leq \\ &\leq \sum_x [p(x) - p(x/y_i)] \log e = 0. \end{aligned}$$

Знак равенства есть тогда и только тогда, когда переменная a выражения (7.13) равна 1, т. е. когда удовлетворяется равенство (7.24). Выражение (7.23) отражает свойство, которое очень важно в случае, когда y_i является символом, принятым на выходе канала с шумом (см. рис. 7.1), а x — какое-либо из возможных сообщений. Тогда выражение (7.23) утверждает, что средняя информация, содержащаяся в принятом символе, относительно переданного сообщения всегда неотрицательна.

В этой связи следует заметить, что полученные результаты подтверждают представление о том, что информация на приемном конце канала никогда не будет отрицательна, если она надлежащим образом оценивается в приемнике.

Контрольные вопросы

1. Что такое энтропия?
2. В чем заключаются основные свойства энтропии?
3. Что такое информационная пропускная способность алфавита?
4. Чем определяется условное среднее значение взаимной информации?

§ 9. ДИСКРЕТНЫЕ ИСТОЧНИКИ СООБЩЕНИЙ. ЭРГОДИЧЕСКИЕ ИСТОЧНИКИ СООБЩЕНИЙ

Дискретный источник сообщений представляет собой объект, состояние которого определяется некоторым физическим процессом, протекающим, как правило, во времени. Информация, генерируемая дискретным источником сообщений, представляет собой последовательность знаков. В источнике последовательные знаки выбираются в соответствии с определенными вероятностями. Источник является входом дискретного передающего устройства. На вход этого устройства поступает одна последовательность знаков, а на выходе образуется другая. Сообщениями, генерируемыми дискретными источниками, могут быть: буквы или цифры; их совокупности, имеющие определенное смысловое содержание; типовые команды или распоряжения; извещения о возможных дискретных состояниях объектов и т. д.

Будем считать, что число различных сообщений конечно, обозначим их символами $x_1, x_2, \dots, x_k, \dots, x_n$. Различными символами могут обозначаться как элементарные сообщения типа «да», «нет», цифры 0, 1, ..., так и более сложные, такие, например, как стандартные тексты, команды и пр.

Считается, что источник дискретных сообщений вырабатывает некоторую последовательность символов x_k , причем порядок следования этих символов случаен и характеризуется некоторой совокупностью вероятностей. С точки зрения теории информации важно установить, какое в среднем количество информации создается таким источником на один символ или в единицу времени. Для этого необходимо выяснить, какие вероятностные показатели могут охарактеризовать рассматриваемый источник. Очевидно, что одних вероятностей появления отдельных символов для описания дискретного источника недостаточно. Рассмотрим следующий пример. Допустим, что передается последовательность из символов x_1, x_2, x_3 и x_4 со следующими вероятностями:

$$\begin{aligned} p(x_1) &= 0,5; & p(x_2) &= 0,25; \\ p(x_3) &= p(x_4) &= 0,125. \end{aligned}$$

При этом известно, что символ x_4 всегда передается после символа x_3 . Ясно, что хотя вероятности передачи символов x_3 и x_4 одинаковы, символ x_4 не несет никакой информации, поскольку, получив символ x_3 , мы достоверно знаем, что следующим будет символ x_4 .

Таким образом, необходимы более детальные вероятностные характеристики источника и, в частности, учет корреляционных связей между передаваемыми символами.

На практике, как правило, корреляционные связи распространяются на конечное число символов. Источники, обладающие таким свойством, называются *эргодическими*. В эргодическом источнике для символов, отстоящих достаточно далеко друг от друга, корреляционная связь отсутствует. Эргодические последовательности обладают свойствами, аналогичными свойствам эргодических функций. Любая достаточно длинная последовательность с вероятностью, близкой к единице, будет типичной, т. е. частота передачи любого символа в этой последовательности отличается от вероятности передачи этого символа на сколь угодно малую величину (частота символа x_k в этом случае определяется как отношение числа символов x_k к общему числу символов в рассматриваемой последовательности).

Таким образом, достаточно длинная эргодическая последовательность, являющаяся частью всей последовательности, вырабатываемой источником сообщения, с вероятностью, близкой к единице, характеризует вероятность появления отдельных символов и корреляционные связи между ними. Примером эргодических сообщений может служить язык. Почти в любой книге частота отдельных букв и различных сочетаний этих букв одинакова, хотя смысловое содержание книг различно. Этот факт позволяет применять математический аппарат в области лингвистики и имеет большое значение при построении систем связи.

§ 10. ЭНТРОПИЯ ЭРГОДИЧЕСКОГО ИСТОЧНИКА ДИСКРЕТНЫХ СООБЩЕНИЙ

Полученное ранее соотношение для вычисления энтропии нельзя применить к эргодическому источнику, поскольку оно выведено для случая независимых сообщений, а следовательно, не учитывает корреляционных связей.

Для дискретного эргодического источника можно найти конечное число характерных состояний v_1, v_2, \dots таких, что условная вероятность появления очередного символа зависит лишь от того, в каком состоянии из этих находится источник. Вырабатывая очередной символ, источник переходит из одного состояния в другое, либо возвращается в исходное состояние. Стохастический процесс такого типа в математике называют цепью Маркова. При отсутствии корреляционной связи в последовательности, вырабатываемой некоторым источником, источник имеет лишь одно характерное состояние, например v_1 . Вероятность появления некоторого символа x_k генерируемой последовательности в момент, когда система находится в этом состоянии, равна $p(x_k)$. Выработав символ x_k , источник возвращается в то же состояние v_1 , графическая интерпретация описанного процесса передана диаграммой перехода, изображенной на рис. 7.2, а.

Состояние источника на диаграмме определено точкой v_1 . Линии со стрелками характеризуют процесс генерации символов. Надписи у стрелок указывают вероятность этого процесса, когда состояние источника нам известно.

При наличии корреляционных связей между двумя соседними символами, вероятность появления некоторого символа x_k зависит лишь от того, какой символ был выработан до этого. Источник, генерирующий n различных символов

$x_1, x_2, x_3, \dots, x_k, \dots, x_n$, в этом случае может иметь n характерных состояний: v_1 — после появления символа x_1 ; v_2 — после появления символа x_2 и т. д.

Диаграмма перехода для такого случая при $n = 3$ приведена на рис. 7.2, б. Для описания такого источника необходимо задать распределение вероятностей $p(x_k)$ и вероятностей переходов $p(x_k/x_i)$ для всех k и i или задать вероятности всех возможных пар символов $p(x_k, x_i)$.

Энтропию для эргодического дискретного источника сообщений будем вычислять в предположении, что нам известно, в каком характерном состоянии находится источник.

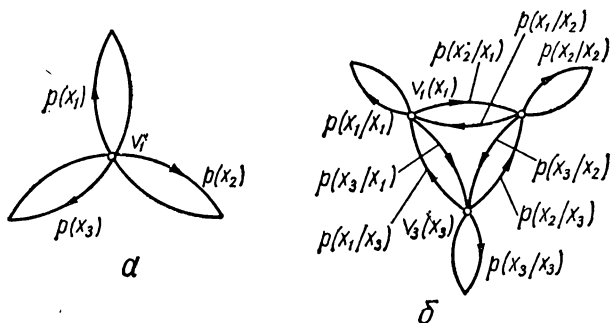


Рис. 7.2. Диаграмма перехода для дискретного источника сообщений.

Например, анализируемый источник сообщений $x_1, x_2, \dots, x_k, \dots, x_n$ имеет характерные состояния $v_1, \dots, v_k, \dots, v_l, \dots, v_n$. При этом $p(v_l/v_k)$ есть вероятность того, что источник, находясь в состоянии v_k , перейдет в состояние v_l при появлении очередного символа. Тогда согласно выражению (7.16)

$$H(X) = - \sum_k p(v_k) \sum_{l/k} p(v_l/v_k) \log(v_l/v_k). \quad (7.25)$$

Здесь $p(v_k)$ — вероятность состояния v_k . Знаки у сумм означают: k — суммирование по всем возможным состояниям, l/k — по всем возможным переходам.

Когда символы источника независимы, есть лишь одно состояние v_1 , вероятность которого $p(v_1) = 1$. С появлением символа x_k источник вновь возвращается в состояние v_1 . При этом $p(v_1/v_1) = p(x_k)$ и, следовательно,

$$H(X) = H(v_1) = - \sum_{k=1}^n p(x_k) \log p(x_k).$$

При наличии корреляционных связей между двумя соседними символами,

$$p(v_k) = p(x_k) \text{ и } p(v_1/v_k) = p(x_1/x_k),$$

из выражения (7.25) получим

$$H(X) = - \sum_{k=1}^n p(x_k) \sum_{l=1}^n p(x_l/x_k) \log p(x_l/x_k),$$

или

$$H(X) = - \sum_{k=1}^n \sum_{l=1}^n p(x_k, x_l) \log p(x_l/x_k). \quad (7.26)$$

Таблица 7.2

$x_k x_l$	$p(x_k x_l)$	$p(x_l/x_k)$	$x_k x_l$	$p(x_k x_l)$	$p(x_l/x_k)$
$x_1 x_1$	$13/32$	$13/16$	$x_3 x_1$	0	0
$x_1 x_2$	$3/32$	$3/16$	$x_3 x_2$	0	0
$x_1 x_3$	0	0	$x_3 x_3$	0	0
$x_1 x_4$	0	0	$x_3 x_4$	$1/8$	1
$x_2 x_1$	$1/32$	$1/8$	$x_4 x_1$	$1/16$	$1/2$
$x_2 x_2$	$1/8$	$1/2$	$x_4 x_2$	$1/32$	$1/4$
$x_2 x_3$	$3/32$	$3/8$	$x_4 x_3$	$1/32$	$1/4$
$x_2 x_4$	0	0	$x_4 x_4$	0	0

Аналогичные соотношения получаются и тогда, когда корреляционные связи распространяются на большее число символов. Рассмотрим два примера на определение энтропии источника.

Пример 1. Источник вырабатывает четыре символа x_1, x_2, x_3, x_4 с вероятностями

$$p(x_1) = p(x_2) = p(x_3) = p(x_4) = \frac{1}{4}.$$

Корреляционные связи отсутствуют. Используя формулу (7.11), получим

$$H(X) = 2 \frac{\partial \text{в.ед.}}{\text{символ}}.$$

Пример 2. Вероятности появления символов источника равны:

$$p(x_1) = \frac{1}{2}; \quad p(x_2) = \frac{1}{4}; \quad p(x_3) = p(x_4) = \frac{1}{8}.$$

При отсутствии корреляционных связей по формуле (7.11) находим

$$H(X) = 1,75 \frac{\text{дв.ед.}}{\text{символ}}.$$

Если же между двумя соседними символами есть корреляционные связи, которые описываются, например, таблицей 7.2, то, используя формулу (7.26), получим

$$H(X) = 0,886 \frac{\text{дв.ед.}}{\text{символ}}.$$

Из таблицы видно, что за символом x_3 в данном источнике всегда следует символ x_4 , а за символом x_1 генерируется либо символ x_1 , либо x_2 . Так как вероятности некоторых пар символов равны нулю, то всего рассматриваемый источник имеет девять характерных состояний.

§ 11. ИЗБЫТОЧНОСТЬ ИСТОЧНИКА СООБЩЕНИЙ

Рассмотренные выше примеры показывают, что при одинаковом количестве различных символов количество информации, приходящееся на одно сообщение, может быть различным в зависимости от статистических характеристик источника.

Энтропия источника максимальна и равна $H_{\max} = \log n$, если символы вырабатываются с равными вероятностями. Если же это не так, т. е. одни символы повторяются реже, другие чаще, то энтропия источника уменьшается, а при появлении дополнительных корреляционных связей между символами энтропия становится еще меньше.

Для того, чтобы выяснить, насколько хорошо в источнике сообщений используются разные символы, вводят параметр, называемый избыточностью

$$R = \frac{H_{\max} - H(X)}{H_{\max}}. \quad (7.27)$$

Здесь $H_{\max} = \log n$ — максимальная энтропия или наибольшее количество информации, которое может приходиться на один символ источника при данном числе n используемых символов. Из выражения (7.27) видно, что при $R = 0$ энтропия источника $H(X) = H_{\max}$, т. е. источник генерирует максимальное количество информации на символ. Если $R = 1$, то $H(X) = 0$, и, следовательно, информация, передаваемая источником, равна нулю.

В общем случае $0 \leq R \leq 1$.

Чем меньше избыточность, тем оптимальнее работает источник, генерируя большее количество информации. Однако следует иметь в виду, что некоторая избыточность бывает полезной для обеспечения надежности работы источника.

Для примеров, рассмотренных выше, избыточность будет следующей: для примера 1 $R = 0$; для примера 2, если нет корреляции,

$$R = \frac{2 - 1,75}{2} = 0,125.$$

При наличии корреляции между двумя соседними символами

$$R = \frac{2 - 0,886}{2} \approx 0,56.$$

§ 12. СКОРОСТЬ СОЗДАНИЯ СООБЩЕНИЙ

Выражение (7.25) позволяет определить количество информации, переносимое в среднем одним символом источника. При работе источника сообщений на его выходе отдельные символы появляются через некоторые интервалы времени. Следовательно, есть смысл говорить о длительности отдельных символов и можно поставить вопрос о количестве информации, вырабатываемой источником в единицу времени.

Определим среднюю длительность символа. Обозначая через $tx_k(v_l/v_k)$ длительность символа x_k , переводящего источник из состояния v_k в состояние v_l ; через $p(v_l/v_k)$ вероятность указанного перевода, среднюю длительность символа определим так:

$$\bar{t}_n = \sum_k p(v_k) \sum_{l/k} \sum_k p x_k(v_l/v_k) tx_k(v_l/v_k).$$

Тогда энтропия источника, приходящаяся на единицу времени, называемая *скоростью создания сообщений*, будет определяться по формуле

$$\bar{H}(X) = \frac{H(X)}{\bar{t}_n} \frac{\text{дв.ед.}}{\text{сек}}. \quad (7.28)$$

Таким образом, для получения большей скорости создания сообщения на выходе источника необходимо не только обеспечить по возможности большую энтропию, но и правильно выбрать длительность различных символов.

§ 13. ПРОПУСКНАЯ СПОСОБНОСТЬ ИНФОРМАЦИОННОГО КАНАЛА

Обозначим через u_T последовательность сообщений, обрабатываемых источником за время T . Соответствующую ей последовательность принятых сообщений обозначим через z_T . Очевидно, что в информационном канале без шумов u_T однозначно определяет z_T , а в канале с шумами при передаче u_T на приемной стороне могут образоваться различные последовательности z_T .

Положим, что $I(z_T, u_T)$ определяет количество информации, содержащееся в последовательностях сообщений на выходе информационного канала о последовательностях на его входе. Ясно, что $I(z_T, u_T)$ зависит от статистических характеристик источника сообщений и характеристик помех, действующих в канале, а также интервала времени T .

Проанализируем предел

$$\lim_{T \rightarrow \infty} \frac{I(z_T, u_T)}{T}. \quad (7.29)$$

При передаче сообщений эргодического источника при $T \rightarrow \infty$ с вероятностью, как угодно близкой к единице, последовательность сообщений источника u_T будет типичной. При соблюдении некоторых условий, таких например, как условие эргодических помех, действующих в канале, последовательность выходных сообщений z_T также будет типичной. Поэтому естественно предположить, что анализируемый предел может являться некоторой характеристикой работы информационного канала, указывающей среднее количество информации, получаемое на выходе за единицу времени, т. е. скорость передачи информации. Обозначив эту характеристику через $\bar{I}(z, u)$, будем иметь

$$\bar{I}(Z, U) = \lim_{T \rightarrow \infty} \frac{I(Z_T, U_T)}{T}. \quad (7.30)$$

Теперь предположим, что известна некоторая совокупность фиксированных ограничений, накладываемых на канал. К фиксированным ограничениям отнесем параметры канала связи, используемый код и т. п.

Введем понятие *пропускной способности информационного канала* C , являющейся максимальным значением ско-

рости передачи информации при заданных фиксированных ограничениях:

$$C = \sup \{ \bar{I}(z, u) \} \frac{\partial \text{в.ед.}}{\text{сек}}$$

$$\alpha_1 \in A_1$$

$$\dots$$

$$\alpha_i \in A_i$$

$$\dots$$

$$\alpha_n \in A_n.$$

Здесь обозначение \sup указывает, что вычисляется верхняя грань. Запись $\alpha_i \in A_i$ говорит о том, что α_i удовлетворяет заданному фиксированному ограничению, т. е. лежит в некоторой области A_i .

Способ отыскания верхней грани зависит от того, какая совокупность фиксированных ограничений задана. Если информационный канал определен полностью, то верхнюю грань следует отыскивать по статистическим характеристикам источника сообщений, т. е. отыскивать распределения вероятностей источника сообщений, корреляционные связи, при которых скорость передачи информации будет наибольшей.

Если совокупность заданных фиксированных ограничений не полностью определяет канал, то отыскиваются оптимальные статистические характеристики источника, при которых скорость передачи информации максимальная.

Аналогично тому, как это было сделано для информационного канала, можно определить пропускную способность канала связи (рис. 7.1)

$$C_{\text{к.с}} = \sup \{ \bar{I}(Y, X) \} \frac{\partial \text{в.ед.}}{\text{сек}}, \quad (7.32)$$

$$\beta_1 \in B_1$$

$$\dots$$

$$\beta_i \in B_i$$

$$\dots$$

$$\beta_n \in B_n,$$

где

$$\bar{I}(Y, X) = \lim_{T \rightarrow \infty} \frac{I(Y_T, X_T)}{T}. \quad (7.33)$$

Канал связи является частью системы связи (информационного канала связи). Рассматривая канал связи в отдельности и определяя его пропускную способность, можно не накладывать никаких фиксированных ограничений на способ кодирования, вид модуляции и т. п. Тем самым создаются возможности использования оптимальных сигналов, при которых скорость передачи информации будет предельно большой. Однако построение реального информационного канала связано с введением тех или иных дополнительных фиксированных ограничений, что приводит к недоиспользованию пропускной способности канала связи. Таким образом, обычно $C_{к.с} > C$.

Контрольные вопросы

1. Какие источники сообщений называют эргодическими?
2. Чем характеризуется энтропия дискретного источника сообщений?
3. Что такое избыточность источника сообщений?
4. Чем характеризуется оптимальность работы источника сообщений?
5. Чем можно обеспечить увеличение скорости создания сообщений?
6. В чем заключается смысл понятия пропускной способности информационного канала?

§ 14. ДИСКРЕТНЫЕ КАНАЛЫ БЕЗ ШУМОВ. ПРОПУСКНАЯ СПОСОБНОСТЬ ДИСКРЕТНЫХ КАНАЛОВ БЕЗ ШУМОВ

При отсутствии шумов в информационном канале, не нарушая общности, можно считать, что $y = x$ и $z = u$. Следовательно, для сообщений, передаваемых за время T ,

$$Y_T = X_T; \quad z_T = u_T.$$

Учитывая свойства количественной меры информации, можно записать

$$I(z_T, u_T) = I(u_T, u_T) = H(u_T)$$

и

$$I(Y_T, X_T) = I(X_T, X_T) = H(X_T).$$

Тогда используя (7.30) и (7.31) получим

$$C = \sup \left\{ \lim_{T \rightarrow \infty} \frac{H(u_T)}{T} \right\}, \quad (7.34)$$

а из (7.32) и (7.33)

$$C_{к.с} = \sup \left\{ \lim_{T \rightarrow \infty} \frac{H(X_T)}{T} \right\}. \quad (7.35)$$

Обозначим через n число всех возможных последовательностей сообщений длительностью T . Из свойств энтропии следует, что $H(u_T)$ будет максимальной, если все возможные последовательности сообщений равновероятны. Это максимальное значение равно $\log n$. Тогда из (7.34) и (7.35) получим

$$C = \lim_{T \rightarrow \infty} \frac{\log n}{T} \quad (7.36)$$

и

$$C_{\text{к.с}} = \lim_{T \rightarrow \infty} \frac{\log n_c}{T}, \quad (7.37)$$

где n_c — число всех возможных последовательностей кодированных сигналов длительностью T .

Выражения (7.36) и (7.37) обычно используются для определения пропускной способности дискретного канала без шумов.

§ 15. ЭФФЕКТИВНОЕ КОДИРОВАНИЕ

Ранее трудно было заметить, что фактическая скорость передачи информации может быть максимальной, равной пропускной способности канала, если статистические характеристики источника сообщений надлежащим образом согласованы со свойствами информационного канала. Для каждого источника сообщений это согласование может быть достигнуто специальным выбором способа кодирования и декодирования сообщений. Такое кодирование сообщений, при котором достигается наилучшее использование пропускной способности канала связи, называется *эффективным*.

Эффективное кодирование должно обеспечивать:

1) при заданной статистике источника сообщений формирование кодированных сигналов с оптимальными статистическими характеристиками, при которых достигается наибольшая скорость передачи информации;

2) возможность декодирования сигналов на приемной стороне, т. е. разделение сигналов отдельных сообщений, опознание этих сигналов и т. п. Способы эффективного кодирования зависят от вида и свойств информационного канала.

По-видимому из всех возможных кодов тот код обеспечивает наибольшую эффективность системы связи, при котором среднее количество кодовых символов, приходящихся на один символ сообщения, будет минимальным. При этом

на передачу сообщения будет затрачено наименьшее число кодовых символов, а следовательно, время и энергия сигнала будут также наименьшими.

Установим минимальное значение среднего числа кодовых символов, приходящихся на один элемент сообщения. Предполагается, что в канале нет помех.

Пусть пропускная способность канала связи — $C_{\text{к.с.}} \frac{\text{дв.ед.}}{\text{сек}}$, а энтропия источника сообщений — $H \frac{\text{дв.ед.}}{\text{элемент сообщ.}}$. При этом не обязательно, чтобы

$$H = H_{\text{max}} = \log n,$$

т. е. предполагается избыточность. Пусть \bar{I}_{max} — наибольшая возможная скорость передачи информации. Тогда на основании определения пропускной способности канала можно записать

$$C_{\text{к.с.}} = H I_{\text{max}}. \quad (7.38)$$

Пусть далее число кодовых символов, приходящееся в среднем на один элемент сообщения

$$l_{\text{cp}} = \sum_{k=1}^n p_k l_k, \quad (7.39)$$

где

p_k — априорная вероятность k -го элемента сообщения; l_k — число кодовых символов, приходящееся в среднем на один элемент сообщения; n — число элементов сообщения в ассортименте источника.

Если за время T передается m элементов сообщения, то скорость передачи кодовых символов W

$$W = \frac{l_{\text{cp}} m}{T} = l_{\text{cp}} \bar{I}, \quad (7.40)$$

где

$$\bar{I} = \frac{m}{T}.$$

Будем рассматривать двоичный код, т. е. код, составленный из двух символов. Для такого кода количество информации, приходящееся в среднем на один кодовый символ, равно одной двоичной единице (при равных априорных вероятностях кодовых символов). Другими словами, наибольшая возможная скорость передачи двоичных символов численно

равна пропускной способности канала, т. е.

$$W_{\max} = C_{\text{к.с.}}$$

Тогда вместо (7.40) можно записать

$$C_{\text{к.с.}} = l_{\text{ср}} \bar{I}_{\max}. \quad (7.41)$$

Анализируя (7.38) и (7.40), получим

$$l_{\text{ср}} = H. \quad (7.42)$$

Следовательно, наибольшая скорость передачи элементов сообщения по каналу связи, а следовательно, и наибольшая эффективность будет при равенстве $l_{\text{ср}}$ и H . В общем случае

$$l_{\text{ср}} \geq H.$$

Из всех кодов наибольшую эффективность дает тот код, для которого справедливо

$$l_{\text{ср}} = H,$$

при этом скорость передачи информации будет равна своему предельному значению — пропускной способности канала связи.

Сравнивая (7.42) с формулой для энтропии, видим, что для обеспечения равенства (7.42) необходимо, чтобы

$$l_k = -\log p_k = \log \frac{1}{p_k}. \quad (7.43)$$

Последнее соотношение определяет правило, позволяющее построить оптимальный код, т. е. максимально краткий код, при котором обеспечивается наибольшая скорость передачи информации.

§ 16. ОСНОВНАЯ ТЕОРЕМА ШЕННОНА ДЛЯ ДИСКРЕТНОГО КАНАЛА БЕЗ ШУМОВ

Основная теорема Шеннона для дискретного канала без шумов дает ответ на вопрос о том, в какой мере скорость передачи информации можно приблизить к пропускной способности информационного канала.

Теорема. Если скорость создания сообщений источником

$$\bar{H}(u) = C_{\text{к.с.}} - \varepsilon, \quad (7.44)$$

где ε может быть как угодно малым, то всегда можно найти такой способ кодирования, который обеспечит передачу всех сообщений, вырабатываемых источником, причем скорость передачи информации будет равна

$$\bar{I} = C_{\text{к.с.}} - \varepsilon.$$

Обратное утверждение заключается в том, что невозможно обеспечить длительную передачу всех сообщений источника, у которого

$$H(u) > C_{\text{к.с.}}$$

Доказательство этой теоремы требует введения множества специальных абстрактных понятий. Детально оно приведено в работах А. Файнштейна по теории информации.

§ 17. ДИСКРЕТНЫЕ КАНАЛЫ С ШУМАМИ

Воздействие различного рода помех на всякий реальный канал связи приводит к искажению передаваемых сообщений, в результате чего, получив сигнал на приемном конце, мы не можем с полной достоверностью утверждать, какое сообщение было передано.

В схеме, показанной на рис. 7.1, в случае воздействия помех нет однозначного соответствия между сигналами на входе и выходе канала связи. При передаче сигнала x_k выходные сигналы y могут принимать различные значения в зависимости от того, каков был шум в момент приема. В общем случае сигнал y может принимать непрерывное множество различных значений. Принятие решения соответствует некоторому разделению всего множества $\{y\}$ на области $Y_1, \dots, Y_k, \dots, Y_m$, так, что если принятый сигнал y принадлежит области Y_k , то делают вывод о том, что был передан сигнал y_k .

Основные соотношения для дискретного канала с шумами в теории информации получены для решающих схем, в которых равномерно ограничена вероятность ошибки при опознании любого переданного сигнала x . В этом случае для вероятности правильного принятия решения $p(y_k/x_k)$, т. е. условной вероятности того, что при передаче сигнала x_k будет принято решение y_k , справедливо соотношение

$$p(y_k/x_k) \geq p_{\text{пр}},$$

где $p_{\text{пр}}$ — вероятность правильного решения, гарантируемая для всех сообщений, т. е. для всех k .

§ 18. ПРОПУСКНАЯ СПОСОБНОСТЬ ДИСКРЕТНОГО КАНАЛА С ШУМАМИ

Полученное ранее соотношение, определяющее пропускную способность дискретного канала связи, справедливо и для дискретных каналов с шумами. Разница между каналами с шумами и рассмотренными каналами без шумов

состоит лишь в способе вычисления количества информации, содержащейся в последовательности исходных сигналов Y_T о входных сигналах X_T , т. е. в вычислении величины $I(Y_T, X_T)$.

Предполагается, что шумы, действующие в канале связи, имеют эргодический характер. Например, при длительной многократной передаче x_k сигналы y_k на выходе канала с вероятностью, как угодно близкой к единице, образуют типичную последовательность. При этом условии выход канала связи можно рассматривать как эргодический источник. Для величины $I(Y_T, X_T)$ можно записать

$$I(Y_T, X_T) = H(Y_T) - H(Y_T/X_T) = H(X_T) - H(X_T/Y_T). \quad (7.45)$$

Для последовательности длительностью T , содержащей M сигналов эргодического источника (канала связи), имеем

$$H(Y_T) = MH(Y), \quad (7.46)$$

где $H(Y)$ — энтропия выходного сигнала канала связи. Величину $H(Y)$ можно подсчитать по формуле (7.25)

$$H(Y) = - \sum_k \sum_{l/k} p(v_k) p(v_l/v_k) \log p(v_l/v_k). \quad (7.47)$$

Здесь v_l и v_k обозначают характерные состояния канала связи.

Для условной энтропии можно записать

$$H(Y_T/X_T) = MH(Y/X). \quad (7.48)$$

Используя (7.25), получим

$$H(Y/X) = \sum_k p(x_k) H(Y/x_k), \quad (7.49)$$

где

$$H(Y/x_k) = - \sum_k \sum_{l/k} p(v_k) p(v_l/v_k, x_k) \log p(v_l/v_k, x_k). \quad (7.50)$$

Здесь $p(v_l/v_k, x_k)$ — условная вероятность перехода выхода канала связи из состояния v_k в состояние v_l при передаче сигнала x_k .

Из выражений (7.45), (7.46) и (7.48) получим

$$I(Y_T; X_T) = MH(Y) - MH(Y/X).$$

Для определения скорости передачи информации учтем, что

$$\lim_{T \rightarrow \infty} \left(\frac{T}{M} \right) = \bar{t}_c,$$

где \bar{t}_c — средняя длительность сигнала одного сообщения. Тогда

$$\bar{I}(Y, X) = \bar{H}(Y) - \bar{H}(Y/X), \quad (7.51)$$

где

$$\left. \begin{aligned} H(Y) &= \frac{H(Y)}{\bar{t}_c}, \\ H(Y/X) &= \frac{H(Y/X)}{\bar{t}_c}. \end{aligned} \right\} \quad (7.52)$$

Аналогично найдем

$$\bar{I}(Y, X) = \bar{H}(Y) - \bar{H}(X/Y). \quad (7.53)$$

В последнем равенстве $\bar{H}(X)$ — скорость создания сообщения на выходе кодирующего устройства; $\bar{H}(X/Y)$ — величина, характеризующая потерю информации, обусловленную наличием помех в канале связи. Из найденных сообщений следует, что

$$C_{\text{к.с.}} = \sup \{ \bar{H}(Y) - \bar{H}(Y/X) \}, \quad (7.54)$$

или

$$C_{\text{к.с.}} = \sup \{ \bar{H}(X) - \bar{H}(X/Y) \}. \quad (7.55)$$

Выражения (7.54) и (7.55) равноправны и дают одинаковые результаты. Использование того или иного выражения зависит от удобства выбора оптимальных статистических характеристик сигналов. При этом следует иметь в виду, что характерные состояния выхода канала связи определяются:

1) наличием фиксированных ограничений относительно последовательности передачи различных сигналов;

2) корреляционными связями между символами, вызываемыми действием шумов. Каналы, в которых на каждый передаваемый сигнал шум воздействует независимо от того, какие сигналы передавались ранее, называются каналами без памяти. В таких каналах шумы не вызывают дополнительных корреляций между символами. В настоящее время основные выводы теории информации получены применительно к каналам без памяти.

§ 19. ОСНОВНАЯ ТЕОРЕМА ШЕННОНА ДЛЯ ДИСКРЕТНОГО КАНАЛА С ШУМАМИ

Для дискретного канала с шумами Шенноном доказана следующая теорема:

Если скорость создания сообщения у источника $\bar{H}(u)$ такая, что

$$\bar{H}(u) = C_{\text{к.с.}} - \varepsilon, \quad (7.56)$$

где ε как угодно мало, то существует способ кодирования, при котором все сообщения, вырабатываемые источником, могут быть переданы, а вероятность ошибочного опознания любого переданного сообщения может быть как угодно малой, т. е.

$$p_n < \eta,$$

где p_n — вероятность неправильного опознания любого переданного сообщения, $\eta > 0$.

Обратное утверждение теоремы состоит в том, что если $\underline{H}(u) > I_c$, то не существует способа кодирования, обеспечивающего передачу всех сообщений с малой вероятностью ошибки.

Доказательство этой теоремы можно также найти в работах А. Файнштейна по теории информации.

Отметим, что фундаментальное значение теоремы состоит в том, что она позволяет, зная предельные значения скорости передачи информации $C_{к.с.}$, оценить эффективность используемых методов кодирования в данной системе передачи информации.

Контрольные вопросы

1. Чем определяется пропускная способность дискретного канала без шумов?
2. Какое кодирование называют эффективным?
3. В чем состоит основное свойство эффективного кода?
4. Сформулируйте правило, позволяющее построить оптимальный код.
5. В чем состоит физический смысл основной теоремы Шеннона для дискретного канала?
6. Чем определяется пропускная способность дискретного канала с шумами?
7. Какими обстоятельствами определяются характерные состояния с шумами?
8. Что такое каналы без памяти?
9. Сформулируйте основную теорему Шеннона для дискретного канала с шумами.

§ 20. ИСТОЧНИКИ НЕПРЕРЫВНЫХ СООБЩЕНИЙ

Источники непрерывных сообщений характеризуются непрерывным изменением во времени или в пространстве физического параметра, который несет ту или иную информацию об источнике сообщений. Для регистрации или передачи на расстояние некоторого параметра u , величина которого непрерывно изменяется (см. рис. 7.1), как прави-

ло, перед поступлением этого параметра в канал связи, происходит преобразование его в электрический сигнал x .

Смысл этого преобразования состоит в установлении некоторого соответствия между u и x в виде функциональной связи $x = f(u)$.

Поскольку всякий реальный преобразователь вносит случайные помехи, нарушающие функциональную связь, соответствие между u и x может характеризоваться лишь условным распределением $W(x/u)$, зависимостью математического ожидания случайной величины X от u , т. е. $M[X] = f(u)$ или дисперсией $D[X]$ при известном u .

Рассмотрим непрерывное сообщение u на некотором отрезке T . Для простоты записи начало рассматриваемого интервала T будем считать совмещенным с началом отсчета времени $t = 0$, поскольку такое совмещение не снижает общности последующих рассуждений. Любую реализацию $u(t) = u_T(t)$ непрерывного сообщения u на рассматриваемом интервале времени можно характеризовать совокупностью параметров, или, другими словами, функцию $u(f)$ на отрезке времени T можно представить вектором в многомерном пространстве.

В качестве координат такого пространства обычно используют совокупность коэффициентов разложения функции $u(t)$ в ряд по некоторой полной системе регулярных функций. При использовании разложения функции $u_T(t)$ в ряд Маклорена такими функциями являются $1, t, t^2, \dots$. В результате разложения

$$u_T(t) = a_0 + a_1 t + a_2 t^2 + \dots + a_k t^k + \dots$$

Коэффициенты $a_0, a_1, a_2, \dots, a_k, \dots$ можно рассматривать как координаты $u_T(t)$.

При использовании в качестве разложения $u_T(t)$ ряда Котельникова координатами является совокупность значений функции в точках опроса

$$u_T(0); \quad u_T(T_0); \quad u_T(2T_0); \quad \dots$$

Можно использовать и другие способы разложения функций $u_T(t)$ и, следовательно, другие способы задания координат.

Важно, чтобы разложение было применимо к любому сообщению $u(t)$. Обозначим совокупность координат, характеризующих данное сообщение, через u_1, u_2, \dots, u_m .

В общем случае можно использовать любую приемлемую систему координат и потому u_1, u_2, \dots могут быть коэф-

фициентами того или иного разложения функции $u(t)$ в ряд и не являться значениями этой функции в точках опроса. Для различных сообщений эти координаты принимают различные значения. В таком случае множество реализаций $u_T(t)$ характеризуется совокупностью случайных величин u_1, u_2, \dots . Если координаты u_1, u_2, \dots могут изменяться непрерывно, то статистическое описание множества сообщений задается совместной плотностью вероятностей $W(u_1, u_2, \dots)$. Следует иметь в виду, что поскольку рассматриваются конечные разложения, указанное представление функции $u_T(t)$ является приближенным.

§ 21. КОЛИЧЕСТВО ИНФОРМАЦИИ, СОДЕРЖАЩЕЕСЯ В ОДНОМ ЗАМЕРЕ \

НЕПРЕРЫВНОЙ СЛУЧАЙНОЙ ВЕЛИЧИНЫ

Пусть преобразователь (рис. 7.3) представляет собой измерительное устройство, что не влияет на общность рассуждений. Определим, какое количество информации о величине u содержится в величине x . Нетрудно убедиться, что если бы измерение было абсолютно точным, то количество информации должно быть бесконечно большим. Действительно, пусть, например, результат измерения x выража-

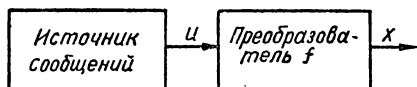


Рис. 7.3. Блок-схема установления соответствия между u и x .

ется числом в двоичной системе счисления. Если u окажется иррациональным, а это всегда возможно, поскольку u непрерывно измеряется в некотором интервале, то,

для того чтобы x точно описывало u , потребуется двоичное число с бесконечно большим числом разрядов. Иначе говоря, если u — непрерывная случайная величина, которая может принимать все возможные значения в некотором интервале $u = u_{\min} \div u_{\max}$, то число этих значений бесконечно велико, и точное определение значения u при измерении соответствует получению бесконечно большого количества информации.

Именно поэтому понятие об энтропии дискретных случайных величин нельзя непосредственно перенести на непрерывные величины, поскольку $H(U) = I(U, U)$ для дискретных величин, а для непрерывных $I(U, U)$ бесконечно велико.

Однако ясно, что в действительности измерить любую случайную величину можно лишь с некоторой степенью точности. Погрешность в измерении приводит к тому, что количество информации о величине u , содержащееся в случайной величине x , становится конечным. В этом можно убедиться на простом примере.

Пусть шкала изменения параметра u :

$$L_u = u_{\max} - u_{\min}. \quad (7.57)$$

Допустим, что измерение u ведется с точностью ε , так что показания x измерительного прибора могут иметь лишь $m = \frac{L_u}{\varepsilon}$ различных значений x_ε . Очевидно, что информация, содержащаяся в одном измерении, будет максимальной, если все результаты равновероятны. Тогда

$$I(X, U)_{\max} = H(X_\varepsilon) = \log_2 m,$$

или

$$I(X, U)_{\max} = \log_2 \frac{L_u}{\varepsilon}. \quad (7.58)$$

Из последнего соотношения следует, что при конечном, хотя и малом, ε величина $I(X, U)_{\max}$ также конечна, если же $\varepsilon \rightarrow 0$, то

$$I(X, U) \rightarrow \infty.$$

Сказанное дает основание распространить соотношения, определяющие $I(X, U)$ для дискретных случайных величин, на непрерывные.

Пусть нам известны:

$W(u)$ — априорная функция распределения случайной величины u ; $W(x/u)$ — условная функция распределения x при известном u . Эта функция характеризует погрешность преобразования. Пусть далее

$$x = u + \delta, \quad (7.59)$$

где δ — погрешность измерения, независимая от u . Очевидно, что

$$W(x/u) = W_\Delta(\delta), \quad (7.60)$$

где $W_\Delta(\delta)$ — плотность распределения погрешностей измерения. Выражение (7.60) можно переписать иначе

$$W(x/u) = W_\Delta(x - u). \quad (7.61)$$

Далее, зная $W(u)$ и $W(x/u)$ найдем плотность совместного распределения вероятностей

$$W(u, x) = W(u) W(x/u) \quad (7.62)$$

и плотность распределения вероятностей случайной величины

$$W(x) = \int_{L_u} W(u, x) du, \quad (7.63)$$

где интегрирование ведется по всей шкале u .

Кроме того,

$$W(u/x) = \frac{W(u, x)}{W(x)}. \quad (7.64)$$

Совокупность определенных распределений вероятностей полностью описывает вероятностные зависимости между u и x . Разбив интервал L_u на конечное число малых интервалов Δu и пронумеровав их, можно совокупность всех значений u внутри i -го интервала представлять одним средним значением u_i , априорная вероятность которого будет $p(u_i) = W(u_i) \Delta u$. Поступая аналогично с величиной x , получаем

$$I(X, U) \approx \sum_k \sum_i W(u_i, x_k) \Delta x \Delta u \log \frac{W(u_i/x_k) \Delta u}{W(u_i) \Delta u}.$$

Умножим числитель и знаменатель дроби на $W(x_k)$

$$I(X, U) \approx \sum_i \sum_k W(u_i, x) \Delta x \Delta u \log \frac{W(u_i/x_k)}{W(u_i) W(x_k)}.$$

Принимая $\Delta x \rightarrow 0$ и $\Delta u \rightarrow 0$, получим

$$I(X, U) = \int_{L_u} \int_{L_x} W(x, u) \log \frac{W(u, x)}{W(u) W(x)} dx du. \quad (7.65)$$

Соотношение (7.65) позволяет подсчитать количество информации, содержащееся в непрерывной случайной величине x о непрерывной величине u .

Отметим несколько важных свойств выражения (7.65):

1) $I(U, X) = I(X, U)$, что следует из равенства

$$W(u, x) = W(x, u);$$

2) $I(X, U) \geq 0$, причем $I(X, U) = 0$, если U, X независимые случайные величины. В последнем случае $W(u, x) = W(u) W(x)$ и выражение, стоящее под знаком логарифма, становится равным единице;

3) значение $I(X, U)$ не зависит от способа отсчета величин u и x , т. е. от выбора системы координат.

§ 22. ЭНТРОПИЯ НЕПРЕРЫВНЫХ СЛУЧАЙНЫХ ВЕЛИЧИН

Учитывая, что

$$W(u, x) = W(u) W(x/u) = W(x) W(u/x),$$

выражение (7.65) можно преобразовать к виду

$$I(X, U) = H(U) - H(U/X), \quad (7.66)$$

где

$$H(U) = - \int_{L_u} W(u) \log W(u) du, \quad (7.67)$$

$$H(U/X) = - \int_{L_u} \int_{L_x} W(x) W(u/x) \log W(u/x) dudx. \quad (7.68)$$

По аналогии со случаем для дискретных случайных величин, $H(U)$ и $H(U/X)$ можно назвать соответственно энтропией и условной энтропией непрерывной случайной величины U . Несмотря на общность свойств энтропии дискретных и непрерывных случайных величин, между ними есть и существенное отличие. Для дискретных случайных величин

$$H(U) = I(U, U).$$

Для непрерывных величин это равенство не справедливо, в чем нетрудно убедиться из сопоставления (7.67) и (7.65). Таким образом, для дискретных случайных величин энтропия определяется как количество информации, получаемое при полной достоверности опыта, а для непрерывных — как некоторая величина, с помощью которой можно определить количество информации по (7.66).

Для непрерывного случая энтропию можно истолковывать как меру неопределенности в выбранной системе координат.

§ 23. КОЛИЧЕСТВО ИНФОРМАЦИИ О НЕПРЕРЫВНОЙ СЛУЧАЙНОЙ ВЕЛИЧИНЕ ПРИ ЗАДАННЫХ ТРЕБОВАНИЯХ К ВЕРНОСТИ ВОСПРОИЗВЕДЕНИЯ

В § 21 было определено количество информации, содержащееся в случайной величине X о случайной величине U при известных вероятностных связях между этими величинами. Однако на практике по x нельзя определить точное значение u , поэтому необходимо предъявить некоторые требования к допущенным ошибкам (к верности воспроиз-

ведения величиной x величины u). Найдём количество информации, содержащееся в случайной величине X о случайной величине U при заданных требованиях к верности воспроизведения.

В общем случае критерий верности воспроизведения может требовать, чтобы плотность совместного распределения вероятностей $W(u, x)$ принадлежала к некоторому классу функций W , т. е. $W(u, x) \in W$. При сопоставлении значений x и u нас могут интересовать величины

$$r(u, x) = |x - u|$$

или

$$r(u, x) = (x - u)^2$$

и т. д.

В таком случае наиболее удобным критерием будет среднее значение функции $r(u, x)$, т. е.

$$q = \int_{L_u} \int_{L_x} W(u, x) r(u, x) du dx. \quad (7.69)$$

Тогда требование к верности воспроизведения можно задать в виде

$$q \leq \varepsilon,$$

где ε — приемлемо малая величина.

Очевидно, что условию (7.69) будет удовлетворять не одна, а некоторое множество функций $W(x/u)$ и $W(u, x)$. Ясно, что наиболее выгодной будет та функция $W(x/u)$, при которой $I(X, U)$ имеет наименьшее значение. Выбор такой выгодной $W(x/u)$ позволяет выполнить заданные требования к верности воспроизведения при получении минимального количества информации.

Наименьшее значение $I(X, U)$, при котором выполняются требования к верности воспроизведения, называется ε -энтропией

$$H_\varepsilon(X) = \inf_{q \leq \varepsilon} \{I(X, U)\},$$

где \inf указывает на то, что отыскивается нижняя грань.

Нижняя грань отыскивается по множеству $W(x/u)$. Таким образом, ε -энтропия равна минимальному количеству информации, которое должно содержать сообщение о случайной непрерывной величине при заданных требованиях к верности воспроизведения.

§ 24. КОЛИЧЕСТВО ИНФОРМАЦИИ, СОДЕРЖАЩЕЕСЯ В ВОСПРОИЗВЕДЕНИИ НЕПРЕРЫВНОГО СООБЩЕНИЯ

Повторяя рассуждения, проверенные при выводе (7.65) для случая, когда U и X — многомерные величины, определим количество информации, содержащееся в воспроизведении x_T сообщений u_T .

Пусть нами выбрана некоторая приемлемая m -мерная система координат. Тогда множество всех возможных сообщений $u_T(t)$ представляется множеством m -мерных векторов (u_1, u_2, \dots, u_m) . Воспроизведение $x_T(t)$ можно представить вектором (x_1, x_2, \dots, x_m) в многомерном пространстве функций, где x_1, x_2, \dots, x_m — координаты $x_T(t)$.

В результате

$$I(X_T; U_T) = \int_{L_{u_1}} \dots \int_{L_{u_m}} \int_{L_{x_1}} \dots \int_{L_{x_m}} W(u_1 \dots u_m; x_1 \dots x_m) \times \\ \times \log \frac{W(u_1 \dots u_m; x_1 \dots x_m)}{W(u_1 \dots u_m) W(x_1 \dots x_m)} \cdot du_1 \dots du_m dx_1 \dots dx_m. \quad (7.70)$$

Совместную плотность распределения вероятностей $W(u_1, \dots, u_m; x_1 \dots x_m)$ можно представить в виде

$$W(u_1 \dots u_m; x_1 \dots x_m) = \\ = W(u_1 \dots u_m) W(x_1 \dots x_m / u_1 \dots u_m). \quad (7.71)$$

Функция распределения координат воспроизведения определяется следующим образом:

$$W(x_1 \dots x_m) = \\ = \int_{L_{u_1}} \dots \int_{L_{u_m}} W(u_1 \dots u_m; y_1 \dots y_m) du_1 \dots du_m. \quad (7.72)$$

Равенство (7.10) приводится к виду

$$I(X_T; U_T) = H(U_T) - H(X_T/U_T). \quad (7.73)$$

Так как основные свойства энтропии для многомерного случая остаются такими же, как и для одномерного, количество информации

$$I(X_T; U_T) = H(X_T) - H(X_T/U_T).$$

Следовательно, в многомерном случае, как и в одномерном, количество информации можно определить при условии, что известна функция распределения множества сообщений

$W(u_1 \dots u_m)$ и условная функция распределения $W(x_1; \dots; x_m/u_1; \dots; u_m)$. Так же как и для одномерного случая ε -энтропия источника непрерывных сообщений определяется из условия

$$H_\varepsilon(U_T) = \inf_{q < \varepsilon} \{I(X_T; U_T)\}.$$

При этом нижняя грань отыскивается по всем возможным условным функциям распределения.

§ 25. НЕПРЕРЫВНЫЕ КАНАЛЫ С ШУМАМИ. ПРОПУСКНАЯ СПОСОБНОСТЬ НЕПРЕРЫВНЫХ КАНАЛОВ

Функциональная схема непрерывного информационного канала отличается от схемы, показанной на рис. 7.1, тем, что в ней вместо кодирующих и декодирующих устройств можно использовать более широкий класс различных преобразователей. Обычно непрерывное сообщение $u(t)$ сначала преобразуется в электрический сигнал $x(t)$, который затем после ряда дополнительных преобразований (модуляция и т. п.) поступает в канал связи. Как правило, сигнал $y(t)$, получаемый на приемном конце, отличается от сигнала $x(t)$.

Это отличие обусловлено наличием собственных шумов аппаратуры и шумов в канале связи. В результате воздействия шумов соответствие между $x(t)$ и $y(t)$ носит вероятностный характер. Для непрерывных каналов с шумами можно использовать понятие о пропускной способности тогда, когда источник сообщений обладает эргодическими свойствами.

Пусть выходные сигналы $x(t)$ и выходные сигналы $y(t)$ являются стационарными, эргодическими и стационарно связанными функциями времени. Возьмем отрезки этих функций на интервале T . Будем полагать также, что значения каждой функции в точках опроса не коррелированы. Определим

$$\begin{aligned} H(Y_T) \quad \text{и} \quad H(Y_T/X_T). \\ H(Y_T) = - \int_{L_{Y_1}} \dots \int_{L_{Y_m}} W(y_1 \dots y_m) \times \\ \times \log W(y_1 \dots y_m) dy_1 \dots dy_m. \end{aligned} \quad (7.74)$$

Так как y_1, \dots, y_m независимые случайные величины, то

$$W(y_1 \dots y_m) = W(y_1) \cdot \dots \cdot W(y_m).$$

Интегрируя, находим

$$H(Y_T) = \sum_{k=1}^m H(Y_k), \quad (7.75)$$

где

$$H(Y_k) = - \int_{L_{y_k}} W(y_k) \log W(y_k) dy_k$$

— энтропия i -го опроса воспроизведения.

Для стационарных и квазистационарных функций $W(y_k)$ для всех k одинаковы. Следовательно,

$$H(Y_1) = H(Y_2) = \dots = H(Y_m) = H(Y)$$

и тогда (7.74) можно записать

$$H(Y_T) = mH(Y). \quad (7.75)$$

Определим условную энтропию

$$H(Y_T/X_T) = \int_{L_{x_1}} \dots \int_{L_{x_m}} \int_{L_{y_1}} \dots \int_{L_{y_m}} W(x_1 \dots x_m; y_1 \dots y_m) \times \\ \times \log W(y_1 \dots y_m/x_1 \dots x_m) dx_1 \dots dx_m, dy_1 \dots dy_m.$$

В силу независимости рассматриваемых случайных величин

$$W(x_1 \dots x_m; y_1 \dots y_m) = W(x_1 y_1) \cdot \dots \cdot W(x_m y_m) \\ \text{и} \quad W(y_1 \dots y_m/x_1 \dots x_m) = W(y_1/x_1) \cdot \dots \cdot W(y_m/x_m).$$

Учитывая эти соотношения, получим

$$H(Y_T/X_T) = \sum_{k=1}^m H(Y_k/X_k),$$

где

$$H(Y_k/X_k) = - \int_{L_{x_k}} \int_{L_{y_k}} W(x_k y_k) \log W(y_k/x_k) dx_k dy_k. \quad (7.77)$$

Для стационарно связанных процессов $W(x_k y_k)$ и $W(y_k/x_k)$ для всех k одинаковы. Поэтому

$$H(Y_1/X_1) \dots H(Y_m/X_m) = H(Y/X)$$

и, следовательно,

$$H(Y_T/X_T) = mH(Y/X), \quad (7.78)$$

Поскольку

$$I(Y_T; X_T) = H(Y_T) - H(Y_T/X_T),$$

учитывая (7.76), (7.78), получим

$$I(Y_T/X_T) = mH(Y) - mH(Y/X).$$

Разделив это равенство на T , получим

$$\bar{I}(Y_T/X_T) = \bar{H}(Y) - \bar{H}(Y/X). \quad (7.79)$$

Соотношение (7.79) определяет скорость создания сообщений на выходе непрерывного канала при сделанных допущениях о характере передаваемых сигналов и выборе частоты опроса. Очевидно, что скорость создания сообщения определяется функцией $W(y_i/x_i)$ и статистикой передаваемых сигналов. Поскольку $W(y_i/x_i)$ определяется свойствами канала, то варьируя $W(x)$ (статистику передаваемых сигналов), можно найти такую функцию распределения входных сигналов, при которой скорость создания сообщений будет наибольшей. Это позволяет получить выражение для пропускной способности непрерывного канала связи в виде соотношения

$$C_{\text{к.с.}} = \sup [\bar{H}(Y) - \bar{H}(Y/X)]. \quad (7.80)$$

§ 26. ОСНОВНАЯ ТЕОРЕМА ШЕННОНА ДЛЯ НЕПРЕРЫВНЫХ КАНАЛОВ

Аналогично теоремам для дискретных каналов для непрерывных каналов справедлива следующая теорема:

Если h -энтропия источника непрерывных сообщений, определяющая количество информации, вырабатываемое источником в единицу времени при заданной оценке вероятности воспроизведения q , равна

$$\bar{H}_h(U) = C_{\text{к.с.}} - \alpha,$$

где α как угодно мало, то существует метод передачи, при котором все сообщения, вырабатываемые источником, могут быть переданы, а верность воспроизведения при этом как угодно близка к q . Обратное утверждение состоит в том, что такая передача невозможна, если

$$\bar{H}_h(U) > C_{\text{к.с.}}$$

Доказательство этой теоремы аналогично доказательству для случая дискретного канала и его можно найти в работах по теории информации. А. Файнштейна. Эта теорема позволяет находить предельную эффективность непрерывных ка-

налов и с этой точки зрения оценивать методы передачи информации, используемые в реальных информационных каналах.

Контрольные вопросы и задания

1. Чем характеризуются источники непрерывных сообщений?
2. В каком случае информация, содержащаяся в одном измерении непрерывного сообщения, максимальна?
3. Чем определяется количество информации, содержащееся в непрерывной случайной величине?
4. Перечислите основные свойства количества информации, содержащегося в непрерывной случайной величине.
5. Как определяется энтропия непрерывной случайной величины?
6. Как определить количество информации, содержащееся в воспроизведении непрерывного сообщения?
7. В каком случае можно использовать понятия о пропускной способности непрерывных каналов с шумами?
8. Как определить скорость создания сообщений для непрерывного канала?
9. Как определить пропускную способность непрерывного канала связи?
10. Как формулируется основная теорема Шеннона для непрерывных каналов?

Глава 8

ЭЛЕМЕНТЫ ТЕОРИИ ИГР

Практические задачи принятия оптимальных решений в сложных динамических системах, где можно выделить две или более антогонистических сторон, преследующих противоположные цели, требуют создания математических методов, которые способны учитывать неопределенные и случайные факторы, сопровождающие непрерывные столкновения сторон.

Первой работой, в которой были сформулированы принципы научного анализа действия в так называемых конфликтных ситуациях, была книга Неймана и Моргенштерна «Теория игр и экономическое поведение», которая вышла в свет в 1944 году.

Ситуация может быть названа конфликтной, если при анализе ее можно выделить определенное количество соперничающих сторон, преследующих противоположные цели, причем результат действия каждой из сторон по направлению к цели зависит от того, какой образ действий выберет противник (противники).

Теория игр определяет рекомендации по рациональному образу действий каждого из противников в ходе конфликтной ситуации.

§ 1. ОСНОВНЫЕ ОПРЕДЕЛЕНИЯ

Формализованную модель конфликтной ситуации называют *игрой*.

Игра отличается от реальной конфликтной ситуации тем, что ведется по определенным правилам.

Если в игре можно выделить две антогонистические стороны, то *игра называется парной*. При большем числе антогонистических сторон игра называется *множественной*.

Для возможности математического анализа игры необходима точная договоренность о *правилах игры*, под которыми понимают систему условий, регламентирующих возможные варианты действий конфликтующих сторон, доступную информацию каждой из сторон о поведении остальных, последовательность чередования отдельных решений, принятых в процессе игры и результат, к которому приводит данная совокупность решений.

Игру можно классифицировать как *игру с нулевой суммой*, если одна из сторон выигрывает, то проигрывает другая.

Отдельные решения, принимаемые в процессе игры, называются *ходами*. При сознательном выборе решения из всех решений, предусмотренных правилами, говорят, что произведен *личный ход*.

Случайным ходом называется выбор решения под воздействием какого-либо механизма случайного выбора (бросание монеты, кости и т. п.).

Игрой с полной информацией называется игра, в которой каждая сторона во время каждого личного хода знает исходы всех предыдущих ходов (как личных, так и случайных). Возможные исходы случайных ходов задаются распределением вероятностей.

Совокупность правил, определяющих однозначно выбор решения при каждом личном ходе отдельного игрока в зависимости от ситуации, сложившейся в процессе игры, называют *стратегией*.

Для того, чтобы понятие стратегии имело смысл, необходимо наличие в игре личных ходов. В противном случае, т. е. в играх, состоящих из одних случайных ходов, стратегии нет.

Игра называется *конечной*, если у каждой из конфликтующих сторон есть только конечное число стратегий, и наоборот, при наличии у каждой из сторон бесконечного числа стратегий игра классифицируется как *бесконечная*.

Если у одной из конфликтующих сторон в запасе n стратегий ($A_1; A_2; \dots; A_j; \dots; A_n$), а у другой — m ($B_1; B_2; \dots; B_i; \dots; B_m$), то в случае, когда игра состоит только из личных ходов, выбор пары стратегий $A_j B_i$ единственным образом определяет исход игры a_{ji} .

Исход игры, в которой кроме личных ходов есть случайные, определяется математическим ожиданием.

Если нам известны значения α_{ji} при каждой паре стратегий $A_j B_i$, то можно составить так называемую *платежную*

матрицу, или матрицу игры, которая для игры ($m \times n$) будет иметь следующий вид:

$\begin{array}{c} B \\ \diagdown \\ A \end{array}$	B_1	B_2		B_i		B_m
A_1	α_{11}	α_{12}		α_{1i}		α_{1m}
A_2	α_{21}	α_{22}		α_{2i}		α_{2m}
\vdots	\vdots		\dots	\vdots	\dots	\vdots
A_j	α_{j1}	α_{j2}		α_{ji}	\dots	α_{jm}
\vdots	\vdots	\vdots	\dots	\vdots	\dots	\vdots
A_n	α_{n1}	α_{n2}	\dots	α_{ni}	\dots	α_{nm}

Сокращенно обозначим эту матрицу в виде (α_{ji}) .

Оптимальной стратегией называется стратегия, которая при многократном повторении игры обеспечивает определенной конфликтующей стороне максимально возможный средний выигрыш.

Рассмотрим элементарный пример игры. (Пример заимствован из книги Е. С. Вентцель «Элементы теории игр»).

Пусть в нашем распоряжении три вида вооружения: A_1 ; A_2 ; A_3 ; у противника три вида самолетов B_1 ; B_2 ; B_3 . Известно, что при применении вооружения A_1 самолеты B_1 ; B_2 ; B_3 поражаются с вероятностями 0,9; 0,4; 0,2; при применении вооружения A_2 — с вероятностями 0,3; 0,6; 0,8; при применении вооружения A_3 — с вероятностями 0,5; 0,7; 0,2. Требуется сформулировать данную конфликтную ситуацию в терминах теории игр.

Решение.

Конфликтную ситуацию можно рассматривать как игру 3×3 с двумя личными ходами и одним случайным. Личный ход противника — выбор типа самолета для участия в бою. Наш личный ход — выбор типа вооружения. Случайный ход — применение вооружения. Этот ход может закончиться поражением или непоражением самолета.

Нашими стратегиями являются три варианта применения типов вооружения. Стратегиями противника — три варианта применения типов самолета. Наш выигрыш равен единице, если самолет поражен, и нулю в противном случае. Среднее значение выигрыша при каждой заданной паре стратегий есть вероятность поражения самолета данным оружием.

Матрица игры имеет вид

A \ B	B		
	B_1	B_2	B_3
A_1	0,9	0,4	0,2
A_2	0,3	0,6	0,8
A_3	0,5	0,7	0,2

Таким образом, под конечной игрой с нулевой суммой двух конфликтующих сторон понимают набор стратегий и соответствующую ему матрицу выигрышей.

Будем считать, что конечная игра с нулевой суммой двух конфликтующих сторон задана, если перечислены все возможные стратегии каждого стратега (хотя бы в виде двух рядов чисел) и дана матрица выигрышей, соответствующая этим стратегиям.

Исследовать, или решить игру — означает найти для каждого стратега наилучшие стратегии в том смысле, что применение их обеспечивает каждому из стратегов наилучший выигрыш из возможных: кроме того, необходимо найти этот наилучший выигрыш.

Рассмотрим игру, заданную следующей матрицей:

$A \backslash B$	B_1	B_2	\dots	B_l	\dots	B_m
A_1	α_{11}	α_{12}	\dots	α_{1l}	\dots	α_{1m}
A_2	α_{21}	α_{22}	\dots	α_{2l}	\dots	α_{2m}
\vdots	\vdots	\vdots	\dots	\vdots	\dots	\vdots
A_j	α_{j1}	α_{j2}	\dots	α_{jl}	\dots	α_{jm}
\vdots	\vdots	\vdots	\dots	\vdots	\dots	\vdots
A_n	α_{n1}	α_{n2}	\dots	α_{nl}	\dots	α_{nm}

Пусть требуется найти оптимальную стратегию стратега (игрока первой из конфликтующих сторон) A .

Проанализируем последовательно каждую из его стратегий, начиная с A_1 .

Если стратег A выбрал из всех своих стратегий стратегию A_j , то он должен рассчитывать на то, что стратег B ответит на нее той из своих стратегий, для которой его (стратега A) выигрыш α_{ji} минимален. Выберем это значение выигрыша, т. е. минимальное из чисел α_{ji} в j -ой строке

$$a_j = \min_i \alpha_{ji}. \quad (8.1)$$

Избегая всякого риска, стратег A должен остановиться на той из стратегий A_j , для которой число a_j является максимальным

$$a = \max_j a_j,$$

или, учитывая (8.1)

$$a = \max_j \min_i \alpha_{ji}. \quad (8.2)$$

Выражение (8.2) определяет нижнюю цену игры (максимальный выигрыш, максимин). Стратегия, соответствующая ниж-

ней цене игры, называется *максиминной стратегией*. Выражение (8.2) определяет тот гарантированный минимум, который получит стратег A , придерживаясь наиболее осторожной из своих стратегий.

Очевидно, стратег B заинтересован в том, чтобы обратить выигрыш стратега A в минимум. Следовательно, он должен проанализировать каждую из своих стратегий с точки зрения максимального выигрыша при этом стратегии

$$b_i = \max_j \alpha_{ji}, \quad (8.3)$$

$$b = \min_i b_i,$$

$$b = \min_i \max_j \alpha_{ji}. \quad (8.4)$$

Величина, определяемая выражением (8.4), называется *верхней ценой игры (минимакс)*. Соответствующая стратегия называется *минимаксной*.

Принцип, по которому каждый из стратегов должен придерживаться своей наиболее осторожной стратегии в теории игр, называется *принципом минимакса*.

Учитывая полученные соотношения (8.1, 8.3), дополним рассматриваемую матрицу столбцом справа, каждый из элементов которого составлен в соответствии с выражением (8.1), и строкой снизу, составленной по формуле (8.3)

$A \backslash B$	B_1	B_2	...	B_l	...	B_m	a_j
A_1	α_{11}	α_{12}	...	α_{1l}	...	α_{1m}	a_1
A_2	α_{21}	α_{22}	..	α_{2l}	...	α_{2m}	a_2
.
.
.
A_j	α_{j1}	α_{j2}	...	α_{jl}	...	α_{jm}	a_j
.
.
.
A_n	α_{n1}	α_{n2}	...	α_{nl}	...	α_{nm}	a_n
b_i	b_1	b_2	...	b_l	...	b_m	

Аналогично дополним матрицу рассмотренного примера

A \ B	B			
	B_1	B_2	B_3	a_j
A_1	0,9	0,4	0,2	0,2
A_2	0,3	0,6	0,8	0,3
A_3	0,5	0,7	0,2	0,2
b_i	0,9	0,7	0,8	

Таким образом, нижняя цена игры рассматриваемого примера $a = 0,3$; верхняя цена игры $b = 0,7$.

Анализируя полученную матрицу, можно сделать вывод о *неустойчивости минимаксных стратегий*. Это означает, что если стратег A применяет свою наиболее осторожную стратегию A_2 , а стратег B — стратегию B_2 (свою наиболее осторожную стратегию), то средний выигрыш равен 0,6.

Как только стратегию B становится известно, что стратег A применяет стратегию A_2 , он может ответить на нее стратегией B_1 и сведет выигрыш к 0,3.

Следовательно, выигрыш при пользовании минимаксными стратегиями является неустойчивым, поскольку зависит от сведений о стратегии антагонистической стороны.

Существуют игры, для которых минимаксные стратегии являются устойчивыми. Для таких игр нижняя цена игры равна верхней и это общее значение называется чистой ценой игры. Элемент матрицы, являющийся одновременно минимальным в своей строке и максимальным в своем столбце, называется *седловой точкой матрицы*. А игра, платежная матрица которой имеет седловую точку, называется *игрой с седловой точкой*.

Седловой точке соответствует пара оптимальных стратегий, совокупность которых называется решением игры.

Таким образом, для игр с седловой точкой существует решение, определяющее пару оптимальных стратегий обеих из конфликтующих сторон.

Если каждая из конфликтующих сторон придерживается своих оптимальных стратегий, то средний выигрыш равен чистой цене игры.

Если одна из конфликтующих сторон придерживается своей оптимальной стратегии, а другая отклоняется от своей, то от этого уклоняющаяся сторона может только проиграть и ни в коем случае не увеличить свой выигрыш.

При рассмотрении конечных игр, не имеющих седловой точки, не может не возникнуть вопроса — нельзя ли гарантировать себе средний выигрыш, больший a , если применять не одну единственную чистую стратегию, а чередовать некоторым случайным образом несколько стратегий.

Смешанной стратегией называется набор вероятностей применения чистых стратегий.

Если, например, один стратег применяет свои чистые стратегии $1, 2 \dots j \dots n$ с вероятностями $p_1; p_2; \dots; p_j; \dots; p_n$, то его фиксированная смешанная стратегия есть этот набор вероятностей, и его можно обозначить буквой P , т. е.

$$P = (p_1; p_2; \dots; p_j; \dots; p_n),$$

где число n показывает количество чистых стратегий данного стратега.

Аналогично для другого стратега фиксированная смешанная стратегия

$$Q = (q_1; q_2; \dots; q_i; \dots; q_m).$$

Так как каждый раз применение одной из чистых стратегий исключает применение другой, то чистые стратегии являются несовместимыми событиями. Кроме того, поскольку есть возможность применения только чистых стратегий, то они являются единственно возможными событиями. Следовательно

$$\sum_{j=1}^n p_j = 1,$$

$$\sum_{i=1}^m q_i = 1.$$

Если один стратег применяет смешанную оптимальную стратегию P , а другой — Q , и матрица игры имеет вид

A \ B	B					
	B_1	B_2	...	B_l	...	B_m
A_1	α_{11}	α_{12}	...	α_{1l}	...	α_{1m}
A_2	α_{21}	α_{22}	...	α_{2l}	...	α_{2m}
...
A_j	α_{j1}	α_{j2}	...	α_{jl}	...	α_{jm}
...
A_n	α_{n1}	α_{n2}	...	α_{nl}	...	α_{nm}

то математическое ожидание выигрыша стратега A

$$M_1 = \sum_{j=1}^n \sum_{i=1}^m \alpha_{ji} p_j q_i = \sum_{i=1}^m q_i \left(\sum_{j=1}^n \alpha_{ji} p_j \right). \quad (8.5)$$

Выражение (8.5) можно преобразовать. В результате получим

$$M_1 = \sum_{j=1}^n p_j \left(\sum_{i=1}^m \alpha_{ji} q_i \right). \quad (8.5, a)$$

— математическое ожидание выигрыша стратега A при использовании обоими стратегами фиксированных смешанных стратегий P и Q соответственно.

Если обозначить

$$P^* = (p_1^*; p_2^*; \dots; p_l^*; \dots; p_n^*)$$

и

$$Q^* = (q_1^*; q_2^*; \dots; q_l^*; \dots; q_m^*)$$

как две произвольные смешанные стратегии соответственно стратегов A и B , то

$$M_2 = \sum_{i=1}^n \sum_{j=1}^m \alpha_{ji} p_j^* q_i = \sum_{i=1}^m q_i \left(\sum_{j=1}^n \alpha_{ji} p_j^* \right) \quad (8.6)$$

будет математическим ожиданием выигрыша стратега A , использующего стратегию P^* , при условии, что стратег B использует стратегию Q .

При использовании стратегом A стратегии P , а стратегом B — стратегии Q^* , математическое ожидание выигрыша стратега A

$$M_3 = \sum_{j=1}^n \sum_{i=1}^m \alpha_{ji} p_j q_i^* = \sum_{i=1}^m q_i^* \left(\sum_{j=1}^n \alpha_{ji} p_j \right). \quad (8.7)$$

Таким образом, если определить M_1 ; M_2 и M_3 выражениями (8.5); (8.6); (8.7) соответственно и, кроме того,

$$M_2 \leq M_1 \leq M_3, \quad (8.8)$$

то P и Q называются *оптимальными смешанными стратегиями* соответственно стратегов A и B .

Из неравенств (8.8), определяемых выражениями (8.5); (8.6); (8.7), можно сделать о свойствах смешанных стратегий следующие выводы, которые здесь приведем без доказательств.

1. Если второй стратег (B) применяет свою оптимальную смешанную стратегию, то математическое ожидание выигрыша первого стратега будет наибольшим тогда, когда первый стратег (A) также применит свою оптимальную смешанную стратегию.

2. Если первый стратег применяет свою оптимальную смешанную стратегию, то математическое ожидание выигрыша второго стратега (выигрыша первого стратега) будет наименьшим тогда, когда второй стратег также применит свою оптимальную стратегию.

3. Если математическое ожидание выигрыша первого стратега (при условии, что он применяет оптимальную смешанную стратегию, а второй — чистую стратегию) больше цены игры, то эта чистая стратегия применяется с вероятностью нуль в оптимальной смешанной стратегии второго стратега.

4. Если математическое ожидание выигрыша первого стратега (при условии, что второй стратег применяет свою оптимальную смешанную стратегию, а первый чистую стратегию) меньше цены игры, то эта чистая стратегия входит

с вероятностью нуль в его оптимальную смешанную стратегию.

5. Каждое из выражений, стоящее в скобках в выражении (8.7), должно быть не меньше цены игры Γ , а каждое из выражений, стоящее в скобках в выражении (8.5, а), должно быть не больше цены игры Γ . Другими словами, справедливы следующие неравенства:

$$\left. \begin{aligned} \sum_{j=1}^n \alpha_{j1} p_j &\geq \Gamma \\ \dots \dots \dots \\ \sum_{j=1}^n \alpha_{jm} p_j &\geq \Gamma \\ \sum_{i=1}^m \alpha_{1i} q_i &\leq \Gamma \\ \dots \dots \dots \\ \sum_{i=1}^m \alpha_{ni} q_i &\leq \Gamma \end{aligned} \right\} \quad (8.9)$$

Таким образом, формально задача нахождения цены игры и оптимальных смешанных стратегий сводится к решению системы неравенств (8.9) с учетом, что

$$\begin{aligned} p_j &\geq 0, & (j = 1, 2, \dots, n); \\ q_i &\geq 0, & (i = 1, 2, \dots, m); \\ \sum_{j=1}^n p_j &= 1; & \sum_{i=1}^m q_i = 1. \end{aligned} \quad (8.10)$$

6. Если ко всем элементам матрицы выигрышей некоторой игры прибавить (или вычесть) одно и то же число, то оптимальные смешанные стратегии не изменятся, а цена игры увеличится или уменьшится на это число.

7. Если каждый элемент матрицы выигрышей умножить на положительное число, то оптимальные смешанные стратегии игры не изменятся, а цена игры умножится на это число. Свойство 5 и 6 оказываются верными и для игр, имеющих седловую точку.

Контрольные вопросы

1. Как в терминах теории игр определить следующую конфликтную ситуацию: два стратега A и B , не глядя друг на друга, кладут на стол по монете вверх лицевой стороной или вверх обратной стороной, по своему усмотрению. Если стратеги выложили монеты одинаковыми сторонами, то стратег A забирает обе монеты, иначе их забирает B .

2. Что такое нижняя и верхняя цена игры?
3. В чем заключается суть принципа минимакса?
4. Что такое оптимальная стратегия?
5. Чем характерна седловая точка платежной матрицы?
6. Что такое смешанная стратегия?
7. Перечислите основные свойства смешанных стратегий.

§ 2. ИГРЫ 2×2

Свойства оптимальных смешанных стратегий, изложенные в предыдущем параграфе, дают возможность находить решение игры с матрицей любого порядка. Для этого необходимо решить линейные неравенства (8.9) и (8.10). Однако, используя свойства 3, 4 и 5, можно применять более простой метод нахождения оптимальных смешанных стратегий и цены игры с матрицей 2×2 .

Пусть игра задана матрицей

A \ B	B	
	B_1	B_2
A_1	α_{11}	α_{12}
A_2	α_{21}	α_{22}

Предположим, что игра не имеет седловой точки и

$$\alpha_{22} > \alpha_{12}.$$

Согласно свойству 5, изложенному в предыдущем параграфе, оптимальные смешанные стратегии игроков A и B

$$P = (p_1; p_2); \quad Q = (q_1; q_2)$$

и цена игры Υ должны удовлетворять следующим неравенствам:

$$\left. \begin{aligned} \alpha_{11}p_1 + \alpha_{21}p_2 &\geq \Upsilon, \\ \alpha_{12}p_1 + \alpha_{22}p_2 &\geq \Upsilon, \\ p_1 &\geq 0; \quad p_2 \geq 0; \quad p_1 + p_2 = 1; \end{aligned} \right\} \quad (8.11)$$

$$\left. \begin{aligned} \alpha_{11}q_1 + \alpha_{12}q_2 &\leq \Upsilon, \\ \alpha_{21}q_1 + \alpha_{22}q_2 &\leq \Upsilon, \\ q_1 &\geq 0; \quad q_2 \geq 0; \quad q_1 + q_2 = 1. \end{aligned} \right\} \quad (8.12)$$

В выражениях (8.11) и (8.12) знаки неравенств нигде нельзя заменить знаками строгих неравенств.

Действительно, пусть

$$\alpha_{12}p_1 + \alpha_{22}p_2 > \Upsilon. \quad (8.13)$$

Тогда согласно свойству 3, изложенному в предыдущем параграфе, получим

$$q_1 = 0 \quad \text{и} \quad q_2 = 1.$$

С другой стороны,

$$\alpha_{22} > \alpha_{12}.$$

А это означает, что α_{22} является седловой точкой. Однако по условию игра не имеет седловой точки и, следовательно, предположение (8.13) неверно.

Если $\alpha_{12} > \alpha_{22}$ и справедливо соотношение (8.13), то α_{12} должно быть седловой точкой.

Таким образом, всякая замена неравенств в выражении (8.11) и (8.12) приводит либо к противоречию, либо к требованию существования седловой точки.

Следовательно, для всякой игры 2×2 оптимальные смешанные стратегии и цена игры должны удовлетворять следующим условиям:

$$\left. \begin{aligned} \alpha_{11}p_1 + \alpha_{21}p_2 &= \Upsilon, \\ \alpha_{12}p_1 + \alpha_{22}p_2 &= \Upsilon, \\ p_1 \geq 0; \quad p_2 \geq 0; \quad p_1 + p_2 &= 1; \end{aligned} \right\} \quad (8.14)$$

$$\left. \begin{aligned} \alpha_{11}q_1 + \alpha_{12}q_2 &= \Upsilon, \\ \alpha_{21}q_1 + \alpha_{22}q_2 &= \Upsilon, \\ q_1 \geq 0; \quad q_2 \geq 0; \quad q_1 + q_2 &= 1. \end{aligned} \right\} \quad (8.15)$$

Если известны вероятности p_1 и p_2 , то из первого уравнения системы (8.14) следует правило нахождения цены игры.

Найдем вероятность p_1 из системы (8.14). Для этого вычтем первое уравнение из второго

$$(\alpha_{12} - \alpha_{11})p_1 + (\alpha_{22} - \alpha_{21})p_2 = 0. \quad (8.16)$$

Отсюда

$$\frac{p_1}{p_2} = \frac{\alpha_{22} - \alpha_{21}}{\alpha_{12} - \alpha_{11}}. \quad (8.17)$$

Учитывая

$$p_1 \geq 0 \quad \text{и} \quad p_2 \geq 0,$$

равенство (8.17) перепишем в следующем виде:

$$\frac{p_1}{p_2} = \frac{|\alpha_{22} - \alpha_{21}|}{|\alpha_{12} - \alpha_{11}|}. \quad (8.18)$$

Далее, выражая

$$p_2 = 1 - p_1 \quad (8.19)$$

и подставляя (8.19) в формулу (8.18), получим

$$\frac{p_1}{1 - p_1} = \frac{|\alpha_{22} - \alpha_{21}|}{|\alpha_{12} - \alpha_{11}|}. \quad (8.20)$$

Из выражения (8.20)

$$p_1 = \frac{|\alpha_{22} - \alpha_{21}|}{|\alpha_{12} - \alpha_{11}| + |\alpha_{22} - \alpha_{21}|}. \quad (8.21)$$

Рассуждая аналогично, найдем

$$p_2 = \frac{|\alpha_{12} - \alpha_{11}|}{|\alpha_{12} - \alpha_{11}| + |\alpha_{22} - \alpha_{21}|}. \quad (8.22)$$

Выполнение подобных операций при анализе системы (8.15) дает возможность находить вероятности q_1 и q_2 .

Докажем, что если стратег A применяет свою оптимальную смешанную стратегию, то математическое ожидание его выигрыша равно цене игры независимо от действия стратега B .

Действительно, математическое ожидание выигрыша стратега A , использующего свою оптимальную смешанную стратегию P ,

$$M = q_1^* (\alpha_{11} p_1 + \alpha_{21} p_2) + q_2^* (\alpha_{12} p_1 + \alpha_{22} p_2).$$

Учитывая (8.14) и (8.15), получим

$$M = q_1^* Y + q_2^* Y = (q_1^* + q_2^*) Y = Y.$$

Таким образом, при любых значениях вероятностей q_1^* и q_2^* величина M остается равной цене игры.

Аналогичные рассуждения можно провести и для случая, когда стратег B применяет свою оптимальную смешанную стратегию.

§ 3. ИГРЫ $2 \times m$ и $n \times 2$

Пусть имеется игра порядка $n \times m$, где n обозначает число стратегий первого стратега (A) и m — число стратегий второго стратега (B).

Если каждый элемент матрицы выигрышей одной строки (столбца) больше соответствующего элемента другой строки (столбца) или равен ему, то говорят, что первая стратегия *доминирует* над второй.

Если первая стратегия (A_1) стратега A доминирует над второй (A_2) и первая стратегия (B_1) стратега B доминирует над второй (B_2), то доминирующая стратегия стратега A лучше и поэтому ее оставляют в матрице выигрышей, а доминирующая стратегия стратега B хуже, поэтому ее вычеркивают из матрицы.

Пусть игра задана следующей матрицей:

A \ B	B			
	B_1	B_2	B_3	B_4
A_1	0	2	3	-1
A_2	-3	3	4	9
A_3	2	4	4	8

В этой игре A_3 доминирует над A_1 и больше нет доминирования стратегий стратега A .

У стратега B следующие соотношения доминирования: B_2 над B_1 ; B_3 над B_2 и B_3 над B_1 .

Здесь интересно отметить, что если доминирует вторая стратегия над первой и третья над второй, то обязательно третья стратегия доминирует и над первой. Это свойство называется *транзитивностью* и справедливо для любых трех стратегий.

Знание соотношения доминирования в случае игр порядка $2 \times m$ и $n \times 2$ достаточно для того, чтобы свести эти игры к играм порядка 2×2 .

Можно показать, что в любой игре порядка $2 \times m$ и $n \times 2$ всегда можно найти не больше двух полезных стратегий для каждого стратега, а остальные стратегии нельзя использовать в оптимальной смешанной стратегии.

Для аналитического решения игры порядка $2 \times m$ или $n \times 2$ необходимо выделить одну или две *активные стратегии* (стратегии, которые входят с положительной вероятностью) из m и n стратегий каждого из стратегов в зависимости от порядка игры.

Для этого нужно исследовать от одной до $\frac{m(m-1)}{2}$ или $\frac{n(n-1)}{2}$ игр порядка 2×2 в зависимости от удачи.

Пусть, например, игра 2×6 задана следующей матрицей выигрышей:

A \ B						
	B_1	B_2	B_3	B_4	B_5	B_6
A_1	5	0	2	5	8	-5
A_2	-4	-1	7	4	-1	8

Эта игра не имеет седловой точки, поэтому для ее решения надо найти оптимальные смешанные стратегии обоих игроков и цену игры.

Выделим сначала доминирующие стратегии стратега B . B_3 , B_4 и B_5 доминируют над стратегией B_2 и никакая другая стратегия не обладает этим свойством. Следовательно, матрицу игры можно преобразовать к следующему виду:

A \ B			
	B_1	B_3	B_4
A_1	5	0	-5
A_2	-4	-1	8

Выделим теперь активные стратегии. Вариантов, подлежащих анализу, может быть не больше трех. Начнем с первых двух стратегий B_1 и B_2 . Матрица игры сводится к игре 2×2

A \ B		
	B_1	B_2
A_1	5	0
A_2	-4	-1

Очевидно, что эта игра имеет седловую точку и чистую цену, равную 0. Следовательно, оптимальные смешанные

стратегии обоих игроков в игре 2×3 соответственно должны быть

$$P = (1, 0); \quad Q = (0, 1, 0).$$

Проверим справедливость полученного решения. Так как оптимальная смешанная стратегия стратега A — A_1 , а оптимальная смешанная стратегия стратега B состоит в применении только чистой стратегии B_2 , то эта игра должна иметь седловую точку (A_1, B_2) . Другими словами, элемент 0 должен быть минимумом первой строки и максимумом второго столбца. Однако, это противоречит действительности (поскольку минимум первой строки равен -5). Таким образом, предположение о том, что первые две стратегии стратега B активны, не оправдалось.

Предположим, что активными являются стратегии B_2 и B_3 , тогда получим игру 2×2 с матрицей

A \ B	B	
	B_2	B_3
A_1	0	-5
A_2	-1	8

Эта игра не имеет седловой точки. Находим, что

$$p_1 = \frac{9}{14}; \quad p_2 = \frac{5}{14}; \quad q_2 = \frac{13}{14}; \quad q_3 = \frac{1}{14};$$

$$\Gamma = -\frac{5}{14}.$$

Учитывая дополнительно, что $q_1 = 0$, проверим справедливость полученного решения. С этой целью для первого столбца матрицы 2×3 составим выражение

$$5p_1 - 4p_2 = 5 \cdot \frac{6}{14} - 4 \cdot \frac{5}{14} = \frac{25}{14}.$$

Таким образом, если стратег B применяет свою первую чистую стратегию, то математическое ожидание выигрыша стратега A равно $\frac{25}{14}$, и это число должно быть большим или равным цене игры. Поскольку

$$\frac{25}{14} > -\frac{5}{14},$$

то решение исходной игры состоит в следующем.

Цена игры равна $\left(-\frac{5}{14}\right)$, оптимальные смешанные стратегии стратега A и стратега B соответственно равны

$$P = \left(-\frac{9}{14}; \frac{5}{14}\right); \quad Q = \left(0; \frac{13}{14}; 0; 0; 0; \frac{1}{14}\right).$$

Заметим, что если бы мы не воспользовались соотношением доминирования, то число пар стратегий в исходной игре, для которых необходимо установить активные они или нет, возросло бы до $\frac{6 \times 5}{2} = 15$.

Изложим графический метод решения игр, который позволяет исследовать игры нагляднее и быстрее.

Хотя известно, что для игр 2×2 графический метод не дает существенного облегчения по сравнению с аналитическим, для уяснения сути метода без громоздких объяснений рассмотрим предварительно матрицу 2×2

A \ B	B	
	B_1	B_2
A_1	2	3
A_2	7	1

Решить игру — значит найти $p_1; p_2; q_1; q_2; \Upsilon$, удовлетворяющие соотношениям:

$$\left. \begin{aligned} 2p_1 + 7p_2 &= \Upsilon, \\ 3p_1 + p_2 &= \Upsilon, \\ p_1 \geq 0; \quad p_2 \geq 0; \quad p_1 + p_2 &= 1; \end{aligned} \right\} \quad (8.23)$$

$$\left. \begin{aligned} 2q_1 + 3q_2 &= \Upsilon, \\ 7q_1 + q_2 &= \Upsilon, \\ q_1 \geq 0; \quad q_2 \geq 0; \quad q_1 + q_2 &= 1. \end{aligned} \right\} \quad (8.24)$$

Или соотношениям:

$$\left. \begin{aligned} 2p_1 + 7(1 - p_1) &= \Upsilon, \\ 3p_1 + (1 - p_1) &= \Upsilon, \\ 0 \leq p_1 \leq 1; \end{aligned} \right\} \quad (8.25)$$

$$\left. \begin{aligned} 2q_1 + 3(1 - q_1) &= \Upsilon, \\ 7q_1 + (1 - q_1) &= \Upsilon, \\ 0 \leq q_1 \leq 1. \end{aligned} \right\} \quad (8.26)$$

Преобразуя последние системы, получим

$$\left. \begin{aligned} -5p_1 + 7 &= \Upsilon, \\ 2p_1 + 1 &= \Upsilon, \\ 0 &\leq p \leq 1; \end{aligned} \right\} \quad (8.27)$$

$$\left. \begin{aligned} -q_1 + 3 &= \Upsilon, \\ 6q_1 + 1 &= \Upsilon, \\ 0 &\leq q_1 \leq 1. \end{aligned} \right\} \quad (8.28)$$

Исследуем системы (8.25) и (8.27).

Из первого уравнения системы (8.25) следует: если $p_1 = 0$, то $\Upsilon = 7$; если $p_1 = 1$, то $\Upsilon = 2$.

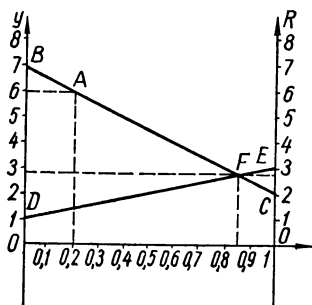


Рис. 8.1. Построение стратегий.

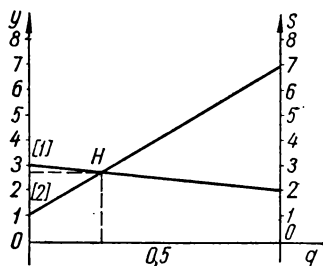


Рис. 8.2. Стратегии в игре 2×2 .

Графические построения иллюстрируются рис. 8.1. Построим точки $B(0; 7)$ и $C(1; 2)$. Соединим полученные точки прямой линией. Это построение соответствует случаю, когда стратег B применяет свою первую чистую стратегию.

Анализируя второе уравнение системы (8.25), построим линию DE . Это построение соответствует случаю, когда стратег B применяет свою вторую чистую стратегию. Точка пересечения F прямых BC и DE соответствует решению системы (8.25) или (8.27).

Таким образом, опуская из точки F перпендикуляр на ось p_1 , получим $p_1 \approx 0,85$ и $p_2 = 1 - p_1 \approx 0,15$. Опуская перпендикуляр из точки F на ось Υ , получим значение цены игры $\Upsilon \approx 2,7$. Если же окажется, что прямые не пересекаются, то игра имеет седловую точку, соответствующую нижней прямой.

Для нахождения q_1 и q_2 рассуждаем аналогично относительно систем (8.26) и (8.28) (рис. 8.2). Получим $q_1 \approx 0,3$; $q_2 \approx 0,7$; $\Upsilon \approx 2,7$.

Может показаться, что графический метод несколько громоздок, однако, если пользоваться полученными правилами как эмпирическими, отвлекаясь от анализа аналитических систем, то это противоречие будет снято. Рассмотрим это на примере игры $2 \times n$. Пусть задана матрица

$A \backslash B$	B_1	B_2	B_3	B_4	B_5	B_6	B_7
A_1	3	-4	2	-1	-3	5	1
A_2	1	2	-1	3	4	0	-3

Для каждой стратегии B построим соответственно прямые с 1 по 7 (рис. 8.3).

Например, чистой стратегии B_5 соответствует пятый столбец. На оси Y откладываем 4, а на оси X — (-3) , получаем прямую 5.

Так как каждая прямая обозначает величину проигрыша стратега B_1 , то, очертив нижнюю границу этих линий (жирная ломаная на рисунке), замечаем, что стратегию B не выгодно применять стратегии $B_1; B_3; B_4; B_5; B_6$, так как соответствующие им линии находятся выше нижней границы. Нижнюю границу образуют только две линии B_2 и B_7 , следовательно, эти стратегии и являются активными.

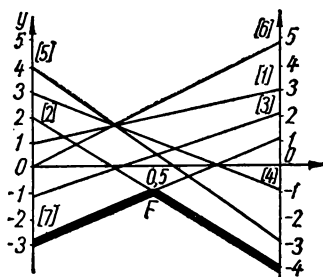


Рис. 8.3. Стратегии в игре $2 \times m$.

Точке пересечения этих линий соответствует $p_1 \approx 0,5$; $p_2 \approx 0,5$; $Y = -1$.

Кроме того, вероятности $q_1; q_3; q_4; q_5; q_6$ равны нулю. Для того, чтобы найти оптимальные стратегии стратега B , надо решить следующую игру 2×2 :

$A \backslash B$	B_1	B_2
A_1	-4	1
A_2	2	-3

Решая ее, получаем

$$p_1 \approx 0,5; \quad p_2 \approx 0,5; \quad q_1 \approx 0,4; \quad q_2 \approx 0,6; \quad \Gamma = -1.$$

Таким образом, решение игры

$$P = \left(\frac{1}{2}; \frac{1}{2} \right); \quad Q = \left(0; \frac{2}{5}; 0; 0; 0; 0; \frac{3}{5} \right); \\ \Gamma = -1.$$

§ 4. ИГРЫ $m \times n$

Для игр $m \times n$, где m и n могут быть любые, метод исследования по существу остается прежним и заключается в последовательном сведении исходной игры к игре меньшего порядка. Однако объем работы при этом может значительно возрасти.

Если игра $m \times n$ задана своей матрицей, которую обозначим буквой W , то ее подматрицей называется такая матрица, которая либо совпадает с W , либо может быть получена из W вычеркиванием некоторых строк и столбцов.

При решении игр порядка $m \times n$ следует иметь в виду, что для всякой конечной прямоугольной игры двух игроков с нулевой суммой и матрицей выигрышей W существует такая квадратная подматрица V , что решение игры с матрицей выигрышей V является одновременно решением исходной игры с матрицей W (в этом случае чистые стратегии игроков, соответствующие вычеркнутым строкам и столбцам при получении подматрицы V из матрицы W , входят в оптимальные смешанные стратегии с вероятностями нуль).

Отсюда вытекает такой *метод решения игр $m \times n$* : последовательно выбирают из матрицы W квадратные подматрицы V_i , решают игры с матрицами выигрышей V_i и проверяют, являются ли эти решения решениями исходной игры с матрицей W . Так делают до тех пор, пока не получат необходимое решение.

Таким образом, идея решения этих игр проста, однако количество вычислений, которое необходимо проделать, может оказаться очень большим.

Полезно бывает проверить с помощью неравенств (8.9) и (8.10) догадку о решении игры.

Прежде всего необходимо выяснить, имеет ли игра седловую точку. Если седловой точки нет, то надо по возможности уменьшить порядок игры, используя соотношение доминирования среди стратегий. Если это приведет к игре $2 \times m$ или $n \times 2$, то решение исходной игры найти сравнительно легко. В противном случае объем работы и метод выбора квадратных подматриц для исследования зависят

от навыка и искусства исследователя и вида матрицы исходной игры.

Пусть выбрана квадратная матрица $n \times n$

$A \backslash B$	B_1	B_2		B_i		B_n
A_1	α_{11}	α_{12}	...	α_{1i}	...	α_{1n}
A_2	α_{21}	α_{22}		α_{2i}	...	α_{2n}
\vdots	\vdots	\vdots	...	\vdots	...	\vdots
A_j	α_{j1}	α_{j2}	...	α_{ji}	...	α_{jn}
\vdots	\vdots	\vdots	...	\vdots	...	\vdots
A_n	α_{n1}	α_{n2}	...	α_{ni}	...	α_{nn}

Следует иметь в виду, что в любом случае решение игры удовлетворяет системам неравенств (8.9) и (8.10), которые преобразуются для матрицы $n \times n$ к виду

$$\left. \begin{aligned} \sum_{j=1}^n \alpha_{j1} p_j &= \gamma, \\ \dots \dots \dots \\ \sum_{j=1}^n \alpha_{jn} p_j &= \gamma, \\ \sum_{j=1}^n p_j &= 1, \end{aligned} \right\} \quad (8.29)$$

$$\left. \begin{aligned} \sum_{i=1}^n \alpha_{1i} q_i &= \gamma, \\ \dots \dots \dots \\ \sum_{i=1}^n \alpha_{ni} q_i &= \gamma, \\ \sum_{i=1}^n q_i &= 1. \end{aligned} \right\} \quad (8.30)$$

(В случае, когда все стратегии активные, неравенства превращаются в равенства).

Таким образом, нахождение решения игры сводится к решению двух систем линейных уравнений (8.29) и (8.30), причем все неизвестные $p_1 \dots p_n$ и $q_1 \dots q_n$ должны быть неотрицательными.

Каждая из систем содержит $n + 1$ неизвестных. Количество неизвестных можно понизить на одну, если каждое последующее уравнение вычесть из предыдущего за исключением последних. Тогда с помощью определителей можно легко найти решение системы линейных уравнений, т. е. уравнений, содержащих неизвестные $(p_1 \dots p_n; \text{ и } q_1 \dots q_n)$ первой степени.

Чтобы удобнее было составлять систему, содержащую n неизвестных из системы, содержащей $(n + 1)$ неизвестных, надо составлять новую матрицу по такому правилу: из каждого столбца матрицы выигрышей вычесть почленно непосредственно следующий за ним столбец и результат записать в виде столбцов новой матрицы.

Если каждый из стратегов имеет, по меньшей мере, одну неактивную стратегию, то при решении преобразованных систем получим, что либо уравнения какой-либо из систем несовместны, либо некоторые из значений неизвестных будут отрицательными. Ни то, ни другое не может служить решением игры.

Из систем (8.29) и (8.30) следует, что тогда, когда у каждого стратега все стратегии активные и кто-нибудь из стратегов применяет свою оптимальную смешанную стратегию, то другой стратег не может играть лучше или хуже. Другими словами, в этом случае при всякой игре другого стратега, первый стратег будет иметь один и тот же средний выигрыш.

§ 5. ПРИБЛИЖЕННЫЕ МЕТОДЫ РЕШЕНИЯ ИГР

При исследовании игровых ситуаций часто нет необходимости в точном решении игры. В таком случае можно воспользоваться приближенными методами решения конечных прямоугольных игр. Здесь мы рассмотрим один из численных методов решения игр — метод итераций.

Идея метода итераций сводится к следующему. Мысленно игру проигрывают много раз, то есть последовательно, в каждой партии игры каждый стратег выбирает такую последовательность своих чистых стратегий, которая обес-

печивает первому стратегу максимальный средний выигрыш, а второму — минимальный средний проигрыш.

Проиграв мысленно несколько партий, вычисляют математическое ожидание обеих выигрышей стратегов, и их среднее арифметическое принимают за цену игры.

Этот метод дает возможность также найти приближенное значение оптимальных смешанных стратегий обоих стратегов. Для этого необходимо подсчитать частоту применения каждой чистой стратегии и принять ее за приближенное значение вероятности использования этой чистой стратегии в оптимальной смешанной стратегии соответствующего стратега.

Можно доказать, что с неограниченным увеличением числа проигранных партий математические ожидания выигрыша первого стратега и проигрыша второго стратега будут стремиться к цене игры, а приближенные значения оптимальных смешанных стратегий в том случае, когда решение игры единственное, будет стремиться к оптимальным смешанным стратегиям каждого стратега.

Вообще говоря, сходимость итерационного алгоритма довольно медленная. Однако этот алгоритм легко автоматизировать, используя средства вычислительной техники, что позволяет получить решение игр при матрицах сравнительно высокого порядка.

Рассмотрим применение этого метода на примере.

Пусть игра задана матрицей

A \ B	B		
	B_1	B_2	B_3
A_1	8	2	4
A_2	4	5	6
A_3	1	7	3

В таблице 8.1. приведены результаты 18 шагов итерационного алгоритма. Здесь: n — номер исследуемой пары ходов; i — номер выбранной стратегии стратега A ; B_1, B_2, B_3 — выигрыш, накопленный за первые n игр стратегом B при стратегиях $B_1; B_2; B_3$. Минимальное из этих значений

Таблица 8.1

n	i	B_1	B_2	B_3	i	A_1	A_2	A_3	r_{\min}	r_{\max}	$r_{\text{ср}}$
1	3	<u>1</u>	7	3	1	<u>8</u>	4	1	1	8	4,50
2	1	9	9	<u>7</u>	3	<u>12</u>	10	4	3,50	6,00	4,75
3	1	17	<u>11</u>	11	2	14	<u>15</u>	11	3,67	5,00	4,33
4	2	21	<u>16</u>	17	2	16	<u>20</u>	18	4,0	5,00	4,50
5	2	25	<u>21</u>	23	2	18	<u>25</u>	25	4,20	5,00	4,60
6	2	29	<u>26</u>	29	2	20	30	<u>32</u>	4,33	5,33	4,82
7	3	<u>30</u>	33	32	1	28	<u>34</u>	33	4,29	4,86	4,57
8	2	<u>34</u>	38	38	1	36	<u>38</u>	34	4,25	4,75	4,50
9	2	<u>38</u>	43	44	1	<u>44</u>	42	35	4,23	4,89	4,56
10	1	46	<u>45</u>	48	2	46	<u>47</u>	42	4,50	4,70	4,60
11	2	<u>50</u>	50	54	1	<u>54</u>	51	43	4,55	4,91	4,72
12	1	58	<u>52</u>	58	2	56	<u>56</u>	50	4,33	4,66	4,49
13	2	62	<u>57</u>	64	2	58	<u>61</u>	57	4,38	4,70	4,54
14	2	66	<u>62</u>	70	2	60	<u>66</u>	64	4,43	4,71	4,56
15	2	70	<u>67</u>	76	2	62	71	<u>71</u>	4,47	4,73	4,60
16	3	<u>71</u>	74	79	1	70	<u>75</u>	72	4,44	4,69	4,56
17	2	<u>75</u>	79	85	1	78	<u>79</u>	73	4,41	4,65	4,53
18	2	<u>79</u>	84	91	1	<u>86</u>	83	74	4,39	4,78	4,53

подчеркнуто. Далее идет номер j стратегии, выбранной противником, и соответствующий выигрыш, накопленный за n игр при стратегиях $A_1; A_2; A_3$, из которых максимальные обозначены чертой сверху. Значения, обозначенные чертой, определяют выбор ответной стратегии.

В последующих столбцах стоят — $\bar{\gamma}_{\min}$ — минимальный средний выигрыш, который равен минимальному накопленному выигрышу, деленному на число игр; $\bar{\gamma}_{\max}$ — максимальный средний выигрыш; $\bar{\gamma}_{\text{ср}}$ — среднее арифметическое, равное $\frac{\bar{\gamma}_{\min} + \bar{\gamma}_{\max}}{2}$. Очевидно, что при увеличении числа

n все три величины $\bar{\gamma}_{\min}; \bar{\gamma}_{\max}; \bar{\gamma}_{\text{ср}}$ будут приближаться к цене игры, причем величина $\bar{\gamma}_{\text{ср}}$ наиболее быстро.

Следует отметить, что для итерационного алгоритма характерно незначительное возрастание объема и сложности вычислений с ростом числа стратегии m и n .

§ 6. МЕТОДЫ РЕШЕНИЯ НЕКОТОРЫХ БЕСКОНЕЧНЫХ ИГР

Бесконечной игрой называется игра, в которой по крайней мере одно из чисел m или n представляет собой бесконечное множество стратегий противников. Тогда стратегии стратегов A и B соответствуют различным значениям непрерывно меняющихся параметров (x и y). Следовательно, игра будет определяться некоторой непрерывной функцией двух аргументов $f = \alpha(x, y)$, называемой *функцией выигрыша*. В дальнейшем анализ игры в общем случае можно свести к анализу функции выигрыша, как это делалось для платежной матрицы в случае конечных игр.

Таким образом, нижняя цена бесконечной игры

$$a_0 = \max_x \min_y \alpha(x, y), \quad (8.31)$$

верхняя цена игры

$$b_0 = \min_y \max_x \alpha(x, y). \quad (8.32)$$

Если

$$a_0 = b_0,$$

говорят, что анализируемая бесконечная игра имеет седловую точку, причем значение f в этой точке и есть цена игры $\bar{\gamma}$. В этом случае анализируемая бесконечная игра имеет решение в области чистых стратегий. Если же $a_0 \neq b_0$, то игра имеет решение только в области смешанных

стратегий, определяемых как некоторые распределения вероятностей для стратегий x и y , которые рассматриваются как случайные величины.

Если задана бесконечная игра с функцией выигрыша $f = \alpha(x, y)$ и стратегиями x и y на отрезках (x_1, x_2) и (y_1, y_2) , то, предполагая что игра не имеет седловой точки, определим цену игры.

Для определения нижней цены игры спроектируем поверхность f на плоскость $xO\alpha$. Нижняя цена игры определяется максимальной ординатой кривой, ограничивающей снизу проекцию f на плоскость $xO\alpha$.

Для определения верхней цены игры нужно спроектировать поверхность f на плоскость $yO\alpha$ и найти минимальную ординату верхней границы проекции f на плоскость $yO\alpha$.

В заключение нужно отметить, что общие подходы к решению бесконечных игр еще мало разработаны.

Контрольные вопросы

1. В чем заключается итерационная процедура нахождения приближенного решения игр?
2. Что такое функция выигрыша?
3. В чем заключается общий подход к решению бесконечных игр?

Глава 9

ГРАФЫ

Графами называются геометрические схемы, представляющие собой системы точек и соединяющих линий.

Учение о графах соединяет исключительную геометрическую наглядность, математическую содержательность и возможность обходиться без громоздкого аппарата.

Теория графов зародилась в XVIII веке в связи с математическими головоломками. Эта тема долгое время рассматривалась как «несерьезная». К ней относились так же, как в первоначальный период развития относились к теории вероятностей, «прикладное» значение которой видели лишь в связи с играми и развлечениями.

В XX веке графами заинтересовались топологи. В дальнейшем выяснилось большое значение теории графов для решения большого числа практических задач (транспортных, составления расписаний, распределения потоков). Теория графов в настоящее время широко применяется в технике, экономике, биологии, психологии.

В математике теория графов рассматривается как ветвь топологии. Она имеет непосредственное отношение также к алгебре и теории чисел.

§ 1. ГРАФ. ПУТИ И КОНТУРЫ

Граф задан или определен, если даны:

- 1) непустое множество X ;
- 2) отображение Γ множества X в X .

Обозначается граф в виде $G = (X, \Gamma)$. В качестве примеров определенных графов могут служить: отношения подчинения в иерархии управления, правила шахматной игры, схемы соединений элементов и узлов в технической системе и т. д.

Элементы множества X обозначим точками плоскости и назовем *вершинами* графа. Пары точек x и y , для которых

$g \in \Gamma x$, соединяются непрерывной линией со стрелкой, направленной от x к y . Эти пары (x, y) назовем *дугами* графа. Множество дуг графа обозначим U , а сами дуги — u, v, w .

Пример. Множество X графа (рис. 9.1) образовано вершинами a, b, c, d, e, x , множество U — дугами (a, b) , (b, a) , (b, x) , (e, e) , (x, x) (e, x) (x, c) , (x, d) . Отображение Γ определяется в виде

$$\begin{aligned}\Gamma x &= \{x, \quad, d\}, \\ \Gamma d &= \emptyset \text{ и т. д.}\end{aligned}\tag{9.1}$$

Граф можно обозначить в виде $G = (X, U)$, так как множество дуг U вполне определяет отображение Γ .

Подграфом графа (X, Γ) назовем граф (A, Γ_A) , где $A \subset X$, а Γ_A определено как

$$\Gamma_A x = \Gamma x \cap A. \tag{9.2}$$

Частичным графом графа (X, Γ) назовем граф вида (X, Δ) , где $\Delta x \subset \Gamma x$ при всех $x \in X$.

Частичным подграфом графа (X, Γ) назовем граф вида (A, Δ_A) , где $A \subset X$ и $\Delta_A x \subset \Gamma x \subset A$.

Рис. 9.1. Ориентированный граф.

В качестве примера рассмотрим граф (X, U) , представляющий собой карту железных дорог страны: X — множество железнодорожных станций, и $(x, y) \in U$, если есть железная дорога (электрифицированная или неэлектрифицированная) между станциями x и y . Карта электрифицированных дорог определяет частичный граф, а карта всех дорог европейской части страны — подграф. Электрифицированные дороги европейской части определяют частичный подграф.

Говорят, что a и b являются *граничными* вершинами дуги $u = (a, b)$, причем a — начало, а b — конец дуги. Две дуги u и v называются смежными, если: 1) они различны; 2) имеют общую граничную точку (независимо от того, является ли эта точка началом или концом дуги u , началом или концом дуги v).

Две вершины x и y смежны, если: 1) они различны; 2) существует дуга, идущая от одной из них к другой.

Дуга u *исходит* из вершины x , если x является началом, но не является концом u . Дуга u *заходит* в x , если x является концом, но не является началом u ; в обоих случаях

дуга u называется *инцидентной* вершине x . Обобщим это понятие.

Если A — данное множество вершин, то говорят, что дуга исходит из A , если

$$u = (a, x), a \in A, x \notin A;$$

множество дуг, исходящих из A , обозначается символом U_A^+ . Аналогично определяется множество U_A^- дуг, заходящих в множество вершин A .

Множество дуг, инцидентных A , обозначается символом

$$U_A = U_A^+ \cup U_A^-.$$

В графе, представленном на рис. 9.1, для $A = \{a, x\}$ имеем

$$U_A^+ = \{(a, b), (x, c), (x, d)\}, U_A^- = \{(b, a), (c, x), (d, x)\}.$$

Путь в графе $G(X, U)$ называется такая последовательность дуг (u_1, u_2, \dots) , когда конец каждой предыдущей дуги совпадает с началом следующей. Путь является *простым*, если в нем никакая дуга не встречается дважды, и составным — в противном случае.

Путь μ , последовательными вершинами которого есть $x_1, x_2, \dots, x_k, x_{k+1}$, можно обозначить символом $\mu = [x_1, x_2, \dots, x_k, x_{k+1}]$; путь, идущий от x_2 к x_k по тем же дугам, что и μ , обозначим

$$\mu[x_2, x_k] = [x_2, x_3, \dots, x_k].$$

Путь, в котором никакая вершина не встречается дважды, называется *элементарным*. Путь этот может быть конечным и бесконечным.

Контур — это конечный путь $\mu = [x_1, x_2, \dots, x_k]$, у которого начальная вершина x_1 совпадает с конечной x_k ; при этом контур называется *элементарным*, если все его вершины различны (за исключением начальной и конечной, которые совпадают). Длина пути $\mu = (u_1, u_2, \dots, u_k)$ есть число $l(\mu) = k$ дуг последовательности; в случае бесконечного пути μ полагаем $l(\mu) = \infty$. Наконец, контур длины 1, образованный дугой вида (x, x) , называется *петлей*.

П р и м е р. Иерархия в структуре автоматизированной системы управления. Пусть X — множество пунктов сбора, обработки и передачи информации и пусть Γx — множество пунктов низшего уровня, непосредственно подчиненных пункту x высшего уровня. Связь любого высшего пункта с любым низшим изобразится в виде пути графа (X, Γ) ;

важно, чтобы граф не имел контуров, так как их наличие может привести к несогласованным, а иногда и противоречивым командам.

Определим некоторые важные категории графов с помощью понятий дуги, пути, контура.

Граф $G(X, U)$ называется *симметрическим*, если

$$(x, y) \in U \rightarrow (y, x) \in U.$$

В симметрическом графе две смежные вершины x и y всегда соединены двумя противоположно ориентированными дугами. Для упрощения изображения в этом случае обычно соединяют обе точки одной непрерывной линией без стрелок.

Граф (X, U) называется *антисимметрическим*, если

$$(x, y) \in U \rightarrow (y, x) \notin U$$

(каждая пара смежных вершин соединена только в одном направлении; петли отсутствуют).

Граф (X, U) называется *полным*, если

$$(x, y) \in U \rightarrow (y, x) \in U$$

(любые две вершины соединены хотя бы в одном направлении).

Граф называется *сильно связным*, если для любых двух вершин x и y ($x \neq y$) существует путь, идущий из x в y .

П р и м е р. Схема коммуникаций. Пусть X — множество людей и $(x, y) \in U$, когда лицо x имеет возможность непосредственно передавать сообщения лицу y . Обычно этот граф симметрический, например, если связь осуществляется с помощью телефона, телеграфа, радио. Но граф может и не быть симметрическим, как в случае ракет или почтовых голубей.

В правильно спроектированной сети коммуникаций каждый человек должен иметь возможность передать сообщение любому другому члену организации непосредственно или через посредников. Таким образом, важно, чтобы граф был сильно связан.

§ 2. ЦЕПИ И ЦИКЛЫ

Ребром графа $G(X, U)$ называется множество из двух элементов x и y , для которых $(x, y) \in U$ или $(y, x) \in U$. Понятие ребра не следует путать с понятием дуги, в котором участвует ориентация. Например, граф, изображенный на рис. 9.1, имеет 8 дуг, но только 6 ребер. Ребро обо-

значают жирной латинской буквой ***u*** или ***v***, а множество ребер — ***V***. Ребро, для которого вершины *x* и *y* граничные, обозначается ***v*** = [*x*, *y*].

Цепь — это последовательность ребер (***v***₁, ***v***₂, ...), в которой у каждого ребра ***v***_{*k*} одна из граничных вершин является также граничной вершиной для ***v***_{*k*-1}, а другая — граничной вершиной для ***v***_{*k*+1}. Цепь называется *простой*, если все ее ребра различны, и *составной* — в противном случае.

Цикл — это конечная цепь, начинающаяся в некоторой вершине *x* и оканчивающаяся в той же вершине *x*. Цикл называется *простым*, если все его ребра различны, и *составным* — в противном случае. Цикл, при обходе которого ни одна вершина не встречается дважды, называется *элементарным*. Граф *связен*, если любые две его различные вершины можно соединить цепью. Сильно связный граф связан, но обратное утверждение неверно.

Обозначим через *C_a* множество, состоящее из данной вершины *a* и всех тех вершин графа, которые могут быть соединены с ней цепью; *компонента связности* (или просто *компонента*) — это подграф, порожденный множеством типа *C_a*.

Теорема о разбиениях. *Различные компоненты графа (X, Γ) образуют разбиение множества X, т. е.:*

- (1) *C_a* ≠ ∅;
- (2) *C_a* ≠ *C_b* ⇒ *C_a* ∩ *C_b* = ∅;
- (3) ∪ *C_a* = X.

Так как *a* ∈ *C_a*, то (1) имеет место. Чтобы доказать (2), предположим

$$C_a \cap C_b \neq \emptyset$$

и покажем, что в этом случае *C_a* = *C_b*.

Пусть *x* ∈ *C_a* ⊂ *C_b*; вершину *x* можно соединить цепями как с *a*, так и с *b*; поэтому *a* можно соединить с *b*, т. е. *b* ∈ *C_a*. Значит, *C_b* ⊂ *C_a*.

Точно так же имеем *C_a* ⊂ *C_b* (в силу симметрии), следовательно, *C_a* = *C_b*.

(3) справедливо потому, что

$$X \supset \bigcup_{a \in X} C_a \supset \bigcup_{a \in X} \{a\} = X,$$

откуда

$$\bigcup C_a = X.$$

Теорема о связности. *Граф связан в том и только в том случае, если он состоит из единственной компоненты. Если*

в графе две различные компоненты C_a и C_b , то он несвязен, так как вершины a и b нельзя соединить цепью.

Если граф несвязен, то найдутся две вершины a и b , которые невозможно соединить цепью, и, значит, C_a и C_b будут различными компонентами.

Граф можно рассматривать либо с учетом, либо без учета ориентации его дуг. В первом случае мы имеем дело с понятиями дуги, пути, контура, сильной связности. Во втором — с понятиями ребра, цепи, цикла, связности. Обычно, если для некоторого понятия, определяемого в терминах дуг графа, имеется параллельное понятие, определяемое в терминах ребер, второе образуется из первого посредством добавления суффикса «оид» (например, центр — центроид).

§ 3. КВАЗИУПОРЯДОЧЕННОСТЬ

Квазипорядок, определяемый графом

Отношение \leq на множество X есть квазипорядок, если имеют место:

(1) рефлексивность: $x \leq x$ (для всех $x \in X$);

(2) транзитивность: $x \leq y, y \leq z \rightarrow x \leq z; x, y, z \in X$.

Примеры квазипорядков: «целое число x делится на y »; « x — вещественное число, большее или равное y »; « x — ситуация, которую предпочитают или считают равносильной ситуации y ».

На графе $G = (X, \Gamma)$ отношение \leq введем следующим образом: для двух вершин x и y пишем $x \leq y$, если $x = y$ или существует путь из x в y (другими словами, если $y \in \Gamma x$).

Это отношение \leq , удовлетворяющее, очевидно, условиям (1) и (2), называется *квазипорядком, определяемым графом G* .

Если $x \leq y$, то говорят, что вершина x *предшествует* вершине y . Отношение $y \leq x$ можно записать также в виде $x \geq y$ и сказать, что вершина x *следует* за вершиной y . Если $x \leq y$ и $y \leq x$, то пишем $x \equiv y$ и говорим, что вершина x эквивалентна вершине y ; в этом случае x и y совпадают или лежат на одном и том же контуре. Здесь \equiv является отношением эквивалентности, т. е. имеют место:

(1) рефлексивность: $x \equiv x$;

(2) транзитивность: $x \equiv y; y \equiv z \rightarrow x \equiv z$;

(3) симметрия: $x \equiv y \rightarrow y \equiv x$.

Если $x \leq y$, но не $x \equiv y$, то пишем $x < y$ и говорим, что вершина x *строго предшествует* вершине y ; пишут также $y > x$ и говорят, что вершина y *строго следует* за x .

Вершина z , следующая за всеми вершинами некоторого множества $B \subset X$, называется *мажорантой* множества B ; Это можно записать так:

$$z \geq b (b \in B) \text{ или } \forall b z \geq b.$$

Если в B имеется элемент, который служит мажорантой B , то этот элемент называется *наибольшим элементом* множества B . Если b и b' — два наибольших элемента множества B , то $b \leq b'$ и $b' \leq b$, следовательно, $b' \equiv b$, т. е. все наибольшие элементы одного и того же множества эквивалентны.

Аналогично, *минорантой* B называется такая вершина z , для которой

$$z \leq b (b \in B) \text{ или } \forall b z \leq b.$$

Миноранта B , принадлежащая B , называется *наименьшим элементом* множества B . На базе этих понятий характеризуются различные категории графов.

Пример 1. Граф без контуров. Если (X, Γ) не содержит контуров, то определяемое им отношение \leq обладает свойствами

$$x \leq y, y \leq x \rightarrow x = y.$$

В этом случае говорят, что \leq есть *отношение порядка*. Наоборот, если отношение \leq , определяемое графом, есть порядок, то граф не имеет контуров.

В качестве примера рассмотрим множество X и некоторое семейство \mathfrak{F} его подмножеств; назовем диаграммой Хассе семейства \mathfrak{F} граф G , вершинами которого служат различные множества семейства \mathfrak{F} , причем из $F \in \mathfrak{F}$ в $F' \in \mathfrak{F}$ идет дуга, если:

- 1) $F \subset \subset F'$;
- 2) не существует такого множества $F'' \in \mathfrak{F}$, что $F \subset \subset F'' \subset \subset F'$.

Квазипорядок, определяемый графом G , есть не что иное, как отношение \subset ; это отношение — порядок: G не содержит контуров. Было бы интересно выяснить при $X = \{1, 2, \dots, n\}$, сколько семейств \mathfrak{F} допускают данную диаграмму Хассе. Эта задача называется *проблемой Рейни*. Для $n > 5$ она еще не решена.

Пример 2. Транзитивный граф. Граф (X, U) называется *транзитивным*, если

$$x \leq y \rightarrow (x, y) \in U.$$

Например, если вершины графа изображают людей, а дуги — иерархическое превосходство (например, старшинство), то граф транзитивный.

Пример 3. Тотальный граф. Отношение квазиупорядка \leq называется тотальным, когда для любой пары (x, y) имеет место или $x \leq y$, или $y \leq x$. Граф является тотальным, если связанный с ним квазиупорядок — тотальный.

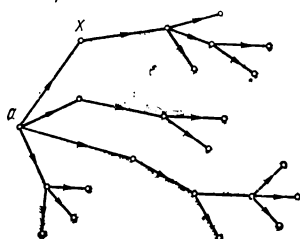


Рис. 9.2. Граф — дерево.

Справедливо утверждение: *всякий сильно связанный граф является тотальным.*

Пример 4. Структурный граф и прадеревья. Рассмотрим отношение порядка \leq , определяемое графом $G(X, \Gamma)$ без контуров. Если множество мажорант некоторого множества $B \subset X$ имеет наименьший элемент c , то c называется *верхней гранью* множества B . Отношение порядка \leq называется структурным (сверху), если каждое множество B допускает верхнюю грань. В этом случае говорят, что граф $G(X, \Gamma)$ является *верхней структурой*. Аналогично определяются понятия *нижней грани* и *нижней структуры*.

Частным случаем структурного графа является так называемое *прадерево с корнем a* , например, рис. 9.2.

§ 4. ИНДУКТИВНЫЙ ГРАФ И БАЗЫ

Множество $B \subset X$ называется базой графа $G(X, \Gamma)$, если оно удовлетворяет следующим двум условиям:

- 1) $b_1 \in B, b_2 \in B, b_1 \neq b_2 \rightarrow b_1 \leq b_2 \wedge b_2 \geq b_1$,
- 2) $x \in B \rightarrow \exists b \in B, b \geq x$.

Пример. Рассмотрим граф рис. 9.3. Множество $B = \{b_1, b_2, \dots\}$ — база этого графа. Подграф, порожденный множеством X/B , не имеет базы.

Существует большое количество задач, связанных с информационными потоками, в которых используется понятие базы графа. Это же понятие участвует во многих задачах теории игр.

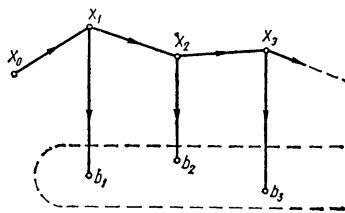


Рис. 9.3. База графа.

Граф называется *индуктивным*, если каждый путь в нем $\mu = [x_1, x_2, \dots]$ допускает мажоранту, т. е. если для каждого μ существует такая вершина z , что

$$z \geq x_n \quad (n = 1, 2, \dots).$$

Пример. Всякий конечный граф индуктивный. Действительно, если путь μ конечный, то мажорантой служит его последняя вершина. Если же μ бесконечный, то по крайней мере одна из вершин этого пути повторяется бесконечно много раз и поэтому является его мажорантой.

§ 5. ТРАНСПОРТНЫЕ СЕТИ

Задача о наибольшем потоке

Транспортной сетью называется конечный граф без петель, каждой дуге u которого отнесено целое число $c(u) > 0$, называемое пропускной способностью дуги u , и у которого:

1) существует одна и только одна такая вершина x_0 , что $\Gamma_{x_0}^{-1} = \emptyset$. Эта вершина называется *входом* сети;

2) существует одна и только одна такая вершина z , что $\Gamma_z = \emptyset$. Эта вершина называется *выходом* сети.

Пусть U_x^+ — множество дуг, заходящих в x , а U_x^- — множество дуг, исходящих из x . Функция $\varphi(u)$ определенная на U и принимающая целочисленные значения, представляет собой поток по этой транспортной сети, если

$$(1) \quad \varphi(u) \geq 0; \quad (u \in U).$$

Из (2) непосредственно следует

$$(2) \quad \sum_{u \in U_x^+} \varphi(u) - \sum_{u \in U_x^-} \varphi(u) = 0; \quad (x \neq x_0, \quad x \neq z);$$

$$(3) \quad \varphi(u) \leq c(u) \quad (u \in U).$$

Функцию $\varphi(u)$ можно интерпретировать как количество некоторого вещества, протекающего (в единицу времени) по дуге $u = (x, y)$ от x к y , причем это количество не превышает пропускной способности дуги, и в каждой вершине x , отличной от входа x_0 и выхода z , количество протекающего вещества равно количеству вытекающего.

$$\sum_{u \in U_z^-} \varphi(u) = \sum_{u \in U_{x_0}^+} \varphi(u) = \varphi_z.$$

Число $\varphi(z)$ — количество протекающего в z вещества, называется *притоком в точке z* , или *величиной потока φ* . Интересной является задача определения наибольшей величины потока по данной транспортной сети.

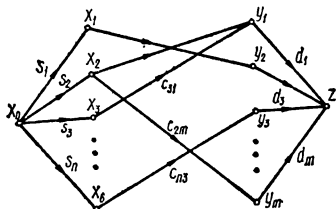


Рис. 9.4. Транспортная сеть.

Пример. Товарооборот между поставщиками и потребителями различных отраслей (рис. 9.4). Предприятия x_1, x_2, \dots производят продукцию, имеющую спрос в пунктах y_1, y_2, \dots . Пусть запас продукции в x_i равен s_i , а потребность в y_j равна d_j . Обозначим через c_{ij} полное количество продукции, которое в состоянии быть перевезено из x_i в y_j установленными транспортными средствами. Возможно ли удовлетворить все потребности? Как организовать перевозки?

Эта задача сводится к задаче о наибольшем потоке. Соединим x_i с y_j дугой пропускной способности c_{ij} , затем соединим вход x_0 с каждой вершиной x_i дугой пропускной способности s_i и, наконец, соединим каждую из вершин y_j с выходом дугой пропускной способности d_j . Если φ — наибольший поток, то $\varphi(x_i, y_j)$ будет означать количество продукции, которое надо перевезти из x_i в y_j , чтобы удовлетворить потребности в наибольшей степени.

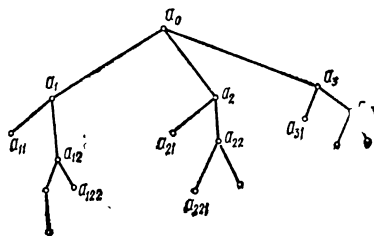


Рис. 9.5. Неориентированный граф — дерево.

§ 6. ДЕРЕВЬЯ И ЛЕСА

Деревом называется связный граф, не содержащий циклов. Из определения дерева вытекает, что для каждой пары его вершин существует единственная соединяющая их цепь (рис. 9.5). Поскольку дерево не имеет циклов, разные цепи (или ветви), выходящие из a_0 , будут изолированы друг от друга. Каждая ветвь такого графа должна иметь последнее, или *конечное, ребро с конечной вершиной*, из которой уже не выходит ни одного нового ребра.

Простейшее дерево имеет только одно ребро. Каждый раз, когда мы добавляем еще одно ребро в конце ветви, прибавляется также и вершина, следовательно, справедливы теоремы:

Теорема 1. *Дерево с n вершинами имеет $n - 1$ ребер.* Рассмотрим граф, состоящий из k связных компонент, каждая из которых представляет собой дерево. Такие графы называются лесами. Для каждой из компонент число ребер на единицу меньше числа вершин. Следовательно, справедлива:

Теорема 2. *Лес, состоящий из k компонент и имеющий n вершин, содержит $n - k$ ребер.*

Процесс сортировки перфокарт, распределения карточек при каталогизации, составление словарей и энциклопедий, распределение оборудования по предприятиям, цехам и участкам и многие аналогичные задачи описываются графами типа деревьев.

Задача о линиях связи

Большое практическое значение имеет задача о средствах сообщения. Имеется некоторое количество пунктов сбора и переработки информации, которые нужно соединить сетью коммуникаций. Для каждой пары пунктов A , B известна стоимость $C(A, B)$ строительства соединяющей их линии. Задача состоит в том, чтобы построить самую дешевую из возможных сетей. В частном случае, когда имеется всего три пункта A , B , C , достаточно построить одну из линий

$$ABC, ACB, BAC,$$

причем если AC — самая дорогостоящая линия, то именно ее и надо исключить, построив линию ABC .

В общем случае граф наиболее дешевой соединяющей сети должен быть деревом, так как если бы он содержал цикл, можно было бы удалить одно из звеньев этого цикла и пункты остались бы соединенными. Следовательно, для соединения n пунктов нужна $n - 1$ линия.

Сеть минимальной стоимости строится по следующему правилу экономичности. Вначале соединяем два пункта с наиболее дешевым соединяющим звеном e_1 . На каждом из следующих шагов добавляем самое дешевое из звеньев e_i , присоединение которого к уже построенным ребрам не образует никакого цикла. Если имеется несколько звеньев

одной и той же стоимости, выбираем любое из них. Каждое дерево, построенное таким образом, назовем экономичным деревом. Его стоимость равна сумме стоимостей отдельных звеньев

$$C(F) = C(\varepsilon_1) + C(\varepsilon_2) + \dots + C(\varepsilon_{n-1}).$$

Установление соответствий

Задача о назначении на должности. Имеется несколько различных вакантных должностей и группа лиц, стремящихся их занять, причем каждый из претендентов обладает достаточной квалификацией для нескольких, но не всех должностей. Требуется определить, можно ли представить каждому из этих людей одну из тех должностей, для которых он подходит?

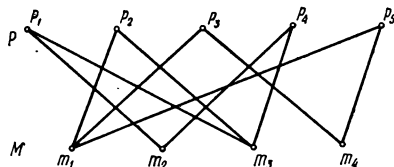


Рис. 9.6. Граф соответствий.

Проиллюстрировать эту задачу можно при помощи графа, имеющего в этом случае специальный вид. Группу претендентов обозначим через M , а множество мест или должностей — P .

Построим граф, проводя ребра (m, p) , соединяющие каждого претендента m из множества M с теми должностями p из множества P , которые он может занять. На таком графе не будет ребер, соединяющих между собой вершины $m \in M$ или соединяющих между собой вершины $p \in P$ (рис. 9.6). Подобные графы, множества вершин которых распадаются на подмножества M и P такие, что никакие две вершины из одного и того же подмножества, не соединенные между собой ребрами, называются *двудольными графами*.

Конечно, не всегда можно найти подходящую работу для каждого претендента. Для этого, во всяком случае, необходимо, чтобы число мест было не меньше числа людей, но и последнее условие еще не является достаточным. Например, группа претендентов состоит из двух электриков и одного бухгалтера. Имеется одна должность для электрика и три места для бухгалтеров. Очевидно, что один из электриков не найдет работы, хотя в этом случае мест больше, чем претендентов, хотя среди последних и есть представители обеих требуемых профессий.

Пусть общее число лиц, желающих получить работу, равно N . Для разрешимости задачи о назначениях на долж-

ности должно выполняться следующее условие. Какую бы группу из k человек, $k = 1, 2, \dots, N$, мы ни взяли, должно найтись по крайней мере k должностей, каждую из которых может занять хотя бы один из k претендентов (т. е. такие k работ, что с каждой из них справится наша группа k лиц).

Круговые соответствия

В различных соревнованиях приходится объединять пары участников. В турах с выбываниями это осуществляется просто.

В случае так называемых круговых турниров эта задача сложнее. Здесь нужно обеспечить обоюдные встречи (попарные) для всех участников, и важно заранее подготовить таблицу. Подобную ситуацию можно изобразить в виде графа N игроков. Каждый из них играет $N - 1$ игру с остальными участниками. Каждую игру представим ребром $[A, B]$. Вся совокупность игр представится при этом полным графом с N вершинами. На рис. 9.7 показан граф для $N = 6$. В каждом туре игроки объединяются в пары. При N четном объединении в пары соответствует выбору $\frac{1}{2} N$ несмежных ребер по одному для каждой из вершин.

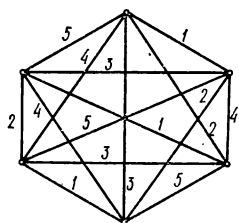


Рис. 9.7. Граф круговых соответствий.

Для следующего тура выбирается новое множество из $\frac{1}{2} N$ ребер и так далее, пока все игры не будут сыграны. При большом числе игроков составление такой таблицы для всех туров весьма трудоемко, если не указан какой-либо систематический метод.

Тензорному исчислению в настоящее время уделяется все большее внимание среди специальных разделов математики, читаемых во вузах. Тензорное исчисление необходимо для изложения механики сплошных сред, кристаллографии, некоторых разделов теоретической физики, электродинамики.

С точки зрения науки об управлении и связи весьма перспективной является идея использования аппарата тензорного исчисления для описания структур и функционирования больших систем, в частности автоматизированных систем управления (АСУ), и построения новой теории информации на основе понятия «информационного поля»¹.

В этой главе кратко изложены элементы тензорной алгебры и тензорного анализа. Все построения проводятся для простоты и наглядности в трехмерном пространстве. Используются только ортогональные системы координат.

В главе не рассматриваются такие вопросы, как приложения тензорного исчисления к дифференциальной геометрии, теории относительности. Для дальнейшего углубления и совершенствования знаний в области тензорного исчисления можно воспользоваться литературой, приведенной в конце книги.

§ 1. ЛИНЕЙНОЕ ПРОСТРАНСТВО

Понятие линейного пространства

Свободным вектором называется направленный отрезок, который можно переносить в пространстве параллельно его первоначальному положению. Обычно свободные векторы обозначают жирными буквами латинского алфавита: ***a, b, ..., x, y, ...***. Можно считать, что все они имеют общую точку *O* — начало координат.

¹ Лапа В. Г. Нестационарное тензорное поле информации. Вестник КПИ, 1971, № 8.

Совокупность всех векторов пространства замкнута относительно операций:

- а) сложение $x + y$;
- б) умножение вектора x на действительное число λ — λx . Замкнутость означает, что при умножении вектора на число снова получается некоторый вектор и при сложении двух векторов — некоторый третий вектор из этой же совокупности.

Свойства операций:

1. $x + y = y + x$.
2. $(x + y) + z = x + (y + z)$.
3. Существует нулевой вектор 0 такой, что $x + 0 = x$.
4. Для каждого вектора x существует *противоположный вектор* $y = -x$ такой, что $x + y = 0$.
5. $1 \cdot x = x$.
6. $\lambda(\mu x) = (\lambda\mu)x$.
7. $(\lambda + \mu)x = \lambda x + \mu x$.
8. $\lambda(x + y) = \lambda x + \lambda y$.

Операции сложения и умножения с указанными выше свойствами могут быть определены не только для совокупности векторов пространства. Существуют и другие множества элементов, на которых определены аналогичные операции. Эти множества называются *линейными* (или *векторными*) *пространствами*. Их обозначим буквой L . Элементы таких пространств будем также называть *векторами*.

П р и м е р ы:

а) Совокупность векторов, лежащих на одной прямой, образует линейное пространство, так как сложение и умножение таких векторов на действительное число приводит нас снова к векторам, лежащим на этой прямой, и свойства 1—8 легко проверяются. Обозначим такое линейное пространство через L_1 .

б) Совокупность векторов, лежащих в одной плоскости, также замкнута по отношению к сложению и умножению на действительное число. Свойства 1—8 для них выполняются, и потому указанная совокупность представляет собой линейное пространство, которое обозначается через L_2 .

в) Совокупность всех векторов 3-мерного пространства также является линейным пространством. Обозначим его через L_3 .

г) Совокупность векторов, лежащих в плоскости xOy , начала которых совпадают с началом координат, а концы лежат в первом квадранте, не образует линейного пространства, так как оказывается незамкнутой относительно умножения на число: при $\lambda < 0$ вектор λx не принадлежит заданному квадранту.

д) Рассмотрим множество, элементами которого являются упорядоченные совокупности n действительных чисел:

$$x = \{x_1, x_2, \dots, x_n\}, \quad y = \{y_1, y_2, \dots, y_n\}, \dots$$

Определим операции сложения и умножения на действительное число λ с помощью равенств

$$x + y = \{x_1 + y_1, x_2 + y_2, \dots, x_n + y_n\},$$

$$\lambda x = \{\lambda x_1, \lambda x_2, \dots, \lambda x_n\}.$$

Такое множество элементов образует линейное пространство. Определенные выше операции удовлетворяют свойствам 1—8. Например, нулевым вектором будет вектор $0 = \{0, 0, \dots, 0\}$, а вектором $-x$ — вектор $\{-x_1, -x_2, \dots, -x_n\}$. Это пространство обозначим L_n .

е) Совокупность всех многочленов степени не выше n

$$P(t) = a_0 + t + \dots + a_n t^n,$$

для которых обычным образом определены сложение и умножение на действительное число, также образует линейное пространство.

ж) Множество непрерывных на отрезке $[a, b]$ функций $f(t)$ также образует линейное пространство $C[a, b]$, если для этих функций естественным образом задать операции сложения и умножения на число.

Линейным подпространством линейного пространства называется непустое подмножество L' векторов из L , которые сами образуют линейное пространство относительно уже введенных в L операций сложения и умножения на число, т. е. такое подмножество L' , для которого из того, что $x \in L'$, $y \in L'$, следует, что $x + y \in L'$, $\lambda x \in L'$.

Простейшими подпространствами пространства L являются подпространство, состоящее из одного нулевого элемента (нулевое подпространство), и все пространство L . Эти подпространства называются *несобственными*.

Суммой двух линейных подпространств L' и L'' линейного пространства L называется совокупность $M = L' + L''$ всех векторов из L , каждый из которых представляется

в виде $x = x' + x''$, где

$$x' \in L' \text{ и } x'' \in L''.$$

Пересечением двух линейных подпространств L' и L'' линейного пространства L называется совокупность

$$N = L' \cap L''$$

всех векторов из L , каждый из которых принадлежит как L' , так и L'' .

Линейная зависимость векторов

Пусть a, b, \dots, e — векторы линейного векторного пространства L , а $\alpha, \beta, \dots, \epsilon$ — действительные числа. Вектор

$$x = \alpha a + \beta b + \dots + \epsilon e$$

называется *линейной комбинацией* векторов a, b, \dots, e а числа $\alpha, \beta, \dots, \epsilon$ — коэффициентами этой линейной комбинации.

Векторы a, b, \dots, e называются *линейно зависимыми*, если найдутся такие действительные $\alpha, \beta, \dots, \epsilon$, не все равные нулю, что

$$\alpha a + \beta b + \dots + \epsilon e = 0.$$

Если же это равенство выполняется только тогда, когда все числа $\alpha, \beta, \dots, \epsilon$ равны нулю, то векторы a, b, \dots, e называются *линейно независимыми*.

Свойства линейно независимых векторов.

а) *Если векторы линейно зависимы, то один из них может быть представлен в виде линейной комбинации остальных, и наоборот, если один из векторов есть линейная комбинация остальных, то векторы линейно зависимы.*

В самом деле, пусть a, b, \dots, e — линейно зависимые векторы. Тогда

$$\alpha a + \beta b + \dots + \epsilon e = 0,$$

где не все коэффициенты равны нулю. Пусть, например, $\alpha \neq 0$. Тогда

$$a = -\frac{\beta}{\alpha} b - \dots - \frac{\epsilon}{\alpha} e,$$

что и доказывает теорему.

Наоборот, если

$$a = mb + \dots + pe,$$

то

$$1 \cdot \mathbf{a} + (-m) \mathbf{b} + \dots + (-p) \mathbf{e} = \mathbf{0},$$

т. е. векторы $\mathbf{a}, \mathbf{b}, \dots, \mathbf{e}$ линейно зависимы.

б) Если некоторые из векторов $\mathbf{a}, \mathbf{b}, \dots, \mathbf{e}$ линейно зависимы, то и вся эта система векторов линейно зависима.

Пусть линейно зависимы векторы \mathbf{a}, \mathbf{b} . Тогда

$$\alpha \mathbf{a} + \beta \mathbf{b} = \mathbf{0},$$

где хотя бы один из коэффициентов α, β отличен от нуля. Но тогда и

$$\alpha \mathbf{a} + \beta \mathbf{b} + 0 \cdot \mathbf{c} + \dots + 0 \cdot \mathbf{e} = \mathbf{0}.$$

Последнее равенство доказывает линейную зависимость векторов $\mathbf{a}, \mathbf{b}, \dots, \mathbf{e}$, так как среди коэффициентов линейной комбинации, стоящей в его левой части, имеются и коэффициенты, отличные от нуля.

в) Если среди векторов $\mathbf{a}, \mathbf{b}, \dots, \mathbf{e}$ имеется хотя бы один нулевой, то эти векторы линейно зависимы.

Пусть, например, $\mathbf{a} = \mathbf{0}$. Тогда

$$\alpha \mathbf{a} + 0 \cdot \mathbf{b} + \dots + 0 \cdot \mathbf{e} = \mathbf{0}, \quad \alpha \neq 0.$$

Рассмотрим примеры линейно зависимых и линейно независимых векторов.

а) Нулевой вектор $\mathbf{0}$ является линейно зависимым, так как $\alpha \cdot \mathbf{0} = \mathbf{0}$ при любом $\alpha \neq 0$ (это следует также из свойства б).

б) Любой вектор $\mathbf{a} \neq \mathbf{0}$ линейно независим, так как $\alpha \mathbf{a} = \mathbf{0}$ только при $\alpha = 0$.

в) Любые два коллинеарных вектора \mathbf{a} и \mathbf{b} линейно зависимы. Действительно, если $\mathbf{a} \neq \mathbf{0}$, то $\mathbf{b} = \lambda \mathbf{a}$ или $\lambda \mathbf{a} + (-1) \mathbf{b} = \mathbf{0}$. Если же $\mathbf{a} = \mathbf{0}$, то эти векторы линейно зависимы в силу свойства б.

г) Два неколлинеарных вектора линейно независимы. В самом деле, предположим обратное: пусть $\alpha \mathbf{a} + \beta \mathbf{b} = \mathbf{0}$, где $\beta \neq 0$. Тогда $\mathbf{b} = -\frac{\alpha}{\beta} \mathbf{a}$. А это означает, что векторы \mathbf{a} и \mathbf{b} коллинеарны.

д) Три компланарных вектора линейно зависимы. Пусть векторы $\mathbf{a}, \mathbf{b}, \mathbf{c}$ компланарны, причем векторы \mathbf{a}, \mathbf{b} не коллинеарны. Тогда вектор \mathbf{c} можно представить в виде $\mathbf{c} = \overrightarrow{OC} = \overrightarrow{OA} + \overrightarrow{OB} = \lambda \mathbf{a} + \mu \mathbf{b}$, что в силу свойства а) означает линейную зависимость векторов $\mathbf{a}, \mathbf{b}, \mathbf{c}$. Если же \mathbf{a} и \mathbf{b} коллинеарны, то они линейно зависимы, а поэтому в силу свойства б) и векторы $\mathbf{a}, \mathbf{b}, \mathbf{c}$ линейно зависимы.

е) любые четыре вектора 3-мерного пространства всегда линейно зависимы. Действительно, если какие-нибудь три вектора линейно зависимы, то согласно свойству б) и все четыре вектора будут линейно зависимы. Если же имеются три линейно независимых вектора \mathbf{a} , \mathbf{b} , \mathbf{c} , то любой четвертый вектор \mathbf{d} может быть представлен в виде линейной комбинации векторов \mathbf{a} , \mathbf{b} , \mathbf{c} (рис. 10.1).

$$\mathbf{d} = \overrightarrow{OD} = \overrightarrow{OP} + \overrightarrow{PD} = \overrightarrow{OB} + \overrightarrow{OA} + \overrightarrow{OC} = \alpha\mathbf{a} + \beta\mathbf{b} + \gamma\mathbf{c},$$

откуда в силу свойства а) следует линейная зависимость векторов \mathbf{a} , \mathbf{b} , \mathbf{c} , \mathbf{d} .

ж) В пространстве L_n линейно независимыми будут векторы

$$\mathbf{e}_1 = \{1, 0, \dots, 0\},$$

$$\mathbf{e}_2 = \{0, 1, \dots, 0\},$$

$$\mathbf{e}_n = \{0, 0, \dots, 1\}.$$

Действительно, рассмотрим их линейную комбинацию

$$\alpha_1\mathbf{e}_1 + \alpha_2\mathbf{e}_2 + \dots + \alpha_n\mathbf{e}_n = \{\alpha_1, \alpha_2, \dots, \alpha_n\}.$$

Эта комбинация равна нулю только в случае

$$\alpha_1 = \alpha_2 = \dots = \alpha_n = 0.$$

Система векторов пространства L_n , состоящая из векторов $\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n$ и произвольного вектора $\mathbf{x} = \{x_1, x_2, \dots, x_n\}$, будет линейно зависимой, так как вектор \mathbf{x} можно представить в виде

$$\mathbf{x} = x_1\mathbf{e}_1 + x_2\mathbf{e}_2 + \dots + x_n\mathbf{e}_n.$$

Размерность и базис линейного пространства

Размерностью линейного пространства называется наибольшее число имеющихся в нем линейно независимых векторов:

L_1 — одномерное линейное пространство — прямая;

L_2 — двумерное линейное пространство — плоскость;

L_n — n -мерное линейное пространство, содержащее n линейно независимых векторов $\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n$. Можно показать, что любые $n + 1$ векторов являются линейно зависимыми.

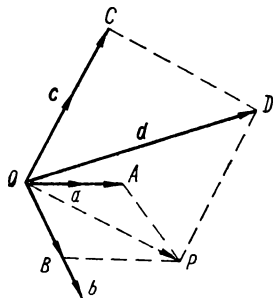


Рис. 10.1. Операции над векторами.

Рассмотрим произвольное n -мерное линейное пространство ($n = 1, 2, 3, \dots$) и выберем в нем любые n линейно независимых векторов e_1, e_2, \dots, e_n . Пусть x — произвольный вектор пространства. Тогда векторы

$$x, e_1, e_2, \dots, e_n$$

будут линейно зависимыми, так как их число превышает размерность пространства. Поэтому найдутся такие числа $\alpha, \alpha_1, \alpha_2, \dots, \alpha_n$, что

$$\alpha x + \alpha_1 e_1 + \alpha_2 e_2 + \dots + \alpha_n e_n = 0.$$

При этом $\alpha \neq 0$, так как иначе векторы e_1, \dots, e_n были бы линейно зависимыми. Тогда x можно представить линейной комбинацией векторов e_1, \dots, e_n :

$$x = -\frac{\alpha_1}{\alpha} e_1 - \frac{\alpha_2}{\alpha} e_2 - \dots - \frac{\alpha_n}{\alpha} e_n.$$

Положим

$$-\frac{\alpha_1}{\alpha} = x_1; \quad -\frac{\alpha_2}{\alpha} = x_2; \quad \dots \quad -\frac{\alpha_n}{\alpha} = x_n.$$

Теперь x представится в виде

$$x = x_1 e_1 + x_2 e_2 + \dots + x_n e_n. \quad (10.1)$$

Легко доказать, что разложение (10.1) единственно. Пусть, например, существует другое разложение

$$x = x'_1 e_1 + x'_2 e_2 + \dots + x'_n e_n.$$

Имеем

$$x_1 e_1 + x_2 e_2 + \dots + x_n e_n = x'_1 e_1 + x'_2 e_2 + \dots + x'_n e_n$$

и

$$(x_1 - x'_1) e_1 + (x_2 - x'_2) e_2 + \dots + (x_n - x'_n) e_n = 0.$$

Поскольку e_1, e_2, \dots, e_n линейно независимы, то

$$x_1 - x'_1 = x_2 - x'_2 = \dots = x_n - x'_n = 0.$$

Любой вектор x может быть единственным образом представлен в виде линейной комбинации линейно независимых векторов e_1, e_2, \dots, e_n .

Совокупность этих векторов называется базисом n -мерного линейного пространства, а числа x_1, x_2, \dots, x_n координатами вектора x в этом базисе.

Любые n линейно независимых векторов n -мерного пространства могут быть приняты за базис пространства.

В частности, на прямой L_1 любой вектор x можно представить в виде

$$x = x_1 e_1,$$

где e_1 — произвольный отличный от нуля вектор этой прямой.

На плоскости L_2 вектор x можно представить в виде

$$x = x_1 e_1 + x_2 e_2,$$

где e_1 и e_2 — любые два неколлинеарных вектора этой плоскости.

Разложение (10.1) можно записать в виде

$$x = \sum_{k=1}^n x_k e_k.$$

Однако и такая запись часто неудобна и ее упрощают, отбрасывая знак суммы $x = x_k e_k$, полагая, что по индексу k , повторяющемуся дважды, производится суммирование от 1 до n . Это правило называется «соглашением о суммировании», оно было предложено А. Эйнштейном. Индекс суммирования k может быть заменен любой другой буквой, так что

$$x_k e_k = x_i e_i = x_\alpha e_\alpha = \dots$$

Очевидно, что при заданном базисе векторы пространств L_1 , L_2 и L_3 вполне определяются своими координатами. Следовательно, эти пространства можно рассматривать как частные виды пространства L_n при $n = 1, 2, 3$.

Дальнейшие изложения, в основном, будем вести, ограничиваясь для наглядности лишь случаем плоскости или обычного 3-мерного пространства. При этом $n = 2$ или 3 и индексы суммирования пробегают соответственно значения 1 и 2 или 1, 2 и 3. Однако большая часть рассуждений справедлива и для общего линейного пространства n измерений.

Напомним, кроме того, известные свойства векторов:

а) если два вектора равны, то равны и их соответствующие координаты;

б) при сложении векторов их соответствующие координаты складываются;

в) при умножении вектора на число каждая его координата умножается на это число.

§ 2. ПРЯМОУГОЛЬНЫЙ БАЗИС В 3-МЕРНОМ ПРОСТРАНСТВЕ

Скалярное произведение векторов

Рассмотрим базис в L_3 , состоящий из трех попарно ортогональных единичных векторов e_1, e_2, e_3 . Такой базис называется *ортонормированным* (прямоугольным), а составляющие его векторы e_1, e_2, e_3 — *ортами*. Разложение произвольного вектора x по ортонормированному базису ничем не отличается от его разложения по произвольному базису и может быть записано в виде

$$x = x_i e_i.$$

Числа x_i теперь называются *прямоугольными* координатами вектора x . Базис e_1, e_2, e_3 называется *правым*, если из конца e_3 поворот на 90° от e_1 к e_2 виден против часовой стрелки. Если же этот поворот от e_1 к e_2 из конца e_3 виден по часовой стрелке, то базис называется *левым*.

Скалярное произведение векторов в аналитической геометрии записывается в виде

$$xy = |x| |y| \cos \varphi,$$

где $|x|$ и $|y|$ — длины заданных векторов, а φ — угол между ними.

Свойства скалярного произведения:

- 1) $xy = yx$;
- 2) $(\lambda x)y = \lambda(x, y)$;
- 3) $(x + y)z = xz + yz$;
- 4) $xx > 0$ при $x \neq 0$.

Скалярные произведения базисных векторов ортонормированного базиса определяются таблицей

	e_1	e_2	e_3
e_1	1	0	0
e_2	0	1	0
e_3	0	0	1

Введем величины δ_{ij} , определяемые равенствами

$$\delta_{ij} = \begin{cases} 1 & \text{при } i = j, \\ 0 & \text{при } i \neq j. \end{cases}$$

Тогда скалярные произведения базисных векторов могут быть записаны в виде

$$e_i e_j = \delta_{ij}.$$

Величины δ_{ij} носят название *симметричных символов Кронекера*.

Пусть $\mathbf{x}_i = x_i \mathbf{e}_i$ и $\mathbf{y}_i = y_i \mathbf{e}_i$ — два произвольных вектора пространства. Тогда

$$\mathbf{x}\mathbf{y} = (x_i \mathbf{e}_i)(y_j \mathbf{e}_j).$$

Из свойств 2) и 3) скалярного произведения получаем

$$\mathbf{x}\mathbf{y} = x_i y_j (\mathbf{e}_i \mathbf{e}_j).$$

Это сумма девяти слагаемых, так как i и j независимо друг от друга пробегают значения 1, 2, 3. Отличными от нуля будут только три из этих слагаемых, так как $\mathbf{e}_i \mathbf{e}_i = 1$ при $i = j$. Поскольку $\mathbf{e}_i \mathbf{e}_j = 0$ при $i \neq j$. Поскольку $\mathbf{e}_i \mathbf{e}_i = 1$, имеем

$$\mathbf{x}\mathbf{y} = x_1 y_1 + x_2 y_2 + x_3 y_3$$

— известное из аналитической геометрии скалярное произведение векторов. Применяя соглашение о суммировании, получим

$$\mathbf{x}\mathbf{y} = x_k y_k.$$

Найдем скалярное произведение произвольного вектора $\mathbf{x} = x_i \mathbf{e}_i$ на базисный вектор \mathbf{e}_k

$$\mathbf{x}\mathbf{e}_k = x_i (\mathbf{e}_i \mathbf{e}_k) = x_i \delta_{ik}.$$

Это сумма трех слагаемых, два из которых при $i \neq k$ равны нулю, так как $\delta_{ik} = 0$ при $i \neq k$. Поскольку $\delta_{kk} = 1$, то $x_i \delta_{ik} = x_k$. Следовательно,

$$\mathbf{x}\mathbf{e}_k = x_k.$$

Прямоугольные координаты вектора \mathbf{x} представляют собой ортогональные проекции этого вектора на соответствующую ось.

Запишем некоторые геометрические факты, вытекающие из определения скалярного произведения:

а) Длина вектора $\mathbf{x} = x_i \mathbf{e}_i$ вычисляется по формуле

$$|\mathbf{x}| = \sqrt{\mathbf{x}\mathbf{x}} = \sqrt{\delta_{ij} x_i x_j}$$

или

$$|\mathbf{x}| = \sqrt{x_i^2};$$

б) Косинус угла φ между векторами $\mathbf{x} = x_i \mathbf{e}_i$ и $\mathbf{y} = y_j \mathbf{e}_j$ вычисляется как

$$\cos \varphi = \frac{\mathbf{x}\mathbf{y}}{|\mathbf{x}| |\mathbf{y}|} = \frac{x_i y_j}{\sqrt{x_i x_i} \sqrt{y_j y_j}}.$$

Поэтому условием ортогональности векторов x и y служит равенство

$$x_i y_i = 0;$$

в) Если a — единичный вектор, то его координаты a_k будут равны косинусам углов, которые этот вектор образует с базисными векторами e_k

$$a_k = a e_k = \cos \alpha_k.$$

Если $a^2 = 1$, то $\cos^2 \alpha_1 + \cos^2 \alpha_2 + \cos^2 \alpha_3 = 1$.

г) Проекция вектора $x = x_i e_i$ на вектор $a = a_i e_i$ определяется формулой

$$\text{Пр}_a x = \frac{ax}{|a|} = \frac{a_i x_i}{\sqrt{a_i a_i}}.$$

Векторное и смешанное произведения векторов

1. Векторным произведением векторов x и y называется вектор z , удовлетворяющий следующим требованиям:

1) длина z равна площади параллелограмма, построенного на векторах x и y , т. е. $|z| = |x| |y| \sin \varphi$, где φ — угол между векторами x и y ;

2) вектор z ортогонален каждому из векторов x и y ;

3) вектор z образует с векторами x и y правую тройку векторов.

Векторное произведение x и y обозначается $x \times y$ и обладает следующими свойствами:

$$1) x \times y = -(y \times x);$$

$$2) (\lambda x) \times y = \lambda (x \times y);$$

$$3) (x + y) \times z = x \times z + y \times z.$$

Построим таблицы векторных произведений ортонормированного базиса пространства L_3 :

Правый				Левый			
	e_1	e_2	e_3		e_1	e_2	e_3
e_1	0	e_3	$-e_2$	e_1	0	$-e_3$	e_2
e_2	$-e_3$	0	e_1	e_2	e_3	0	$-e_1$
e_3	e_2	$-e_1$	0	e_3	$-e_2$	e_1	0

Чтобы записать векторные произведения базисных векторов в одной форме для любого ортонормированного базиса, введем величину ε , которая равна $+1$, если базис $\{e_1, e_2, e_3\}$ правый, и -1 , если базис левый.

Затем введем величины ε_{ijk} , определяемые равенствами

$$\varepsilon_{123} = \varepsilon_{231} = \varepsilon_{312} = \varepsilon,$$

$$\varepsilon_{213} = \varepsilon_{132} = \varepsilon_{321} = -\varepsilon$$

и равные нулю, если какие-нибудь два из индексов i, j, k равны между собой. Эти величины также зависят от выбора базиса. Их называют *кососимметричными символами Кронекера*. При помощи величин ε_{ijk} векторные произведения базисных векторов всегда, при любой ориентации базиса, можно записать в виде

$$e_i \times e_j = \varepsilon_{ijk} e_k,$$

где в правой части, как обычно, производится суммирование по k . Например,

$$e_1 \times e_2 = \varepsilon_{12k} e_k = \varepsilon_{121} e_1 + \varepsilon_{122} e_2 + \varepsilon_{123} e_3 = \varepsilon_{123} e_3,$$

откуда

$$e_1 \times e_2 = \varepsilon e_3$$

для правой системы $e_1 \times e_2 = e_3$,

а для левой $e_1 \times e_2 = -e_3$,

что соответствует нашим таблицам.

Пусть $x = x_i e_i$ и $y = y_j e_j$ — два произвольных вектора.

Тогда

$$x \times y = (x_i e_i) \times (y_j e_j).$$

Отсюда, пользуясь свойствами 2), 3) векторного произведения, получим

$$x \times y = x_i y_j (e_i \times e_j) = \varepsilon_{ijk} x_i y_j e_k,$$

где в правой части суммирование происходит по i, j, k . Отбросив в правой части этого равенства слагаемые, равные нулю, запишем

$$x \times y = \varepsilon \{ (x_2 y_3 - x_3 y_2) e_1 + (x_3 y_1 - x_1 y_3) e_2 + (x_1 y_2 - x_2 y_1) e_3 \},$$

или в виде определителя третьего порядка

$$x \times y = \varepsilon \begin{vmatrix} e_1 & e_2 & e_3 \\ x_1 & x_2 & x_3 \\ y_1 & y_2 & y_3 \end{vmatrix}.$$

Обозначим векторное произведение $\mathbf{x} \times \mathbf{y} = \mathbf{z}$. Тогда координаты z_k вектора \mathbf{z} запишутся в виде

$$z_k = \varepsilon_{kij} x_i y_j$$

(так как $\varepsilon_{ijk} = \varepsilon_{kij}$), или более подробно:

$$z_1 = \varepsilon(x_2 y_3 - x_3 y_2),$$

$$z_2 = \varepsilon(x_3 y_1 - x_1 y_3),$$

$$z_3 = \varepsilon(x_1 y_2 - x_2 y_1).$$

При $\varepsilon = 1$ эти формулы совпадают с хорошо известными из аналитической геометрии (где рассматривались лишь правые базисы) формулами для координат векторного произведения.

Введенное нами определение векторного произведения независимо от выбора базиса, поэтому оно является обычным вектором. Тем самым мы избавляемся от необходимости рассматривать аксиальные векторы.

2. Смешанное произведение векторов определяется формулой

$$(\mathbf{x}, \mathbf{y}, \mathbf{z}) = (\mathbf{x} \times \mathbf{y}) \mathbf{z}$$

и равно объему параллелепипеда, построенного на векторах \mathbf{x} , \mathbf{y} и \mathbf{z} , взятому со знаком плюс, если векторы \mathbf{x} , \mathbf{y} , \mathbf{z} образуют правую тройку, и со знаком минус — в противном случае.

Смешанное произведение векторов обладает следующими свойствами:

- 1) $(\mathbf{x}, \mathbf{y}, \mathbf{z}) = -(\mathbf{y}, \mathbf{x}, \mathbf{z})$;
- 2) $(\mathbf{x}, \mathbf{y}, \mathbf{z}) = (\mathbf{y}, \mathbf{z}, \mathbf{x}) = (\mathbf{z}, \mathbf{x}, \mathbf{y})$;
- 3) $(\lambda \mathbf{x}, \mathbf{y}, \mathbf{z}) = \lambda (\mathbf{x}, \mathbf{y}, \mathbf{z})$;
- 4) $(\mathbf{x} + \mathbf{y}, \mathbf{z}, \mathbf{u}) = (\mathbf{x}, \mathbf{z}, \mathbf{u}) + (\mathbf{y}, \mathbf{z}, \mathbf{u})$.

Легко проверить, что для смешанных произведений базисных векторов справедливы формулы

$$(\mathbf{e}_i, \mathbf{e}_j, \mathbf{e}_k) = \varepsilon_{ijk}.$$

Действительно,

$$(\mathbf{e}_i, \mathbf{e}_j, \mathbf{e}_k) = (\mathbf{e}_i \times \mathbf{e}_j) \mathbf{e}_k = \varepsilon_{lje} (\mathbf{e}_l \mathbf{e}_k).$$

В правой части этого равенства производится суммирование по индексу l . Но скалярное произведение $\mathbf{e}_l \mathbf{e}_k$ будет отлично от нуля только тогда, когда $l = k$. Поэтому в рассматриваемой сумме останется только один, отличный

от нуля член $\varepsilon_{ijk}(\mathbf{e}_k \mathbf{e}_k)$. И так как $\mathbf{e}_i \mathbf{e}_i = 1$, то мы и получим доказываемую формулу.

Теперь рассмотрим три произвольных вектора $\mathbf{x} = x_i \mathbf{e}_i$, $\mathbf{y} = y_j \mathbf{e}_j$ и $\mathbf{z} = z_k \mathbf{e}_k$. Их смешанное произведение запишется в виде

$$(\mathbf{x}, \mathbf{y}, \mathbf{z}) = (x_i \mathbf{e}_i, y_j \mathbf{e}_j, z_k \mathbf{e}_k).$$

Пользуясь свойствами 3) и 4) смешанного произведения, мы можем раскрыть скобки, стоящие в правой части этого равенства. Тогда получим

$$(\mathbf{x}, \mathbf{y}, \mathbf{z}) = x_i y_j z_k (\mathbf{e}_i, \mathbf{e}_j, \mathbf{e}_k) = \varepsilon_{ijk} x_i y_j z_k,$$

где индексы i, j, k независимо друг от друга принимают значения 1, 2, 3, и по ним производится суммирование. Следовательно, в правой части этого равенства стоит сумма, содержащая $3^3 = 27$ слагаемых. Но из этих слагаемых только 6 будут отличными от нуля, так как в остальных слагаемых у величин ε_{ijk} будут повторяющиеся индексы. Поэтому в подробной записи предыдущая сумма примет вид

$$(\mathbf{x}, \mathbf{y}, \mathbf{z}) = \varepsilon (x_1 y_2 z_3 + x_2 y_3 z_1 + x_3 y_1 z_2 - x_2 y_1 z_3 - x_3 y_2 z_1 - x_1 y_3 z_2).$$

Это выражение можно записать в виде определителя третьего порядка

$$(\mathbf{x}, \mathbf{y}, \mathbf{z}) = \varepsilon \begin{vmatrix} x_1 & x_2 & x_3 \\ y_1 & y_2 & y_3 \\ z_1 & z_2 & z_3 \end{vmatrix}.$$

3. Рассмотрим двойное векторное произведение $\mathbf{x} \times (\mathbf{y} \times \mathbf{z})$ трех векторов $\mathbf{x}, \mathbf{y}, \mathbf{z}$ и докажем, что имеет место соотношение

$$\mathbf{x} \times (\mathbf{y} \times \mathbf{z}) = \mathbf{y}(\mathbf{xz}) - \mathbf{z}(\mathbf{xy}). \quad (10.2)$$

Если векторы \mathbf{y} и \mathbf{x} коллинеарны, то легко проверить, что как левая, так и правая часть равенства (10.2) будут равны нулю.

Теперь предположим, что \mathbf{y} и \mathbf{z} неколлинеарны, и пусть $\mathbf{u} = \mathbf{x} \times (\mathbf{y} \times \mathbf{z})$. Вектор \mathbf{u} ортогонален вектору $\mathbf{y} \times \mathbf{z}$ и поэтому лежит в плоскости π определяемой \mathbf{y} и \mathbf{z} , т. е.

$$\mathbf{u} = \lambda \mathbf{y} + \mu \mathbf{z}. \quad (10.3)$$

Обозначим через \mathbf{z}^* вектор, лежащий в плоскости π и получающийся из \mathbf{z} поворотом на 90° по часовой стрелке, если смотреть из конца вектора $\mathbf{y} \times \mathbf{z}$. Векторы \mathbf{z}^* , \mathbf{z} и $\mathbf{y} \times \mathbf{z}$ образуют правую ортогональную тройку векторов.

Теперь

$$uz^* = \lambda(yz^*), \quad (10.4)$$

с другой стороны,

$$uz^* = [x \times (y \times z)]z^* = [(y \times z) \times z^*]x.$$

Пусть $(y \times z) \times z^* = v$, тогда вектор v имеет то же направление, что и вектор z , и, так как векторы $y \times z$ и z^* ортогональны, $|v| = |y \times z| |z^*|$, откуда

$$|v| = |y| |z|^2 \sin(y, z) = |y| |z|^2 \cos(y, z^*) = |z|(yz^*),$$

где y, z — угол между y и z . Поэтому

$$v = (yz^*)z.$$

Теперь

$$uz^* = (yz^*)(xz),$$

и, сравнивая это соотношение с равенством (10.4), получаем $\lambda = xz$. Если умножить (10.3) скалярно на x , получим

$$\lambda(xy) + \mu(xz) = 0,$$

откуда $\mu = -xy$. Формула (10.2) доказана.

§ 3. ПРЕОБРАЗОВАНИЕ

ОРТОНОРМИРОВАННОГО БАЗИСА.

ОСНОВНАЯ ЗАДАЧА ТЕНЗОРНОГО ИСЧИСЛЕНИЯ

1. Пусть в пространстве L_3 , кроме ортонормированного базиса $\{e_1, e_2, e_3\}$ с началом в 0, задан другой ортонормированный базис $\{e_{1'}, e_{2'}, e_{3'}\}$ с тем же началом 0 (рис. 10.2). Векторы нового базиса $\{e_{1'}, e_{2'}, e_{3'}\}$ сами могут быть разложены по векторам старого базиса.

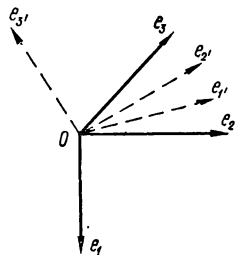


Рис. 10.2. Базисы линейного пространства.

Обозначим через $\gamma_{i'i}$ коэффициент при e_i в разложении вектора $e_{i'}$ по векторам старого базиса. Тогда разложения векторов $e_{i'}$ будут иметь вид

$$e_{1'} = \gamma_{1'1}e_1 + \gamma_{1'2}e_2 + \gamma_{1'3}e_3,$$

$$e_{2'} = \gamma_{2'1}e_1 + \gamma_{2'2}e_2 + \gamma_{2'3}e_3,$$

$$e_{3'} = \gamma_{3'1}e_1 + \gamma_{3'2}e_2 + \gamma_{3'3}e_3.$$

Более кратко эти три равенства можно записать в виде

$$e_{i'} = \gamma_{i'i}e_i. \quad (10.5)$$

Умножим скалярно каждое из равенств (10.5) на каждый из векторов e_i . Тогда, учитывая, что $e_i e_j = \delta_{ij}$, получим

$$e_{i'} e_i = \gamma_{i'i},$$

и так как $e_{i'}$ и e_i — единичные векторы, то

$$e_{i'} e_i = \cos(e_{i'}, e_i).$$

Поэтому

$$\gamma_{i'i} = \cos(e_{i'}, e_i). \quad (10.6)$$

С другой стороны, можно записать разложение векторов старого базиса по векторам $e_{i'}$ нового. Обозначим через $\gamma_{ii'}$ коэффициент при $e_{i'}$ в разложении e_i по векторам нового базиса.

Получим

$$e_1 = \gamma_{11'} e_{1'} + \gamma_{12'} e_{2'} + \gamma_{13'} e_{3'},$$

$$e_2 = \gamma_{21'} e_{1'} + \gamma_{22'} e_{2'} + \gamma_{23'} e_{3'},$$

$$e_3 = \gamma_{31'} e_{1'} + \gamma_{32'} e_{2'} + \gamma_{33'} e_{3'},$$

или сокращенно

$$e_i = \gamma_{ii'} e_{i'} \quad (i, i' = 1, 2, 3). \quad (10.7)$$

Теперь, если каждое из равенств (10.7) умножить скалярно на каждый из векторов $e_{i'}$, то получим

$$e_i e_{i'} = \cos(e_i, e_{i'}) = \gamma_{ii'}. \quad (10.8)$$

Равенства (10.6) и (10.8) показывают, что

$$\gamma_{ii'} = \gamma_{i'i}. \quad (10.9)$$

Числа $\gamma_{i'i}$ можно записать в виде таблицы

$$\Gamma = \begin{pmatrix} \gamma_{1'1} & \gamma_{1'2} & \gamma_{1'3} \\ \gamma_{2'1} & \gamma_{2'2} & \gamma_{2'3} \\ \gamma_{3'1} & \gamma_{3'2} & \gamma_{3'3} \end{pmatrix}.$$

Это квадратная матрица. Число строк (столбцов) — порядок матрицы. Γ — представляет собой квадратную матрицу третьего порядка и называется *матрицей перехода* от старого базиса к новому.

Аналогично числа $\gamma_{ii'}$ образуют матрицу

$$\Gamma^{-1} = \begin{pmatrix} \gamma_{11'} & \gamma_{12'} & \gamma_{13'} \\ \gamma_{21} & \gamma_{22'} & \gamma_{23'} \\ \gamma_{31'} & \gamma_{32'} & \gamma_{33'} \end{pmatrix}.$$

Это матрица перехода от нового базиса к старому (обозначение Γ^{-1} показывает, что эта матрица обратного

перехода). Коротко матрицы Γ и Γ^{-1} записывают в виде

$$\Gamma = (\gamma_{i' i}), \quad \Gamma^{-1} = (\gamma_{i i'}).$$

Равенство

$$\gamma_{i' i} = \gamma_{i i'}$$

показывает, что матрица Γ^{-1} получается из матрицы Γ , если в последней стробки заменить столбцами.

Для элементов этих двух матриц справедливо

$$\begin{aligned} \gamma_{i' k} \gamma_{j' k} &= \gamma_{k i'} \gamma_{k j'} = \delta_{i' j'}, \\ \gamma_{i k} \gamma_{j k'} &= \gamma_{k' i} \gamma_{k' j} = \delta_{i j}. \end{aligned} \quad (10.10)$$

Действительно,

$$\gamma_{i' k} \gamma_{j' k} = \gamma_{i' 1} \gamma_{j' 1} + \gamma_{i' 2} \gamma_{j' 2} + \gamma_{i' 3} \gamma_{j' 3} = \mathbf{e}_{i'} \mathbf{e}_{j'} = \delta_{i' j'}.$$

Аналогично доказывается второе равенство.

Равенства (10.10) означают, что для матриц Γ и Γ^{-1} сумма произведений элементов какой-нибудь строки (столбца) на соответствующие элементы другой строки (столбца) равна нулю, а сумма квадратов элементов любой строки (столбца) равна единице. Матрицы, элементы которых обладают такими свойствами, называются ортогональными. Мы доказали, что *переход от одного ортонормированного базиса к другому в L_3 задается ортогональной матрицей.*

Пусть дана произвольная ортогональная матрица $\Gamma = (\gamma_{i' i})$. Векторы $\mathbf{e}_{i'}$, определяемые формулами (10.5), в силу свойств ортогональной матрицы будут попарно ортогональными и единичными. Поэтому *всякая ортогональная матрица служит матрицей перехода от одного ортонормированного базиса к другому.*

Рассмотрим определитель ортогональной матрицы Γ

$$|\Gamma| = |\gamma_{i' i}| = \begin{vmatrix} \gamma_{1' 1} & \gamma_{1' 2} & \gamma_{1' 3} \\ \gamma_{2' 1} & \gamma_{2' 2} & \gamma_{2' 3} \\ \gamma_{3' 1} & \gamma_{3' 2} & \gamma_{3' 3} \end{vmatrix}.$$

Поскольку строки определителя $|\Gamma|$ составлены из координат векторов $\mathbf{e}_{1'}$, $\mathbf{e}_{2'}$, $\mathbf{e}_{3'}$ относительно базиса $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$, то $|\Gamma|$ равен смешанному произведению векторов $\mathbf{e}_{1'}$, $\mathbf{e}_{2'}$, $\mathbf{e}_{3'}$

$$|\Gamma| = (\mathbf{e}_{1'}, \mathbf{e}_{2'}, \mathbf{e}_{3'}).$$

Абсолютная величина этого смешанного произведения равна единице, так как она равна объему куба, построенного на векторах $\mathbf{e}_{1'}$, $\mathbf{e}_{2'}$, $\mathbf{e}_{3'}$.

Следовательно, определитель любой ортогональной матрицы равен ± 1 , причем знак плюс или минус зависит от того, какую ориентацию имеют базисы $\{e_1, e_2, e_3\}$ и $\{e_{1'}, e_{2'}, e_{3'}\}$ — одинаковую или противоположную. В первом случае базис $\{e_1, e_2, e_3\}$ можно совместить с $\{e_{1'}, e_{2'}, e_{3'}\}$ путем поворота вокруг 0, во втором случае кроме поворота необходимо осуществить отражение базиса $\{e_1, e_2, e_3\}$ относительно некоторой плоскости, проходящей через 0.

Запишем формулы преобразования ортонормированного базиса для плоскости. Это преобразование представляет собой либо чистый поворот на некоторый угол α вокруг начала координат 0, либо поворот на угол α с последующим отражением относительно некоторой прямой, проходящей через начало координат. В первом случае формулы имеют вид

$$\begin{aligned} e_{1'} &= \cos \alpha e_1 + \sin \alpha e_2, \\ e_{2'} &= -\sin \alpha e_1 + \cos \alpha e_2 \end{aligned}$$

и матрица Γ имеет вид

$$\Gamma = \begin{pmatrix} \cos \alpha & \sin \alpha \\ -\sin \alpha & \cos \alpha \end{pmatrix}.$$

Определитель этой матрицы равен единице.

Во втором случае формулы преобразования базиса выглядят так:

$$\begin{aligned} e_{1'} &= \cos \alpha e_1 + \sin \alpha e_2, \\ e_{2'} &= \sin \alpha e_1 - \cos \alpha e_2, \end{aligned}$$

и определитель матрицы $\Gamma = \begin{pmatrix} \cos \alpha & \sin \alpha \\ \sin \alpha & -\cos \alpha \end{pmatrix}$ будет равен -1 .

2. Пусть в L_3 задан вектор x . Он представляет собой некоторый геометрический или физический объект, заданный по величине и направлению (сила, скорость, ускорение, напряженность поля и т. п.). Этот реально существующий объект не зависит от того, в какой системе координат мы его рассматриваем. Любые действия или вычисления, проводимые непосредственно над векторами, можно всегда физически истолковать.

Кроме исчисления, связанного с собственно векторами, важную роль играет координатный метод. Его применение позволяет изучать геометрические образы не непосредственно,

а достаточно хорошо развитыми методами алгебры (в аналитической геометрии) и анализа (в дифференциальной геометрии). При этом легко получаются результаты, непосредственное доказательство которых иногда очень громоздко или вообще невозможно.

Однако при применении координатного метода мы с каждым вектором \mathbf{x} связываем его координаты x_1, x_2, x_3 , которые зависят не только от самого вектора, но и от базиса. Ортонормированные базисы можно выбирать различными способами: например, выбрав один базис и поворачивая его вокруг начала, можно получить из него другие.

При применении координатного метода получаются данные, отражающие не только геометрическую картину, но и произвольность выбора координатной системы. Например, сами координаты вектора, конечно, зависят от координатной системы, но сумма их квадратов (квадрат длины вектора) уже не должна зависеть от выбора системы координат.

Инвариантные свойства — свойства объектов, не зависящие от выбора системы координат. Только такие свойства и представляются интересными для изучения.

Основная задача тензорного исчисления заключается в том, чтобы научиться отделять результаты, относящиеся к самим геометрическим объектам, от того, что привнесено случайным выбором координатной системы.

Выясним, как преобразуются координаты вектора \mathbf{x} при переходе от ортонормированного базиса $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$ с началом в 0 к другому ортонормированному базису $\{\mathbf{e}_{1'}, \mathbf{e}_{2'}, \mathbf{e}_{3'}\}$ с тем же началом 0. Запишем разложения вектора \mathbf{x} в каждом из этих двух базисов

$$\mathbf{x} = x_i \mathbf{e}_i; \quad \mathbf{x} = x_{i'} \mathbf{e}_{i'}.$$

Справедливо

$$x_i \mathbf{e}_i = x_{i'} \mathbf{e}_{i'}.$$

Заменяя векторы \mathbf{e}_i по (10.7), имеем

$$x_i \gamma_{ii'} \mathbf{e}_{i'} = x_{i'} \mathbf{e}_{i'},$$

откуда, в силу линейной независимости векторов $\mathbf{e}_{i'}$, следует, что

$$x_{i'} = x_i \gamma_{ii'}.$$

Учитывая, что $\gamma_{ii'} = \gamma_{i'i}$, запишем полученное равенство в виде

$$x_{i'} = \gamma_{i'i} x_i. \quad (10.11)$$

Эти формулы дают выражение новых координат вектора \mathbf{x} через старые. Если в равенстве $x_i \mathbf{e}_i = x_{i'} \mathbf{e}_{i'}$ заменить векторы $\mathbf{e}_{i'}$ по формулам (10.5), то получим выражение старых координат вектора \mathbf{x} через новые

$$x_i = \gamma_{ii'} x_{i'}. \quad (10.12)$$

Выясним, какие из проведенных рассмотрений имеют инвариантный характер, т. е. не зависят от выбора системы координат. Начнем со скалярного произведения векторов. В пространстве L_3 оно было определено геометрически, поэтому инвариантность его очевидна. Покажем, что полученное в § 2 выражение скалярного произведения через координаты перемножаемых векторов в некотором ортонормированном базисе обладает свойствами инвариантности. Это важно, потому что в L_4 скалярное произведение векторов в ортонормированном базисе определяют как сумму произведений одноименных координат, и там доказательство его инвариантности уже не может носить геометрического характера и должно быть обязательно проведено аналитически, например так, как это сейчас будет сделано в L_3 .

Ранее доказано, что скалярное произведение векторов \mathbf{x} и \mathbf{y} , имеющих в базисе $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$ координаты x_i и y_i , а в базисе $\{\mathbf{e}_{1'}, \mathbf{e}_{2'}, \mathbf{e}_{3'}\}$ — координаты $x_{i'}$, $y_{i'}$, можно записать в виде $x_i y_i$ в первом базисе и $x_{i'} y_{i'}$ — во втором.

Покажем тождественность этих выражений. Пользуясь формулами (10.10) и (10.11), получим

$$x_{i'} y_{i'} = \gamma_{i'i} x_i \gamma_{i'j} y_j = \delta_{ij} x_i y_j = x_i y_i.$$

Из инвариантности формулы для вычисления скалярного произведения немедленно следует инвариантность формул для вычисления длины вектора и косинуса угла между двумя векторами, так как эти величины выражаются через скалярное произведение векторов.

Прежде чем доказать инвариантность формул для векторного и смешанного произведений векторов через координаты сомножителей, посмотрим, как изменяются компоненты кососимметричного символа Кронекера при переходе к новому базису. В новом базисе мы получим

$$\varepsilon_{i'j'k'} = (\mathbf{e}_{i'}, \mathbf{e}_{j'}, \mathbf{e}_{k'}),$$

и так как

$$\mathbf{e}_{i'} = \gamma_{i'i} \mathbf{e}_i; \quad \mathbf{e}_{j'} = \gamma_{j'j} \mathbf{e}_j; \quad \mathbf{e}_{k'} = \gamma_{k'k} \mathbf{e}_k,$$

то

$$\varepsilon_{i'j'k'} = \gamma_{i'i} \gamma_{j'j} \gamma_{k'k} \varepsilon_{ijk}.$$

В частности,

$$\varepsilon_{1'2'3'} = \gamma_{1'i} \gamma_{2'j} \gamma_{3'k} \varepsilon_{ijk}.$$

Но в этой сумме отличными от нуля будут только шесть членов

$$\begin{aligned} \varepsilon_{1'2'3'} = & (\gamma_{1'1} \gamma_{2'2} \gamma_{3'3} + \gamma_{1'2} \gamma_{2'3} \gamma_{3'1} + \gamma_{1'3} \gamma_{2'1} \gamma_{3'2} - \\ & - \gamma_{1'2} \gamma_{2'1} \gamma_{3'3} - \gamma_{1'3} \gamma_{2'2} \gamma_{3'1} - \gamma_{1'1} \gamma_{2'3} \gamma_{3'2}). \end{aligned}$$

Стоящее в скобках выражение ε_{123} представляет собой определитель матрицы Γ , так что

$$\varepsilon_{1'2'3'} = |\Gamma| \varepsilon_{123}$$

или, если обозначить через ε' значение величины ε в новом базисе,

$$\varepsilon' = |\Gamma| \varepsilon.$$

Эта формула показывает, что если ориентация базиса не меняется на противоположную, то не меняется и величина ε , если же ориентация базиса меняется на противоположную, то величина ε меняет знак, т. е. эта формула согласуется с определением величины ε .

Пусть теперь $\mathbf{z} = \mathbf{x} \times \mathbf{y}$. Тогда в старом базисе

$$z_k = \varepsilon_{ijk} x_i y_j, \quad (10.13)$$

а в новом базисе

$$z_{k'} = \varepsilon_{i'j'k'} x_{i'} y_{j'}. \quad (10.14)$$

Покажем инвариантность этой формулы, т. е. покажем, что формула (10.13) переходит в (10.14) при преобразовании базиса. В самом деле, подставляя выражения

$$x_i = \gamma_{ii'} x_{i'}, \quad y_j = \gamma_{jj'} y_{j'}, \quad z_k = \gamma_{kk'} z_{k'}$$

в первую формулу, получим

$$\gamma_{kk'} z_{k'} = \varepsilon_{ijk} \gamma_{ii'} \gamma_{jj'} x_{i'} y_{j'}.$$

Умножим эти соотношения на $\gamma_{kl'}$ и просуммируем по индексу k . Так как

$$\gamma_{kk'} \gamma_{kl'} = \delta_{k'l'},$$

то

$$z_{e'} = \varepsilon_{ijk} \gamma_{ii'} \gamma_{jj'} \gamma_{kl'} x_{i'} y_{j'},$$

а поскольку

$$\varepsilon_{ijk} \gamma_{ii'} \gamma_{jj'} \gamma_{kl'} = \gamma_{i'i} \gamma_{j'j} \gamma_{l'k} \varepsilon_{ijk} = \varepsilon_{i'j'l'},$$

полученная формула совпадает с формулой (10.13).

Аналогично можно доказать, что вычислительная формула для смешанного произведения также остается инвариантной при преобразовании базиса, т. е.

$$\varepsilon_{ijk}x_iy_jz_k = \varepsilon_{i'j'k'}x_{i'}y_{j'}z_{k'}.$$

Заметим, что инвариантность формулы для смешанного произведения следует из того, что

$$(x, y, z) = (x \times y) z,$$

а векторное и скалярное произведение векторов, как мы только что доказали, выражаются инвариантными формулами.

Контрольные вопросы и задания

1. Что такое смешанное произведение векторов?
2. Сформулируйте основные свойства смешанного произведения.
3. Параллелепипед построен на векторах a, b, c . Найдите площади его диагональных сечений.
4. Что такое двойное векторное произведение?
5. Сформулируйте основные свойства смешанного произведения.
6. Охарактеризуйте двойное векторное произведение.
7. Что представляют собой элементы матриц перехода от одного базиса к другому?
8. Что такое ортогональные матрицы?
9. Что такое инвариантность?
10. Сформулируйте основную задачу тензорного исчисления.

§ 4. ПОЛИЛИНЕЙНЫЕ ФОРМЫ И ТЕНЗОРЫ

Линейные формы

1. Рассмотрим простейшие скалярные функции одного или нескольких векторных элементов. В линейном пространстве L задана скалярная функция $\varphi = \varphi(x)$ векторного аргумента x , если каждому вектору x пространства L поставлено в соответствие некоторое число φ . Эта функция носит название линейной функции, или линейной формы, если она обладает следующими двумя свойствами:

- 1) $\varphi(x + y) = \varphi(x) + \varphi(y)$;
- 2) $\varphi(\lambda x) = \lambda \varphi(x)$.

Некоторые примеры линейных функций:

а) Обозначим через $\text{Пр}_e x$ величину проекции вектора x на ось e . $\text{Пр}_e x$ является *линейной формой вектора x* , так

как из аналитической геометрии известно, что

$$\text{Пр}_e(\mathbf{x} + \mathbf{y}) = \text{Пр}_e \mathbf{x} + \text{Пр}_e \mathbf{y}, \quad \text{Пр}_e(\lambda \mathbf{x}) = \lambda \text{Пр}_e \mathbf{x}.$$

б) Пусть \mathbf{a} постоянный, а \mathbf{x} — переменный векторы пространства L . Тогда их скалярное произведение $\varphi = \mathbf{a}\mathbf{x}$ является линейной формой вектора \mathbf{x} . Действительно, в силу свойств скалярного произведения векторов

$$\mathbf{a}(\mathbf{x} + \mathbf{y}) = \mathbf{a}\mathbf{x} + \mathbf{a}\mathbf{y} \quad \text{и} \quad \mathbf{a}(\lambda \mathbf{x}) = \lambda(\mathbf{a}\mathbf{x}).$$

в) Так как координата x_i вектора \mathbf{x} пространства L_3 по отношению к ортонормированному базису $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$ может быть представлена в виде $x_i = \mathbf{e}_i\mathbf{x}$, то она также является линейной формой вектора \mathbf{x} .

г) Пусть \mathbf{a} и \mathbf{b} — два неколлинеарных вектора пространства L_3 . Тогда смешанное произведение $(\mathbf{a}, \mathbf{b}, \mathbf{x})$ является линейной формой вектора \mathbf{x} , так как в силу свойств смешанного произведения

$$(\mathbf{a}, \mathbf{b}, \mathbf{x} + \mathbf{y}) = (\mathbf{a}, \mathbf{b}, \mathbf{x}) + (\mathbf{a}, \mathbf{b}, \mathbf{y})$$

и

$$(\mathbf{a}, \mathbf{b}, \lambda \mathbf{x}) = \lambda(\mathbf{a}, \mathbf{b}, \mathbf{x}).$$

Теперь найдем выражение линейной формы $\varphi = \varphi(\mathbf{x})$ в ортонормированном базисе $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$.

Так как

$$\mathbf{x} = x_i \mathbf{e}_i$$

и функция φ линейная, то

$$\varphi(\mathbf{x}) = \varphi(x_i \mathbf{e}_i) = x_i \varphi(\mathbf{e}_i).$$

Обозначим числа $\varphi(\mathbf{e}_i) = a_i$, тогда линейная форма φ запишется в виде

$$\varphi(\mathbf{x}) = a_i x_i. \quad (10.15)$$

Это выражение — однородный многочлен первой степени от переменных x_i , поэтому линейная функция и называется линейной формой. Коэффициенты a_i в этом выражении зависят от выбора базиса.

2. Рассмотрим преобразование коэффициентов линейной формы $\varphi = \varphi(\mathbf{x})$ при переходе от ортонормированного базиса $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$ к новому ортонормированному базису $\{\mathbf{e}_{1'}, \mathbf{e}_{2'}, \mathbf{e}_{3'}\}$. При таком преобразовании

$$\mathbf{e}_{i'} = \gamma_{i'i} \mathbf{e}_i,$$

где $\Gamma = (\gamma_{i'i})$ — матрица перехода от старого базиса к новому. В новом базисе форма φ запишется в виде

$$\varphi = a_{i'} x_{i'},$$

где $x_{i'}$ — новые координаты вектора \mathbf{x} , а коэффициенты a_i вычисляются по формулам

$$a_{i'} = \varphi(\mathbf{e}_{i'}) = \varphi(\gamma_{i'i}\mathbf{e}_i) = \gamma_{i'i}\varphi(\mathbf{e}_i) = \gamma_{i'i}a_i.$$

Следовательно, коэффициенты линейной формы φ при переходе от старого базиса к новому изменяются по закону

$$a_{i'} = \gamma_{i'i}a_i.$$

Сравнивая эти формулы с формулами (10.11), видим, что закон изменения коэффициентов линейной формы при переходе к новому базису в точности совпадает с законом изменения координат вектора. Теперь легко увидеть, что

$$a_i\mathbf{e}_i = a_{i'}\mathbf{e}_{i'},$$

поэтому коэффициенты a_i линейной формы φ являются координатами некоторого вектора

$$\mathbf{a} = a_i\mathbf{e}_i.$$

Формула показывает, что саму *линейную форму* $\varphi = \varphi(\mathbf{x})$ всегда можно записать в виде скалярного произведения векторов \mathbf{a} и \mathbf{x}

$$\varphi(\mathbf{x}) = \mathbf{a}\mathbf{x}.$$

Для выяснения геометрического смысла вектора \mathbf{a} рассмотрим поверхности уровня линейной формы φ . Эти поверхности определяются уравнением $\varphi = c$, или

$$\mathbf{a}\mathbf{x} = c.$$

Это уравнение семейства параллельных плоскостей, для которых вектор \mathbf{a} является нормальным вектором. Следовательно, вектор \mathbf{a} представляет собой общий нормальный вектор к плоскостям, являющимся поверхностями уровня формы φ .

Билинейные формы

1. Скалярная функция $\varphi = \varphi(\mathbf{x}, \mathbf{y})$ двух векторных аргументов \mathbf{x} и \mathbf{y} называется *билинейной функцией*, или *билинейной формой*, если она линейна по каждому своему аргументу, т. е. если:

- 1) $\varphi(\mathbf{x}_1 + \mathbf{x}_2, \mathbf{y}) = \varphi(\mathbf{x}_1, \mathbf{y}) + \varphi(\mathbf{x}_2, \mathbf{y})$;
- 2) $\varphi(\lambda\mathbf{x}, \mathbf{y}) = \lambda\varphi(\mathbf{x}, \mathbf{y})$;
- 3) $\varphi(\mathbf{x}, \mathbf{y}_1 + \mathbf{y}_2) = \varphi(\mathbf{x}, \mathbf{y}_1) + \varphi(\mathbf{x}, \mathbf{y}_2)$;
- 4) $\varphi(\mathbf{x}, \lambda\mathbf{y}) = \lambda\varphi(\mathbf{x}, \mathbf{y})$.

Рассмотрим примеры билинейных форм:

а) Скалярное произведение xu векторов x и y является билинейной формой, так как оно обладает всеми перечисленными выше свойствами.

б) Пусть a — постоянный вектор, а x и y — переменные векторы. Легко проверить, что смешанное произведение (a, x, y) также является билинейной формой.

в) Пусть $\varphi(x)$ и $\psi(y)$ — линейные формы переменных векторов x и y . Их произведение $f(x, y) = \varphi(x)\psi(y)$ является билинейной формой, так как

$$\begin{aligned} f(x_1 + x_2, y) &= \varphi(x_1 + x_2)\psi(y) = \varphi(x_1)\psi(y) + \\ &+ \varphi(x_2)\psi(y) = f(x_1, y) + f(x_2, y); \\ f(\lambda x, y) &= \varphi(\lambda x)\psi(y) = \lambda\varphi(x)\psi(y) = \lambda f(x, y), \end{aligned}$$

и аналогично для другого аргумента.

2. Отнесем теперь линейное пространство L_3 к прямоугольному базису $\{e_1, e_2, e_3\}$ и найдем выражение билинейной формы $\varphi = \varphi(x, y)$ в этой системе координат. Имеем

$$x = x_i e_i, \quad y = y_j e_j,$$

и так как функция φ линейна относительно обеих своих переменных, то

$$\varphi(x, y) = \varphi(x_i e_i; y_j e_j) = x_i y_j \varphi(e_i e_j).$$

Обозначим значения билинейной формы φ от базисных векторов через a_{ij}

$$\varphi = a_{ij} x_i y_j$$

или подробнее

$$\begin{aligned} \varphi &= a_{11}x_1y_1 + a_{12}x_1y_2 + a_{13}x_1y_3 + a_{21}x_2y_1 + a_{22}x_2y_2 + \\ &+ a_{23}x_2y_3 + a_{31}x_3y_1 + a_{32}x_3y_2 + a_{33}x_3y_3. \end{aligned}$$

Это выражение линейно относительно двух рядов переменных (x_1, x_2, x_3) и (y_1, y_2, y_3) .

Коэффициенты билинейной формы можно записать в виде таблицы

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix},$$

которая, как указывалось выше, является квадратной матрицей третьего порядка.

Назовем ее *матрицей билинейной формы* φ .

Таким образом, в пространстве L_3 билинейной форме соответствует в каждом базисе определенная матрица третьего порядка.

Запишем в координатной форме рассмотренные выше билинейные формы и найдем их матрицы.

а) Билинейная форма xy в ортонормированном базисе записывается в виде

$$xy = x_1y_1 + x_2y_2 + x_3y_3.$$

Следовательно, ее матрица выглядит так:

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} = (\delta_{ij}).$$

б) Рассмотрим билинейную форму (a, x, y) .

В координатной форме имеем

$$(a, x, y) = \varepsilon_{kij} a_k x_i y_j.$$

Поэтому матрица коэффициентов этой формы имеет вид

$$(\varepsilon_{kij} a_k) = \varepsilon \begin{pmatrix} 0 & a_3 & -a_2 \\ -a_3 & 0 & a_1 \\ a_2 & -a_1 & 0 \end{pmatrix}.$$

в) В ортонормированном базисе (e_1, e_2, e_3) линейные формы $\varphi(x)$ и $\psi(y)$ можно записать в виде

$$\varphi = a_i x_i; \quad \psi = b_j y_j.$$

Билинейная форма $f(x, y) = \varphi(x) \psi(y)$ теперь имеет вид

$$f = (a_i x_i) (b_j y_j) = a_i b_j x_i y_j.$$

Матрица этой билинейной формы выглядит как

$$A = (a_i b_j) = \begin{pmatrix} a_1 b_1 & a_1 b_2 & a_1 b_3 \\ a_2 b_1 & a_2 b_2 & a_2 b_3 \\ a_3 b_1 & a_3 b_2 & a_3 b_3 \end{pmatrix}.$$

3. Рассмотрим, как преобразуются коэффициенты билинейной формы $\varphi = \varphi(x, y)$ при преобразовании базиса. В новом базисе $\{e_{1'}, e_{2'}, e_{3'}\}$ эту билинейную форму запишем в виде

$$\varphi = a_{i'j'} x_{i'} y_{j'},$$

где

$$a_{i'j'} = \varphi(e_{i'}, e_{j'}).$$

Но при переходе к новому базису

$$e_{i'} = \gamma_{i'i} e_i,$$

поэтому, используя основные свойства билинейной формы, получим

$$a_{i'j'} = \Phi(\gamma_{i'i} e_i, \gamma_{j'j} e_j) = \gamma_{i'i} \gamma_{j'j} \Phi(e_i, e_j) = \gamma_{i'i} \gamma_{j'j} a_{ij}.$$

Таким образом, при переходе к новому базису коэффициенты билинейной формы преобразуются по закону

$$a_{i'j'} = \gamma_{i'i} \gamma_{j'j} a_{ij}. \quad (10.16)$$

Сравнивая эти формулы с формулами преобразования коэффициентов линейной формы, видим, что обе эти группы формул устроены аналогично.

Докажем теперь обратное: если элементы a_{ij} матрицы A при преобразовании базиса пространства L_3 преобразуются по закону (10.16), то этой матрице отвечает билинейная форма.

Пусть $\{e_1, e_2, e_3\}$ и $\{e_{1'}, e_{2'}, e_{3'}\}$ — два базиса в пространстве L_3 , а x и y — два его произвольных вектора, тогда

$$x = x_i e_i = x_{i'} e_{i'}, \quad y = y_i e_i = y_{i'} e_{i'}.$$

Рассмотрим билинейное выражение $\Phi = a_{ij} x_i y_j$.

Для доказательства того, что это выражение действительно является билинейной формой в пространстве L_3 , следует доказать, что оно не меняется при преобразовании базиса, т. е. что его величина зависит только от выбора векторов x и y , но не зависит от выбора базиса.

После преобразования имеем

$$\Phi' = a_{i'j'} x_{i'} y_{j'}.$$

Следовательно, необходимо доказать, что $\Phi = \Phi'$.

В самом деле, из соотношений (10.15) и (10.16) следует, что

$$\begin{aligned} \Phi' &= a_{i'j'} x_{i'} y_{j'} = \gamma_{i'i} \gamma_{j'j} a_{ij} \gamma_{i'k} \gamma_{j'l} x_k y_l = \\ &= \gamma_{i'i} \gamma_{j'j} \gamma_{i'k} \gamma_{j'l} a_{ij} x_k y_l. \end{aligned}$$

В силу свойств ортогональной матрицы

$$\gamma_{i'i} \gamma_{i'k} = \delta_{ik},$$

$$\gamma_{j'j} \gamma_{j'l} = \delta_{jl}.$$

Поэтому

$$\Phi' = \delta_{ik} \delta_{jl} a_{ij} x_k y_l,$$

но

$$\delta_{ik}x_k = x_i; \quad \delta_{jl}y_l = y_j,$$

вследствие чего

$$\varphi' = a_{ij}x_i y_j = \varphi,$$

что и требовалось доказать.

Полилинейные формы. Общее определение тензора

1. Рассмотрим теперь в линейном пространстве L_3 скалярную функцию от p векторных аргументов — функцию

$$\varphi = \varphi(x, y, z, \dots, w).$$

Эта функция называется полилинейной функцией, или полилинейной формой, если она линейна по каждому из своих аргументов, т. е. если для каждого из аргументов выполнены условия

$$1) \quad \varphi(x, y, z_1 + z_2, \dots, w) = \varphi(x, y, z_1, \dots, w) + \\ + \varphi(x, y, z_2, \dots, w);$$

$$2) \quad \varphi(x, y, \lambda z, \dots, w) = \lambda \varphi(x, y, z, \dots, w).$$

Число аргументов p называется степенью полилинейной формы φ . Форма φ называется также p -линейной формой.

Рассмотренные ранее линейные формы — частные случаи полилинейных форм. Это формы первой степени, 1-линейные формы. Билинейные формы также частный случай — 2-линейные формы и т. д. Рассмотрим еще несколько примеров полилинейных форм степени больше двух.

а) Смешанное произведение векторов (x, y, z) является трилинейной формой, так как для каждого из ее аргументов выполняются условия 1) и 2).

б) Произведение трех линейных форм $\alpha(x)$, $\beta(y)$, $\gamma(z)$ представляет собой трилинейную форму. Действительно, если

$$\varphi(x, y, z) = \alpha(x)\beta(y)\gamma(z),$$

то

$$\begin{aligned} \varphi(x_1 + x_2, y, z) &= \alpha(x_1 + x_2)\beta(y)\gamma(z) = \\ &= [\alpha(x_1) + \alpha(x_2)]\beta(y)\gamma(z) = \alpha(x_1)\beta(y)\gamma(z) + \\ &+ \alpha(x_2)\beta(y)\gamma(z) = \varphi(x_1, y, z) + \varphi(x_2, y, z), \\ \varphi(\lambda x, y, z) &= \alpha(\lambda x)\beta(y)\gamma(z) = \lambda\alpha(x)\beta(y)\gamma(z) = \\ &= \lambda\varphi(x, y, z), \end{aligned}$$

и аналогично для других аргументов.

2. Рассмотрим запись полилинейной формы $\varphi(x, y, \dots, w)$, зависящей от P векторных аргументов, в координатном виде.

Для определенности рассмотрим трилинейную форму $\varphi = \varphi(x, y, z)$. Каждый из векторов x, y, z можно разложить по базису $\{e_1, e_2, e_3\}$:

$$x = x_i e_i; \quad y = y_j e_j; \quad z = z_k e_k.$$

Обозначения для индексов суммирования выбираются различными для удобства дальнейших выкладок.

Так как форма φ линейна по всем аргументам, имеем

$$\varphi(x, y, z) = \varphi(x_i e_i, y_j e_j, z_k e_k) = x_i y_j z_k \varphi(e_i, e_j, e_k),$$

где $\varphi(e_i, e_j, e_k)$ — значения формы φ от векторов базиса.

Обозначим эти значения через a_{ijk} , тогда форму φ запишем в виде $\varphi(x, y, z) = a_{ijk} x_i y_j z_k$.

Следовательно, трилинейная форма записывается как однородный полином третьей степени, линейный относительно трех рядов переменных (x_1, x_2, x_3) , (y_1, y_2, y_3) и (z_1, z_2, z_3) . Этот многочлен содержит $3^3 = 27$ слагаемых и столько же коэффициентов a_{ijk} . Совокупность этих коэффициентов можно представить кубической матрицей третьего порядка.

Аналогично, для 4-линейной формы $\varphi(x, y, z, u)$, зависящей от четырех векторных аргументов, получим

$$\varphi = a_{ijkl} x_i y_j z_k u_l,$$

где

$$a_{ijkl} = \varphi(e_i, e_j, e_k, e_l).$$

Многочлен, при помощи которого записывается эта форма, имеет 3^4 слагаемых и столько же коэффициентов a_{ijkl} .

Аналогично, p -линейная форма

$$\varphi = \varphi(x, y, \dots, w),$$

зависящая от p аргументов, запишется в базисе $\{e_1, e_2, e_3\}$ в виде

$$\varphi = a_{ijk\dots m} x_i y_j z_k \dots w_m, \quad (10.17)$$

где

$$a_{ijk\dots m} = \varphi(e_i, e_j, e_k, \dots, e_m).$$

Коэффициенты $a_{ijk\dots m}$ этой формы имеют p индексов, каждый из которых может принимать 3 значения. Всего такая полилинейная форма имеет 3^p коэффициентов.

3. Введенные в рассмотрение полилинейные формы определены независимо от выбора системы координат. Значения этих форм зависят только от значений их векторных аргументов. Например, для формы $\varphi = \varphi(\mathbf{x}, \mathbf{y}, \mathbf{z})$ — только от значений векторов $\mathbf{x}, \mathbf{y}, \mathbf{z}$, но не зависят от того, в каком базисе рассматриваются эти векторы. Следуя терминологии, введенной ранее, можно сказать, что полилинейные формы определены инвариантным способом. При переходе к новому базису координаты векторов меняются. При этом должны меняться и коэффициенты полилинейных форм (поскольку сама форма должна оставаться инвариантной). Совокупность коэффициентов инвариантной полилинейной формы представляет собой очень важный геометрический объект.

О п р е д е л е н и е. Геометрический (или физический) объект, который определяется совокупностью коэффициентов $a_{ijk\dots m}$ полилинейной формы $\varphi(\mathbf{x}, \mathbf{y}, \mathbf{z}, \dots, \mathbf{w})$, записанной в некотором ортонормированном базисе, называется *ортгональным тензором*. Сами числа $a_{ijk\dots m}$ называются *компонентами* или *координатами* этого тензора.

Поскольку никаких других тензоров, кроме ортгональных, мы рассматривать не будем, то всюду далее они будут называться просто тензорами.

Говорят, что тензор $a_{ijk\dots m}$ определяется линейной формой $\varphi = \varphi(\mathbf{x}, \mathbf{y}, \mathbf{z}, \dots, \mathbf{w})$. Коэффициенты $a_{ijk\dots m}$ формы φ степени p вычисляются по формуле (10.17) и имеют p индексов. Поэтому тензор, соответствующий полилинейной форме степени p , называют тензором валентности p .

Если форма φ задана в пространстве L_3 , то каждый из индексов тензора может принимать независимо от других индексов значения 1, 2 и 3. Поэтому тензор валентности p в 3-мерном пространстве имеет 3^p компонент. На плоскости такой тензор имеет 2^p компонент, в линейном пространстве L_n — n^p компонент.

Таким образом, совокупность коэффициентов a_i линейной формы $\varphi = \varphi(\mathbf{x})$ представляет собой тензор валентности 1. Поскольку скалярное произведение произвольного постоянного вектора \mathbf{a} на переменный вектор \mathbf{x} представляет собой линейную форму, то совокупность координат a_i произвольного вектора \mathbf{a} также представляет собой тензор валентности 1.

Точно так же совокупность коэффициентов a_{ij} билинейной формы $\varphi = \varphi(\mathbf{x}, \mathbf{y})$, образующая матрицу $A = (a_{ij})$,

представляет собой тензор валентности 2. В частности, таким тензором будет совокупность симметричных символов Кронекера δ_{ij} , так как они являются коэффициентами билинейной формы $\varphi = xy$. Этот тензор называется *единичным тензором*.

Еще один пример тензора представляет совокупность кососимметричных символов Кронекера ε_{ijk} — они являются коэффициентами трилинейной формы $\varphi = (x, y, z)$. Валентность этого вектора равна трем. Тензор ε_{ijk} называется *дискриминантным тензором*.

Скалярная величина, не зависящая от выбора ортонормированного базиса пространства, называется *тензором нулевой валентности*. Тензор нулевой валентности называют также *инвариантом*, так как его единственная компонента не меняет своего значения при преобразовании базиса.

Два тензора называются равными, если тождественно равны определяющие их полилинейные формы. Равные тензоры имеют одинаковую валентность, и их соответствующие компоненты попарно равны в любой системе координат. В самом деле, тождество

$$\varphi(x, y, z, \dots, w) = \psi(x, y, z, \dots, w)$$

в координатной форме можно записать в виде

$$a_{ijk\dots m}x_iy_jz_k\dots w_m = b_{ijk\dots m}x_iy_jz_k\dots w_m,$$

откуда непосредственно следует

$$a_{ijk\dots m} = b_{ijk\dots m}.$$

Тензор называется *нулевым*, если определяющая его полилинейная форма

$$\varphi = \varphi(x, y, z, \dots, w),$$

тождественно равна нулю. Все компоненты такого тензора равны нулю.

4. При переходе к новому базису координаты векторов, являющихся аргументами полилинейной формы, меняются по определенному закону. Поэтому коэффициенты полилинейной формы также будут изменяться совершенно определенным образом. Этот закон преобразования компонент тензора устанавливается следующей теоремой.

Теорема. Для того чтобы совокупность величин $a_{ijk\dots m}$, зависящая от выбора базиса, была тензором, необходимо и достаточно, чтобы при переходе ортонормированного базиса $\{e_i\}$ к такому же базису $\{e'_i\}$ она изменялась по зако-

ну

$$a_{i'j'k'...m'} = \gamma_{i'} \gamma_{j'} \gamma_{k'} \dots \gamma_{m'} a_{ijk...m}. \quad (10.18)$$

Докажем сначала необходимость условия теоремы. Пусть $a_{ijk...m}$ — тензор. Тогда величины $a_{ijk...m}$ представляют совокупность коэффициентов полилинейной формы $\varphi = \varphi(x, y, z, \dots, w)$:

$$a_{ijk...m} = \varphi(e_i, e_j, e_k, \dots, e_m).$$

В новом базисе коэффициенты этой формы вычисляются по аналогичным формулам

$$a_{i'j'k'...m'} = \varphi(e_{i'}, e_{j'}, e_{k'}, \dots, e_{m'}).$$

Но векторы $e_{i'}$ нового базиса выражаются через векторы e_i старого по формулам

$$e_{i'} = \gamma_{i'i} e_i,$$

поэтому

$$a_{i'j'k'...m'} = \varphi(\gamma_{i'i} e_i, \gamma_{j'j} e_j, \gamma_{k'k} e_k, \dots, \gamma_{m'm} e_m).$$

И так как форма φ полилинейная, то

$$a_{i'j'k'...m'} = \gamma_{i'i} \gamma_{j'j} \gamma_{k'k} \dots \gamma_{m'm} \varphi(e_i, e_j, \dots, e_m).$$

С учетом (10.17) очевидно, что (10.18) доказано.

Теперь докажем достаточность. Пусть величины $a_{ijk...m}$ при переходе к новому базису преобразуются по формулам (10.18). Рассмотрим p векторов x, y, z, \dots, w . Их разложения по старому и новому базисам можно записать в виде

$$x = x_i e_i = x_{i'} e_{i'}, \quad y = y_j e_j = y_{j'} e_{j'},$$

$$z = z_k e_k = z_{k'} e_{k'}, \quad w = w_m e_m = w_{m'} e_{m'}.$$

Чтобы доказать, что система величин $a_{ijk...m}$ образует тензор, нужно доказать, что выражение

$$\varphi = a_{ijk...m} x_i y_j z_k \dots w_m$$

является полилинейной формой, т. е. что оно зависит только от выбора векторов x, y, z, \dots, w и не зависит от выбора базиса. После преобразования базиса это выражение перейдет в

$$\varphi' = a_{i'j'k'...m'} x_{i'} y_{j'} z_{k'} \dots w_{m'}.$$

Подставляя сюда значения коэффициентов $a_{i'j'k'...m'}$ из соотношения (10.18) и $x_{i'}$ из соотношения (10.13), а

$y_{i'}, z_{k'}, \dots, w_{m'}$ из аналогичных соотношений, получим

$$\begin{aligned}\Phi' &= \gamma_{i'i} \gamma_{j'j} \gamma_{k'k} \dots \gamma_{m'm} a_{ijk\dots m} \gamma_{i'p} x_p \gamma_{j'q} y_q \gamma_{k'r} z_r \dots \gamma_{m's} w_s = \\ &= (\gamma_{i'i} \gamma_{i'p}) (\gamma_{j'j} \gamma_{j'q}) (\gamma_{k'k} \gamma_{k'r}) \dots (\gamma_{m'm} \gamma_{m's}) \times \\ &\quad \times a_{ijk\dots m} x_p y_q z_r \dots w_s.\end{aligned}$$

В силу свойств ортогональных матриц

$$\gamma_{i'i} \gamma_{i'p} = \delta_{ip}; \quad \gamma_{j'j} \gamma_{j'q} = \delta_{jq}, \quad \dots, \quad \gamma_{m'm} \gamma_{m's} = \delta_{ms},$$

поэтому

$$\Phi' = a_{ijk\dots m} \delta_{ip} x_p \delta_{jq} y_q \dots \delta_{ms} w_s = a_{ijk\dots m} x_i y_j z_k \dots w_m = \Phi,$$

что и требовалось доказать.

§ 5. АЛГЕБРАИЧЕСКИЕ ОПЕРАЦИИ НАД ТЕНЗОРАМИ

1. Сложение тензоров. Пусть $\Phi = \Phi(x, y, z, \dots, w)$ и $\Psi = \Psi(x, y, z, \dots, w)$ — две полилинейные формы от одних и тех же векторных аргументов одной и той же степени p .

Их суммой $\Phi + \Psi$, как легко увидеть, будет полилинейная форма той же степени p . Суммой тензоров $a_{ijk\dots m}$ и $b_{ijk\dots m}$ валентности p , определяемых полилинейными формами Φ и Ψ , назовем тензор $c_{ijk\dots m}$, определяемый формой $\Phi + \Psi$. Так как

$$\Phi + \Psi = (a_{ijk\dots m} + b_{ijk\dots m}) x_i y_j z_k \dots w_m,$$

то компоненты тензора $c_{ijk\dots m}$ связаны с компонентами тензоров $a_{ijk\dots m}$ и $b_{ijk\dots m}$ соотношениями

$$c_{ijk\dots m} = a_{ijk\dots m} + b_{ijk\dots m}.$$

2. Умножение тензора на скаляр. Произведение $\lambda\Phi$ полилинейной формы Φ степени p на действительное число λ является полилинейной формой той же степени p . Произведением тензора $a_{ijk\dots m}$ валентности p , определяемого формой Φ , на число λ , называется тензор $b_{ijk\dots m}$ той же валентности, определяемой формой $\lambda\Phi$. Так как

$$\lambda\Phi = (\lambda a_{ijk\dots m}) x_i y_j z_k \dots w_m,$$

то

$$b_{ijk\dots m} = \lambda a_{ijk\dots m}.$$

Из сказанного выше следует, что совокупность полилинейных форм степени p , так же как и совокупность тен-

зоров валентности p , образует линейное пространство размерности 3^p . Такое пространство называют p -кратным тензорным произведением линейного пространства L_3 . Базисом этого пространства могут служить, например, 3^p p -линейных форм вида

$$\varphi_{ijk\dots m} = x_i y_j z_k \dots \omega_m.$$

3. У м н о ж е н и е т е н з о р о в. Пусть φ и ψ — две полилинейные формы соответственно степеней p и q от различных векторных аргументов. Тогда их произведение $\varphi\psi$ будет полилинейной формой степени $p + q$. Например, если

$$\varphi = \varphi(x, y, z) —$$

трилинейная, а

$$\psi = \psi(u, v) —$$

билинейная форма, то их произведение

$$\varphi(x, y, z) \cdot \psi(u, v)$$

будет полилинейной формой степени пять.

Формы φ и ψ определяют тензоры соответственно валентностей p и q .

Назовем *произведением тензоров*, определяемых формами φ и ψ , тензор, определяемый их произведением $\varphi \cdot \psi$.

Так как форма $\varphi\psi$ имеет степень $p + q$, то произведением тензоров валентности p и q является тензор валентности $p + q$. Например, формы

$$\varphi(x, y, z) = a_{ijk} x_i y_j z_k$$

и

$$\psi(u, v) = b_{lm} u_l v_m$$

определяют соответственно тензоры a_{ijk} и b_{lm} валентностей 3 и 2, а их произведение

$$\varphi(x, y, z) \psi(u, v) = (a_{ijk} b_{lm}) x_i y_j z_k u_l v_m —$$

тензор $a_{ijk} b_{lm}$ валентности 5, который является произведением тензоров a_{ijk} и b_{lm} .

Таким образом, компоненты произведения двух тензоров представляют собой произведения каждой компоненты первого тензора на каждую компоненту второго.

4. С в е р т ы в а н и е т е н з о р а. Пусть $\varphi = \varphi(x, y, z, \dots, \omega)$ — полилинейная форма степени p . Подставим в нее вместо каких-либо двух аргументов, например x и y ,

базисные векторы e_i и e_j

$$\varphi_{ij} = \varphi(e_i, e_j, z, \dots, w).$$

Эти выражения являются линейными функциями векторных аргументов

$$z, \dots, w,$$

но они не являются линейными формами, так как зависят еще от выбора базиса. Найдем, как выражения φ_{ij} изменяются при преобразованиях базиса пространства L_3 . Если обозначить

$$\varphi_{i'j'} = \varphi(e_{i'}, e_{j'}, z, \dots, w),$$

то, так как

$$e_{i'} = \gamma_{i'i} e_i, \quad e_{j'} = \gamma_{j'j} e_j,$$

имеем

$$\begin{aligned} \varphi_{i'j'} &= \varphi(\gamma_{i'i} e_i, \gamma_{j'j} e_j, z, \dots, w) = \\ &= \gamma_{i'i} \gamma_{j'j} \varphi(e_i, e_j, z, \dots, w) = \gamma_{i'i} \gamma_{j'j} \varphi_{ij}. \end{aligned}$$

Положим теперь, что $i' = j'$, и сложим три получающихся при этом равенства. Тогда

$$\varphi_{i'i'} = \gamma_{i'i} \gamma_{i'j} \varphi_{ij}.$$

Но по свойству ортогональных матриц

$$\gamma_{i'i} \gamma_{i'j} = \delta_{ij}$$

и

$$\varphi_{i'i'} = \delta_{ii} \varphi_{ii} = \varphi_{ii}.$$

Эти соотношения показывают, что выражение φ_{ii} , которое линейно зависит от векторных аргументов z, \dots, w , не зависит от выбора базиса и, следовательно, является полилинейной формой от z, \dots, w . Степень этой полилинейной формы равна $p - 2$, так как число векторных аргументов, от которых она зависит, на две единицы меньше числа аргументов, от которых зависит форма φ .

Запишем теперь исходную форму φ в координатах

$$\varphi = \varphi(x, y, z, \dots, w) = a_{ijk\dots m} x_i y_j z_k, \dots, w_m.$$

Если положить здесь

$$x = e_i, \quad y = e_j,$$

то будем иметь

$$x_i = 1, \quad x_p = 0 \quad \text{при} \quad p \neq i;$$

$$y_j = 1, \quad y_q = 0 \quad \text{при} \quad q \neq j.$$

Поэтому выражения φ_{ij} примут вид

$$\varphi_{ij} = \varphi(e_i, e_j, z, \dots, w) = a_{ijk\dots m} z_k \dots w_m.$$

Отсюда вытекает, что

$$\varphi_{ii} = a_{iik\dots m} z_k \dots w_m.$$

Следовательно, компоненты тензора $b_{k\dots m}$ валентности $p-2$, определяемого формой φ_{ii} , выражаются через компоненты тензора $a_{ijk\dots m}$, определяемого исходной формой φ , по формулам

$$b_{k\dots m} = a_{iik\dots m},$$

или, более подробно, по формулам

$$b_{k\dots m} = a_{11k\dots m} + a_{22k\dots m} + a_{33k\dots m}.$$

Операция получения тензора

$$b_{k\dots m}$$

из тензора $a_{ijk\dots m}$ называется *свертыванием* тензора $a_{ijk\dots m}$ по индексам i и j .

Точно так же можно определить свертывание тензора

$$a_{ijk\dots m}$$

по любой другой паре индексов.

При свертывании тензора его валентность понижается на две единицы. Например, при свертывании двухвалентного тензора a_{ij} мы получим тензор a_{ii} нулевой валентности, т. е. инвариант.

Этот инвариант называется *следом тензора* a_{ij} и обозначается

$$a_{ii} = Sp(a_{ij}) \text{ или } tr(a_{ij}).$$

Spur (нем.), *trace* (англ.).

5. **С в е р т ы в а н и е п р о и з в е д е н и я т е н - з о р о в.** Рассмотрим два произвольных тензора

$$a_{ijk} \text{ и } b_{lm}$$

валентностей 3 и 2 и образуем их произведение $a_{ijk}b_{lm}$ — пятивалентный тензор. Теперь свернем полученный тензор, например, по индексам k и e . В результате получим тензор

$$a_{ijk}b_{km} = a_{ijl}b_{lm} + b_{ij2}b_{2m} + a_{ij3}b_{3m}$$

валентности 3. Такая операция называется *свертыванием* тензоров a_{ijk} и b_{lm} по индексам k и l .

Операция свертывания двух тензоров состоит в их умножении и свертывании полученного в результате умножения

тензора по индексам, принадлежащим разным сомножителям.

В результате свертывания тензоров валентности p и q получается тензор валентности $p + q - 2$.

По существу, операция свертывания уже встречалась.

а) Например, скалярное произведение векторов $\mathbf{x} = x_i \mathbf{e}_i$ и $\mathbf{y} = y_i \mathbf{e}_i$, которое вычисляется по формуле

$$\mathbf{x}\mathbf{y} = x_i y_i,$$

представляет собой результат свертывания одновалентных тензоров x_i и y_i — координат векторов \mathbf{x} и \mathbf{y} .

б) Линейная форма $\phi(\mathbf{x}) = a_i x_i$ является результатом свертывания тензоров a_i и x_i ;

в) Билинейная форма $\phi(\mathbf{x}, \mathbf{y}) = a_{ij} x_i y_j$ является результатом свертывания тензора a_{ij} с тензором x_i и последующего свертывания $a_{ij} x_i$ с тензором y_j и т. д.

Весьма простой характер носит свертывание произвольного тензора с единичным тензором δ_{ij} . Например,

$$a_{ijk} \delta_{kl} = a_{ijl} \delta_{1l} + a_{ij2} \delta_{2l} + a_{ij3} \delta_{3l} = a_{ijl},$$

так как δ_{kl} отлично от нуля только при $k = l$.

Как видно из примеров, свертывание тензоров можно производить не только по одной паре индексов, а по любому количеству r таких пар. В результате этого свертывания получается новый тензор, валентность которого на $2r$ единиц меньше суммы валентностей исходных тензоров.

Теорема (обратный тензорный признак). Пусть в каждом ортонормированном базисе задана совокупность 3^{p+q} чисел $a_{i_1 \dots i_p j_1 \dots j_q}$ такая, что при свертывании ее с произвольным тензором $t_{j_1 \dots j_q}$ валентности q снова получается тензор валентности p . Тогда исходная система чисел является тензором валентности $p + q$.

Докажем эту теорему для частотного случая, когда $p = 3$, $q = 2$ и заданная система чисел имеет вид a_{ijklm} . По условию теоремы величины

$$s_{ijk} = a_{ijklm} t_{lm}$$

образуют тензор, если только t_{lm} — тензор.

Пусть $t_{lm} = \mathbf{u}_l \mathbf{v}_m$ — произведение векторов \mathbf{u}_l и \mathbf{v}_m . Тогда

$$s_{ijk} = a_{ijklm} \mathbf{u}_l \mathbf{v}_m.$$

Свернем это выражение с произвольными векторами $\mathbf{x}_i, \mathbf{y}_j, \mathbf{z}_k$:

$$s_{ijk} \mathbf{x}_i \mathbf{y}_j \mathbf{z}_k = a_{ijklm} \mathbf{x}_i \mathbf{y}_j \mathbf{z}_k \mathbf{u}_l \mathbf{v}_m.$$

Так как s_{ijk} — тензор, то выражение в левой части этого равенства представляет скалярную функцию. Правая часть этого выражения линейно зависит от координат векторов x, y, z, \dots, w , поэтому эта скалярная функция является полилинейной формой степени пять. Следовательно, числа a_{ijklm} , являющиеся коэффициентами этой полилинейной формы, образуют тензор валентности пять. Так же доказывается эта теорема в общем случае.

6. Перестановка индексов тензора. Пусть $\varphi = \varphi(x, y, z, \dots, w)$ — полилинейная форма и $a_{ijk\dots m}$ — определяемый ею тензор, так что

$$\varphi = a_{ijk\dots m} x_i y_j z_k \dots w_m.$$

Рассмотрим форму ψ , которая получается из формы φ путем перестановки некоторых ее аргументов. Например:

$$\psi(x, y, z, \dots, w) = \varphi(y, z, x, \dots, w).$$

Если обозначить через $b_{ijk\dots m}$ тензор, определяемый формой ψ , то последнее соотношение перепишем в виде

$$b_{ijk\dots m} x_i y_j z_k \dots w_m = a_{ijk\dots m} y_i z_j x_k \dots w_m.$$

Если поменять индексы суммирования в правой части и учесть, что это соотношение является тождеством, получим

$$b_{ijk\dots m} = a_{jki\dots m}.$$

Тензор $b_{ijk\dots m}$ отличается от тензора $a_{ijk\dots m}$ только другой нумерацией своих компонент. Операция, состоящая в перенумеровании компонент тензора $a_{ijk\dots m}$, называется перестановкой индексов тензора.

Заметим, что тензоры $a_{ijk\dots m}$ и $b_{ijk\dots m}$ — существенно различные тензоры, так как их соответствующие компоненты (компоненты с одинаковыми индексами), вообще говоря, не равны между собой.

§ 6. ТЕНЗОРНЫЙ АНАЛИЗ

Тензорное поле

Все введенные ранее операции над тензорами выполнялись в одной точке. Можно, однако, считать, что все эти операции осуществляются в некоторой области V евклидова пространства E_3 (или во всем E_3), если считать, что тензоры одинаковы во всех точках области V .

В связи с переходом от тензорной алгебры к тензорному анализу будем рассматривать *тензорные поля*. Для них определяется дополнительная операция — дифференцирование.

Говорят, что в области $V \subset E$ задано *тензорное поле*, если каждой точке $M \in V$ поставлен в соответствие тензор одной и той же валентности. Этот тензор поля в общем случае меняется от точки к точке.

Примеры тензорных полей

1). Скалярное поле — поле тензора нулевой валентности. Учитывая инвариантность такого тензора для задания скалярного поля надо задать инвариант

$$\varphi = \varphi(M) \quad \text{или} \quad \varphi = (x_1, x_2, x_3)$$

в каждой точке $M \in V$. Конкретные скалярные поля: поле температур, поле плотностей неоднородной среды, поле давлений газа, поле количества информации.

2). Векторное поле — поле тензора первой валентности

$$a_i = a_i(x_1, x_2, x_3). \quad (10.19)$$

Из (10.19) очевидно, что векторное поле задается тремя функциями от трех аргументов. Примерами являются: поле вектора скорости или ускорения, движущегося потока (жидкости, газа, информации), поле гравитационных сил и т. д.)

3). Поле двухвалентного тензора

$$a_{ij} = a_{ij}(x_1, x_2, x_3) \quad (10.20)$$

задается уже девятью функциями от трех аргументов. Примеры: поле напряжений и поле деформаций твердого тела.

Операции тензорного анализа

Операции тензорной алгебры естественно переносятся на тензорные поля. Считается, что эти операции производятся над тензором поля в каждой точке $M \in V$.

Пусть, например, заданы поля

$$a_{ijk} = a_{ijk}(M), \quad (10.21)$$

$$b_{ijk} = b_{ijk}(M), \quad (10.22)$$

$$c_{ij} = c_{ij}(M). \quad (10.23)$$

Складывая компоненты (10.21) и (10.22) в каждой точке $M \in V$, получим в V новое тензорное поле

$$d_{ijk}(M) = a_{ijk}(M) + b_{ijk}(M), \quad (10.24)$$

представляющее собой сумму полей (10.21) и (10.22). Перемножая в каждой точке $M \in V$ компонент (10.22) и (10.23), получим новое тензорное поле

$$f_{ijkem}(M) = a_{ijk}(M) c_{lm}(M) \quad (10.25)$$

— произведение соответствующих полей.

Точно так же можно рассматривать операции свертывания тензорных полей и перестановки индексов в данном тензорном поле.

Основной операцией тензорного анализа является операция дифференцирования, которая вводится следующим образом.

Пусть в области $V \subset E_3$ задано поле тензора

$$a_{ijk} = a_{ijk}(M). \quad (10.26)$$

Определим характер изменения этого тензора при переходе из $M(x_1, x_2, x_3)$ в бесконечно близкую к M точку M' . Разложение вектора $dx = MM'$ по базисным векторам имеет вид

$$dx = dx_i e_i. \quad (10.27)$$

В новом ортонормированном базисе $e_{i'} = \gamma_{i'i} e_i$ координаты dx преобразуются по формулам

$$dx_{i'} = \gamma_{i'i} dx_i. \quad (10.28)$$

Поскольку $OM' = OM + MM'$, компоненты $x'_{i'}$ точки M' получим в виде

$$x'_{i'} = x_i + dx_i.$$

Через Δa_{ijk} обозначим приращения компонентов тензора a_{ijk} при переходе из точки M в точку M' . В предположении, что компоненты являются дифференцируемыми функциями от координат точки M , главные части приращений могут быть записаны в виде

$$da_{ijk} = \frac{da_{ijk}}{dx_e} dx_e. \quad (10.29)$$

Совокупность величин da_{ijk} образует тензор третьей валентности. В самом деле, при переходе к базису $\{e_1, e_2, e_3\}$ компоненты тензора a_{ijk} преобразуются по формулам

$$a_{i'j'k'} = \gamma_{j'j} \gamma_{i'i} \gamma_{k'k} a_{ijk}. \quad (10.30)$$

После почленного дифференцирования, учитывая, что $\gamma_{i'i}$ постоянны (не зависят от положения точки M , так как являются косинусами углов между векторами старого и нового

базисов), получим

$$da_{i'j'k'} = \gamma_{i'i} \gamma_{j'j} \gamma_{k'k} da_{ijk}. \quad (10.31)$$

Таким образом, величины da_{ijk} при замене базиса преобразуются по тензорному закону.

Назовем *абсолютным дифференциалом* тензора поля a_{ijk} тензор с координатами da_{ijk} .

Выражение (10.29) показывает, что при свертывании величин $\frac{da_{ijk}}{dx_l}$ с координатами произвольного вектора dx_l получается тензор da_{ijk} . Из этого следует, что величины $\frac{da_{ijk}}{dx_l}$ образуют тензор четвертой валентности в точке M .

Так как эти построения можно проводить в любой точке $M \in V$, получим в V новое тензорное поле, которое назовем *абсолютной произвольной* тензорного поля a_{ijk} . В обозначении $\frac{da_{ijk}}{dx_l} = a_{ijk,l}$ для абсолютной производной тензора поля a_{ijk} в качестве добавочного индекса e на последнем месте ставится индекс той координаты x_e точки M , по которой производится дифференцирование, причем этот индекс от остальных отделится запятой. В новых обозначениях (10.29) имеет вид

$$da_{ijk} = a_{ijk,e} dx_e. \quad (10.32)$$

Абсолютный дифференциал тензора a_{ijk} представляет собой результат свертывания тензора dx_l и абсолютной производной $a_{ijk,l}$ этого тензора.

В общем случае, *совокупность всех частных производных первого порядка от компонент данного тензорного поля по координатам x_e той точки, в которой рассматривается тензорное поле, образует тензор валентности на единицу большей, чем валентность исходного тензорного поля — абсолютную производную данного поля. Результат свертывания этой абсолютной производной по индексу дифференцирования с координатами вектора dx представляет собой абсолютный дифференциал заданного векторного поля.*

Сделаем некоторые обобщения.

1. В прямоугольной декартовой системе координат координаты абсолютного дифференциала и абсолютной производной тензорного поля совпадают с обычными дифференциалами и частными производными компонент исходного поля. Поэтому правила абсолютного дифференцирования те же, что и для обычного дифференцирования.

2. Абсолютный дифференциал и абсолютную производную второго и более высоких порядков можно построить аналогично тому, как были построены абсолютный дифференциал и абсолютная производная первого порядка. Вторую абсолютную производную тензора d_{ijk} для рассмотренного тензорного поля образуют частные производные

$$\frac{da_{ijk}}{dx_m dx_l} = a_{ijk, l_m}. \quad (10.33)$$

Результат свертывания второй абсолютной производной (тензора пятой валентности) с дифференциалами dx_e и dx_m дает выражение второго абсолютного дифференциала тензора a_{ijk}

$$a^2 a_{ijk} = a_{ijk, l_m} dx_e dx_m. \quad (10.34)$$

В общем случае, если произвольное тензорное поле дифференцируемо не менее p раз, абсолютный дифференциал порядка p от этого поля представляет собой тензорное поле той же валентности, что и исходное поле, а его абсолютная производная порядка p — тензорное поле, валентность которого на p единиц больше валентности исходного поля.

3. Пусть функции, определяющие тензорное поле, имеют непрерывные частные производные $(n+1)$ -го порядка в точке $M(x_1, x_2, x_3)$ и ее окрестности. Можно разложить эти функции в окрестности точки M в ряд Тейлора:

$$\begin{aligned} a_{ijk}(x_1 + \Delta x_1, x_2 + \Delta x_2, x_3 + \Delta x_3) = \\ = a_{ijk}(x_1, x_2, x_3) + da_{ijk}(x_1, x_2, x_3) + \\ + \frac{1}{2!} d^2 a_{ijk}(x_1, x_2, x_3) + \dots + \frac{1}{n!} d^n a_{ijk}(x_1, x_2, x_3) + \\ + \frac{1}{(n+1)!} d^{n+1} a_{ijk}(x_1 + \theta_{ijk} \Delta x_1, x_2 + \theta_{ijk} \Delta x_2, x_3 + \theta_{ijk} \Delta x_3), \end{aligned}$$

а в развернутом виде

$$\begin{aligned} a_{ijk}(x_1 + \Delta x_1, x_2 + \Delta x_2, x_3 + \Delta x_3) = a_{ijk}(x_1, x_2, x_3) + \\ + a_{ijk, l_1}(x_1, x_2, x_3) \Delta x_{l_1} + \frac{1}{2} a_{ijk, l_1 l_2}(x_1, x_2, x_3) \Delta x_{l_1} \Delta x_{l_2} + \\ + \dots + \frac{1}{n!} a_{ijk, l_1 l_2 \dots l_n}(x_1, x_2, x_3) \Delta x_{l_1} \dots \Delta x_{l_n} + \\ + \frac{1}{(n+1)!} a_{ijk, l_1 l_2 \dots l_{n+1}}(x_1 + \theta_{ijk} \Delta x_1, x_2 + \theta_{ijk} \Delta x_2, \\ x_3 + \theta_{ijk} \Delta x_3) \Delta x_{l_1} \dots \Delta x_{l_n} \Delta x_{l_{n+1}}. \end{aligned} \quad (10.36)$$

В выражениях (10.35) и (10.36) $\Delta x_i = dx_i$ и $0 < \theta_{ijk} < 1$, причем θ_{ijk} в общем случае различны для различных наборов i, j, k . Коэффициенты в каждой группе членов формулы (10.36) являются тензорами.

Дифференцирование скалярного поля. Производная поля тензора нулевой валентности

$$\varphi = \varphi(x_1, x_2, x_3) \quad (10.37)$$

в соответствии с общим правилом имеет вид

$$\varphi_i = \frac{d\varphi}{dx_i}. \quad (10.38)$$

Это тензор первой валентности, определяющий векторное поле, называемое *градиентом скалярного поля*

$$\text{grad } \varphi = \varphi_{,i} e_i. \quad (10.39)$$

Инвариантность полученного поля одновалентного тензора очевидна. Градиент скалярного поля в данной точке M — это вектор, в направлении которого скалярное поле возрастает с наибольшей скоростью и модуль которого равен этой наибольшей скорости.

Дифференцирование векторного поля. Абсолютная производная поля первой валентности

$$a_i = a_i(x_1, x_2, x_3) \quad (10.40)$$

равна

$$a_{i,k} = \frac{da_i}{dx_k}.$$

Это тензор второй валентности $a_{i,k}$, называемый *градиентом векторного поля*. Запишем выражение, связывающее абсолютный дифференциал da_i и абсолютную производную $a_{i,k}$

$$da_i = a_{i,k} dx_k, \quad (10.41)$$

где dx_k — координаты вектора $d\mathbf{x} = \overline{MM'}$.

Тензор второй валентности $a_{i,k}$ порождает линейное преобразование

$$\mathbf{y} = \mathbf{A}\mathbf{x}, \quad (10.42)$$

а в координатной форме

$$y_i = a_{i,k} x_k. \quad (10.43)$$

Тогда (10.41) переписывается в виде

$$d\mathbf{a} = \mathbf{A}(\mathbf{M}) d\mathbf{x}. \quad (10.44)$$

Так как $da \approx a(M') - a(M)$, $dx = \overline{MM'}$, то с точностью до бесконечно малых высшего порядка линейное преобразование $A(M)$, действуя на вектор бесконечно малого смещения $\overline{MM'} = dx$, дает соответствующее приращение векторного поля $a(M)$

$$a(M') - a(M) = \Delta a(M) \approx A(M) dx. \quad (10.45)$$

Выражение (10.45) показывает, что линейное преобразование $A(M)$ определяет главную линейную часть приращения векторного поля a в точке M . След линейного преобразования $A(M)$

$$Sp A(M) = a_{i,i} = \operatorname{div} a \quad (10.46)$$

представляет собой инвариант, который называется *дивергенцией* или дифференциальным инвариантом поля a . Поле этого инварианта — скалярное поле, определенное в области V .

Рассмотрим вектор z , координаты которого получаются при свертывании тензора $a_{i,j}$ с дискриминантным тензором ε_{ijk} . Положим, что

$$z_i = -\varepsilon_{ijk} a_{j,k}. \quad (10.47)$$

Вектор z называется *ротором* векторного поля a

$$z = \operatorname{rot} a. \quad (10.48)$$

Расписав (10.47) для координат вектора z , получим

$$\begin{aligned} z_1 &= (a_{3,2} - a_{2,3}) \varepsilon, \\ z_2 &= (a_{1,3} - a_{3,1}) \varepsilon, \\ z_3 &= (a_{2,1} - a_{1,2}) \varepsilon, \end{aligned} \quad (10.49)$$

где ε равен $+1$ в правой и -1 в левой системах координат.

Компоненты вектора $\operatorname{rot} a$ с точностью до множителя ε совпадают с компонентами удвоенного альтернированного тензора $a_{i,k}$. Таким образом, с полем $a(M)$, определенным в области V , инвариантно связывается поле вектора $\operatorname{rot} a$, определенное в той же области V . Если в области V $\operatorname{div} a = 0$, векторное поле $a = a(M)$ называется *соленоидальным*. Если в области V $\operatorname{rot} a = 0$, векторное поле называется *безвихревым*.

Найдем дивергенцию векторного поля, являющегося градиентом скалярного поля $\varphi(M)$

$$a = \operatorname{grad} \varphi = \varphi_{,i} e_i = \frac{d\varphi}{dx_i} e_i. \quad (10.50)$$

Учитывая, что $a_{i,k} = \varphi_{,ik}$, имеем

$$\operatorname{div} \mathbf{a} = \varphi_{,ii} = \frac{d^2 \varphi}{dx_1^2} + \frac{d^2 \varphi}{dx_2^2} + \frac{d^2 \varphi}{dx_3^2}. \quad (10.51)$$

Оператор

$$\Delta = \frac{d^2}{dx_1^2} + \frac{d^2}{dx_2^2} + \frac{d^2}{dx_3^2}$$

называется *лапласианом* (оператором Лапласа). При помощи лапласиана дивергенцию векторного поля φ можно записать в виде

$$\operatorname{div} (\operatorname{grad} \varphi) = \Delta \varphi. \quad (10.52)$$

Если скалярное поле $\varphi (M)$ удовлетворяет условию

$$\Delta \varphi = 0, \quad (10.53)$$

то такое поле называется *гармоническим* (или Лапласовым).

Уравнение

$$\frac{d^2 \varphi}{dx_1^2} + \frac{d^2 \varphi}{dx_2^2} + \frac{d^2 \varphi}{dx_3^2} = 0,$$

которому удовлетворяет функция $\varphi (x_1, x_2, x_3)$, определяющая такое поле, называется уравнением Лапласа, а сама функция $\varphi (x_1, x_2, x_3)$ носит название гармонической.

Нестационарное тензорное поле

Рассмотренные поля характерны тем, что их тензоры зависят от положения точки в пространстве, но не зависят от момента времени, в который рассматривается это поле. Такие поля называются *стационарными* тензорными полями.

Нестационарными тензорными полями будем называть такие, для которых тензор поля зависит не только от положения точки в пространстве, но и от времени.

Для трехвалентного тензорного поля, например, имеем

$$a_{ijk} = a_{ijk} (x_1, x_2, x_3, t). \quad (10.54)$$

Скорость изменения тензорного поля во времени в некоторой неподвижной точке M описывается частными производными

$\frac{da_{ijk}}{dt}$, которые образуют тензоры той же валентности, что и исходное поле.

Пусть нестационарное тензорное поле a_{ijk} описывает некоторое свойство среды, частицы которой находятся в движении. Определим, как изменятся компоненты тензора a_{ijk} , связанные с определенной частицей, при ее движении.

Траекторию движения частицы опишем в виде

$$x_i = x_i(t). \quad (10.55)$$

При этом скорость изменения компонент тензора a_{ijk} , связанных с частицей, равна

$$\frac{da_{ijk}}{dt} = \frac{da_{ijk}}{dt} + \frac{da_{ijk}}{dx_l} \frac{dx_l}{dt}. \quad (10.56)$$

Учитывая, что $\frac{da_{ijk}}{dx_l} = a_{ijk,l}$, а $\frac{dx_l}{dt} = v_l$ — компоненты скорости частицы движущейся материальной среды, перепишем (10.56) в виде

$$\frac{da_{ijk}}{dt} = \frac{da_{ijk}}{dt} + a_{ijk,l} v_l. \quad (10.57)$$

Первый член правой части (10.57) описывает изменение компонент тензора a_{ijk} в неподвижной точке M , а второй член связан с движением частицы в пространстве (переносный член).

Выражения типа (10.57), очевидно, справедливы для нестационарных тензорных полей любой валентности. Для скалярного поля $\varphi = \varphi(M, t)$ имеем

$$\frac{d\varphi}{dt} = \frac{d\varphi}{dt} + \mathbf{v} \operatorname{grad} \varphi. \quad (10.58)$$

Для нестационарного векторного поля $\mathbf{a} = \mathbf{a}(M, t)$

$$\frac{da_i}{dt} = \frac{da_i}{dt} + a_{i,k} v_k. \quad (10.59)$$

Формула (10.59) равносильна выражению

$$\frac{d\mathbf{a}}{dt} = \frac{d\mathbf{a}}{dt} + \mathbf{A}(M) \mathbf{v}, \quad (10.60)$$

где $\mathbf{A}(M)$ — линейное преобразование, определяемое тензором $a_{i,k}$.

Контрольные вопросы и задания

1. Дайте определение тензорного поля.
2. Приведите примеры тензорных полей различной валентности.
3. Что такое абсолютный дифференциал и абсолютная производная тензорного поля?
4. Охарактеризуйте задачу дифференцирования тензорного поля произвольной валентности.
5. В чем состоит тензорная сущность градиента, дивергенции, ротора?
6. Дайте определение нестационарного тензорного поля.
7. Охарактеризуйте физическую сущность и приведите примеры нестационарных тензорных полей.

ЛИТЕРАТУРА

К главе 1

С о б о л е в В. И. Лекции по дополнительным главам математического анализа. М., Наука, 1968.

В книге излагаются элементы общей теории множеств, теории точечных множеств на прямой и плоскости, основы теории метрических пространств и множеств в них. Дается построение интеграла по абстрактным множествам. Приведены основные сведения о функциях с ограниченной вариацией и абсолютно непрерывных функциях, включая дифференциальные свойства таких функций. Рассматриваются линейные нормированные пространства и простейшие свойства операторов, действующих в них.

К е м е н и Д., С н е л л Д., Т о м п с о н Д. Введение в конечную математику. М., «Мир», 1965.

Переведенная с английского языка книга известных американских математиков посвящена вопросам дискретной математики. Во второй главе излагаются элементы теории множеств. Даны основные понятия, определения, соотношения между множествами, операции. В книге содержится большое количество примеров и упражнений.

Г р а д ш т е й н И. С. Прямая и обратная теоремы. М., «Наука», 1965.

В первой главе в доступной для широкого читателя форме освещаются элементы теории множеств.

У с п е н с к и й В. А. Лекции о вычислимых функциях. М., Физматгиз, 1960.

Основным вопросам теории множеств посвящен § 2.

К главе 2

Н о в и к о в П. С. Элементы математической логики. М., Физматгиз, 1959.

Систематически изложены классические исчисления высказываний и предикатов, а также формальная система арифметики. Освещаются различные вопросы, относящиеся к основаниям математики. Книга предназначена для читателя, впервые знакомящегося с математической логикой, не требует никакой специальной подготовки.

Я г л о м И. М. Необыкновенная алгебра. М., Наука, 1968.

Брошюра из серии «Популярные лекции по математике». В ней освещены основные понятия, относящиеся к «алгебрам Буля», играю-

щим большую роль в математической логике и всех направлениях современной математики, связанной с электронными вычислительными машинами и кибернетикой.

Градштейн И. С. Прямая и обратная теоремы. М., «Наука», 1965.

Глава вторая посвящена элементам математической логики.

Гильберт Д. и Аккерман В. Основы теоретической логики. Пер. с нем. М., Издательство иностранной литературы, 1947. Учебник по математической логике.

Клини С. К. Введение в метаматематику. М., Изд-во иностр. лит-ры, 1957. Монография по математической логике. В конце книги приведена обширная библиография.

Беркли Э. Символическая логика и разумные машины. М., Изд-во иностр. лит-ры, 1961.

Книга написана видным американским популяризатором кибернетики. В ней излагаются элементы математической логики и вопросы ее применения к синтезу машин, моделирующих некоторые операции человеческого мышления. Большое количество примеров. Книга представляет интерес для инженеров, работающих в области связи, автоматики и телемеханики.

Подлипенский В. С. Бесконтактные логические схемы автоматики. К., «Наукова думка», 1965.

Справочное руководство, позволяющее инженерам, техникам, научным работникам, студентам овладеть основами алгебры логики, эффективными методами синтеза экономичных бесконтактных логических схем. В книге приведены расчетные соотношения и таблицы, существенно облегчающие проектирование логических устройств. Много примеров. Дана обширная библиография отечественных и зарубежных работ по математической логике и ее инженерным приложениям.

К главе 3

Глушков В. М. Введение в кибернетику. К., Изд-во АН УРСР, 1964.

В книге собран и обобщен материал, необходимый для построения таких разделов кибернетики, как теория электронных цифровых машин, теория дискретных автоматов, теория самоорганизующихся систем, автоматизация мыслительных процессов и другие. Первая глава содержит изложенные теории алгоритмов.

Успенский В. А. Лекции о вычислимых функциях. М., Физматгиз, 1960.

Понятие вычислимой функции тесно связано с понятием алгоритма и является одним из центральных в математике и кибернетике. Книга дает систематическое изложение теории вычислимых функций.

Трахтенберг Б. А. Алгоритмы и машинное решение задач. М., Гостехиздат, 1957.

Брошюра в популярной форме знакомит с численными и логическими алгоритмами.

Айзерман М. А. и др. Логика, автоматы, алгоритмы. М., «Наука», 1967.

Глушков В. М. Синтез цифровых автоматов. М., Физматгиз, 1962.

Кобринский Н. Е., Трахтенберг Б. А. Введение в теорию конечных автоматов. М., Физматгиз, 1962.

К главе 4

Пугачев В. С. Теория случайных функций. М., Физматгиз, 1962.

Систематически изложена теория вероятностей и теория случайных функций. Рассмотрена общая теория линейных систем, методы исследования точности линейных и нелинейных, одномерных и многомерных систем. Особое внимание уделено современной статистической теории обнаружения и воспроизведения сигналов в присутствии помех. Применение изложенных методов наглядно показано на примерах.

Вентцель Е. С. Теория вероятностей. М., Физматгиз, 1962.

Один из лучших учебников по теории вероятностей, случайных величин и случайных функций.

Давенпорт В. Б. и Рут В. Л. Введение в теорию случайных сигналов и шумов. М., Изд-во иностр. лит-ры, 1960.

Учебное руководство по вероятностным методам. Рассчитано на студентов старших курсов, аспирантов и инженеров радиотехнических, электротехнических и радиофизических специальностей. Главы 2—6 представляют собой руководство по теории вероятностей и случайных процессов.

Гнеденко Б. В. и Хинчин А. Я. Элементарное введение в теорию вероятностей. М., Физматгиз, 1961.

Яглом И. М. и Яглом А. М. Вероятность и информация. М., Физматгиз, 1960.

Феллер В. Введение в теорию вероятностей и ее приложения. М., Изд-во иностр. лит-ры, 1952.

Кемени Д. и др. Введение в конечную математику. М., «Мир», 1965.

Глава IV посвящена теории вероятностей. Излагаются элементы теории марковских процессов.

К главе 5

Дунин-Барковский И. В., Смирнов Н. В. Курс теории вероятностей и математической статистики. М., «Наука», 1965.

Крамер Г. Математические методы статистики. М., Изд-во иностр. лит-ры, 1948.

Линник Ю. В. Метод наименьших квадратов и основы теории обработки наблюдений. М., Физматгиз, 1958.

Лукомский Я. И. Теория корреляции. М., Физматгиз, 1953.

Систематически изложена теория статистического измерения формы связи (регрессионный анализ) и тесноты связи. Книга изобилует большим количеством конкретных примеров, доведенных до числовых решений. Одно из лучших руководств по регрессионному анализу.

К главе 6

Харкевич А. А. Спектры и анализ. М., Гостехиздат, 1949. Монография, посвященная гармоническому анализу процессов.

Хеннан Э. Анализ временных рядов. М., «Наука», 1964.

Натансон И. П. Краткий курс высшей математики. М., «Наука», 1968.

В главе XI § 4 посвящен теории гармонических колебаний.

Рассмотрены свободные и вынужденные колебания, явление резонанса. В главе XIII в § 3 изложена теория рядов Фурье.

Заездный А. М. Гармонический синтез в радиотехнике и в электросвязи. М.—Л., ГЭИ, 1961.

Серебренников М. Г. и Первозванский А. А. Выявление скрытых периодичностей. М., «Наука», 1965.

К главе 7

Колмогоров А. И. Теория передачи информации. М., Изд-во АН СССР, 1956.

Шеннон К. Э. Статистическая теория передачи электрических сигналов. М., Изд-во иностр. лит-ры, 1953.

Файнштейн А. Основы теории информации. М., Изд-во иностр. лит-ры, 1960.

В книге дано систематическое изложение теории информации.

Бриллюэн Л. Наука и теория информации. М., Физматгиз, 1962.

Голдман С. Теория информации. М., Издательство иностранной литературы, 1957.

К главе 8

Вентцель Е. С. Элементы теории игр. М., Физматгиз, 1961.

Вентцель Е. С. Введение в исследование операций. М., «Сов. радио», 1964.

Книга посвящена вопросам рациональной организации целенаправленной деятельности. Большое внимание уделено описанию и исследованию конфликтных ситуаций. Двенадцатая глава полностью посвящена теории игр.

Льюис Р. Д., Райфа Х. Игры и решения. М., Издательство иностранной литературы, 1961.

Монография охватывает широкий круг вопросов теории игр и решений, теории полезности. Написана в форме, доступной для широкого читателя.

Вильямс Дж. Д. Совершенный стратег или Букварь по теории стратегических игр. М., «Сов. радио», 1960.

Элементарное изложение теории игр. Книга содержит большое количество примеров. Написана в живой, увлекательной форме. Может быть рекомендована для первого знакомства с предметом.

К главе 9

Оре О. Графы и их применение. М., «Мир», 1965.

Книга из популярной серии «Современная математика». В основной части доступная. Удачно подобраны примеры: простые интересные и практически важные.

Берж К. Теория графов и ее применение. М., ИЛ, 1962.

Фундаментальная монография, освещающая вопросы теории графов и ее применение.

Дынкин Е. Б. и Успенский В. А. Математические беседы. М.—Л., Гостехиздат, 1952.

Кроуэлл Р., Фокс Р. Введение в теорию узлов. М., «Мир», 1967.

Гроссман И., Магнус В. Группы и их графы. М., «Мир», 1971.

К главе 10

Рашевский П. К. Риманова геометрия и тензорный анализ. М., «Наука», 1964.

Фундаментальная монография по теории пространств, векторной алгебре и тензорному исчислению. Некоторые главы посвящены математическому аппарату теории относительности.

Ш и р о к о в П. А. Тензорное исчисление. Изд-во Казанского университета, 1961.

Учебник по векторному и тензорному исчислению для вузов.

А к и в и с М. А., Г о л ь д б е р г В. В. Тензорное исчисление. М., «Наука», 1969.

Книга из серии «Избранные главы высшей математики для инженеров и студентов втузов». Кроме изложения основ тензорного исчисления приведено большое количество задач и упражнений по использованию аппарата в физике, технике.

Б е р л е е н к о А. И., Т а р а п о в И. Е. Векторный анализ и начала тензорного исчисления, М., «Высшая школа», 1966.

СОДЕРЖАНИЕ

Введение	6
Глава 1. Элементы теории множеств	11
§ 1. Основные определения. Способы задания множеств	11
§ 2. Линейные точечные множества	14
§ 3. Соотношения между множествами	15
§ 4. Эквивалентные множества. Мощность множества	17
§ 5. Основные теоремы	21
§ 6. Функции. Отношения. Способы задания функций	26
§ 7. Изоморфизм	30
Глава 2. Элементы математической логики	33
§ 1. Высказывания. Их истинность и ложность	34
§ 2. Связь высказываний. Символы логических связей	35
§ 3. Эквивалентность. Заменяемость основных связей	37
§ 4. Нормальная форма логических выражений	40
§ 5. Всегда истинные и всегда ложные высказывания	40
§ 6. Предикаты	41
§ 7. Операции навешивания кванторов	44
Глава 3. Алгоритмы	46
§ 1. Численные и логические алгоритмы	47
§ 2. Эмпирические свойства алгоритмов	51
§ 3. Элементы теории алгоритмов. Алфавитные операторы и алгоритмы	53
§ 4. Слова в ассоциативном исчислении	56
§ 5. Эквивалентные алгоритмы. Нормальный алгоритм Маркова	59
§ 6. Алгоритмически неразрешимые проблемы	61
§ 7. Сведение любого алгоритма к численному. Метод Гёделя	62
Глава 4. Вероятности. Случайные величины. Случайные функции	67
§ 1. Элементы теории вероятностей. Основные понятия и определения	67
§ 2. Основные теоремы	72
§ 3. Случайные величины	77
Функция распределения	79
Функция плотности вероятности	82
Числовые характеристики	83
§ 4. Некоторые законы распределения	88

§ 5. Системы случайных величин	92
Функция распределения	93
Функция плотности вероятности	95
Зависимые и независимые случайные величины	98
Числовые характеристики системы случайных величин	99
Система произвольного числа случайных величин	101
§ 6. Случайные функции. Случайные процессы	103
Законы распределения	104
Характеристики случайных функций	105
Элементарные операции над случайными функциями	109
Определение характеристик случайных функций из опыта	111
§ 7. Методы теории случайных функций в исследовании систем	112
Оператор системы	114
Линейные и нелинейные операторы и системы	115
Стационарные и нестационарные системы	120
Линейные преобразования случайных функций	121
Сложение случайных функций	125
§ 8. Комплексные случайные функции	128
§ 9. Канонические разложения случайных функций	131
Определение канонического представления случайных функций	131
Каноническое разложение комплексной случайной функции	135
Линейные преобразования случайных функций, заданных каноническим разложением	136
§ 10. Стационарные случайные функции	139
Определение стационарной случайной функции	141
Дифференцирование стационарной случайной функции	143
Спектральное разложение стационарной случайной функции	144
Свойство эргодичности стационарных случайных процессов	149
§ 11. Нестационарные случайные функции	151
§ 12. Конечные случайные процессы. Вероятностные последовательности	159
Три типа конечных случайных последовательностей	162
Марковские цепи	165
 Глава 5. Статистический анализ	 168
§ 1. Основные задачи	168
Определение закона распределения по статистическим данным	169
Проверка правдоподобия гипотез	169
Определение неизвестных параметров распределения	169
§ 2. Ряды распределения и их характеристики	170
Построение рядов распределения	170
Графическое изображение рядов распределения	174
Числовые характеристики рядов распределения	177
Некоторые свойства статистических параметров	178
Основные свойства средней арифметической	180
Выравнивание рядов распределения	182
§ 3. Статистическое измерение связи	188
Корреляционная зависимость	189
§ 4. Исследование формы связи. Эмпирическая линия регрессии	191
Теоретическая линия регрессии	197
Выбор и обоснование типа кривой регрессии	200
Расчет параметров уравнения регрессии	202

	Функциональная корреляция	214
	Функциональные средние	217
§ 5.	Исследование тесноты связи	220
	Эмпирическое корреляционное отношение	221
	Теоретическое корреляционное отношение	224
	Коэффициент корреляции	226
	Модификации формулы коэффициента корреляции	228
	Коэффициент корреляции при нелинейной зависимости	232
	Определение прямой регрессии по основным статистическим параметрам	233
	Сопряженные показатели корреляции	234
	Корреляция многих переменных. Уравнение множественной регрессии	238
	Уравнение чистой регрессии	242
	Коэффициент множественной корреляции	243
	Эмпирические меры тесноты связи	244
§ 6.	Элементы теории ошибок	250

Глава 6. Теория спектров 255

§ 1.	Гармонический анализ	255
§ 2.	Метод Эйлера — Фурье для определения коэффициентов ряда Фурье	259
§ 3.	Ряд Фурье в комплексной форме	261
§ 4.	Ряд Фурье и наименьшая среднеквадратичная ошибка	262
§ 5.	Интеграл Фурье	264
§ 6.	Признаки сходимости интеграла Фурье	268
§ 7.	Комплексная форма интеграла Фурье	269
§ 8.	Преобразование Фурье	271
§ 9.	Спектр амплитуд и спектр фаз	273
§ 10.	Основные теоремы о спектрах	277
§ 11.	Текущий и мгновенный спектры	282
§ 12.	Модуляция. Спектры модулированных колебаний	285
§ 13.	Перенос спектра	296
§ 14.	Детектирование. Преобразование спектров при детектировании	299
§ 15.	Спектр суммы периодических функций. Спектры суммы и разности двух сдвинутых во времени колебаний	301
§ 16.	Спектры некоторых сигналов	303

Глава 7. Элементы теории информации 314

§ 1.	Общие положения	314
§ 2.	Модель системы связи Шеннона	315
§ 3.	Информация	317
§ 4.	Измерение взаимной информации	320
§ 5.	Измерение количества собственной информации	322
§ 6.	Свойства количества информации	322
§ 7.	Энтропия	324
§ 8.	Условная средняя взаимная информация	327
§ 9.	Дискретные источники сообщений. Эргодические источники сообщений	328
§ 10.	Энтропия эргодического источника дискретных сообщений	330
§ 11.	Избыточность источника сообщений	333

§ 12. Скорость создания сообщений	334
§ 13. Пропускная способность информационного канала	335
§ 14. Дискретные каналы без шумов. Пропускная способность дискретных каналов без шумов	337
§ 15. Эффективное кодирование	338
§ 16. Основная теорема Шеннона для дискретного канала без шумов	340
§ 17. Дискретные каналы с шумами	341
§ 18. Пропускная способность дискретного канала с шумами	341
§ 19. Основная теорема Шеннона для дискретного канала с шумами	343
§ 20. Источники непрерывных сообщений	344
§ 21. Количество информации, содержащееся в одном замере непрерывной случайной величины	346
§ 22. Энтропия непрерывных случайных величин	
§ 23. Количество информации о непрерывной случайной величине при заданных требованиях к верности воспроизведения	349
§ 24. Количество информации, содержащееся в воспроизведении непрерывного сообщения	351
§ 25. Непрерывные каналы с шумами. Пропускная способность непрерывных каналов	352
§ 26. Основная теорема Шеннона для непрерывных каналов	354

Глава 8. Элементы теории игр 356

§ 1. Основные определения	356
§ 2. Игры 2×2	367
§ 3. Игры $2 \times m$ и $n \times 2$	369
§ 4. Игры $m \times n$	376
§ 5. Приближенные методы решения игр	378
§ 6. Методы решения некоторых бесконечных игр	381

Глава 9. Графы 383

§ 1. Граф. Пути и контуры	383
§ 2. Цепи и циклы	386
§ 3. Квазиупорядоченность	388
§ 4. Индуктивный граф и базы	390
§ 5. Транспортные сети	391
§ 6. Деревья и леса	392

Глава 10. Тензорное исчисление 396

§ 1. Линейное пространство	396
§ 2. Прямоугольный базис в 3-мерном пространстве	404
§ 3. Преобразование ортонормированного базиса. Основная задача тензорного исчисления	410
§ 4. Полилинейные формы и тензоры	417
§ 5. Алгебраические операции над тензорами	428
§ 6. Тензорный анализ	433

Л и т е р а т у р а 442

Лапа Валентин Григорьевич

Математические основы кибернетики

Учебное пособие для студентов
электроприборостроительных специальностей вузов

Издательское объединение «Вища школа»
Главное издательство

Редактор Ж. Г. Д а в и д е н к о
Литредактор А. П. К о в а л ь ч у к
Обложка художника В. В. Т е р е щ е н к о
Художественный редактор С. П. Д у х л е н к о
Технические редакторы Л. Ф. В о л к о в а,
И. И. К а т к о в а
Корректор Л. А. К р ю к о в а

Сдано в набор 13/III 1974 г. Подписано к печати 3/VI 1974 г.
Формат бумаги 84×108¹/₃₂. Бумага тип. № 3. Физ. печ. л. 14,125.
Усл. печ. л. 23,73. Уч.-изд. л. 22,03. Тираж 16 000. Изд. № 2091.
БФ 31422. Цена 76 коп.

Главное издательство издательского объединения «Вища школа»,
252054, Киев, 54, Гоголевская, 7

Отпечатано с матриц Главного предприятия республиканского
производственного объединения «Поліграфкнига» Госкомиздата
УССР, Киев, ул. Довженко, 3 на Белоцерковской книжной фаб-
рике, ул. К Маркса, 4. Зак. 960.

76 коп.

