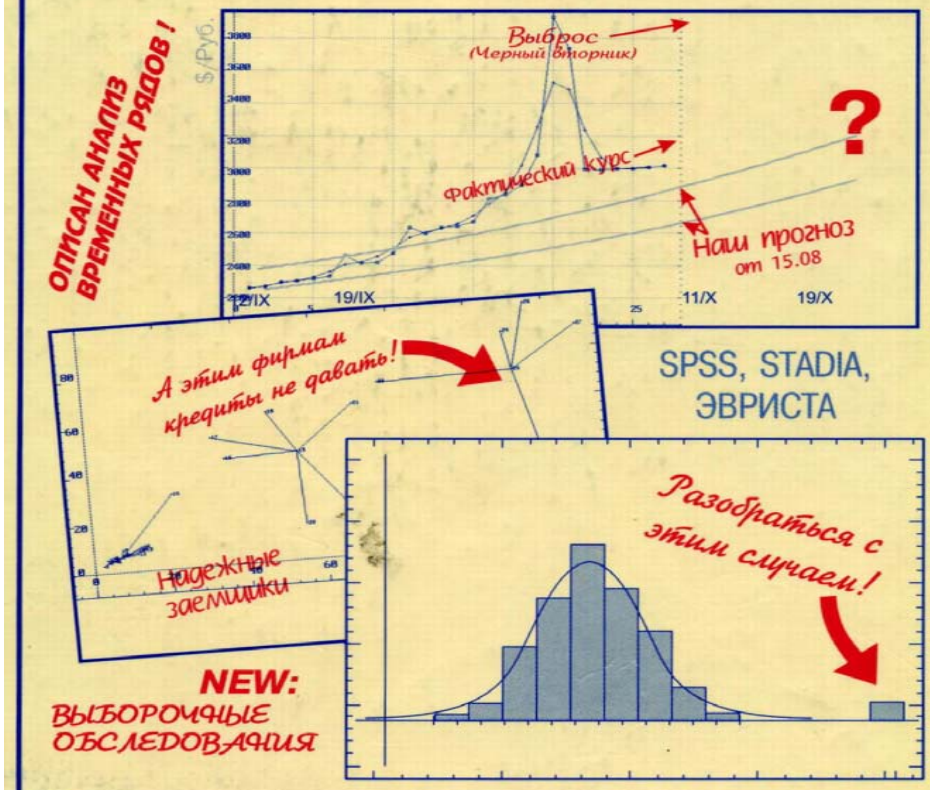
Ю.Н.ТЮРИН, А.А.МАКАРОВ

АНАЛИЗ ДАННЫХ НА КОМПЬЮТЕРЕ

3-е ИЗДАНИЕ

Под редакцией В.Э.Фигурнова



Оглавление

Благодарности

В первую очередь, мы отдаем дань признательности и восхищения нашим учителям А.И.Колмогорову и Б.В.Гнеденко. Вместе с другими выдающимися учеными они были основателями школы математической статистики в СССР и учителями многих исследователей, внесших вклад в развитие отечественной статистики.

Мы также хотели бы поблагодарить всех наших коллег, чья помощь и участие в той или иной мере и форме способствовали написанию этой книги. К сожалению, невозможно упомянуть всех, и мы персонально скажем лишь о некоторых.

Мы глубоко признательны Д.С.Шмерлингу, который был инициатором создания этой книги, и чей интерес и постоянное внимание нас всегда поддерживали.

Мы благодарны нашему редактору В.Э.Фигурнову, который внес в текст много улучшений. Он также провел литературное и техническое редактирование, выполнил компьютерную верстку. Мы благодарны С.А. Айвазяну, М.В. Болдину, В.Н. Тутубалину за многие обсуждения, советы и поддержку. Общение с этими и другими выдающимися учеными помогало нашему усовершенствованию в статистической теории и практике.

Мы выражаем глубокую признательность фирмам НПО «Информатика и Компьютеры», «ИнфоСтрой», «Центр Статистических Исследований МГУ», «Статистические системы и сервис», «СтатДиалог», «ТАНДЕМ» за предоставленные для ознакомления версии пакетов: STADIA, STATGRAPHICS, ЭВРИСТА, SPSS, SYSTAT, МЕЗОЗАВР, Статистик-Консультант.

Глава 1

Основные понятия прикладной статистики

Цель этой главы — познакомить читателя с основными понятиями теории вероятностей и статистики, на которые опирается анализ данных изменчивой (случайной) природы. Не стремясь к строгому формальному изложению, мы расскажем о случайных событиях и случайных величинах, об их характеристиках: распределении вероятностей, математическом ожидании, дисперсии и т.д. Будут введены наиболее распространенные понятия описательной статистики, используемые при обработке данных, такие как генеральная совокупность, выборка, выборочная функция распределения, медиана, квантили, гистограмма и др. В конце главы мы опишем, как можно вычислить соответствующие характеристики на компьютере.

1.1. Случайная изменчивость

Статистика изучает числа, чтобы обнаружить в них закономерности. Все мы хорошо знакомы с закономерными явлениями и закономерными изменениями, они составляют главный объект научных исследований. Например, исследователя могут интересовать вопросы типа: как изменяется давление в жидкости с изменением глубины? С какой скоростью движутся падающие тела? Как будет проходить химическая реакция, если мы определенным образом изменим температуру, давление и концентрации участвующих в реакции веществ и т.п. Знание законов природы позволяют нам ответить на подобные вопросы, не производя реальных опытов, т.е. заранее. Например, мы можем точно вычислить, какие вещества и в какой пропорции образуются при той или иной химической реакции, или предсказать, когда в данной местности произойдет следующее солнечное затмение.

Но отнюдь не во всех ситуациях интересующий нас результат полностью и жестко определяется влияющими на него факторами. Например, мы не можем указать, сколько часов будет светить электрическая лампочка или как долго будет служить телевизионный приемник. Невозможно предвидеть число посетителей магазина и количество товаров, которое они купят, каков будет результат бросания игральных костей и т.д. Ответы на подобные вопросы можно получить, только проведя

соответствующие испытания. Часто явления (ситуации), в которых результат полностью определяется влияющими на него факторами, называются *детерминированными* или *закономерными*, а те, в которых это не выполняется — *недетерминированными* или *стохастическими*.

Идея случайности. Для описания явлений с неопределенным исходом (как в повседневной жизни, так и в науке) используется *идея случайности*. Согласно этой идее, результат явления с неопределенным исходом как бы определяется неким случайным испытанием, случайным экспериментом, случайным выбором. Иначе говоря, считается, что для выбора исхода в неопределенной ситуации природа словно бы бросает кости. Вопрос о том, насколько применим такой подход к явлениям окружающего мира, решается не путем его логического обоснования, а по результатам практического применения.

Замечание. Вопросы о том, существует ли случайность «на самом деле», о происхождении случайного и соотношении закономерного и случайного являются дискуссионными философскими темами. Действительно, закономерные изменения, как подчеркивает само их название, порождены определенными причинами, которые могут быть названы, указаны и изучены. Отыскивая эти причины, мы исходим из убеждения, что если нечто изменилось, так это потому, что изменилось что-то другое, и это другое служит причиной первому. Когда же изменения происходят при полной неизменности условий, в которых протекает явление, мы объясняем это случайностью. Но поскольку полной неизменности условий на практике достичь невозможно, сохраняется логическая возможность отрицать наличие в природе случайности и объяснять неопределенность результатов эксперимента воздействием неизвестных нам и неучтенных факторов. Мы не будем входить в эти философские споры и будем рассматривать проблемы случайности чисто технически, принимая этот подход лишь как модель для описания непредсказуемой изменчивости, дабы на его основе получать количественные выводы и рекомендации для практики.

Случайная изменчивость. Мы все хорошо знаем, что такое закономерность. Например, при формулировке законов природы мы говорим, что если одна величина принимает такое-то значение, то другая примет такое-то. Случайная изменчивость нам знакома в меньшей степени, а потому о ней надо поговорить подробнее. Для начала лучше взять такой пример, где случайная изменчивость действует отдельно от закономерной, так сказать, «в чистом виде».

Рассмотрим пример, заимствованный из книги А.Хальда. В таблице 1.1 приведены размеры головок 200 заклепок, изготовленных станком (который делает их тысячами). Все контролируемые условия, в которых работал станок, оставались неизменными. В то же время диаметры головок раз от разу несколько изменялись. Характерная черта случайных колебаний — эти изменения выглядят бессистемными, хаотичными. Действительно, если бы в этих изменениях мы смогли обнаружить

какую-либо закономерность, у нас появились бы основания, чтобы искать ответственную за эту закономерность причину, тем самым изменчивость не была бы чисто случайной. Если бы, скажем, с течением времени размер головки заклепки проявил тенденцию к увеличению, мы могли бы попытаться связать это, например, с износом инструмента.

Таблица 1.1

Диаметры 200 головок заклепок, мм											
13.39	13.43	13.54	13.64	13.40	13.55	13.40	13.26	13.42	13.50	13.32	13.31
13.28	13.52	13.46	13.63	13.38	13.44	13.52	13.53	13.37	13.33	13.24	13.13
13.53	13.53	13.39	13.57	13.51	13.34	13.39	13.47	13.51	13.48	13.62	13.58
13.57	13.33	13.51	13.40	13.30	13.48	13.40	13.57	13.51	13.40	13.52	14.56
13.40	13.34	13.23	13.37	13.48	13.48	13.62	13.35	13.40	13.36	13.45	13.48
13.29	13.58	13.44	13.56	13.28	13.59	13.47	13.46	13.62	13.54	13.20	13.38
13.43	13.36	13.56	13.51	13.47	13.40	13.29	13.20	13.46	13.44	13.42	13.29
13.41	13.39	13.50	13.48	13.53	13.34	13.45	13.42	13.29	13.38	13.45	13.50
13.55	13.33	13.32	13.69	13.46	13.32	13.32	13.48	13.29	13.25	13.44	13.60
13.43	13.51	13.43	13.38	13.24	13.28	13.58	13.31	13.31	13.45	13.43	13.44
13.34	13.49	13.50	13.38	13.48	13.43	13.37	13.29	13.54	13.33	13.36	13.46
13.23	13.44	13.38	13.27	13.66	13.26	13.40	13.52	13.59	13.48	13.46	13.40
13.43	13.26	13.50	13.38	13.43	13.34	13.41	13.24	13.42	13.55	13.37	13.41
13.38	13.14	13.42	13.52	13.38	13.54	13.30	13.18	13.32	13.46	13.39	13.35
13.34	13.37	13.50	13.61	13.42	13.32	13.35	13.40	13.57	13.31	13.40	13.36
13.28	13.58	13.58	13.38	13.26	13.37	13.28	13.39	13.32	13.20	13.43	13.34
13.33	13.33	13.31	13.45	13.39	13.45	13.41	13.45				

Обсуждение случайной изменчивости не обязательно начинать с такого специального примера. Каждому известны более простые опыты, в которых результат определяется случаем: раздача игральных карт или костей домино, бросание игральных костей, монет и т.д. У всех этих примеров есть общая черта — непредсказуемость результатов для действий, проводящихся в неизменных условиях.

Закономерность и случайность. В большинстве явлений присутствуют оба вида изменчивости — и закономерная, и случайная, и для нахождения закономерностей нам приходится «отсеивать» мешающие случайные факторы. Например, при внесении удобрений на пшеничное поле мы не можем точно предсказать, какова будет урожайность на этом поле, поскольку она зависит от множества причин, которые мы считаем случайными (от погодных условий, нашествий вредителей, болезней растений и т.д.). Однако с помощью методов статистического анализа мы все же можем определить степень влияния на урожайность внесения удобрений и применения других агротехнических приемов. Для этого могут потребоваться многолетние тщательно спланированные эксперименты, с помощью которых влияние агротехнических приемов оценивается на фоне мешающих факторов.

Итак, статистический подход к изучению явлений природы состоит в мысленном разделении наблюдаемой изменчивости на две части — обусловленные закономерными и случайными причинами, и выявлению закономерной изменчивости на фоне случайной. Например, в табл. 1.2 и на рис. 1.1 отображено изменение урожайности зерновых (в центнерах с гектара) в СССР за 45 лет, с 1945 по 1989 год. Данные предоставлены А.И.Манелля, которому авторы выражают глубокую признательность.

Таблица 1.2

Урожайность зерновых культур в СССР с 1945 по 1989 гг.
(в центнерах с гектара в первоначально оприходованном весе)

Год	Урожайность	Год	Урожайность	Год	Урожайность
1945	5.6	1960	10.9	1975	10.9
1946	4.6	1961	10.7	1976	17.5
1947	7.3	1962	10.9	1977	15.0
1948	6.7	1963	8.3	1978	18.5
1949	6.9	1964	11.4	1979	14.2
1950	7.9	1965	9.5	1980	14.9
1951	7.4	1966	13.7	1981	12.6
1952	8.6	1967	12.1	1982	15.2
1953	7.8	1968	14.0	1983	15.9
1954	7.7	1969	13.2	1984	14.4
1955	8.4	1970	15.6	1985	16.2
1956	9.9	1971	15.4	1986	18.0
1957	8.4	1972	14.0	1987	18.3
1958	11.1	1973	17.6	1988	17.0
1959	10.4	1974	15.4	1989	18.8

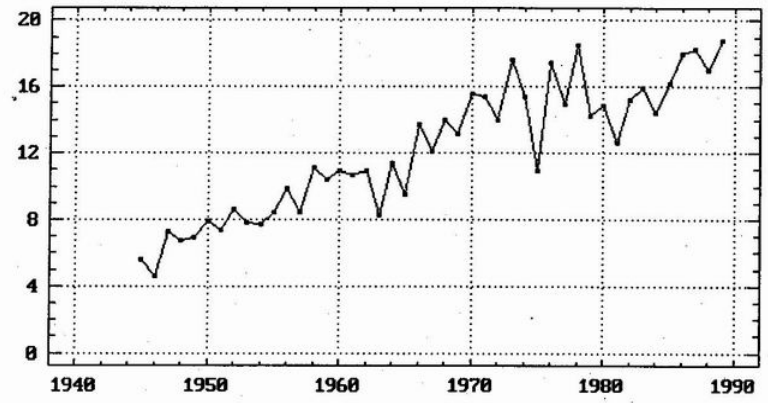


Рис. 1.1. Урожайность всех зерновых культур в СССР с 1945 по 1989 гг. (ц/га)

Хорошо видно, что урожайность, в целом, возрастала (по-видимому, за счет улучшения агротехники и внесения минеральных удобрений). Ее рост и составляет закономерную часть картины. В то же время видны

и значительные колебания урожайности в разные годы, по-видимому, за счет погодных условий и иных факторов, изменения которых мы считаем случайными. Методы *математической статистики* позволяют в подобных ситуациях оценивать параметры имеющихся закономерностей, проверять те или иные гипотезы об этих закономерностях и т.д. В последующих главах этой книги мы рассмотрим, как решаются многие из подобных задач.

Однако случайности могут не только мешать нам постигать закономерности — они способны и сами порождать их. Рассмотрим, например, газ в некотором сосуде (скажем, воздух в комнате). Поведение каждой молекулы газа носит случайный характер, но вся совокупность этих молекул ведет себя вполне закономерно, подчиняясь хорошо известным законам физики. Так, давление газа на каждую единицу площади поверхности сосуда строго постоянно (колебания проявляются только для очень сильно разреженных газов), а объем газа, его давление и температура связаны друг с другом уравнением Менделеева-Клапейрона. Аналогично, выбор времени для телефонных звонков каждый человек осуществляет сам, но нагрузка на телефонную станцию (АТС), распределение интервалов между звонками различных абонентов и т.д. подчиняются вполне определенным закономерностям. Изучением закономерностей, которые порождаются случайными событиями, занимается наука *теория вероятностей*.

1.2. События и их вероятности

Хотя результаты эксперимента (наблюдений, опыта), зависящего от случайных факторов, нельзя предсказать, все же разные возможные его исходы и связанные с ними события имеют неодинаковые шансы на появление. Количественное описание правдоподобия отдельных исходов и событий основывается на понятии вероятности. Предполагается, что каждому событию, возможному в данном случайном испытании, может быть приписана числовая мера его правдоподобия, называемая его *вероятностью*. Если, скажем, A есть случайное событие, то его вероятность обычно обозначается через $P(A)$. (Буква P — начальная в латинском слове «вероятность».) Вероятность *невозможного* события (которое никогда не происходит) принимается равной 0, а вероятность *достоверного* события (которое происходит всегда) принимается равной 1. Поэтому для любого события A: $0 \leqslant P(A) \leqslant 1$.

Свойства вероятности просты, естественны и, в общем, известны каждому. Однако перед тем, как рассказывать о них, необходимо дать некоторые определения, касающиеся случайных событий.

Случайные события.

Объединением, или суммой событий A и B называют событие C, которое состоит в том, что происходит хотя бы одно из событий A и B. (C происходит тогда и только тогда, когда происходит либо A, либо B, либо оба вместе.) Обозначение:

$$C = A \cup B, \qquad \text{или} \qquad C = A + B.$$

Пересечением, или произведением событий A и B называют событие C, которое состоит в том, что происходят оба события A и B. Обозначение:

$$C = A \cap B, \qquad \text{или} \qquad C = AB.$$

Отрицанием события A называют такое событие, которое состоит в том, что A не происходит. Обозначение для него $\overline{A}$.

Событие, которое при нашем случайном испытании обязательно происходит, называют *достоверным*; которое не может произойти — *невозможным*. Вероятность достоверного события равна 1; вероятность невозможного события равна 0.

Если события A и B не могут произойти одновременно (т.е. если AB — невозможное событие), их называют *несовместимыми*. Несовместимы, например, события A и $\overline{A}$. В то же время $A + \overline{A}$ — событие достоверное.

Например, при бросании игральной кости:

- событие, состоящее в том, что в результате бросания кости выпадет 1, 2, 3, 4, 5 или 6 очков, является достоверным;
- событие, состоящее в том, что результате бросания кости выпадет 7 очков, является невозможным;
- объединением события A, состоящего в том, что в результате бросания кости выпадет меньше 4 очков, и события B, состоящего в том, что в результате бросания кости выпадет 3 или 6 очков, будет событие $A + B$, состоящее в том, что в результате бросания кости выпадет 1, 2, 3 или 6 очков;
- пересечение AB событий A и B состоит в том, что в результате бросания кости выпадет 3 очка;
- отрицание события A, обозначаемое $\overline{A}$, состоит в том, что в результате бросания кости выпадет 4, 5 или 6 очков.

Свойства вероятности. Теперь свойства вероятности перечислить просто:

1. $0 \leqslant P(A) \leqslant 1$ для любого события A;
2. $P(A+B) = P(A) + P(B)$, если события A и B несовместимы, а в общем случае $P(A+B) = P(A) + P(B) - P(AB)$;

3. Вероятность достоверного события равна 1, а невозможного события — нулю.

Для полного описания случайного эксперимента нужно указать все его возможные исходы и их вероятности. Например, бросание игральной кости, имеющей форму куба, приводит к выпадению одной из ее шести граней. Это шесть элементарных исходов, т.е. неразложимых на более простые. Если кость, как говорят, правильная, то эти шесть исходов равноправны и поэтому должны иметь равные вероятности. Следовательно, вероятность каждого из них равна 1/6. Вероятности остальных (составных) событий может быть вычислена из приведенных выше свойств вероятности. Например, вероятность $P(B)$ события B, состоящего в том, что в результате бросания кости выпадет 3 или 6 очков, равна 1/3. Действительно, это событие является объединением двух несовместимых событий: «выпало 3 очка» и «выпало 6 очков», вероятность каждого из которых равна 1/6. Аналогично, вероятность $P(A)$ события A, состоящего в том, что в результате бросания кости выпадет меньше 4 очков, равна 1/2.

Не будем далее развивать эту тему, оставив ее теории вероятностей. Но все же нам придется ввести еще два важных понятия — независимости событий и условной вероятности.

Независимость событий.

Определение 1. *События A и B называются независимыми, если*

$$P(AB) = P(A)\,P(B).$$

На практике независимость событий обычно устанавливается не с помощью проверки этого равенства, а из условий опыта и других содержательных соображений. При этом указанное соотношение можно использовать для вычисления вероятности событий AB через вероятности событий A и B. Понятие независимости очень существенно для теории вероятностей. То, насколько в своей математической форме понятие независимости соответствует нашим интуитивным представлениям, лучше всего разобрать с помощью понятия *условной вероятности*.

Условная вероятность. Для простоты мы рассмотрим, как можно определить понятие условной вероятности в случайном испытании с конечным числом исходов. Пусть Ω — совокупность всех таких исходов, ω обозначает произвольный элементарный исход, $P(\omega)$ — его вероятность. Любые события A и B в этом опыте представляют собой некоторые подмножества Ω, поскольку они состоят из элементарных исходов. Обозначим через $P(A \mid B)$ условную вероятность события A

при условии, что произошло событие B. Достаточно определить условную вероятность для элементарных исходов ω. Те исходы ω, которые не входят в B, невозможны при наступлении события B, поэтому для них следует положить условную вероятность равной нулю:

$$P(\omega \mid B) = 0, \quad \text{если} \quad \omega \notin B.$$

Для исходов ω, входящих в B, сумма их вероятностей $\sum_{\omega \in B} P(\omega)$ равна $P(B)$, а сумма их условных вероятностей должна быть равна единице. Действительно, $\sum_{\omega \in B} P(\omega \mid B)$ равна $P(B \mid B)$. Но при наступления B событие B является достоверным, поэтому согласно свойству 3 вероятностей $P(B \mid B)$ равно 1. Чтобы это условие выполнялось, естественно положить для $\omega \in B$:

$$P(\omega \mid B) = P(\omega)/P(B).$$

Теперь мы можем определить условную вероятность для любого события A.

Определение. *Условная вероятность события A при условии B есть*

$$P(A \mid B) = \sum_{\omega \in A} P(\omega \mid B).$$

Из этого определения легко вывести, что:

$$P(A \mid B) = \frac{P(AB)}{P(B)}.$$

Это соотношение в общем случае (когда число элементарных исходов не обязательно конечно) и принимают за определение условной вероятности. Из него легко следует известная формула умножения вероятностей:

$$P(AB) = P(A \mid B)P(B).$$

Заметим, что равноправие событий A и B позволяет написать также, что $P(AB) = P(B \mid A)P(A)$.

С помощью понятия условной вероятности мы можем дать другое определение независимости событий.

Определение 2. *Событие A не зависит от события B, если*

$$P(A \mid B) = P(A).$$

Иначе говоря, событие A не зависит от события B, если вероятность события A не зависит от того, произошло или нет событие B. Нетрудно показать, что два определения независимости события A от B, данные выше, эквивалентны. Так же можно показать, что если A не зависит

от B, то и B не зависит от A. Единственная оговорка, которую надо добавить к сказанному, — что условную вероятность можно определять таким образом, лишь если $P(B) > 0$.

1.3. Измерения вероятности

Раз мы ввели понятие вероятности как количественное выражение для правдоподобия случайного события, нам необходим метод ее численного выражения. Здесь возможны два пути — умозрения и прямого измерения.

Умозрительный способ определения численного значения вероятности зиждется, в основном, на понятии равновозможности тех или иных исходов эксперимента. Мы уже прибегали к помощи этого соображения при обсуждении бросания игральной кости. Основная область приложения этого принципа — случайный выбор и азартные игры. Поэтому принцип равновозможности исходов эксперимента имеет ограниченное применение. Кроме того, выводы из этого принципа всегда относятся к некому идеальному случайному опыту, и то, насколько им подчиняется реальный эксперимент, само зачастую нуждается в проверке.

Измерение вероятности события отличается от измерения других физических величин. Для массы, скорости, температуры и большинства других физических величин есть специальные приборы, позволяющие выразить их числом (что и означает измерить). К сожалению, для вероятности такого прибора нет. Все же прямое измерение вероятности возможно, оно основано на независимых повторениях случайного эксперимента.

Пусть в определенном случайном эксперименте нас интересует вероятность некоторого события A. Допустим, что мы можем многократно осуществлять этот эксперимент в неизменных условиях, так что от опыта к опыту $P(A)$ не меняется. Проведем N таких повторений (иногда говорят — *реализаций*) этого опыта. Число N не должно зависеть от исходов отдельных опытов; например, оно может быть назначено заранее. Подсчитаем число тех опытов из N, в которых событие A произошло. Обозначим это число через $N(A)$. Рассмотрим отношение $N(A)/N$ — частоту события A в N повторениях опыта. *Оказывается, частота $N(A)/N$ приблизительно равна $P(A)$, если число повторений N велико.*

Указанная связь между частотой события и его вероятностью составляет содержание теоремы Бернулли, о которой подробнее мы будем говорить в главе 4. Там будет дана ее точная формулировка и доказательство. Кроме того, важен и вопрос о достигаемой точности приближения

частоты к вероятности, в частности, о числе опытов, необходимых для получения заранее указанной точности. Этому второму вопросу должно предшествовать прояснение содержания статистической точности, которое реализуется через посредство *доверительных интервалов*. Об этом речь пойдет в главе 5.

Итак, задав вопрос об измерении вероятностей, мы столкнулись с неприятной неожиданностью — это измерение оказалось, во-первых, непростым с чисто физической точки зрения (многократное повторение опыта), а во-вторых, сопряженным с довольно сложными и новыми понятиями.

Особо надо подчеркнуть, что описанные выше опыты должны происходить независимо друг от друга в неизменных условиях, чтобы вероятность события сохранялась постоянной. При большом числе повторений опытов соблюсти это требование зачастую оказывается нелегко. Даже небольшие отклонения от статистической устойчивости могут оказать воздействие на результаты, особенно при высоких требованиях к точности выводов. Не говоря уже о том, что повторения опытов, да еще многократные, далеко не всегда возможны.

1.4. Случайные величины. Функции распределения

В случайных экспериментах нас часто интересуют такие величины, которые имеют числовое выражение. Например, у каждого человека имеется много числовых характеристик: рост, возраст, вес и т.д. Если мы выбираем человека случайно (например, из группы или из толпы), то случайными будут и значения указанных характеристик. Чтобы подчеркнуть то обстоятельство, что измеряемая по ходу опыта численная характеристика зависит от его случайного исхода и потому сама является случайной, ее называют *случайной величиной*.

Случайной величиной, в частности, является упомянутое выше число очков, выпадающее при бросании игральной кости. Случайна сумма очков, выпавших при бросании двух игральных костей (а также их разность, произведение и т.д.). Случайной величиной надо считать диаметр головки заклепки, изготавливаемой станком (см. табл. 1.1 выше, где приведены значения, которые приняла эта случайная величина в 200 опытах).

Часто говорят, что случайная величина реализуется во время опыта. Если употребить это слово, то можно также сказать, что табл. 1.1 дает 200 *реализаций* этой случайной величины.

Каждая случайная величина задает *распределение вероятностей* на множестве своих значений. Если ξ — случайная величина, принимающая значения из X, то мы можем задать распределение вероятностей P_ξ на X следующим образом:

$$P_\xi(A) = P(\xi \in A).$$

Чтобы дать полное математическое описание случайной величины, надо указать множество ее значений и соответствующее случайной величине распределение вероятностей на этом множестве.

Виды случайных величин. В практических задачах обычно используются два вида случайных величин — *дискретные* и *непрерывные*, хотя бывают и такие случайные величины, которые не являются ни дискретными, ни непрерывными. Рассмотрим сначала дискретные случайные величины.

Дискретные случайные величины обладают тем свойством, что мы можем перечислить (перенумеровать) все их возможные значения. Таким образом, для задания распределения вероятностей, порожденных дискретными случайными величинами, надо только указать вероятности каждого возможного значения этой случайной величины. Например, число очков, выпавших при бросании игральной кости, — это дискретная случайная величина, так как она может принимать только 6 значений: 1, 2, 3, 4, 5 или 6. Для определения вероятностей любых событий, связанных с этой случайной величиной, нам надо только указать вероятности каждого из этих значений.

Определение. *Случайную величину называют дискретной, если множество ее возможных значений конечно, либо счетно.*

Напомним, что множество называется счетным, если его элементы можно перенумеровать натуральными числами.

Каждое возможное значение дискретной случайной величины имеет положительную вероятность (иногда, впрочем, допускают, что некоторые значения могут иметь нулевые вероятности, особенно когда рассматривают не одно, а несколько дискретных распределений одновременно). Чтобы полностью описать дискретное распределение вероятностей, надо указать все значения, вероятности которых положительны (точнее, могут быть положительны), и вероятности этих значений.

Пример. При бросании двух игральных костей сумма выпавших очков может принимать значения от 2 до 12. При этом для правильных костей, бросаемых независимо, вероятность получить в сумме 2 очка равна $1/6 \times 1/6 = 1/36$, получить 3 очка — равна $2/36$ и так далее. Распределение вероятностей суммы выпавших очков определяется следующей таблицей 1.3.

Таблица 1.3

значения	2	3	4	5	6	7	8	9	10	11	12
вероятности	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

Однако не все случайные величины могут быть описаны так просто, как дискретные случайные величины. Например, время службы электрической лампочки может, в принципе, принимать любое значение от нуля до бесконечности (как хорошо известно, это множество не является счетным). И если предполагается, что лампочка была в начале исправна, то вероятность того, что время ее службы будет в точности равно некоторому значению, будет равна нулю. Ненулевыми будут вероятности только сложных событий: например, что время службы лампочки — от одного до двух месяцев. Для подобных (так называемых *непрерывных*) случайных величин мы не можем задать их распределение путем указания вероятностей каждого возможного значения, так как все эти вероятности равны нулю. При описании таких случайных величин используются другие средства. В частности, если значениями случайной величины являются вещественные числа, то распределение случайной величины полностью определяется ее *функцией распределения*.

Функция распределения. Пусть ξ обозначает случайную величину, принимающую вещественные значения, x — вещественное число.

Определение. *Функцией распределения $F(x)$ случайной величины ξ называют* $F(x) = P(\xi \leqslant x)$.

Ясно, что функция $F(x)$ монотонно возрастает с ростом x (точнее сказать, не убывает, потому что могут существовать участки, на которых она постоянна). У дискретной случайной величины функция распределения ступенчатая, она возрастает скачком в тех точках, вероятности которых положительны. Это точки разрыва $F(x)$. На рис. 1.2 приведен график функции распределения для описанной выше случайной величины — суммы очков, выпавшей при бросании двух игральных костей.

Непрерывные случайные величины. Для случайной величины, принимающей вещественные значения, то свойство, что вероятность любого отдельного ее значения равна нулю, может легко быть выражено через функцию распределения.

Определение. *Случайную величину, принимающую вещественные значения, называют непрерывной, если непрерывна ее функция распределения.*

Непрерывным в этом случае называют и соответствующее распределение вероятностей. Для непрерывного распределения вероятность каждого отдельного значения случайной величины равна нулю. На этом

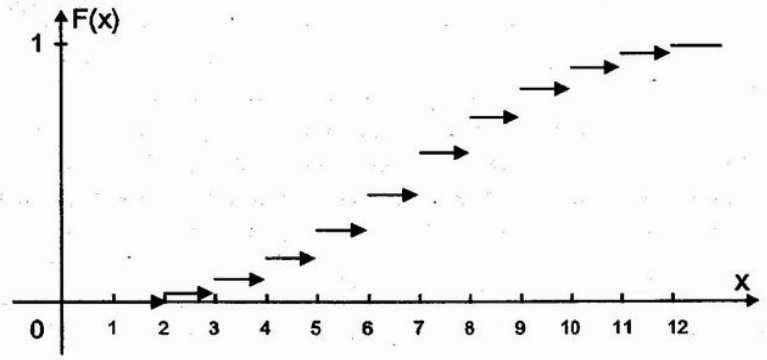


Рис. 1.2. График функции распределения суммы очков, выпавших на двух игральных костях

и основано противопоставление непрерывных и дискретных распределений — ведь для последних вся единичная вероятность распределена конечными положительными порциями. Для непрерывных же она как бы «разлита» по области определения случайной величины (в данном случае — по прямой).

Плотность вероятности. Нагляднее всего непрерывную случайную величину можно представить тогда, когда ее функция распределения не только непрерывна, но и дифференцируема (за исключением, может быть, конечного числа точек). В этом случае вероятности связанных с данной случайной величиной событий можно выразить через посредство так называемой функции *плотности вероятности*. Есть две эквивалентных формы определения плотности: интегральная и дифференциальная. Сначала мы дадим определение в интегральной форме.

Определение. *Функция $p(t)$ называется плотностью вероятности в точке t (иногда — плотностью случайной величины ξ), если для любых чисел a, b (пусть $a < b$)*

$$P(a < \xi < b) = \int_a^b p(x)\,dx.$$

Эквивалентная дифференциальная форма определения плотности звучит так: для любого $\Delta > 0$ и любого[1] действительного t

$$P(t < \xi < t + \Delta) = p(t)\,\Delta + o(\Delta),$$

[1] Если говорить точно — любого, за исключением множества меры нуль. Предыдущее (интегральное) определение показывает, что функция плотности может быть произвольно изменена на любом множестве нулевой меры, все равно удовлетворяя определению. Практически, разумеется, используют наиболее регулярную и простую из возможных функций плотности.

где $o(\Delta)$ — малая (точнее, бесконечно малая) по сравнению с Δ величина.

Наглядное содержание второго из этих определений состоит в том, что вероятность, приходящаяся на малый отрезок, оказывается приблизительно пропорциональной длине этого отрезка, причем коэффициент пропорциональности равен значению функции плотности вероятности в некоторой точке этого отрезка.

Функция распределения и плотность связаны соотношениями:

$$F(x) = \int_{-\infty}^{x} p(t)\,dt, \qquad p(x) = F'(x).$$

(для почти всех x — с теми же оговорками, что были сделаны выше).

Как правило, для приложений достаточно двух вышеописанных типов распределений — дискретного и непрерывного, точнее, имеющего плотность. Хотя можно встретиться с распределениями, представляющими собой смесь двух этих типов, и даже с более сложными. В главе 2 мы подробнее познакомимся с некоторыми важными для приложений законами вероятностей на числовой прямой.

Примеры. Покажем на примерах различные типы функций распределения и их свойства. Пусть случайная величина ξ может принимать только значения 0 и 1 с вероятностями, соответственно, p и $1-p$ (причем $0 \leqslant p \leqslant 1$). В этом случае функция распределения имеет вид:

$$F(x) = \begin{cases} 0, & \text{если} \quad x < 0; \\ p, & \text{если} \quad 0 \leqslant x < 1; \\ 1, & \text{если} \quad x \geqslant 1. \end{cases}$$

График этой функции изображен на рис. 1.3.

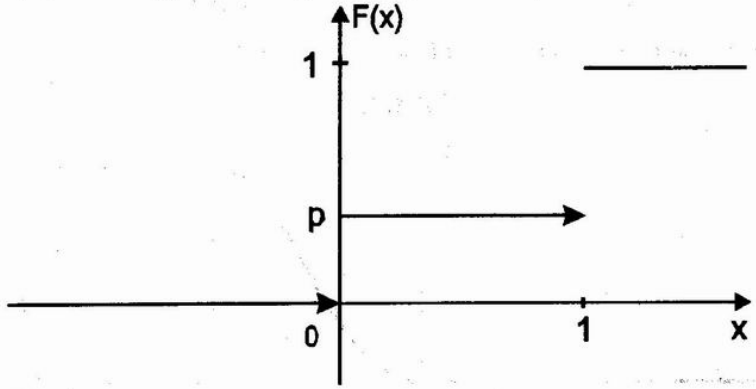


Рис. 1.3. График функции распределения, сосредоточенного в двух точках.

Рассмотрим функцию распределения случайной величины более общего вида. Пусть случайная величина ξ принимает конечное число значений $a_1, \ldots, a_n$,

причем $P(\xi = a_k) = p_k \geqslant 0$, $\left(\sum_{k=1}^{n} p_k = 1\right)$. График функции этого дискретного распределения изображен на рис. 1.4. (Для удобства предположим, что возможные значения занумерованы в порядке возрастания.)

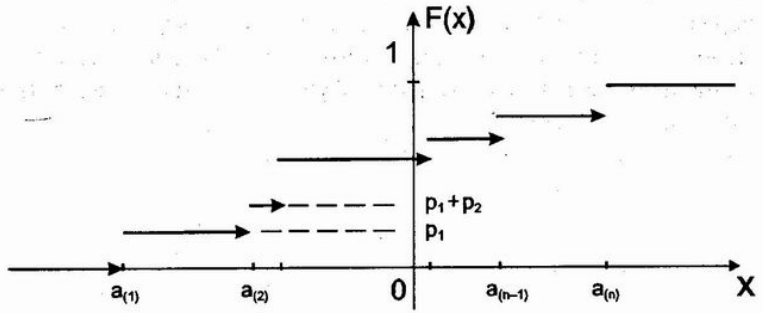


Рис. 1.4. График функции дискретного распределения

Рассмотрим пример непрерывного распределения вероятностей. Пусть функция плотности $p(t)$ равна

$$p(t) = \begin{cases} 0, & \text{если} \quad t < 0; \\ 6t(1-t), & \text{если} \quad 0 \leqslant t < 1; \\ 0, & \text{если} \quad t \geqslant 1. \end{cases}$$

(Легко проверить, что в данном случае $\int_{-\infty}^{+\infty} p(t)\,dt = 1$, $p(t) \geqslant 0$, так что функция $p(t)$ может быть плотностью случайной величины). Функция распределения в этом примере равна

$$F(x) = \begin{cases} 0, & \text{для} \quad x \leqslant 0; \\ -2x^3 + 3x^2, & \text{для} \quad 0 \leqslant x \leqslant 1; \\ 1, & \text{для} \quad x \geqslant 1. \end{cases}$$

График этой функции приведен на рис. 1.5.

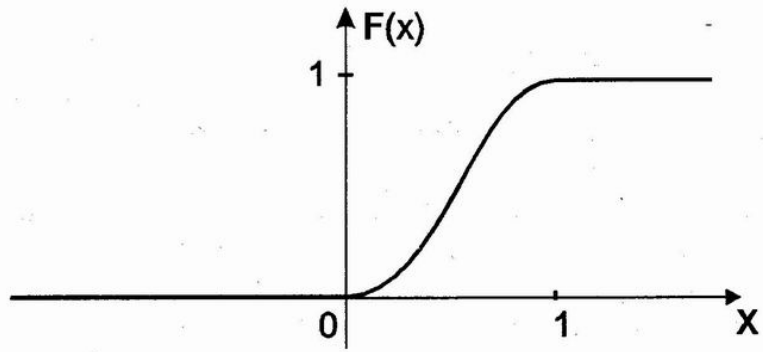


Рис. 1.5. Пример непрерывной функции распределения

В приведенных примерах можно заметить, что $F(x) \to 0$ при $x \to -\infty$ и $F(x) \to 1$, при $x \to +\infty$, и что $F(x)$ — неубывающая функция. Это общие свойства всех функций распределения.

Если в точке x функция распределения $y = F(x)$ имеет скачок, то величина этого скачка равна вероятности, сосредоточенной в точке x, т.е. вероятности события $\xi = x$. Если же точка x — точка непрерывности функции $y = F(x)$, и более того, $F(x)$ имеет производную в этой точке, то график $F(x)$ в точке x имеет касательную, тангенс угла наклона которой равен плотности $p(x)$ в этой точке.

1.5. Числовые характеристики распределения вероятностей

Числовые характеристики распределения вероятностей полезны тем, что помогают составить наглядное представление об этом распределении. Наиболее часто употребляемыми характеристиками случайной величины (и соответствующего распределения вероятностей) служат *моменты* и *квантили*. Ниже мы их определим, но надо сделать оговорку: универсальные (пригодные для любых случайных величин) определения этих характеристик требуют весьма сложного математического аппарата (они основаны на теории меры, интеграла Лебега-Стилтьеса и т.д.), поэтому мы приводить их не будем. Вместо этого мы дадим более простые определения для дискретных и для непрерывных случайных величин.

Начнем с так называемого первого момента случайной величины ξ, называемого также *математическим ожиданием*, или *средним значением* ξ. Его обозначают через $M\xi$ или $E\xi$.

Определение. *Для дискретной случайной величины ξ со значениями $x_1, x_2, \ldots$, имеющих вероятности $p_1, p_2, \ldots$*

$$M\xi = \sum_k x_k p_k.$$

Если число возможных значений ξ конечно, то $M\xi$ всегда существует и не зависит от способа нумерации этих значений. В том случае, если число возможных значений ξ счетно, необходимо, чтобы сумма ряда $\sum_k x_k p_k$ не зависела от нумерации значений x, то есть, чтобы этот ряд сходился абсолютно ($\sum_k |x_k| p_k < \infty$).

Определение. *Для непрерывной случайной величины ξ с плотностью $p(x)$,*

$$M\xi = \int_{-\infty}^{\infty} x\, p(x)\, dx,$$

причем интеграл должен сходиться абсолютно.

Как говорилось выше, приведенные определения $M\xi$ не являются исчерпывающими, поскольку пригодны не для всех видов случайных величин. Общее определение математического ожидания выглядит следующим образом:

$$M\xi = \int x\, dP_\xi(x),$$

где $P_\xi(x)$ — распределение вероятностей, порожденное случайной величиной ξ. Приведенные выше формулы для дискретного и непрерывного распределений являются частными случаями этого выражения. Мы не будем пользоваться общим определением, так как это потребует множества математических знаний (о том, что такое $dP(x)$, в каком смысле понимается интеграл и т.д.).

Заметим, что существуют распределения вероятностей без математического ожидания и с такими случайными величинами иногда приходится сталкиваться на практике. Простой пример: пусть случайная величина ξ принимает значения $1^1, 2^2, \ldots, n^n, \ldots$ с вероятностями $2^{-1}, 2^{-2}$ и т.д. Тогда эта случайная величина не имеет математического ожидания.

Свойства математического ожидания. Перечислим без доказательства основные свойства математического ожидания.

1. Математическое ожидание постоянной равно этой постоянной.
2. Математическое ожидание суммы случайных величин равно сумме их математических ожиданий, т.е.

$$M(\xi + \eta) = M\xi + M\eta.$$

3. Математическое ожидание произведения случайной величины на константу равно произведению этой константы на математическое ожидание случайной величины, т.е.

$$Ma\xi = aM\xi.$$

(другими словами, постоянный множитель можно выносить за знак математического ожидания).

Полезно иметь ввиду следующее геометрическое толкование математического ожидания. Пусть $F(x)$ — функция распределения случайной величины ξ. Тогда $M\xi$ равно разности площадей, заключенных ме-

жду осью ординат, прямой $y = 1$ и кривой $y = F(x)$ в интервале $(0, +\infty)$ и между осью абсцисс, кривой $y = F(x)$ и осью ординат в промежутке $(-\infty, 0)$ (см. рис. 1.6). Это правило позволяет во многих случаях находить математическое ожидание почти без вычислений, используя различные свойства функции распределения.

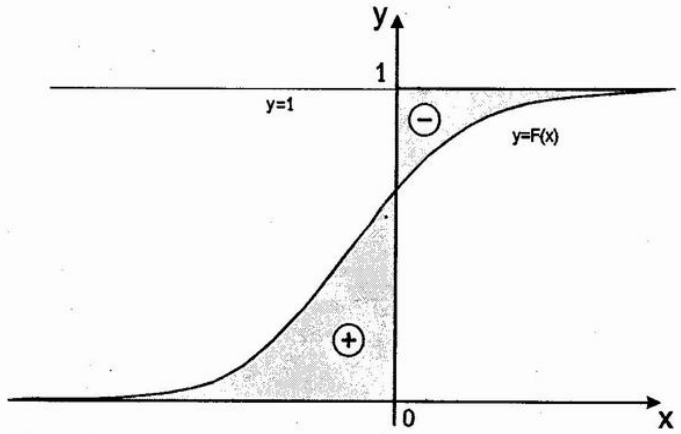


Рис. 1.6. Геометрическая интерпретация математического ожидания

Кроме среднего значения случайной величины, которое в определенном смысле характеризует центр распределения вероятностей, представляет интерес и разброс случайной величины относительно этого центра. Для характеристики (количественного описания) данного разброса в теории вероятностей используют *второй центральный момент* случайной величины. В русскоязычной литературе его называют *дисперсией* и обычно обозначают через $D\xi$.

Определение. *Дисперсией $D\xi$ случайной величины ξ называется величина*

$$D\xi = M(\xi - M\xi)^2, \qquad \textit{или} \qquad D\xi = M\xi^2 - (M\xi)^2.$$

Дисперсия, так же как и математическое ожидание, существует не для всех случайных величин (не для всех распределений вероятностей).

Если необходимо, чтобы показатель разброса случайной величины выражался в тех же единицах, что и значение этой случайной величины, то вместо $D\xi$ используют величину $\sqrt{D\xi}$, которая называется *средним квадратическим отклонением*, или стандартным отклонением случайной величины ξ.

Свойства дисперсии. Из свойств дисперсии отметим следующие:

1. Дисперсия постоянной равна нулю.

2. Для любой неслучайной постоянной a

$$D(\xi + a) = D(\xi), \qquad D(a\xi) = a^2 D(\xi).$$

Моменты. Кроме первого и второго моментов, при описании случайных величин иногда используются и другие моменты: третий, четвертый и т.д. Мы дадим их определения отдельно для дискретных и для непрерывных случайных величин.

Определение. *Для дискретной случайной величины ξ со значениями $x_1, x_2, \ldots$, имеющих вероятности $p_1, p_2, \ldots$, k-ым моментом $M\xi^k$ называется величина $M\xi^k = \sum_i x_i^k p_i$, а k-ым центральным моментом называется величина $\sum_i (x_i - M\xi)^k p_i$. Для непрерывной случайной величины с плотностью $p(x)$, k-ым моментом называется величина $\int_{-\infty}^{\infty} x^k p(x)\,dx$, а k-ым центральным моментом называется величина $M(\xi - M\xi)^k = \int_{-\infty}^{\infty} (x - M\xi)^k p(x)\,dx$.*

Чтобы приведенные формулы имели смысл, требуется, чтобы суммы и интегралы сходились абсолютно. Так же, как математическое ожидание и дисперсия, моменты существуют не для всех случайных величин.

Асимметрия и эксцесс. В отличие от обычных моментов, центральные моменты не меняются при прибавлении к случайной величине постоянного слагаемого, то есть они не зависят от выбора начала отсчета в шкале измерения случайной величины. Но от выбранной единицы измерения зависимость остается: если, скажем, случайную величину начать измерять не в метрах, а в сантиметрах, то значения центральных моментов также изменятся. Иногда это бывает неудобно. В таких случаях, чтобы устранить подобное влияние, моменты тем или иным способом *нормируют*, например, деля их на соответствующую степень среднего квадратического отклонения. В результате получается безразмерная величина, не зависящая от выбора начала отсчета и единиц измерения исходной случайной величины.

Чаще всего из нормированных моментов используются *асимметрия* и *эксцесс* — соответственно третий и четвертый нормированные центральные моменты. Для случайной величины ξ:

$$\text{асимметрия} = \frac{M(\xi - M\xi)^3}{(D\xi)^{3/2}}, \qquad \text{эксцесс} = \frac{M(\xi - M\xi)^4}{(D\xi)^2}.$$

Принято считать, что асимметрия в какой-то степени характеризует несимметричность распределения случайной величины, а эксцесс — степень выраженности «хвостов» распределения, т.е. частоту появления удаленных от среднего значений. Иногда значения асимметрии и эксцесса используют для проверки гипотезы о том, что наблюденные данные (выборка) принадлежат заданному семейству распределений, например нормальному (см. п. 2.4). Так, для любого нормального распределения асимметрия равна нулю, а эксцесс — трем.

Квантили. Для случайных величин, принимающих вещественные значения, часто используются такие характеристики, как *квантили*.

Определение. *Квантилью x_p случайной величины, имеющей функцию распределения $F(x)$, называется решение x_p уравнения $F(x) = p$.*

Величину x_p часто называется p-квантилью или квантилью уровня p распределения $F(x)$. Среди квантилей чаще всего используются *медиана* и *квартили* распределения.

Медианой называется квантиль, соответствующая значению $p = 0.5$. **Верхней квартилью** называется квантиль, соответствующая значению $p = 0.75$. **Нижней квартилью** называется квантиль, соответствующая значению $p = 0.25$.

В описательной статистике (см. ниже) нередко используют *децили*, т.е. квантили уровней $0.1, 0.2, \ldots, 0.9$. Знание децилей позволяет неплохо представлять поведение графика $y = F(x)$ в целом.

Отметим, что уравнение $F(x) = p$, определяющее p-квантили, для некоторых значений p, $0 < p < 1$, может не иметь решений либо иметь неединственное решение. Для соответствующей случайной величины ξ это означает, что некоторые p-квантили не существуют, а некоторые определены неоднозначно.

1.6. Независимые и зависимые случайные величины

Введем очень важное понятие *независимости* случайных величин. Это понятие не менее важно, чем понятие независимости событий, и тесно с ним связано. Говоря описательно, случайные величины ξ и η независимы, если независимы любые два события, которые выражаются по отдельности через ξ и η.

Для случайных величин, принимающих вещественные значения, мы можем дать следующее определение.

Определение. *Случайные величины ξ и η независимы, если*

$$P(AB) = P(A)\,P(B),$$

для любых событий $A = (a_1 < \xi < a_2)$ и $B = (b_1 < \eta < b_2)$, где числа a_1, a_2, b_1 и b_2 могут быть произвольными.

Нам незачем стремиться к большей математической аккуратности в определении независимости случайных величин, поскольку на практике

им пользоваться приходится редко. Дело в том, что независимость случайных величин обеспечивается скорее схемой постановки опытов, нежели проверкой математических соотношений. В этом вновь проглядывает аналогия с независимостью событий.

Для независимых случайных величин можно пополнить список свойств математического ожидания и дисперсии:

$$M\xi\eta = M\xi\, M\eta,$$
$$D(\xi + \eta) = D\xi + D\eta,$$

если случайные величины ξ и η независимы и указанные моменты существуют.

Ковариация. Для зависимых случайных величин часто желательно знать степень их зависимости, связи друг с другом. Таких характеристик можно придумать много, но наиболее употребительны из них *ковариация* и *корреляция*.

Определение. *Ковариацией* $\operatorname{cov}(\xi, \eta)$ *случайных величин ξ и η называют*

$$\operatorname{cov}(\xi, \eta) = M(\xi - M\xi)(\eta - M\eta),$$

если указанное математическое ожидание существует.

Легко видеть, что верна и другая формула:

$$\operatorname{cov}(\xi, \eta) = M\xi\eta - M\xi\, M\eta.$$

Поэтому для независимых случайных величин ковариация равна нулю. Обратное, естественно, неверно: равенство нулю ковариации не означает независимости случайных величин (придумайте пример!). Кроме того, ковариация вообще может не существовать (так же как и математические ожидания). Так что обращение в нуль ковариации признаков не является достаточным для их независимости, а только необходимым (и то лишь если ковариация существует).

Из других свойств ковариации отметим, что

$$\operatorname{cov}(A\xi + a,\, B\eta + b) = AB \operatorname{cov}(\xi, \eta),$$

если A, B, a, b — постоянные (неслучайные) величины.

Корреляция. Использование ковариации в качестве меры связи случайных переменных неудобно, так как величина ковариации зависит от единиц измерения, в которых измерены случайные величины. При переходе к другим единицам измерения (например, от метров к сантиметрам) ковариация тоже изменяется, хотя степень связи случайных переменных, естественно, остается прежней. Поэтому в качестве

меры связи признаков обычно используют другую числовую величину, называемую *коэффициентом корреляции*.

Определение. *Коэффициентом корреляции случайных величин* ξ *и* η *(обозначение* $\operatorname{corr}(\xi, \eta)$*, либо* $\rho(\xi, \eta)$*, либо просто* ρ*) называют*

$$\rho = \frac{\operatorname{cov}(\xi, \eta)}{\sqrt{D\xi}\sqrt{D\eta}}.$$

Заметим, что для существования коэффициента корреляции необходимо (и достаточно) существование дисперсий $D\xi > 0$, $D\eta > 0$.

Отметим следующие свойства коэффициента корреляции:

1. Модуль коэффициента корреляции не меняется при линейных преобразованиях случайных переменных: $|\rho(\xi,\eta)| = |\rho(\xi',\eta')|$, где $\xi' = a_1 + b_1\xi$, $\eta' = a_2 + b_2\eta$, a_1, b_1, a_2, b_2 — произвольные числа.
2. $|\rho(\xi,\eta)| \leqslant 1$
3. $|\rho(\xi,\eta)| = 1$ тогда и только тогда, когда случайные величины ξ и η *линейно связаны*, т.е. существуют такие числа a, b, что
$$P(\eta = a\xi + b) = 1.$$
4. Если ξ и η статистически независимы, то $\rho(\xi,\eta) = 0$. Уже отмечалось, что обратное заключение, вообще говоря, неверно. Об этом мы еще будем говорить.

Свойства 1 и 4 проверяются непосредственно. Докажем свойства 2 и 3 (при желании читатель может эти доказательства пропустить). Пусть t — переменная величина в смысле математического анализа. Рассмотрим дисперсию случайной величины $D(\eta - t\xi)$ как функцию переменной t. По свойствам дисперсии $D(\eta - t\xi) = t^2 D\xi - 2t\operatorname{cov}(\xi, \eta) + D\eta$, т.е. она представляется квадратным трехчленом от t. Этот квадратный трехчлен неотрицателен, поскольку дисперсия всегда неотрицательна. Поэтому его дискриминант $[\operatorname{cov}(\xi,\eta)]^2 - D\xi\, D\eta \leqslant 0$, а это и означает, что $|\rho(\xi,\eta)| \leqslant 1$ (свойство 2).

Для доказательства свойства 3 заметим, что при $|\rho(\xi,\eta)| = 1$ дискриминант приведенного выше квадратного трехчлена обращается в 0, а поэтому при некотором t_0 значение $D(\eta - t_0\xi)$ равно нулю. Равенство нулю дисперсии означает, что эта случайная величина постоянна, т.е. для некоторого c вероятность $P(\eta - t_0\xi = c)$ равна единице, что и требовалось доказать.

Итак, корреляция случайных величин принимает значения от -1 до 1 и может быть равна ± 1, только если эти величины линейно зависят друг от друга. Значения корреляции, близкие к -1 или 1, указывают, что зависимость случайных величин друг от друга почти линейная. Значения ковариации, близкие к нулю, означают, что связь между случайными величинами либо слаба, либо не носит линейного характера. Подробнее о связи между случайными величинами мы расскажем в главе 9.

1.7. Случайный выбор

Значительная часть статистики связана с описанием больших совокупностей объектов. Если интересующая нас совокупность слишком многочисленна, либо ее элементы малодоступны, либо имеются другие причины, не позволяющие изучать сразу все ее элементы, прибегают к изучению какой-то части этой совокупности. Эта выбранная для полного исследования группа элементов называется *выборкой* или *выборочной совокупностью*, а все множество изучаемых элементов — *генеральной совокупностью*. Естественно стремиться сделать выборку так, чтобы она наилучшим образом представляла всю генеральную совокупность, то есть была бы, как говорят, *репрезентативной*. Как этого добиться? Если генеральная совокупность нам мало известна или совсем неизвестна, не удается предложить ничего лучшего, чем чисто случайный выбор. Дадим его определение, начав со случайного выбора одного объекта.

Определение. *Выбор одного объекта называют чисто случайным (или, как иногда говорят, простым случайным), если все объекты имеют равные вероятности оказаться выбранными.*

Если речь идет о выборе одного объекта из N, это означает, что для каждого элемента вероятность выбора равна $1/N$.

Определение. *Выбор* n *объектов из* N *называют чисто случайным (или простым случайным), если все наборы из* n *объектов имеют одинаковые вероятности быть выбранными.*

Чисто случайный выбор n объектов (иногда говорят — *случайную выборку объема* n можно получить, извлекая из генеральной совокупности по одному объекту последовательно и чисто случайно.

Нарушение принципов случайного выбора порой приводило к серьезным ошибкам. Стал знаменитым своей неудачей опрос, проведенный американским журналом «Литературное обозрение» относительно исхода президентских выборов в США в 1936 году.

Кандидатами на этих выборах были Ф.Д.Рузвельт и А.М.Ландон. В качестве генеральной совокупности редакция журнала использовала телефонные книги. Отобрав случайно 4 миллиона адресов, она разослала по всей стране открытки с вопросом об отношении к кандидатам в президенты. Затратив большую сумму на рассылку и обработку открыток, журнал объявил, что на предстоящих выборах президентом США с большим перевесом будет избран Ландон. Результат выборов оказался противоположным этому прогнозу.

Здесь были совершены сразу две ошибки — во-первых, телефонные книги сами по себе дают не репрезентативную выборку из населения страны, хотя бы потому, что абоненты — в основном зажиточные главы семейств. Во-вторых, прислали ответы не все, а люди, не только достаточно уверенные в своем мнении, но и привыкшие отвечать на письма, т.е. в значительной части представители

делового мира, которые и поддерживали Ландона. Если бы редакция критически подошла к своей работе, она поняла бы, что методика опроса страдает изъянами.

Явление, подобное только что описанному, когда выборка представляет не всю генеральную совокупность, а лишь какой-то ее слой, какую-то ее часть, называется *смещением выборки*. Смещение — один из основных источников ошибок при использовании выборочного метода.

Однако для тех же самых президентских выборов социологи Дж.Гэллап и Э.Роупер правильно предсказали победу Рузвельта, основываясь только на 4 тысячах анкет. Причиной этого успеха, прославившего его авторов, было не только правильное составление выборки. Они учли, что общество распадается на социальные группы, которые более однородны, в том числе по своим политическим взглядам. Поэтому выборка из слоя может быть относительно малочисленной с тем же результатом точности. Имея результаты обследования по слоям, можно характеризовать общество в целом. Сейчас такая методика является общепринятой.

В главе 15 мы кратко расскажем о том, как случайный выбор используется при практическом проведении выборочных обследований. Более подробно с этим кругом вопросов можно познакомиться по книге [54].

1.8. Выборки и их описание

1.8.1. Что такое выборка

В предыдущем параграфе мы использовали слово «выборка» для описания результата случайного выбора нескольких объектов из некоторой заданной генеральной совокупности. В этом смысле слово «выборка» используется, когда мы говорим «социологический опрос произведен на выборке из 2000 человек (респондентов)». Но в математической литературе слово «выборка» гораздо чаще используется в другом смысле. Дадим его определение.

Определение. *Выборкой называют последовательность независимых одинаково распределенных случайных величин.*

Именно в этом значении слово «выборка» употребляется в статистических задачах естествознания и в этом значении оно будет встречаться далее в этой книге.

Замечание. Происхождение данного значения слова «выборка» связано с давними ассоциациями всякого случайного испытания со случайным выбором из некоей совокупности. Если эта совокупность является конечной (как это и бывает на практике), то последовательные результаты случайных выборов из нее не являются независимыми, поскольку каждое изъятие элемента из совокупности изменяет эту совокупность. Конечно, для обширных совокупностей извлечение одного или нескольких элементов мало изменяет вероятности выбора, но все же они не остаются постоянными в процессе выбора. В связи с этим иногда говорят о *бесконечных генеральных совокупностях* (популяциях) и о случайном выборе из них. Это образное выражение может сделать более наглядным представление о независимых случайных величинах.

1.8.2. Выборочные характеристики

Перечисленные в параграфе 1.4 характеристики случайной величины существенно опираются на знание закона ее распределения $F(x)$. Для практических задач такое знание — редкость. Здесь закон распределения обычно неизвестен, в лучшем случае он известен с точностью до некоторых неизвестных параметров. Как же тогда получить сведения о распределении случайной величины и его характеристиках? Это становится возможным, когда имеются независимые многократные повторения опыта, в котором мы измеряем значения интересующей нас случайной величины.

Предположим, что наблюдения над случайной величиной ξ можно повторять независимо и в неизменных условиях, получая ее независимые реализации $x_1, x_2, \ldots, x_n$. Тогда $x_1, x_2, \ldots, x_n$ будут независимыми одинаково распределенными случайными величинами, то есть *выборкой*. Зная величины $x_1, x_2, \ldots, x_n$, мы можем построить приблизительные значения для функции распределения и других характеристик случайной величины ξ. Это и позволяет нам изучать свойства случайных величин, не зная их законов распределения.

Замечание. Мы уже встречались с идеей независимых повторений случайного опыта в неизменных условиях, когда обсуждали измерения вероятностей событий. Возвращение к этой идее не удивительно, поскольку для описания распределения случайной величины ξ мы как раз и должны уметь указывать вероятности всех событий, выражаемых через ξ.

Расскажем о том, как по имеющейся выборке можно получить приближенные значения для характеристик случайных величин. Начнем с функции распределения случайной величины.

Эмпирическая функция распределения.

Определение. *Выборочной (эмпирической) функцией распределения случайной величины ξ, построенной по выборке $x_1, x_2, \ldots, x_n$, называется функция $F_n(x)$, равная доле таких значений x_i, что $x_i \leqslant x$, $i = 1, \ldots, n$.*

Иначе говоря, $F_n(x)$ есть частота события $x_i \leqslant x$ в ряду $x_1, x_2, \ldots, x_n$.

Для построения выборочной функции распределения удобно от выборки $x_1, \ldots, x_n$ перейти к вариационному ряду $x_{(1)}, \ldots, x_{(n)}$.

Определение. *Вариационным рядом называют выборку, перенумерованную в порядке возрастания.*

Так, $x_{(1)}$ обозначает наименьшее из чисел $x_1, \ldots, x_n$, $x_{(2)}$ — наименьшее из оставшихся после удаления $x_{(1)}$ и т.д. В частности, $x_{(n)}$ обозначает наибольшее из $x_1, \ldots, x_n$. При $x < x_{(1)}$, по определению, $F_n(x) = 0$, в точке $x_{(1)}$ функция $F_n(x)$ совершает скачок, равный $1/n$, и остается постоянной до значения $x_{(2)}$, и т.д. Таким образом, выборочная функция распределения является ступенчатой с точками скачков $x_{(1)}, \ldots, x_{(n)}$, причем величина каждого скачка равна $1/n$ (рис. 1.7).

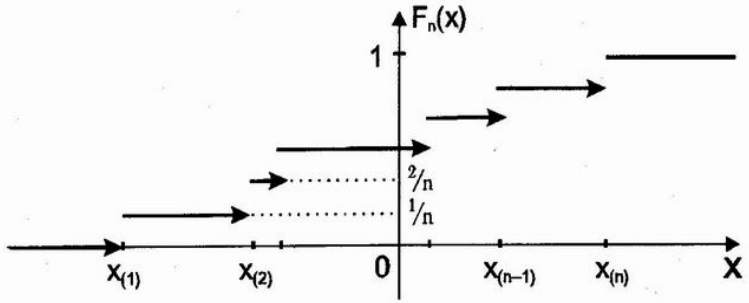


Рис. 1.7. Общий вид эмпирической функции распределения

Видно, что график эмпирической функции распределения напоминает график дискретного распределения вероятностей. Это не случайно: эмпирическую функцию выборки $x_1, \ldots, x_n$ можно рассматривать как функцию распределения вероятностей, где каждому значению x_i, $i = 1, \ldots, n$, приписана вероятность $1/n$. Иногда поэтому вместо эмпирической (или выборочной) функции распределения употребляют название «функция распределения выборки».

Связь между эмпирической функцией распределения и функцией распределения (иногда, чтобы подчеркнуть разницу, говорят о теоретической функции распределения, что не вполне правильно, ибо никакой теории здесь нет) основана на уже упомянутой теореме Бернулли. Она такая же, как связь между частотой события и его вероятностью. Для любого числа x значение $F_n(x)$ представляет собой частоту события $(\xi \leqslant x)$ в ряду из n независимых повторений. Поэтому $F_n(x) \to F(x)$ при $n \to \infty$.

Установлено, что выборочная функция распределения с ростом объема выборки n равномерно по x аппроксимирует теоретическую функцию распределения $F(x)$ случайной величины ξ, т.е. величина $\sup_x |F_n(x) - F(x)|$ стремится к нулю при $n \to \infty$ с вероятностью 1.

Выборочные характеристики. На указанном выше свойстве выборочной функции распределения основаны многие методы математической статистики. Замена функции распределения $F(x)$ на ее выборочный аналог $F_n(x)$ в определении математического ожидания, дисперсии, медианы и т.п. приводят к *выборочному среднему*, *выборочной дисперсии*, *выборочной медиане* и т.д. Покажем, как действует это правило и чему равны соответствующие выборочные характеристики.

В случае математического ожидания, используя в качестве функции распределения случайной величины ξ выборочную функцию $F_n(x)$ мы подразумеваем, что некая случайная величина может принять значения $x_{(1)}, \ldots, x_{(n)}$, каждое с вероятностью $1/n$. Воспользовавшись формулой для определения математического ожидания для дискретной случайной величины приходим к следующему определению.

Средним значением выборки (выборочным средним), *или выборочным аналогом математического ожидания, называется величина*

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

Аналогично,

Дисперсией выборки (выборочной дисперсией), *или выборочным аналогом дисперсии, называется величина*

$$\frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^2.$$

Однако в статистике чаще в качестве выборочной дисперсии используют

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2.$$

поскольку математическое ожидание величины s^2 равно дисперсии ξ, т.е. $Ms^2 = D\xi$.

Выборочной квантилью *называется решение уравнения*

$$F_n(x) = p.$$

В частности, *выборочная медиана* есть решение уравнения

$$F_n(x) = 0.5.$$

Замечание. Решение уравнения $F_n(x) = 0.5$ при четном $n = 2k$ определено не однозначно. Действительно, для каждого x из промежутка $x_{(k)} \leqslant x < x_{(k+1)}$ $F(x) = 0.5$. В этом случае условились определить выборочную медиану как $\frac{x_{(k)} + x_{(k+1)}}{2}$. При нечетном $n = 2k + 1$ решение уравнения

$F_n(x) = 0.5$ не существует, так как выборочная функция распределения принимает только значения из множества $\left\{\frac{i}{2k+1},\ i = 0, 1, \ldots, 2k+1\right\}$. В связи с этим выборочную медиану определяют как $x_{(k+1)}$, ибо в этой точке $F_n(x)$ переходит через $1/2$. Выборочная медиана разбивает выборку пополам: слева и справа от нее оказывается одинаковое число элементов выборки. Заметим, что при больших значениях n: $F_n\left(x_{(k+1)}\right) = \frac{(k+1)}{2k+1} \to \frac{1}{2}$.

Важным свойством выборочных характеристик является то, что все они сходятся к соответствующим теоретическим характеристикам при растущих объемах выборки n. Характер этой сходимости будет рассмотрен в главах 4 и 5, когда речь пойдет о законе больших чисел и о построении статистических оценок различных параметров распределения.

Выборочные ковариация и корреляция. Если в каждом наблюдении мы регистрируем значения не одной, а двух (или нескольких) случайных величин одновременно, мы получаем в результате двумерную (или многомерную) выборку. Для таких выборок тоже можно говорить о числовых характеристиках, например, о ковариации или корреляции компонент этой выборки.

Коэффициентом корреляции двумерной выборки $(x_1, y_1), \ldots, (x_n, y_n)$, или *выборочным коэффициентом корреляции* называют величину

$$r = \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^n (x_i - \overline{x})^2}\sqrt{\sum_{i=1}^n (y_i - \overline{y})^2}}.$$

(Иногда ее называют коэффициентом корреляции К.Пирсона.) Аналогично определяется *выборочная ковариация*, она равна $\frac{1}{n}\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})$.

1.8.3. Ранги и ранжирование

Ранги. Во многих случаях имеющиеся в нашем распоряжении числовые данные (например, значения элементов выборки) носят в той или иной мере условный характер. Например, эти данные могут быть тестовыми баллами, экспертными оценками, данными о вкусовых или политических предпочтениях опрошенных людей и т.д. Анализ таких данных требует особой осторожности, поскольку многие предпосылки классических статистических методов (например, предположения о каком-либо конкретном, скажем нормальном, законе распределения) для них не выполняются. Твердую основу для выводов здесь дают только соотношения между наблюдениями типа «больше-меньше», так как они не меняются при изменении шкалы измерений. Например, при анализе анкет с данными о симпатиях избирателей к политическим деятелям

мы можем сказать, что политик, получивший больший балл в анкете, более симпатичен отвечавшему на вопросы человеку (респонденту), чем политик, получивший меньший балл. Но на сколько (или во сколько раз) он более симпатичен, сказать нельзя, так как для предпочтений нет объективной единицы измерения.

В подобных случаях (которые мы будем более подробно рассматривать в последующих главах), имеет смысл вообще отказаться от анализа конкретных значений данных, а исследовать только информацию об из взаимной упорядоченности. Для этого от исходных числовых данных осуществляют переход к их *рангам*.

Определение. *Рангом наблюдения называют тот номер, который получит это наблюдение в упорядоченной совокупности всех данных — после их упорядочения по определенному правилу (например, от меньших значений к большим или наоборот).*

Чаще всего упорядочение чисел (набор которых составляют упомянутые выше данные) производят по величине — от меньших к большим. Именно такое упорядочение и связанное с ним ранжирование (присвоение рангов) мы будем иметь в виду в дальнейшем.

Пример. Пусть выборка состоит из чисел 6, 17, 14, 5, 12. Тогда рангом числа 6 оказывается 2, рангом 17 будет 5 и т.д.

Определение. *Процедура перехода от совокупности наблюдений к последовательности их рангов называется ранжированием. Результат ранжирования называется ранжировкой.*

Статистические методы, в которых мы делаем выводы о данных на основании их рангов, называются ранговыми. Они получили широкое распространение, так как надежно работают при очень слабых предположениях об исходных данных (не требуя, например, чтобы эти данные имели какой-либо конкретный закон распределения). В последующих главах этой книги мы рассмотрим применение ранговых методов в наиболее распространенных практических задачах.

Средние ранги. Трудности в назначении рангов возникают, если среди элементов выборки встречаются совпадающие. (Так часто бывает, когда данные регистрируются с округлением.) В этом случае обыкновенно используют *средние ранги*.

Средние ранги вводятся так. Предположим, что наблюдение x_i имеет ту же величину, что и некоторые другие из общего числа n наблюдений. (Эту совокупность одинаковых наблюдений из набора $x_1, \ldots, x_n$ называют *связкой*; количество таких одинаковых наблюдений в данной связке называют ее размером.) Средний ранг x_i в ранжировке

наблюдений $x_1, \ldots, x_n$ есть среднее арифметическое тех рангов, которые были бы назначены x_i и всем остальным элементам связки, если бы одинаковые наблюдения оказались различны.

В качестве примера рассмотрим выборку 6, 17, 12, 6, 12. Ее ранжировка равна $1\frac{1}{2}$, 5, $3\frac{1}{2}$, $1\frac{1}{2}$, $3\frac{1}{2}$.

1.8.4. Методы описательной статистики

В практических задачах мы обычно имеем совокупность наблюдений $x_1, x_2, \ldots, x_n$, на основе которых требуется сделать те или иные выводы. Часто этих наблюдений много — несколько десятков, сотен или тысяч, так что возникает задача компактного описания имеющихся наблюдений. В идеале таким описанием могло бы быть утверждение, что $x_1, x_2, \ldots, x_n$ являются выборкой, то есть независимыми реализациями случайной величины ξ с известным законом распределения $F(x)$. Это позволило бы теоретически провести расчеты всех необходимых исследователю характеристик наблюдаемого явления.

Однако далеко не всегда мы можем утверждать, что $x_1, x_2, \ldots, x_n$ являются независимыми и одинаково распределенными случайными величинами. Во-первых, это не так-то просто проверить (для подтверждения этого требуются значительные объемы наблюдений и специальные, порой многочисленные, тесты). А во-вторых, часто заведомо известно, что это не так. Поэтому для компактного описания совокупности наблюдений $x_1, x_2, \ldots, x_n$ используют другие методы — методы описательной статистики.

Определение. *Методами описательной статистики принято называть методы описания выборок $x_1, x_2, \ldots, x_n$ с помощью различных показателей и графиков.*

Полезность методов описательной статистики состоит в том, что несколько простых и довольно информативных статистических показателей способны избавить нас от просмотра сотен, а порой и тысяч, значений выборки.

Показатели описательной статистики. Описывающие выборку показатели можно разбить на несколько групп.

1. *Показатели положения* описывают положение данных на числовой оси. Примеры таких показателей — минимальный и максимальный элементы выборки (первый и последний член вариационного ряда), верхний и нижний квартили (они ограничивают зону, в которую попадают 50% центральных элементов выборки). Наконец, сведения о середине совокупности могут дать выборочное среднее значение, выборочная медиана и другие аналогичные характеристик.
2. *Показатели разброса* описывают степень разброса данных относительно своего центра. К ним в первую очередь относятся: дисперсия выборки, стандартное отклонение, размах выборки (разность между максимальным и минимальным элементами), межквартильный размах (разность между верхней и нижней квартилью), коэффициент эксцесса и т.п. По сути дела, эти показатели говорят, насколько кучно основная масса данных группируется около центра.
3. *Показатели асимметрии.* Третья группа показателей отвечает на вопрос о симметрии распределения данных около своего центра. К ней можно отнести: коэффициент асимметрии, положение выборочной медианы относительно выборочного среднего и относительно выборочных квартилей, гистограмму и т.д.
4. *Показатели, описывающие закон распределения.* Наконец, четвертая группа показателей описательной статистики дает представление собственно о законе распределения данных. Сюда относятся графики гистограммы и эмпирической функции распределения, таблицы частот.

Применение показателей описательной статистики. Из перечисленных выше характеристик на практике по традиции чаще всего используются выборочное среднее, медиана и дисперсия (или стандартное отклонение). Однако для получения более точных и достоверных выводов мы настоятельно рекомендуем внимательно изучать и другие из перечисленных выше характеристик, а так же обращать внимание на условия получения выборочных совокупностей.

Особое внимание следует обратить на наличие в выборке *выбросов* — грубых (ошибочных), сильно отличающихся от основной массы, наблюдений. Дело в том, что даже одно или несколько грубых наблюдений способны сильно исказить такие выборочные характеристики, как среднее, дисперсия, стандартное отклонение, коэффициенты асимметрии и эксцесса. Проще всего обнаружить такие наблюдения с помощью перехода от выборки к ее вариационному ряду или гистограммы с достаточно большим числом интервалов группировки (см. ниже). Подозрение о присутствии таких наблюдений может возникнуть, если выборочная медиана заметно отличается от выборочного среднего, хотя в целом совокупность симметрична; если положение медианы сильно несимметрично относительно минимального и максимального элементов выборки, и т.д.

Замечание. Наличие выбросов, то есть грубых (ошибочных) наблюдений, может не только сильно исказить значения выборочных показателей — выборочного среднего, дисперсии, стандартного отклонения и т.д., — но и привести к многим другим ошибочным выводам. Дело в том, что большинство традиционных статистических методов весьма чувствительно к отклонениям от условий применимости метода. К сожалению, интенсивно развивающиеся в последние два десятилетия статистические методы, устойчивые к выбросам и другим отклонениям, еще не получили широкого распространения на практике, за исключением ранговых процедур для наиболее стандартных задач. Отчасти причиной здесь является значительная вычислительная сложность этих методов, из-за чего их применение невозможно без использования специальных компьютерных программ.

1.8.5. Наглядные методы описательной статистики

Рассмотренные выше вопросы и понятия дают первое представление о теоретических и выборочных характеристиках случайных величин. С различной степенью подробности и строгости этот материал изложен во многих учебниках по теории вероятностей и математической статистике, выбор которых должен определяться направленностью интересов и уровнем математической подготовки читателя.

Группировки. Нередко (для облегчения регистрации или при невысокой точности измерений) данные группируют, т.е. числовую ось разбивают на промежутки и для каждого промежутка указывают число n_j элементов выборки $x_1, \ldots, x_n$, которые в него попали (здесь j — номер промежутка). Ясно, что $\sum_j n_j = n$.

В этом случае в качестве выборочного среднего и дисперсии используют следующие величины. Пусть $t_1, t_2, \ldots$ — центры (середины) выбранных промежутков. Тогда вместо выборочного среднего $\overline{x}$ используют величину $\overline{t}$:

$$\overline{x} \simeq \overline{t} = \frac{\sum_j t_j n_j}{n} = \sum_j t_j \frac{n_j}{n},$$

а в качестве выборочной дисперсии s^2

$$s^2 \simeq \frac{1}{n-1} \sum_j (t_j - \overline{t})^2 n_j.$$

Приведем ниже еще несколько полезных приемов описательной статистики для работы с выборкой. В качестве примера рассмотрим данные из таблицы 1.1, в которой приведены результаты измерения диаметров 200 головок заклепок. Здесь случайная величина — диаметр

изготовляемой заклепки, приведенные 200 значений — ее независимые реализации.

Точечная диаграмма. Данные, собранные в таблицу, трудно обозреть. Они нуждаются в наглядном представлении. Одной из форм такого наглядного представления служит *точечная диаграмма*: табличные данные отмечаются точками на числовой шкале. Если некоторое число встречается в таблице несколько раз, его представляют соответствующим количеством точек. Точечная диаграмма для данных таблицы 1.1 приведена на рис. 1.8.

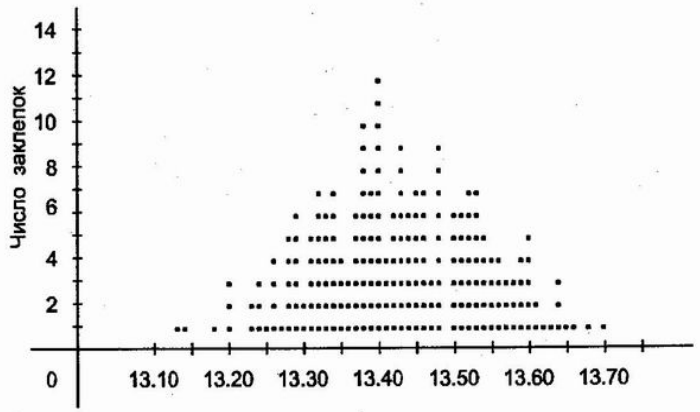


Рис. 1.8. Точечная диаграмма. Распределение диаметров 200 головок заклепок, выраженных в мм.

Эта диаграмма удобна в том случае, когда одно и то же значение случайной величины повторяется в выборке несколько раз. В противном случае точечная диаграмма сводится к последовательности точек на оси абсцисс. Во всех случаях точечная диаграмма помогает построить график выборочной функции распределения.

Гистограмма. Более наглядное описание данных достигается путем группировки наблюдений в классы. Под группировкой, или классификацией, мы будем понимать некоторое разбиение интервала, содержащего все n наблюденных результатов $x_1, \ldots, x_n$ на m интервалов, которые будем называть *интервалами группировки*. Длины интервалов обозначим через $\Delta_1, \ldots, \Delta_m$, а середины интервалов группировки — через $t_1, \ldots, t_m$.

Число наблюдений n_{ij} в j-м интервале группировки равно количеству x_i, $i = 1, \ldots, n$, удовлетворяющих неравенству

$$|x_i - t_j| < \frac{1}{2}\Delta_j.$$

Определим величину $h_j = n_j/n$, которая означает частоту попадания наблюдений в j-ый интервал группировки. Для того, чтобы избавиться от влияния размера интервала группировки на h_j, вводится величина $f_j = h_j/\Delta_j$.

Определение. *Графическое изображение зависимости частоты попадания элементов выборки от соответствующего интервала группировки называется гистограммой выборки.*

Подчеркнем, что в качестве ординаты здесь берется не сама частота, а частота, деленная на длину интервала группировки. Если все интервалы группировки имеют одинаковую длину, деление на Δ обычно опускают и n_j или h_j используют как ординаты, как это показано на нескольких рисунках ниже. На рис. 1.9 приведена гистограмма выборки при длине интервала группировки, равной 0.01 мм. Ординатой на этом рисунке является число заклепок в каждом интервале группировки.

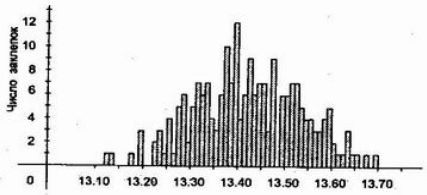


Рис. 1.9. Гистограмма. Длина интервала группировки равна 0.01 мм

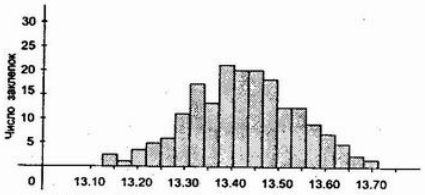


Рис. 1.10. Гистограмма. Длина интервала группировки равна 0.03 мм

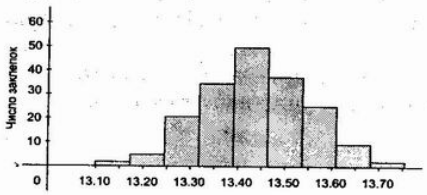


Рис. 1.11. Гистограмма. Длина интервала группировки равна 0.07 мм

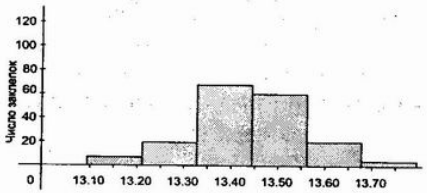


Рис. 1.12. Гистограмма. Длина интервала группировки равна 0.11 мм

Отметим, что согласно определению площадь каждого столбца гистограммы равна (точнее, пропорциональна) частоте попадания наблюдений в данный интервал группировки.

Ясно, что величина интервала группировки существенно влияет на общий вид гистограммы. Если длина интервала группировки мала, то влияние случайных колебаний начинает преобладать, так как каждый интервал содержит при этом лишь небольшое число наблюдений. Этот эффект хорошо виден на рис. 1.9. На рис. 1.10—1.12 приведены гистограммы выборки при длине интервала группировки, равной 0.03, 0.07 и 0.11 мм соответственно. Из приведенных рисунков видно, что

чем больше величина интервала группировки, тем более скрадываются характерные черты распределения.

Если группированное распределение должно являться основой для последующих вычислений, то, как правило, все интервалы группировки должны быть небольшими и иметь одну и ту же длину.

Пример. О пользе наглядных приемов описательной статистики красноречиво говорит следующий пример, относящийся еще к началу XX века. Мы изложим его, следуя Р.Фишеру (одному из создателей современной математической статистики).

... Иоханес Шмидт из Карлсберговской лаборатории в Копенгагене был не только ихтиологом, но и неутомимым биостатистиком. Он развивал идею, что рыбы одного вида распадаются на относительно изолированные сообщества. Между этими группами он находил статистические различия по числу позвонков или лучей плавников. Для доказательства этого он строил гистограммы распределений числа позвонков (лучей плавников) для каждой из групп и сравнивал их между собой. Причиной различий сообществ рыб служит то, что эти сообщества не смешиваются при размножении: каждая группа мечет икру в своем месте. Часто такие различия были заметны даже между стаями рыб одного вида, обитавшими в одном фьорде.

Однако для угрей Шмидт не смог найти никаких статистических различий между выборками, выловленными даже в очень далеких друг от друга местах — будь то различные части Европейского материка, Азорские острова, Нил или Исландия. Шмидт решил, что угри всех различных речных систем составляют одно сообщество, а значит, они должны иметь общее место размножения.

Через некоторое время это предположение подтвердилось в ходе экспедиции исследовательского судна «Диана». Одним из главных успехов этого плавания была поимка личинок угря в некотором ограниченном районе Западной Атлантики — Саргассовом море. Выяснилось, что все угри, независимо от своего «места жительства», отправляются выводить потомство только в Саргассово море.

1.9. Методы описательной статистики в пакетах STADIA и SPSS

1.9.1. Пакет STADIA

В пакете STADIA довольно полно представлены методы описательной статистики, все они собраны воедино в разделе пакета «Параметрические тесты» меню Статистические методы. Проиллюстрируем их работу на рассмотренных выше примерах. При этом мы будем рассматривать версию пакета STADIA 6.0 для Windows — интерфейс более поздних версий этого пакета не отличается от версии 6.0.

Более подробное изложение свойств этих и многих других распределений можно найти в [19], [65], [77], [87] и [111].

2.1. Биномиальное распределение

Область применения. Биномиальное распределение — это одно из самых распространенных дискретных распределений, оно служит вероятностной моделью для многих явлений. Оно возникает в тех случаях, когда нас интересует, сколько раз происходит некоторое событие в серии из определенного числа независимых наблюдений (опытов), выполняемых в одинаковых условиях. Поясним сказанное на примере.

Рассмотрим какое-либо массовое производство. Даже во время его нормальной работы иногда изготавливаются изделия, не соответствующие стандарту, т.е. дефектные. Обозначим долю дефектных изделий через p, $0 < p < 1$. Какое именно произведенное изделие окажется негодным, сказать заранее (до его изготовления) невозможно. Для описания подобной ситуации обычно используется следующая математическая модель:

а) каждое изделие с вероятностью p может оказаться дефектным (с вероятностью $q = 1 - p$ оно соответствует стандарту); эта вероятность для всех изделий одинакова;

б) появление как дефектных, так и стандартных изделий происходит независимо друг от друга. Это значит, что в нормальном процессе производства появление бракованного изделия не влияет на возможность появления брака в дальнейшем. Нарушение этого условия означает сбой нормального технологического режима.

Последовательность независимых испытаний, в которых результатом каждого из испытаний может быть один из двух исходов (например, успех и неудача), и вероятность «успеха» (или «неудачи») в каждом из испытаний одна и та же, называется *схемой испытаний Бернулли*. Поэтому мы можем перефразировать вышесказанное так: в нормальных условиях технологический процесс производства математически представляется схемой испытаний Бернулли.

Для чего же на производстве требуется подсчитывать число дефектных изделий? Как правило, это делается для контроля технологического процесса. При массовом производстве сплошная проверка качества изготовленных изделий обычно неоправданна. Поэтому для контроля качества из произведенной продукции наудачу отбирают определенное количество изделий (в дальнейшем — n), проверяют их, регистрируют

найденное число бракованных изделий (в дальнейшем — X) и в зависимости от значения X принимают то или иное решение о состоянии производственного процесса. Теоретически X может принимать любые целые значения от 0 до n включительно, но, конечно, вероятности этих значений различны. Для того, чтобы делаемые по значению X выводы были обоснованными, требуется знать распределение случайной величины X. Если выполняются приведенные выше условия схемы испытаний Бернулли, то распределение X является *биномиальным распределением*, и вероятности значений X можно получить очень просто.

Пронумеруем в произвольном порядке n проверяемых изделий (например, в порядке их поступления на контроль). Будем обозначать исход испытания каждого изделия нулем или единицей (ноль — нормальное изделие, единица — дефектное), и будем записывать итоги проверки партии из n изделий в виде последовательности из n нулей и единиц. Событие $(X = k)$, или, другими словами «среди n испытаний изделий оказалось k бракованных, а остальные $(n-k)$ — годные» — это совокупность всех последовательностей, содержащих в любом порядке k единиц и $(n - k)$ нулей. Вероятность того, что в результате проверки будет получена любая из таких последовательностей, равна $p^k(1-p)^{n-k}$, а число таких последовательностей — $C_n^k = \frac{n!}{k!(n-k)!}$. Поэтому, согласно свойствам вероятностей, описанным в п. 1.2, вероятность события $(X = k)$ равна:

$$P(X = k) = C_n^k p^k (1-p)^{n-k} = \left(\frac{n!}{k!(n-k)!}\right) p^k q^{n-k}.$$

Определение. *Случайная величина X имеет биномиальное распределение с параметрами n и p, если она принимает значения $0, 1, \ldots, n$ с вероятностями:*

$$P(X = k) = C_n^k p^k (1-p)^{n-k} \qquad k = 0, 1, \ldots, n.$$

Параметр p обычно называют вероятностью «успеха» в испытании Бернулли. В приведенном выше примере «успех» соответствует обнаружению бракованной детали. Распределение называется биномиальным, потому что вероятности $P(X = k)$ являются слагаемыми бинома Ньютона:

$$1^n = [p + (1-p)]^n = \sum_{k=0}^{n} C_n^k p^k (1-p)^{n-k} = \sum_{k=0}^{n} P(X = k).$$

Чтобы подчеркнуть зависимость $P(X = k)$ от p и n, вероятность $P(X = k)$ обычно записывают в виде:

$$P(X = k \mid n, p).$$

Свойства. Математическое ожидание и дисперсия случайной величины, имеющей биномиальное распределение, равны:

$$MX = np, \qquad DX = np(1-p).$$

Эти выражения легко получить с помощью следующего полезного приема. Введем для каждого отдельного испытания Бернулли случайную величину ξ, которая может принимать только два значения: 1, если испытание закончилось успехом, и 0, если неудачей. Если дать номера $1, 2, \ldots$ отдельным испытаниям, то те же номера надо присвоить и соответствующим им случайным величинам ξ: $\xi_1, \xi_2, \ldots$ Тогда X можно представить в виде: $X = \xi_1 + \xi_2 + \cdots + \xi_n$, причем случайные слагаемые в данной формуле статистически независимы и одинаково распределены. Для любого k от 1 до n выполняется $M\xi_k = p$, $D\xi_k = p(1-p)$, поэтому, согласно свойствам математического ожидания и дисперсии из п. 1.5: $MX = nM\xi$, $DX = nD\xi$, что и приводит к указанным выше выражениям.

На рис. 2.1 показаны вероятности $P(X = k)$ при $n = 10$ для различных значений p (p = 0.1, 0.2, 0.4 и 0.5).

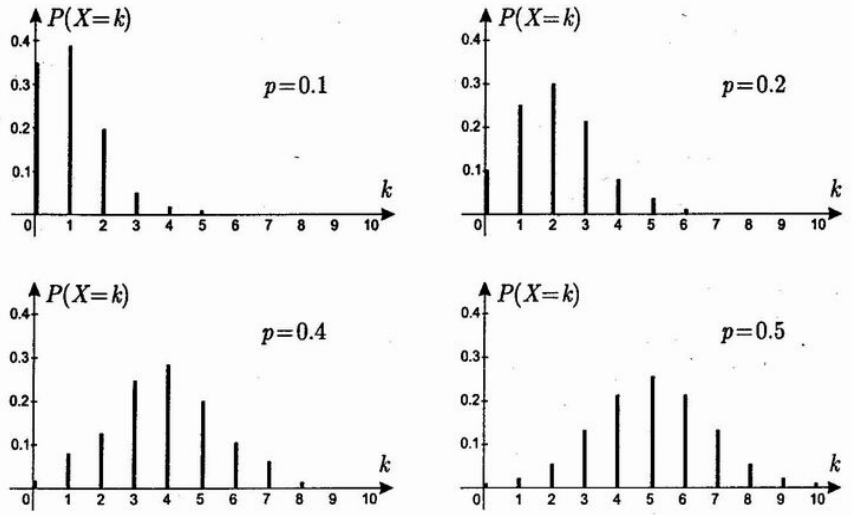


Рис. 2.1. Вид биномиального распределения для различных значений p при $n = 10$

Связь с другими распределениями. Биномиальное распределение тесно связано с многими другими распределениями. Ниже мы укажем наиболее часто используемые из этих связей. Описание других можно найти в [19], [111].

1. Биномиальное распределение с параметрами n и p может быть аппроксимировано нормальным распределением со средним np и стандартным отклонением $(np(1-p))^{1/2}$, если только выполняются условия $np(1-p) > 5$ и $0.1 \leqslant p \leqslant 0.9$. При условии $np(1-p) > 25$ эту аппроксимацию можно применять независимо от значения p.

2. Биномиальное распределение с параметрами n и p может быть аппроксимировано распределением Пуассона со средним np при условии, что $p < 0.1$ и n достаточно велико.

Таблицы. Для биномиального распределения, как и для других распределений вероятностей, есть два типа таблиц.

В таблицах первого типа приводятся вероятности $P(X = k)$ при различных значениях p и n. Например, в [19] приведены таблицы $P(X = k \mid n, p)$ (с пятью десятичными знаками) для n от 5 до 30, с шагом по n, равным 5 (краткое обозначение: $n = 5(5)30$), и $p = 0.01; 0.02(0.02); 0.10(0.10); 0.50$. Последнее выражение для p означает, что в таблицах есть значения для $p = 0.01$, для $p = 0.02$, далее p изменяется с шагом 0.02 до 0.10 и со значения $p = 0.1$ оно изменяется с шагом 0.1 до 0.5.

В таблицах второго типа даны значения накопленных вероятностей биномиального распределения, т.е. значения

$$P(X \leqslant k \mid n, p) = \sum_{m=0}^{k} P(X = m \mid n, p).$$

Например, в [77], $P(X \leqslant k \mid n, p)$ даны для $n = 1(1)25$, $p = 0.005(0.005); 0.02(0.01); 0.10(0.05); 0.30(0.10); 0.50$, для $k = 0(1)n$.

В описаниях таблиц обычно можно найти указания, как поступать, если интересующие нас значения n и/или p в данных таблицах отсутствуют (см., например, [19]).

Замечание. Значения вероятностей $P(X = k)$ биномиального распределения с параметром $p > 0.5$ легко получить, зная соответствующие вероятности при $p < 0.5$. Действительно, если вероятность «успеха» $p > 0.5$, то вероятность «неудачи» $q = 1 - p < 0.5$. Поменяв названия «успех» и «неудача» одно на другое, мы сведем случай $p > 0.5$ к $p < 0.5$. Другими словами:

$$P(X = k \mid n, p) = P(X = n - k \mid n, 1-p).$$

Это свойство учитывается при составлении статистических таблиц биномиального распределения.

2.2. Распределение Пуассона

Область применения. Распределение Пуассона играет важную роль в ряде вопросов физики, теории связи, теории надежности, теории массового обслуживания и т.д. — словом, всюду, где в течение определенного времени может происходить случайное число каких-то событий (радиоактивных распадов, телефонных вызовов, отказов оборудования, несчастных случаев и т.п.).

Рассмотрим наиболее типичную ситуацию, в которой возникает распределение Пуассона. Пусть некоторые события могут происходить в случайные моменты времени, а нас интересует число появлений таких событий в промежутке времени от 0 до T. (Например, это могут быть помехи в канале связи, появления метеоритов, дорожные происшествия и т.п.) Сделаем следующие предположения.

1. Пусть вероятность появления события за малый интервал времени длины Δ примерно пропорциональна Δ, т.е. равна $a\Delta + o(\Delta)$, здесь $a > 0$ — параметр задачи, отражающий среднюю частоту событий.
2. Если в интервале времени длины Δ уже произошло одно событие, то условная вероятность появления в этом же интервале другого события стремится к 0 при $\Delta \to 0$.
3. Количества событий, происшедших на непересекающихся интервалах времени, независимы как случайные величины.

В этих условиях можно показать, что случайное число событий, происшедших за время от 0 до T, распределено по закону Пуассона с параметром $\lambda = aT$.

Определение. *Случайная величина ξ, которая принимает только целые, неотрицательные значения $0, 1, 2, \ldots$, имеет закон распределения Пуассона с параметром $\lambda > 0$, если*

$$P(\xi = k \mid \lambda) = \frac{\lambda^k}{k!} e^{-\lambda} \qquad \text{для } k = 0, 1, 2, \ldots$$

Свойства. Математическое ожидание и дисперсия случайной величины, имеющей распределение Пуассона с параметром λ, равны:

$$M\xi = \lambda, \qquad D\xi = \lambda.$$

Эти выражения несложно получить прямыми вычислениями. Имеем:

$$M\xi = \sum_{k=0}^{\infty} kP(\xi = k \mid \lambda) = \sum_{k=0}^{\infty} k \frac{\lambda^k}{k!} e^{-\lambda} = \lambda \sum_{k=1}^{\infty} \frac{\lambda^{(k-1)}}{(k-1)!} e^{-\lambda}$$

$$= \lambda \sum_{n=0}^{\infty} \frac{\lambda^n}{n!} e^{-\lambda} = \lambda.$$

Здесь была осуществлена замена $n = k - 1$ и использован тот факт, что $\sum_{n=0}^{\infty} \frac{\lambda^n}{n!} = e^{\lambda}$. Аналогично можно вычислить дисперсию случайной величины ξ.

На рис. 2.2 показаны значения вероятностей $P(\xi = k \mid \lambda)$ для различных значений k и λ.

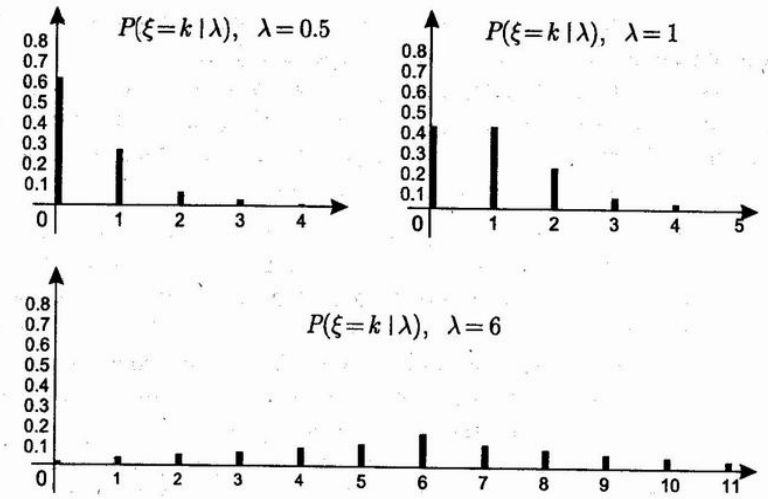


Рис. 2.2. Вид распределения Пуассона для различных значений k и λ

Связь с другими распределениями. 1. Выше уже указывалась связь между распределением Пуассона и биномиальным. Остановимся на этом вопросе более подробно.

При большом n и малом p действует приближенное соотношение:

$$C_n^k p^k (1-p)^{n-k} \simeq \frac{\lambda^k}{k!} e^{-\lambda}, \quad k = 0, 1, 2, \ldots$$

где $\lambda = np$. Этот факт можно сформулировать в виде предельного утверждения: при всяком k, $(k = 0, 1, 2, \ldots)$

$$\lim_{\substack{n\to\infty \\ p\to 0}} C_n^k p^k (1-p)^{n-k} = \frac{\lambda^k}{k!} e^{-\lambda}, \text{ если существует } \lim_{\substack{n\to\infty \\ p\to 0}} np = \lambda > 0.$$

2. При $\lambda > 9$ распределение Пуассона может быть аппроксимировано нормальным распределением со средним λ и дисперсией λ.

3. Сумма n независимых случайных величин, имеющих пуассоновские распределения с параметрами $\lambda_1, \lambda_2, \ldots, \lambda_n$ соответственно, имеет также распределение Пуассона с параметром

$$\lambda = \lambda_1 + \lambda_2 + \cdots + \lambda_n.$$

Таблицы. Таблицы распределения Пуассона при различных значениях даны, например, в [19], [65], [77], а также в других сборниках таблиц и монографиях.

Дадим описание таблиц, приведенных в [19] для $P(\xi = k \mid \lambda)$. При этом значение λ изменяется от 0.1 (0.1) 15.0, а значение k изменяется с единичным шагом в таких пределах, где $P(\xi = k \mid \lambda) > 5 \cdot 10^{-7}$. Там

же указано, как вычислять значение $P(\xi = k \mid \lambda)$ с помощью таблиц функции распределения χ^2, о которой речь пойдет ниже.

Более подробные таблицы распределения Пуассона даны в [65], где λ изменяется до 205. Отметим, что при больших значениях λ для вычисления $P(\xi = k \mid \lambda)$ можно использовать приближенную формулу

$$P(\xi = k \mid \lambda) \sim \frac{1}{\sqrt{\lambda}} \varphi\left(\frac{k-\lambda}{\sqrt{\lambda}}\right),$$

где φ — плотность нормального распределения с параметрами 0 и 1.

Наряду с таблицами для $P(\xi = k \mid \lambda)$ составлены и таблицы накопленной вероятности распределения Пуассона, т.е. таблицы для

$$P(\xi \leqslant k \mid \lambda) = \sum_{m=0}^{k} P(\xi = m \mid \lambda).$$

В [77] приведены таблицы $P(\xi \leqslant k \mid \lambda)$ для $\lambda = 0.01\ (0.01);\ 1\ (0.05);\ 5\ (0.1);\ 10\ (0.5);\ 20\ (1);\ 30\ (5);\ 50$ с точностью до $0.5 \cdot 10^{-4}$.

2.3. Показательное распределение

Область применения. Укажем две области применения статистических методов, в которых показательное распределение играет базовую роль.

К первой из них относятся задачи связанные с данными типа «времени жизни». Понимать этот термин следует достаточно широко. В медико-биологических исследованиях под ним может подразумеваться продолжительность жизни больных при клинических исследованиях, в технике — продолжительности безотказной работы устройств, в психологии — время, затраченное испытуемым на выполнение тестовых задач, и т.д. Подробное изложение обработки подобных данных дано в [55].

Второй областью активного использования показательного распределения являются задачи массового обслуживания. Здесь речь может идти об интервалах времени между вызовами «скорой помощи», телефонными звонками или обращениями клиентов и т.д. В условиях модели п. 2.2, в которой речь шла о появлении в случайные моменты неких событий и которую мы использовали для иллюстрации распределения Пуассона, длина интервала времени между появлениями последовательных событий имеет показательное распределение.

Определение. *Положительная случайная величина X имеет показательное распределение с параметром $\theta > 0$, если ее плотность задана формулой*

$$p(x, \theta) = \theta e^{-\theta x} \qquad (x \geqslant 0).$$

Показательное распределение часто называют еще экспоненциальным. Параметр θ в ряде прикладных областей именуют «отношением риска». Иногда вместо параметра θ используют параметр $b = 1/\theta$, тогда функция плотности записывается в виде:

$$p(x, b) = \frac{1}{b} e^{-x/b} \qquad (x \geqslant 0).$$

Свойства. Математическое ожидание и дисперсия случайной величины X, распределенной по показательному закону с параметром θ, равны

$$MX = 1/\theta, \qquad DX = 1/\theta^2.$$

Первое из этих соотношений придает параметру θ ясный вероятностный смысл: $1/\theta$ — это среднее время службы изделия, среднее время между вызовами и т.д.

На рис. 2.3. приведен графический вид плотности показательного распределения с параметром θ.

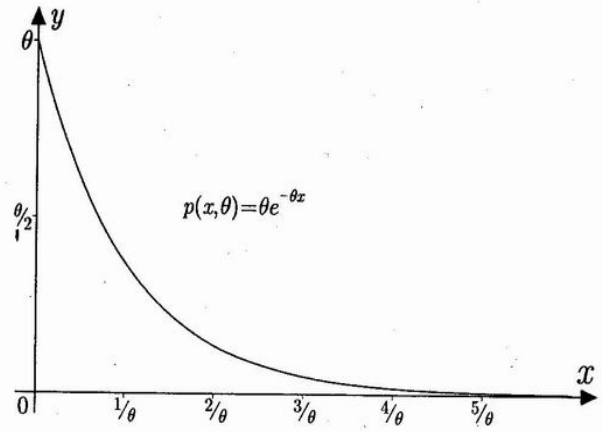


Рис. 2.3. Плотность показательного распределения с параметром θ

Функция показательного распределения, т.е. $P(X < x)$, равна

$$F(x, \theta) = \begin{cases} 1 - e^{-\theta x}, & \text{для } x \geqslant 0; \\ 0, & \text{для } x < 0. \end{cases}$$

Показательное распределение среди всех других выделяется, как иногда говорят, отсутствием «памяти», т.е. отсутствием последействия.

Это подразумевает следующее: для показательно распределенной случайной величины X (и только для такой)

$$P(X \geqslant s+t \mid X \geqslant t) = P(X \geqslant s)$$

для любых s, $t \geqslant 0$. Поясним смысл этой формулы на примере. Пусть X — время службы некоего изделия, и оно подчиняется экспоненциальному распределению. Тогда для изделия, прослужившего время t, вероятность прослужить дополнительное время s совпадает с вероятностью прослужить то же время s для нового (только начавшего работу) изделия. Как видим, это соотношение как бы исключает износ и старение. Поэтому в статистических моделях срока службы, если мы хотим учесть старение, приходится привлекать различного рода обобщения показательного распределения.

Связь с другими распределениями. Показательное распределение является частным случаем гамма-распределения, распределения Вейбулла и некоторых других. Подробную информацию на эту тему можно получить в [111].

Таблицы. Функция показательного распределения достаточно проста, поэтому специальные таблицы для этого распределения не нужны. Значения функции показательного распределения можно вычислить с помощью калькулятора.

2.4. Нормальное распределение

Область применения. Нормальное распределение относится к числу наиболее распространенных и важных, оно часто используется для приближенного описания многих случайных явлений, например, для случайного отступления фактического размера изделия от номинального, рассеяния снарядов при артиллерийской стрельбе и во многих других ситуациях, в которых на интересующий нас результат воздействует большое количество независимых случайных факторов, среди которых нет сильно выделяющихся.

Замечание. Использованию нормального распределения для приближенного описания распределений случайных величин не препятствует то обстоятельство, что эти величины обычно могут принимать значения только из какого-то ограниченного интервала (скажем, размер изделия должен быть больше нуля и меньше километра), а нормальное распределение не сосредоточено целиком ни на каком интервале. Дело в том, что вероятность больших отклонений нормальной случайной величины от центра распределения настолько мала, что ее практически можно считать равной нулю.

Определение. *Случайная величина ξ имеет нормальное распределение вероятностей с параметрами a и σ^2 (краткое обозначение: $\xi \sim N(a, \sigma^2)$), если ее плотность распределения задается формулой:*

$$\varphi(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-a)^2}{2\sigma^2}}, \qquad -\infty < x < +\infty.$$

Смысл параметров нормального распределения наглядно показан на рис. 2.4.

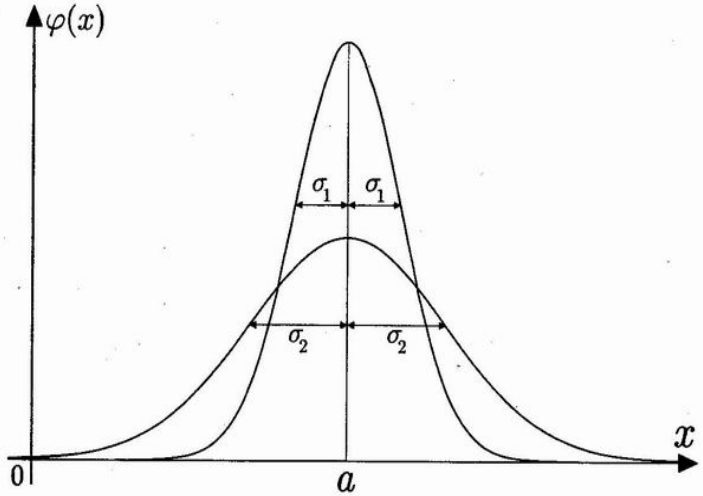


Рис. 2.4. Плотность нормального распределения со средним a и различными значениями дисперсии σ^2

Отметим, что $\varphi(x)$ стремится к нулю при $x \to -\infty$ и $x \to +\infty$. График функции $\varphi(x)$ симметричен относительно точки a. При этом в точке a функция $\varphi(x)$ достигает своего максимума, который равен $1/(\sqrt{2\pi}\sigma)$.

Параметр a характеризует положение графика функции на числовой оси (параметр положения). Параметр $\sigma (\sigma > 0)$ характеризует степень сжатия или растяжения графика плотности (параметр масштаба). Как видим, вся совокупность нормальных распределений представляет собой двухпараметрическое семейство.

Свойства. Математическое ожидание и дисперсия случайной величины ξ, распределенной как $N(a, \sigma^2)$, равны

$$M\xi = a, \qquad D\xi = \sigma^2.$$

Медиана нормального распределения равна a, так как плотность распределения симметрична относительно точки $x = a$.

Особую роль играет нормальное распределение с параметрами $a = 0$ и $\sigma = 1$, т.е. распределение $N(0,1)$, которое часто называют *стандартным* нормальным распределением. Плотность стандартного нормального распределения есть

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

Функция распределения стандартного нормального распределения равна

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-y^2/2}\, dy.$$

Функцию $\Phi(\cdot)$ часто называют функцией Лапласа. Отметим, что $\Phi(x) = 1 - \Phi(-x)$, поэтому достаточно знать значения функции $\Phi(x)$ для $x \geqslant 0$. Это свойство функции $\Phi(x)$ используется при составлении таблиц.

Функцию произвольного нормального распределения $N(a, \sigma^2)$ можно легко выразить через $\Phi(\cdot)$. Для этого следует заметить, что если ξ распределена по закону $N(a, \sigma^2)$, то её линейная функция $X = (\xi - a)/\sigma$ подчиняется стандартному нормальному распределению. Поэтому

$$P(\xi < x) = P\left(X < \frac{x-a}{\sigma}\right) = \Phi\left(\frac{x-a}{\sigma}\right).$$

Эта формула позволяет вычислять вероятности событий, связанных с произвольными нормальными случайными величинами, с помощью таблиц стандартного нормального распределения.

Аналогичным образом, легко показать, что если ξ распределена по нормальному закону, скажем, $N(a, \sigma^2)$, то случайная величина $k\xi + b$ (линейная функция ξ) имеет нормальное распределение $N(a + b, k^2\sigma^2)$.

Напомним, что площадь фигуры, ограниченная графиком функции плотности распределения, осью абсцисс и отрезками двух вертикальных прямых, $x = b$, $x = c$, есть вероятность попадания случайной величины в интервал (b, c). В связи с этим полезно представить, как распределяются доли площадей между кривой $\varphi(x)$ и осью абсцисс (см. рис. 2.5). Более подробный анализ показывает, что случайная величина $N(0,1)$ с вероятностью, примерно равной 0.94 попадает в интервал $(-2, 2)$, и с вероятностью, примерно равной 0.9973 — в интервал $(-3, 3)$. Отсюда для произвольной нормально распределенной случайной величины можно сформулировать правило, именуемое в литературе *правилом трех сигм*. А именно, нормальная случайная величина $N(a, \sigma^2)$ с вероятностью 0.9973 попадает в интервал $(a - 3\sigma, a + 3\sigma)$.

Таблицы. Для функции $\Phi(x)$ и ее производной, т.е. для плотности стандартного нормального распределения, существуют многочисленные

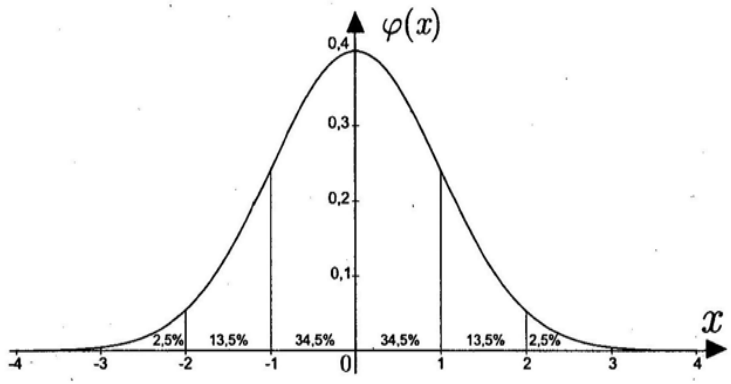


Рис. 2.5. Примерное распределение площадей под кривой функции плотности стандартного нормального распределения

таблицы разной степени подробности. Так, в [19] указаны значения $\Phi(x)$ с шестью значащими цифрами для $x = 0.000\ (0.001)\ 3.000$ и с пятью значащими цифрами для $x = 3.00\ (0.01)\ 5.00$ (в данном случае значащими называются все разряды десятичной дроби начиная с первого, отличного от девятки, например, если $\Phi(x) = 0.99976737$, то значащими цифрами считаются 76737).

Для статистических применений часто оказываются полезными таблицы, представляющие накопленную нормальную вероятность, отсчитываемую справа, т.е. таблицы, в которых в зависимости от x указаны значения $P(\xi \geqslant x) = 1 - \Phi(x)$. Например, в [115] дана таблица $P(\xi \geqslant x)$ для $x = 0.00\ (0.01)\ 3,5$ с четырьмя значащими цифрами. Как будет показано в гл. 5, таблицы подобного вида более удобны в статистической практике, чем таблицы для $\Phi(x)$.

В большинстве сборников также приводятся таблицы квантилей стандартного нормального распределения. Они позволяют по заданному значению вероятности p, $0 < p < 1$, находить точку x, такую, что $P(\xi < x) = p$. Последнее бывает часто необходимо при проверке статистических гипотез.

2.5. Двумерное нормальное распределение

Область применения. Двумерное нормальное распределение используется при описании совместного распределения двух случайных переменных (двух признаков). В этой ситуации двумерное нормальное распределение является столь же важным, как одномерное нормальное распределение для описания одного случайного признака. Обсуждение

двумерного нормального распределения начнем с обсуждения многомерных распределений вообще.

Многомерные распределения. В главе 1 мы установили, что для непрерывной одномерной случайной величины ξ ее функция плотности вероятности, скажем, $p(x)$ полностью задает распределение случайной величины: для любых чисел a, b $(a < b)$

$$P(a < \xi < b) = \int_a^b p(x)\,dx\,.$$

Аналогичным образом можно задать закон распределения случайной величины, принимающей значения не на числовой прямой, а на плоскости, в трехмерном пространстве, на сфере и т.д. Надо только иметь соответствующую функцию плотности $p(x)$. Тогда для любого множества X его вероятность $P(X)$ равна

$$P(X) = \int_X p(x)\,dx\,,$$

где интегрирование производится соответственно по области X в плоскости, трехмерном пространстве, сфере и т.д.

Двумерное нормальное распределение. В качестве примера определим двумерное нормальное распределение на плоскости. Пусть η_1 и η_2 — независимые случайные величины, имеющие стандартное нормальное распределение. Тогда двумерная случайная величина $\eta = (\eta_1, \eta_2)$ имеет *стандартное двумерное нормальное распределение*. Его плотность $p(x, y)$ равна:

$$p(x, y) = \frac{1}{2\pi}\,exp\left(-\frac{x^2+y^2}{2}\right).$$

Для одномерного случая все нормальные распределения могут быть получены как линейные преобразования стандартного нормального распределения: если $\xi \sim N(a, \sigma^2)$, то ξ можно представить в виде $\xi = a + \sigma\eta$, где случайная величина η имеет стандартное нормальное распределение. Аналогичным образом можно определить двумерные нормальные распределения — это те распределения, которые можно получить из стандартного двумерного распределения линейным преобразованием. По определению, случайная величина $\xi = (\xi_1, \xi_2)$ имеет двумерное нормальное распределение, если ее можно представить в виде

$$\begin{cases} \xi_1 = a_1 + b_1\eta_1 + c_1\eta_2 \\ \xi_2 = a_2 + b_2\eta_1 + c_2\eta_2 \end{cases}$$

где a_1, b_1, c_1, a_2, b_2, c_2 — некоторые вещественные числа. Заметим, что согласно свойствам нормальных случайных величин, компоненты двумерной нормальной случайной величины, т.е. ξ_1 и ξ_2, являются нормальными (одномерными) случайными величинами. Разумеется, случайные величины ξ_1 и ξ_2 могут быть зависимыми. Ниже мы покажем, что ξ_1 и ξ_2 зависимы тогда и только тогда, когда их ковариация (или корреляция) не равна нулю.

Аналогичным образом можно определить и многомерные нормальные распределения.

Частные (маргинальные) плотности. Если $\xi = (\xi_1, \xi_2)$ — двумерная случайная величина, то ее компоненты ξ_1 и ξ_2 — тоже случайные величины. Можно показать, что если ξ имеет плотность $p(x, y)$, то ξ_1 и ξ_2 тоже непрерывные случайные величины, имеющие плотности $p_1(x)$ и $p_2(y)$ (называемые *частными плотностями*), и эти плотности выражаются формулами:

$$p_1(x) = \int_{-\infty}^{\infty} p\,(x, y)\,dy, \qquad p_2(y) = \int_{-\infty}^{\infty} p\,(x, y)\,dx.$$

Характеристики многомерных распределений. Чаще всего в качестве характеристик многомерных распределений используются те или иные функции от компонент (координат) многомерных случайных величин, имеющих данное распределение. Например, для двумерной случайной величины $\xi = (\xi_1, \xi_2)$ мы можем рассматривать ее математическое ожидание $M\xi = (M\xi_1, M\xi_2)$ и вторые центральные моменты:

$$\sigma_{11} = D\xi_1, \qquad \sigma_{22} = D\xi_2, \qquad \sigma_{12} = \sigma_{21} = \operatorname{cov}(\xi_1, \xi_2)\,.$$

Если $\xi = (\xi_1, \xi_2)$ имеет плотность $p(x, y)$, то эти моменты, естественно, выражаются в виде интегралов от плотности. Например,

$$M\xi_1 = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} x\,p(x, y)\,dx\,dy\,.$$

Двумерная нормальная плотность. Укажем формулы для плотности двумерного нормального распределения. Пусть $\xi = (\xi_1, \xi_2)$ — двумерная нормальная случайная величина. Формула для плотности будет выглядеть проще, если мы от ξ перейдем к случайной величине $\eta = (\eta_1, \eta_2)$, где $\eta_1 = (\xi_1 - a_1)/\sqrt{\sigma_{11}}$, $\eta_2 = (\xi_2 - a_2)/\sqrt{\sigma_{11}}$, где $a_1 = M\xi_1$, $a_2 = M\xi_2$, а σ_{11} и σ_{22} были определены выше. Тогда η_1 и η_2 — случайные величины, имеющие стандартное нормальное распределение. Пусть их корреляция (она же ковариация) равна ρ. Легко видеть, что $\rho = \sigma_{12}/\sqrt{\sigma_{11}\sigma_{22}}$ — то же самое, что величина корреляции исходных случайных величин ξ_1 и ξ_2. Тогда можно показать, что функция плотности $p(x_1, x_2)$ двумерной случайной величины $\eta = (\eta_1, \eta_2)$ равна:

$$p(x_1, x_2) = \frac{1}{2\pi\sqrt{1-\rho^2}}\exp\left\{-\frac{x_1^2 - 2\rho x_1 x_2 + x_2^2}{2(1-\rho^2)}\right\}.$$

Для исходной двумерной случайной величины $\xi = (\xi_1, \xi_2)$ плотность вероятности в точке (x_1, x_2) равна

$$p(x_1, x_2) = \frac{1}{2\pi\sqrt{\sigma_{11}\sigma_{22}(1-\rho^2)}} \times \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{(x_1-a_1)^2}{\sigma_{11}} - 2\rho\frac{(x_1-a_1)(x_2-a_2)}{\sqrt{\sigma_{11}\sigma_{22}}} + \frac{(x_2-a_2)^2}{\sigma_{22}}\right]\right\}.$$

Практически это выражение используют редко.

2.6. Распределения, связанные с нормальным

Область применения. При операциях с нормальными случайными величинами, которые приходится проводить при анализе данных, возникает несколько новых видов распределений (и соответствующих им случайных величин). В первую очередь, это распределение Стьюдента, χ^2 и F-распределения. Эти распределения играют очень важную роль в прикладном и теоретическом анализе. Так, при выяснении точности и достоверности статистических оценок используются процентные точки распределений Стьюдента и хи-квадрат. Распределение статистик многих критериев, использующихся для проверки различных предположений, хорошо приближается этими распределениями.

2.6.1. Распределение хи-квадрат

Определение. Пусть случайные величины $\xi_1, \xi_2, ..., \xi_n$ — независимы, и каждая из них имеет стандартное нормальное распределение $N(0,1)$. Говорят, что случайная величина χ^2_n, определенная как:

$$\chi^2_n = \xi_1^2 + \cdots + \xi_n^2,$$

имеет распределение хи-квадрат с n степенями свободы. Для обозначения этого распределения также обычно используется выражение χ^2_n.

Ясно, что χ^2_n (для любого $n \geqslant 1$) с вероятностью 1 принимает положительные значения. Функция плотности χ^2_n в точке $x (x > 0)$ равна

$$\frac{1}{2^{n/2}} \frac{1}{\Gamma(n/2)} x^{n/2-1} e^{x/2}.$$

где $\Gamma(\cdot)$ есть гамма-функция. На практике эта плотность распределения непосредственно используется редко.

Заметим, что показательное распределение с параметром $\theta = 1/2$ из параграфа 2.3 — это распределение χ^2 с двумя степенями свободы.

На рис. 2.6 изображены функции плотности распределения хи-квадрат с различным числом степеней свободы.

Свойства. Нетрудно убедиться, что математическое ожидание и дисперсия случайной величины χ^2_n равны:

$$M\chi^2_n = n, \qquad D\chi^2_n = 2n.$$

Таблицы. Для случайной величины χ^2_n составлены разнообразные таблицы (см. [19], [65], [77]). Чаще всего они содержат значения p-квантилей случайных величин χ^2_n, $n = 1, 2, ..., m$ (если вероятность

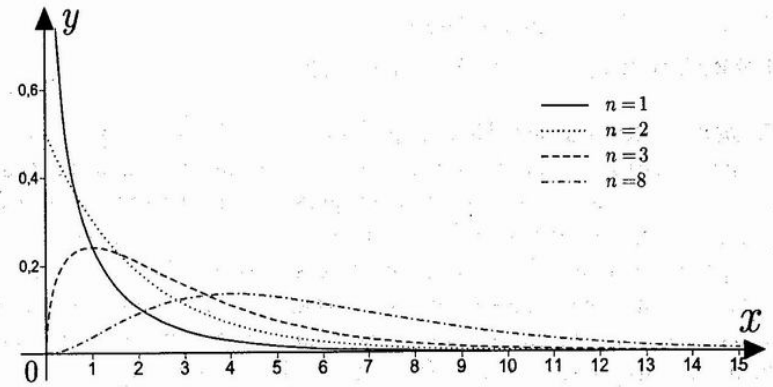


Рис. 2.6. Функции плотности распределения хи-квадрат с различным числом степеней свободы n

выражена в процентах, их называют процентными точками и, соответственно, говорят о таблицах процентных точек). Аргумент p, $0 < p < 1$, при этом пробегает тот или иной набор значений.

2.6.2. Распределение Стьюдента

Определение. Пусть случайные величины $\xi_0, \xi_1, \ldots, \xi_n$ — независимы, и каждая из них имеет стандартное нормальное распределение $N(0,1)$. Введем случайную величину

$$t_n = \frac{\xi_0}{\sqrt{\frac{1}{n}\sum_{i=1}^n \xi_i^2}}.$$

Ее распределение называют распределением Стьюдента. Саму случайную величину часто называют стьюдентовской дробью, стьюдентовым отношением и т.п. Число n, $n = 1, 2, ...$ называют числом степеней свободы распределения Стьюдента.

Плотность распределения Стьюдента в точке x равна

$$\frac{\Gamma((n+1)/2)}{\Gamma(n/2)\sqrt{\pi n}} \left(1 + \frac{x^2}{n}\right)^{-(n+1)/2}.$$

Из определения видно, что плотность симметрична относительно $x = 0$. Это обстоятельство используют при составлении таблиц.

На рис. 2.7 изображены функции плотности распределения Стьюдента с различным числом степеней свободы.

Свойства. Можно показать, что:

$$Mt_n = 0, \qquad Dt_n = \frac{n}{n-2}.$$

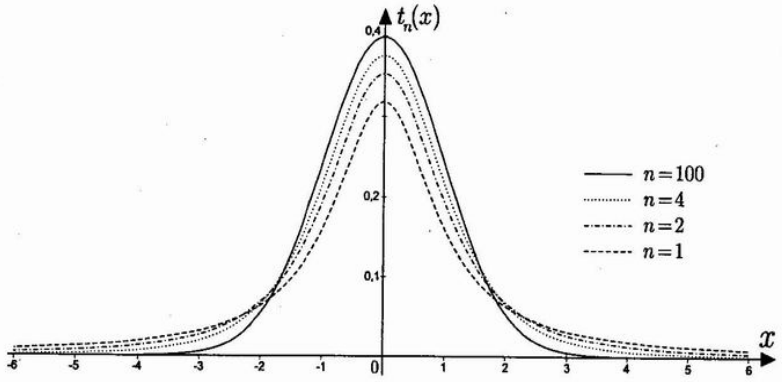


Рис. 2.7. Функции плотности распределения Стьюдента с различным числом степеней свободы n

Таблицы. В сборниках обычно приводятся таблицы процентных точек для последовательных $n = 1, 2, \ldots$ вплоть до некоторого значения. При больших n обычно рекомендуют использовать таблицы стандартного нормального распределения, иногда с поправками.

2.6.3. F-распределение

Определение. Пусть $\eta_1, \ldots, \eta_m$; $\xi_1, \ldots, \xi_n$ (где m, n — натуральные числа) обозначают независимые случайные величины, каждая из которых распределена по стандартному нормальному закону $N(0,1)$. Говорят, что случайная величина $F_{m,n}$, определенная как

$$F_{m,n} = \frac{\frac{1}{m}\left(\eta_1^2 + \cdots + \eta_m^2\right)}{\frac{1}{n}\left(\xi_1^2 + \cdots + \xi_n^2\right)},$$

имеет F-распределение с параметрами m и n. Натуральные числа m, n называют числами степеней свободы. F-распределение иногда называют еще распределением дисперсионного отношения (смысл этого названия станет ясен в гл. 6).

Плотность $F_{m,n}$ выражается довольно сложной формулой, которая редко непосредственно используется на практике, поэтому мы ее приводить не будем.

На рис. 2.8 изображены функции плотности F-распределения с различным числом степеней свободы.

Свойства. Математическое ожидание и дисперсия случайной величины $F_{m,n}$ равны:

$$MF_{m,n} = \frac{n}{n-2} \quad \text{для } n > 2, \quad DF_{m,n} = \frac{2n^2(m+n-2)}{m(n-2)^2(n-4)} \quad \text{для } n > 4.$$

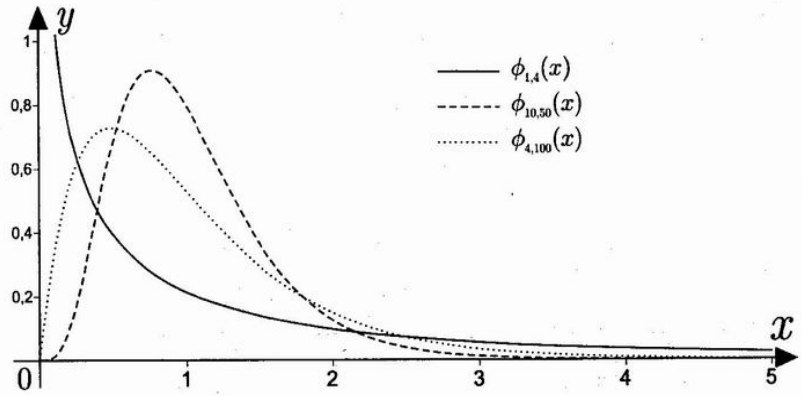


Рис. 2.8. Функции плотности F -распределения с различным числом степеней свободы

Таблицы. Семейство F-распределений зависит от двух натуральных параметров m и n, в связи с чем даже таблицы процентных точек занимают большой объем. Ради экономии места они часто публикуются в сжатом виде, поэтому при их практическом использовании приходится прибегать к дополнительным вычислениям и интерполяции.

2.7. Законы распределения вероятностей в пакетах STADIA и SPSS

Статистические пакеты могут предоставлять обширную справочную информацию по различным семействам вероятностных распределений, наглядно иллюстрируя их свойства и заменяя статистические таблицы.

2.7.1. Пакет STADIA

Пакет предоставляет возможность работать с пятью дискретными и восемью непрерывными распределениями вероятностей, приведенными ниже на рис. 2.9. Доступ к ним осуществляется из раздела **Распределения и частоты** меню блока статистических методов (см. рис. 1.17). Разберем несколько примеров.

Пример 2.1к. Построим графики плотности распределения вероятностей нормального распределения с параметрами $a = 0, \sigma^2 = 1$; $a = 0, \sigma^2 = 4$; $a = 2, \sigma^2 = 1$.

Выбор процедуры. В меню блока **Статистические методы** (рис. 1.17) щелкнем мышью кнопку [T = Вычисление вероятностей] или нажмем клавишу [T]. На экране появится меню выбора **Функция вероятности распределения** (рис. 2.9), в котором нужно выбрать пункт **6=нормальное**.

Глава 11

Временные ряды: теоретические основы

11.1. Введение

Что такое временной ряд. Временной ряд — это последовательность чисел; его элементы — это значения некоторого протекающего во времени процесса. Они измерены в последовательные моменты времени, обычно через равные промежутки.

Как правило, составляющие временной ряд числа — *элементы временного ряда*, — нумеруют в соответствии с номером момента времени, к которому они относятся (например, x_1, x_2, x_3 и т.д.). Таким образом, порядок следования элементов временного ряда весьма существен.

Почти в каждой области знания встречаются явления, которые важно изучать в развитии во времени или пространстве. И почти всегда в закономерное течение явления вмешивается случай в виде случайных импульсов, случайных помех, случайных ошибок и т.д. Поэтому изучение временных рядов — это составная часть прикладной статистики (и довольно важная ее часть).

Расширения понятия временного ряда. Понятие временного ряда часто толкуют расширительно. Например, одновременно могут регистрироваться несколько характеристик упомянутого процесса. В этом случае говорят о *многомерных временных рядах*. Если измерения производятся непрерывно, говорят о временных рядах с непрерывным временем, или *случайных процессах*. Наконец, текущая переменная может иметь не временной, а какой-нибудь иной характер, например пространственный (тогда говорят о *случайных полях*).

Примеры временных рядов. Данные типа временных рядов широко распространены в самых различных областях человеческой деятельности. В экономике это ежедневные цены на акции, курсы валют, еженедельные и месячные объемы продаж, годовые объемы производства и т.п. В метеорологии типичными временными рядами являются ежедневная температура, месячные объемы осадков, в гидрологии — периодически измеряемые уровни воды в реках. В технике временные ряды возникают в результате отслеживания различных параметров технологических процессов.

На рис. 11.1 приведены примеры различных временных рядов (для наглядности последовательные измерения, составляющие временной ряд, на графиках соединены линиями).

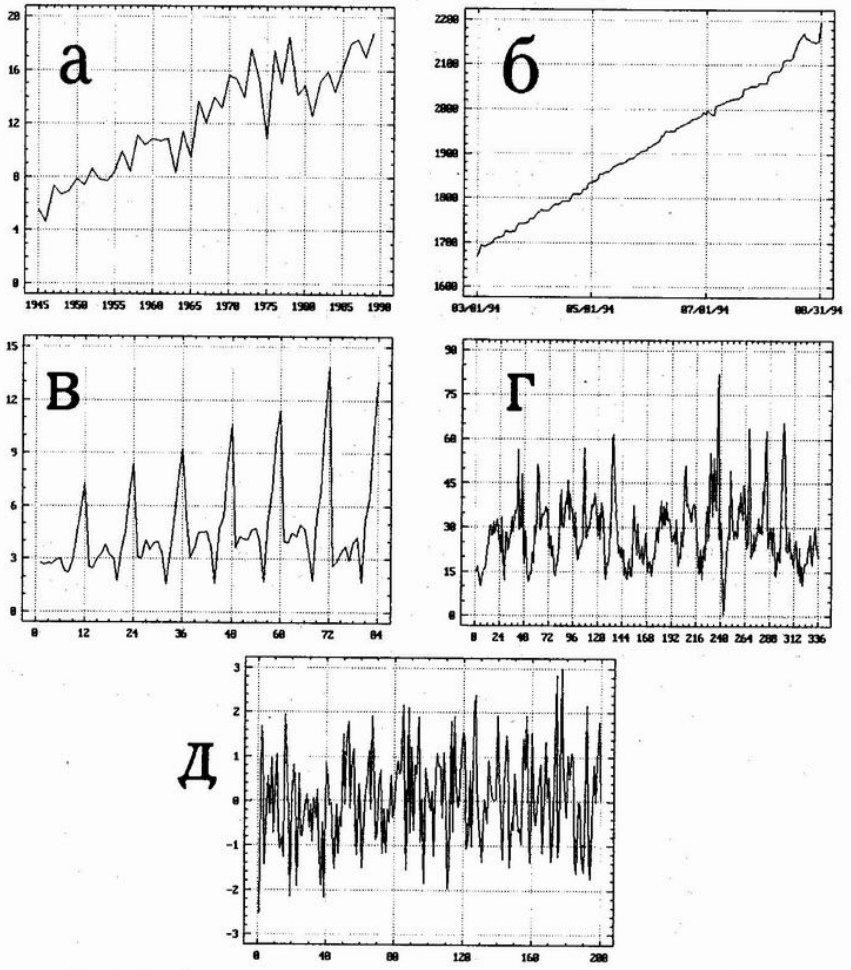


Рис. 11.1. а) – урожайность зерновых культур в СССР с 1945 по 1989 гг. в ц/га; б) – курс доллара на торгах ММВБ с 1.03 по 31.08 1994 г. в рублях; в) – ежемесячные продажи шампанского за 7 последовательных лет; г) – среднечасовая нагрузка телекоммуникационного канала Москва – Париж в течение 2-х недель (в Кбит/сек); д) – гауссовский «белый шум» с параметрами 0 и 1

Видно, что поведение временных рядов может быть весьма различным. Так, динамика урожайности зерновых в СССР (ряд а)) имеет

скорее всего линейный тренд, отклонения от которого можно считать независимыми случайными величинами. Курс доллара на торгах ММВБ весной-летом 1994 г. (ряд б)) также содержит линейный тренд, однако отклонения от него имеют более сложную статистическую структуру, чем в предыдущем случае. График ежемесячных продаж шампанского (ряд в)) содержит явно повторяющиеся годовые циклы с возрастающей амплитудой. Среднечасовая загрузка телекоммуникационного канала Москва-Париж в течение двух недель (одна неделя равна 168 часам) февраля 1996 г. (ряд г)) имеет ясные суточные циклы. Кроме суточных, этот ряд содержит и недельные циклы, но на приведенном графике они заметны мало, так как недостаточно длителен интервал наблюдения. Ряд д) создан датчиком нормальных случайных чисел на компьютере и служит примером чисто случайного процесса без внутренних закономерностей и зависимостей.

Измерение значений временного ряда. Чаще всего значения временного ряда получаются непосредственной записью значений некоторого процесса через определенные промежутки времени. Например, если ежесуточно в определенное время записывать показания термометра, то получится временной ряд со значениями температуры в том месте, в котором находится термометр.

Иногда значения элементов временного ряда получаются накапливанием некоторых данных за определенный промежуток времени (например, суммарное число посетителей магазина за день), усреднением (средняя температура за день) и т.д.

11.2. Анализ временных рядов и его разделы

Анализ временных рядов. Исследование временных рядов отличается от других задач анализа данных как кругом представляющих интерес вопросов, так и методами, применяемыми для исследования. Поэтому наука об исследовании временных рядов — *анализ временных рядов*, — образует самостоятельную и весьма обширную область статистики.

Разделы анализа временных рядов. Временные ряды, возникающие в различных предметных областях, имеют различную природу, поэтому для их изучения оказались эффективными разные методы. Исследователи придумывали и развивали многочисленные методы анализа, подходящие для изучения временных рядов в разных предметных областях. И в результате анализ временных рядов превратился в довольно

разветвленную науку. Вот только некоторые из видов временных рядов, исследование которых можно рассматривать как самостоятельный раздел теории анализа временных рядов:

- *стационарные случайные процессы*, то есть последовательности случайных величин, вероятностные свойства которых не изменяются во времени. Стационарные случайные процессы широко применяются в радиотехнике, теории связи, механике жидкости и газа, океанологии, метеорологии и т.д.;
- *диффузионные процессы* возникли при изучении процесса диффузии, то есть взаимопроникновения различных жидкостей или газов. Эти процессы используются при построении моделей непрерывных процессов, в которых существенна случайная составляющая;
- *точечные процессы* используют для описания таких явлений, как поступление вызовов или заявок на обслуживание, моментов несчастных случаев, стихийных и техногенных катастроф, каких-либо приметных явлений и т.п. Они широко применяются в таких разделах статистики, как теория очередей, теория массового обслуживания и т.д.

Всех этих обширных разделов анализа временных рядов в данной книге мы касаться не будем. Вместо этого мы хотим рассказать о тех прикладных аспектах анализа временных рядов, которые полезны и важны при решении практических задач в экономике, финансах, а также в различных гуманитарных науках. В частности, мы расскажем о методах подбора математической модели для описания временного ряда, об изучении взаимозависимостей временных рядов, выявления в них периодических и других составляющих, прогнозировании поведения временных рядов и т.д.

Как мы будем рассказывать об анализе временных рядов. В этой главе (главе 11) мы обсудим основные понятия статистической теории временных рядов. Мы расскажем о структуре временных рядов; о вероятностных предпосылках для их анализа; о детерминированных компонентах временных рядов и о причинах, их порождающих; о корреляционной структуре ряда; о случайной составляющей временного ряда и ее описании и т.д. В главе 12 обсуждаются вопросы прикладного анализа временных рядов, а в главе 13 мы опишем, как различные практические задачи анализа временных рядов решаются с помощью статистических пакетов Эвриста и SPSS. Наконец, в главе 14 мы вновь возвратимся к теории и расскажем о некоторых математических моделях временных рядов, важных для прикладного анализа.

Мы будем стараться вести изложение на уровне, доступном широкому читателю. Наш рассказ мы сопроводим обсуждением примеров (как тех, что привели на рис. 11.1, так и многих других).

Литература. Те читатели, которые захотят глубже изучить теоретические или прикладные аспекты анализа временных рядов, могут обратиться к литературе. По большинству разделов анализа временных рядов существует множество книг разной степени подробности и математической сложности. Одни из этих книг рассчитаны на математиков, другие — на практических работников и инженеров. Укажем некоторые наиболее известные из этих книг:

- в области теории временных рядов — [6], [97], [51], [123], [127], [25];
- в области общего прикладного анализа — [13], [14], [17], [52], [87], [11];
- в области прикладного спектрального анализа — [37], [75];
- в области радиотехнических приложений — [62], [31];
- в области экономических приложений — [130], [67] и др.

11.3. Цели, этапы и методы анализа временных рядов

Цели анализа временных рядов. При практическом изучении временных рядов исследователь на основании наблюденного отрезка временного ряда (конечной длины) должен сделать выводы о свойствах этого ряда и о вероятностном механизме, порождающем этот ряд. Чаще всего при изучении временных рядов ставятся следующие цели:

- краткое (сжатое) описание характерных особенностей ряда;
- подбор статистической модели (моделей), описывающей временной ряд;
- предсказание будущих значений на основе прошлых наблюдений;
- управление процессом, порождающим временной ряд.

На практике эти и подобные цели достижимы далеко не всегда и далеко не в полной мере. Часто этому препятствует недостаточный объем наблюдений (недостаточная длительность); еще чаще — изменяющаяся с течением времени статистическая структура временного ряда. Из-за этих изменений значение прошлых наблюдений обесценивается, и они уже не помогают предвидеть будущее.

Стадии анализа временных рядов. Обычно при практическом анализе временных рядов последовательно проходят следующие этапы:

- графическое представление и описание поведения временного ряда;
- выделение и удаление закономерных составляющих временного ряда, зависящих от времени: тренда, сезонных и циклических составляющих;
- выделение и удаление низко- или высокочастотных составляющих процесса (фильтрация);
- исследование случайной составляющей временного ряда, оставшейся после удаления перечисленных выше составляющих;
- построение (подбор) математической модели для описания случайной составляющей и проверка ее адекватности;
- прогнозирование будущего развития процесса, представленного временным рядом;
- исследование взаимодействий между различными временными рядами.

Методы анализа временных рядов. Для решения указанных выше (а также многих других) задач исследователями предложено большое количество различных методов. Отметим из них наиболее распространенные:

- *корреляционный анализ* позволяет выявить существенные периодические зависимости и их *лаги* (задержки) внутри одного процесса (автокорреляция) или между несколькими процессами (кросскорреляция);
- *спектральный анализ* позволяет находить периодические и квазипериодические составляющие временного ряда;
- *сглаживание и фильтрация* предназначены для преобразования временных рядов с целью удаления из них высокочастотных или сезонных колебаний;
- модели *авторегрессии и скользящего среднего* оказываются особенно полезными для описания и прогнозирования процессов, проявляющих однородные колебания вокруг среднего значения;
- *прогнозирование* позволяет на основе подобранной модели поведения временного ряда предсказывать его значения в будущем.

Как уже говорилось, в этой книге мы расскажем не обо всех из этих методов, а лишь о тех, которые наиболее важны для экономических и гуманитарных наук.

11.4. Детерминированная и случайная составляющие временного ряда

Следуя основной идее статистики, при анализе временного ряда видимую его изменчивость стараются разделить на закономерную и случайную составляющие. Закономерные изменения членов временного ряда следуют какому-то определенному правилу и поэтому предсказуемы. Эта составляющая x_t может быть вычислена при каждом t как некоторая функция от текущего момента t. Эта функция может зависеть, помимо t, также от некоторого набора параметров. Когда эти параметры неизвестны, их приходится оценивать по имеющимся наблюдениям — как, например, бывает в случае регрессии.

Изменчивость, оставшаяся необъясненной, иррегулярна и хаотична. Для ее описания необходим статистический подход (за неимением лучшего).

Определение. *Под закономерной (детерминированной) составляющей временного ряда $x_1, \ldots, x_n$ мы будем понимать числовую последовательность $d_1, \ldots, d_n$, элементы которой d_t вычисляются по определенному правилу как функция времени t.*

Детерминированная составляющая часто отражает действия каких-либо определенных факторов или причин. Так, у временных рядов из различных областей техники детерминированная составляющая обычно обязана своим возникновением действиям физических законов или условиям эксплуатации оборудования. Например, если значения временного ряда соответствуют положениям маятника в определенные моменты времени, то в качестве детерминированной компоненты ряда можно взять решение дифференциального уравнения движения маятника в эти моменты времени. В экономических и многих других приложениях математические модели изучаемых процессов нам обычно не известны, так что тенденции, отраженные в поведении временного ряда, нам приходится выявлять по наблюдаемым значениям временного ряда. Например, при изучении данных о месячном производстве молока в России (рис. 11.2) мы можем пытаться описать закономерную часть данного временного ряда в виде комбинации линейной функции (в течение последних лет производство молока, нашедшее отражение в статистической отчетности, постепенно уменьшалось) и периодической функции с периодом 12 месяцев. Эта периодическая компонента отражает влияние времени года на производство молока.

Для многих рядов в экономике и социальных науках причины, порождающие их закономерные составляющие, могут не быть столь ясными.

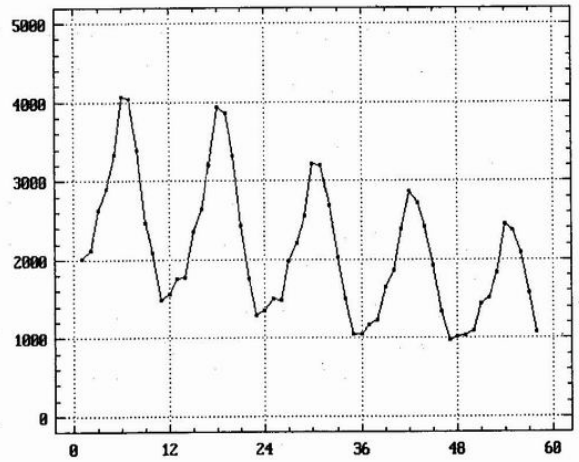


Рис. 11.2. Ежемесячное производство молока в России с 01.1992 по 10.1996 (в тыс. тонн)

Тем не менее, их совокупное влияние может быть устойчивым в течение достаточно длительных промежутков времени. Это обеспечивает возможность прогноза для подобных временных рядов. Если мы полностью выявим закономерную составляющую в поведении временного ряда, то оставшаяся часть выглядит хаотично и непредсказуемо. Ее обычно именуют иррегулярной, или *случайной компонентой* временного ряда. Обозначим эту случайную компоненту через $\varepsilon_1, \ldots, \varepsilon_t, \ldots, \varepsilon_n$.

Для описания и анализа случайной компоненты временных рядов обычно используют понятия и методы теории вероятностей и математической статистики.

Аддитивная и мультипликативная модели. Формы разложения (декомпозиции) временного ряда на детерминированную и случайную компоненты могут различаться. Укажем наиболее простых из них.

Определение. *Аддитивной моделью временного ряда называется представление ряда в виде суммы детерминированной и случайной компонент, а именно:*

$$x_t = d_t + \varepsilon_t \quad \textit{при} \quad t = 1, \ldots, n \quad \textit{или} \quad X = D + E. \qquad (11.1)$$

Определение. *Мультипликативной моделью временного ряда называется представление ряда в виде произведения детерминированной и случайных компонент, а именно:*

$$x_t = d_t \times \varepsilon_t \quad \textit{при} \quad t = 1, \ldots, n \quad \textit{или} \quad X = D \times E$$

Мультипликативные модели часто бывают удобны при анализе экономических временных рядов.

Если в приведенном выше соотношении перейти к логарифмам, то мы вновь получим формулу (11.1) — но не для самих x_t, а для их логарифмов.

$$\log(x_t) = \log(d_t) + \log(\varepsilon_t) \qquad \text{при} \qquad t = 1, \ldots, n$$

Указанное соотношение объясняет распространенность логарифмических шкал при анализе экономических временных рядов.

11.5. Тренд, сезонная и циклическая компоненты

Способы описания детерминированных компонент временного ряда сильно зависят от области приложений. При выборе модели детерминированной компоненты должны прежде всего учитываться содержательные соображения, то есть те объективные факторы и закономерности, которые приводят к ее формированию.

В экономических (и многих других) приложениях в детерминированной компоненте временного ряда d_t обычно выделяют три составляющих части: тренд tr_t, сезонную компоненту s_t и циклическую компоненту c_t. Для простоты изложения мы рассмотрим только аддитивную модель временного ряда. Мы можем записать:

$$d_t = tr_t + s_t + c_t, \qquad \text{при} \quad t = 1, \ldots, n.$$

Замечание. В последнее время к указанным трем компонентам все чаще добавляют еще одну компоненту, именуемую интервенцией. Под интервенцией понимают существенное кратковременное воздействие на временной ряд. Примером интервенции могут служить события «черного вторника», когда курс доллара за день вырос почти на тысячу рублей. С анализом интервенций можно познакомиться в [11].

Термины «тренд», «сезонная компонента» и «циклическая компонента» не имеют однозначных общепринятых определений. Чаще всего расхождения относятся к определению тренда и циклической компоненты и связаны с различными традициями в разных науках. Мы определим их в виде, наиболее часто используемом в экономических приложениях.

Тренд. Анализ временного ряда обычно начинается с выделения именно этой компоненты. Ее присутствие или отсутствие наглядно показывает график временного ряда. Выделение тренда позволяет перейти к дальнейшей идентификации других компонент ряда.

Определение. *Трендом временного ряда tr_t при $t = 1, \ldots, n$ называют плавно изменяющуюся, не циклическую компоненту, описывающую чистое влияние долговременных факторов, эффект которых сказывается постепенно.*

В экономике к таким факторам можно отнести:

- изменение демографических характеристик популяции, включая рост населения, изменение структуры возрастного состава, изменение географического расселения и т.д.;
- технологическое и экономическое развитие;
- рост потребления и изменение его структуры.

Действие этих и им подобных факторов происходит постепенно, поэтому их вклад исследователи предпочитают описывать с помощью гладких кривых, просто задающихся в аналитическом виде. Мы опишем некоторые модели тренда в следующем параграфе.

Сезонная компонента. Сезонная компонента отражает присущую миру и человеческой деятельности повторяемость процессов во времени. Она часто присутствует в экономических, метеорологических и других временных рядах. Сезонная компонента чаще всего служит главным источником краткосрочных колебаний временного ряда, так что ее выделение заметно снижает вариацию остаточных компонент.

Определение. *Сезонная компонента s_t временного ряда при $t = 1, \ldots, n$ описывает поведение, изменяющееся регулярно в течение заданного периода (года, месяца, недели, дня и т.п.). Она состоит из последовательности почти повторяющихся циклов.*

Типичным примером сезонного эффекта является объем продаж в декабре каждого года в преддверии Рождества и нового года. А пик объема продаж товаров для школьников приходится на начало нового учебного года. Объем перевозок пассажиров городским транспортом имеет два пика — утром и вечером, причем период вечернего пика продолжительней, а сам пик менее высокий. Сезонные эффекты присущи многим сферам человеческой активности: многие виды продукции имеют сезонный характер производства, потребление товаров также имеет ярко выраженную сезонность. На графике месячных объемов продаж шампанского в течение 7 лет (рис. 11.1в) видно, что пик реализации приходится на декабрь, а спад на жаркие летние месяцы.

В некоторых временных рядах сезонная компонента может иметь плавающий или изменяющийся характер. Классическим примером подобного эффекта является праздник Пасхи, сроки которого изменяются из года в год. Поэтому локальный пик объемов междугородных перевозок во время пасхальных каникул является плавающим сезонным эффектом.

Главная идея подхода к анализу сезонных компонент заключается в переходе от сравнения всех значений временного ряда между собой к сравнению значений через определенный период времени. Это позволяет заметно снизить оценку вариации временного ряда около своего среднего значения. Так, при изучении динамики месячных объемов продаж за несколько лет данные декабря одного года обычно сравнивают с данными декабря предыдущего года, а не с данными других месяцев рассматриваемого года. Методы анализа сезонных эффектов и выделения сезонных компонент рассмотрены в пункте 12.3.2.

Циклическая компонента занимает как бы промежуточное положение между закономерной и случайной составляющими временного ряда. Если тренд — это плавные изменения, проявляющиеся на больших временных промежутках, если сезонная компонента — это периодическая функция времени, ясно видимая, когда ее период много меньше общего времени наблюдений, то под циклической компонентой обычно подразумевают изменения временного ряда, достаточно плавные и заметные для того, чтобы не включать их в случайную составляющую, но такие, которые нельзя отнести ни к тренду, ни к периодической компоненте.

Определение. *Циклическая компонента c_t временного ряда описывает длительные периоды относительного подъема и спада. Она состоит из циклов, которые меняются по амплитуде и протяженности.*

Изучение циклической компоненты полезно для прогнозирования (особенно краткосрочного).

Замечание. Выделение в экономических временных рядах циклических компонент связано с тем, что экономическая активность не растет (или спадает) постоянными темпами. Она состоит из периодов относительных подъемов и спадов. Считается, что причиной циклических изменений в экономических показателях является взаимодействие спроса и предложения. Играют роль и другие факторы: рост и истощение ресурсов, увеличение размеров капитала, используемого в бизнесе, продолжительно действующие неблагоприятные (либо благоприятные) для тех или иных отраслей сельского хозяйства погодные условия, изменения в правительственной финансовой и налоговой политике и т.п. Влияние всех этих факторов приводит к тому, что циклическую компоненту крайне трудно идентифицировать формальными методами, исходя только из данных изучаемого ряда. Поэтому для ее анализа обычно приходится привлекать дополнительную информацию в виде других временных рядов, которые оказывают влияние на изучаемый ряд, например, учитывать информацию типа налоговых льгот, перенасыщенности рынка и т.п.

Методы определения циклической компоненты в экономических временных рядах и связанные с ней индексы деловой активности относятся к эконометрии и выходят за рамки материала, рассматриваемого в данной книге. Более подробно об этом можно прочесть в [135].

11.6. Модели тренда

Простейшие модели тренда. Приведем модели трендов, наиболее часто используемые при анализе экономических временных рядов, а также во многих других областях. Во-первых, это простая линейная модель

$$tr_t = b_0 + b_1 \cdot t, \tag{11.2}$$

которая, несмотря на свою простоту, оказывается полезной во многих реальных задачах. Если нелинейный характер тренда очевиден, то может подойти одна из следующих моделей:

- полиномиальная: $tr_t = b_0 + b_1 t + b_2 t^2 + \cdots + b_n t^n$, где значение степени полинома n в практических задачах редко превышает 5;
- логарифмическая: $tr_t = \exp(b_0 + b_1 t)$. Эта модель чаще всего применяется для данных, имеющих тенденцию сохранять постоянные темпы прироста;
- логистическая: $tr_t = \frac{a}{1+b \cdot e^{-ct}}$;
- Гомперца: $\log(tr_t) = a - b \cdot r^t$, где $0 < r < 1$.

Две последние модели задают кривые тренда S-образной формы. Они соответствуют процессам с постепенно возрастающими темпами роста в начальной стадии и постепенно затухающими темпами роста в конце. Необходимость подобных моделей обусловлена невозможностью многих экономических процессов продолжительное время развиваться с постоянными темпами роста или по полиномиальным моделям, в связи с их довольно быстром ростом (или уменьшением).

Первое представление о возможном характере тренда дает графическое представление временного ряда. Так, график роста урожайности зерновых культур (рис. 11.1а) позволяет предположить наличие линейного тренда в этом временном ряде. Аналогичное предположение очевидно справедливо и для ряда роста курса доллара весной и летом 1994 г. (рис. 11.1б).

При прогнозировании тренд используют в первую очередь для долговременных прогнозов. Точность краткосрочных прогнозов, основанных только на подобранной кривой тренда, как правило, недостаточна. Методы выделения и удаления тренда подробно рассматриваются в пункте 12.3.1, а также в главе 8.

О временных рядах в технических приложениях. В технических приложениях мы часто знаем физические законы или технические ха-

рактеристики механизмов, генерирующих исследуемые временные ряды. Это, разумеется, существенно облегчает исследование. Мы рассмотрим только один тип моделей временных рядов, часто используемый в технических приложениях — *полигармоническую модель*. О других моделях временных рядов, возникающих в технических приложениях, Вы можете узнать в [62], [31].

Полигармоническая модель. Простейший вариант полигармонической модели временного ряда — это косинусоидальная модель:

$$x_t = a\cos(\omega t + \theta) + \varepsilon_t \qquad (11.3)$$

Здесь детерминированная компонента представляет собой косинусоидальную функцию с амплитудой a, частотой ω, периодом $2\pi/\omega$ и фазой θ. Величины a, ω и θ в выражении (11.3) являются константами.

Комментарий. В технических приложениях часто рассматриваются модели типа (11.3), в которых амплитуда a является случайной величиной с нулевым средним или фаза θ является случайной равномерно распределенной величиной в интервале $(0, 2\pi)$. Такой подход, превращающий процесс (11.3) и ему подобные в стационарные процессы, часто обусловлен необходимостью обоснования возможности применения методов исследования стационарных процессов к процессам типа (11.3).

Круг данных, описываемых чисто косинусоидальной моделью (11.3), невелик. Во-первых, часто встречаются периодические зависимости, которые описываются не косинусоидальной, а более сложной функцией. Во-вторых, обычно в изучаемом процессе можно выделить не одну, а несколько периодических компонент с разными периодами.

Как известно из математического анализа, любую гладкую периодическую функцию $G(t)$ с периодом p (то есть функцию, для которой $G(t + kp) = G(t)$ для любого целого k) можно представить в виде ряда Фурье:

$$G(t) = \sum_{j=1}^{p} a_j \cos(j\omega t + \theta_j), \qquad (11.4)$$

где $\omega = 2\pi/p$ называется основной (Найквистовой) частотой, a_j, θ_j — некоторые параметры. Частоты $j\omega$ называются гармониками основной частоты.

Функцию, являющуюся суммой нескольких периодических функций с разными периодами, можно задать в виде $G(t) = \sum_k G_k(t) = \sum_{j,k} a_{jk}\cos(j\omega_k t + \theta_{jk})$. Таким образом, получаем следующее обобщение модели (11.3).

Определение. *Говорят, что временной ряд описывается полигармонической моделью, если он представлен в виде:*

$$x_t = \sum_{j,k} a_{jk}\cos(j\omega_k t + \theta_{jk}) + \varepsilon_t \qquad (11.5)$$

где $\omega_k = 2\pi/p_k$, а ε_t является белым шумом (см. п. 11.7).

Пример ряда, описываемого полигармонической моделью, приведен на рис. 11.3.

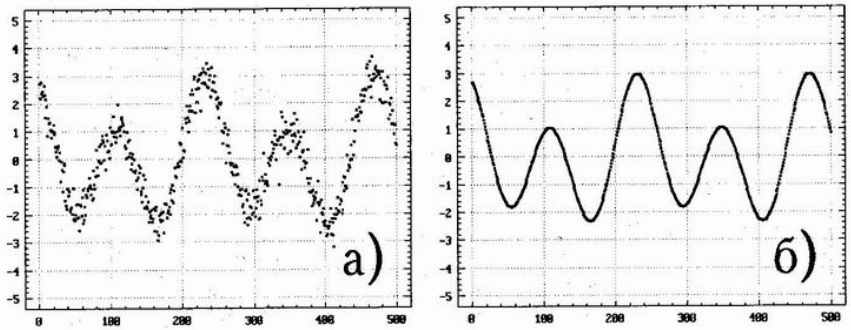


Рис. 11.3. а) — 500 значений ряда, описываемого полигармонической моделью $x_t = 2\cos(\frac{\pi}{60}t + \frac{\pi}{6}) + \cos(\frac{\pi}{120}t) + \varepsilon_t$, где ε_t — белый шум с дисперсией 0.16; б) — детерминированная компонента этого ряда

Если периоды p_k известны, то для определения величин a_{jk} и θ_{jk} можно использовать методы линейного регрессионного анализа. Если периоды p_k не известны, для их определения используют методы спектрального анализа. Мы практически не будем затрагивать этот вопрос в настоящей главе, так как он требует достаточно высокой математической подготовки читателей, и ограничимся лишь кратком рассказом об одном из его простейших случаев — периодограмме (см. п. 11.10). Наиболее полный обзор современного состояния методов прикладного спектрального анализа на русском языке дан в [75]. Этим методам посвящено много различной специальной литературы: [37], [52], [87], [51], [123], [62], [31].

Исторически анализ временных рядов из различных областей деятельности, включая экономику, начинался в конце XIX и начале XX века именно с подбора полигармонических моделей для их описания. Однако с середины XX века стали появляться более простые модели и методы анализа временных рядов, включая линейные параметрические модели типа авторегрессии-скользящего среднего, на которых и будет в основном сосредоточено наше внимание.

11.7. Модели случайной компоненты

Прежде чем перейти к вопросам практического анализа временных рядов, кратко остановимся на математических основаниях этого анализа. При первом чтении этот параграф можно пропустить (тогда к нему время от времени придется возвращаться впоследствии).

Случайные процессы. Практический опыт показывает, что обычно временной ряд не удается полностью описать одной лишь детерминированной компонентой. В нем, как правило, присутствует и нерегулярная, случайная компонента. Ее поведение нельзя точно предсказать заранее. Для ее описания приходится привлекать понятия из теории вероятностей.

Для описания нерегулярной компоненты и всего временного ряда в целом используют понятие *случайного (стохастического) процесса* или случайной последовательности (как процесса от целочисленного аргумента). Ниже будут приведены некоторые сведения из теории случайных процессов, необходимые для понимания процедур прикладного анализа временных рядов. Более подробное изложение математической теории случайных процессов можно найти в [25], [97], [51], [123].

Определение. *Случайным процессом $X(t)$, заданном на множестве T, называют функцию от t, значения которой при каждом $t \in T$ является случайной величиной.*

Выделяются случайные процессы с непрерывным временем (когда T — интервал на числовой оси, например) и с дискретным временем (когда T — натуральный ряд или его часть, например). Последние чаще называют случайными последовательностями.

Если T — конечное множество, то случайный процесс — это просто совокупность случайных величин. Для статистического описания такой совокупности надо указать распределение вероятностей в конечномерном пространстве. Для этого можно использовать многомерную функцию распределения или плотности, если распределение непрерывное.

Если T — бесконечное множество, то для описания бесконечной совокупности случайных величин (которые в этом случае и составляют случайный процесс) применяется следующая конструкция.

Определение. *Говорят, что случайный процесс $X(t)$ задан, если для каждого t из T определена функция распределения величины $X(t)$:*

$$F_t(x) = P(X(t) \leqslant x), \tag{11.6}$$

для каждой пары элементов t_1, t_2 из T определена функция распределения двумерной случайной величины $(X(t_1), X(t_2))$

$$F_{t_1,t_2}(x_1, x_2) = P(X(t_1) \leqslant x_1, X(t_2) \leqslant x_2), \tag{11.7}$$

и вообще для любого конечного числа элементов $t_1, t_2, \ldots, t_n$ из множества T определена n-мерная функция распределения величины $(X(t_1), X(t_2) \ldots, X(t_n))$

$$F_{t_1,t_2,\ldots,t_n}(x_1, x_2, \ldots, x_n) = P(X(t_1) \leqslant x_1, X(t_2) \leqslant x_2, \ldots, X(t_n) \leqslant x_n) \tag{11.8}$$

При этом распределения (11.6)–(11.8) должны быть согласованы в том смысле, что «старшие» распределения определяют «младшие». Например,

$$F_{t_1,t_2,\ldots,t_n}(x_1, x_2, \ldots, x_n) = \lim_{x_{n+1}\to\infty} F_{t_1,t_2,\ldots,t_n,t_{n+1}}(x_1, x_2, \ldots, x_n, x_{n+1})$$

Функции (11.6)–(11.8) называют конечномерными распределениями случайного процесса.

На практике общее определение случайного процесса используется редко. Чаще случайные процессы задают с помощью предположений типа независимости приращений, марковского свойства траекторий и т.д. Примеры подобных определений будут даны чуть позже.

Гауссовские случайные процессы. Важным классом случайных процессов являются *нормальные (гауссовские) случайные процессы*. Все конечномерные распределения этих процессов являются нормальными. (Определения одномерного и двумерного нормального распределений даны в пунктах 2.4 и 2.5. Аналогичным образом можно определить многомерные нормальные распределения.) Для полного описания нормальных случайных процессов достаточно указать его двумерные распределения. Если эти распределения должным образом согласованы, то с их помощью можно задать любые конечномерные распределения вида (11.8). Это обстоятельство играет важную роль в прикладном анализе гауссовских процессов, позволяя ограничиться исследованием математического ожидания и корреляционной функции процесса.

Белый шум. Математически простейшей моделью случайной компоненты временного ряда является последовательность независимых случайных величин. Независимость двух случайных величин была определена ранее (смотри, например, п. 1.6). Аналогично определяется и независимость произвольного числа случайных величин. С помощью функций распределения независимость последовательности случайных величин определяется так:

Определение. *Пусть T — множество типа $t = 0, 1, 2, \ldots$ или $t = 0, \pm 1, \pm 2, \ldots$ Случайный процесс $X(t)$ называется последовательностью независимо распределенных случайных величин, если для любых наборов чисел $t_1, t_2, \ldots, t_n$*

$$F_{t_1,t_2,\ldots,t_n}(x_1, x_2, \ldots, x_n) = F_{t_1}(x_1) \cdot F_{t_2}(x_2) \cdot \ldots \cdot F_{t_n}(x_n) \tag{11.9}$$

Из (11.9) следует, что для последовательности независимых случайных величин все ее конечномерные распределения определяются с помощью одномерных распределений (11.6).

Определение. *Белым шумом называют временной ряд (случайный процесс) с нулевым средним, если составляющие его случайные величины $X(t)$ независимы и распределены одинаково (при всех t).*

Это так называемый *белый шум в узком смысле*. Белый шум в широком смысле будет определен позже, после определения свойства стационарности в широком смысле. В определение белого шума часто включают предположение о нормальности распределения величин $X(t)$. Другими словами, *гауссовский белый шум* — это последовательность независимых нормально распределенных случайных величин со средним 0 и общей дисперсией (скажем, σ^2).

Последовательности независимых случайных величин далеко не всегда адекватно описывают случайные компоненты временных рядов. Теорией и практикой для описания случайных последовательностей выработаны и более сложные модели. Некоторые из них мы упомянем ниже, а более подробно рассмотрим в дальнейшем.

Процессы скользящего среднего. Для этих процессов часто употребляют аббревиатуру MA — от английского moving average (движущееся среднее). Это сокращение стандартно используется в англоязычной литературе и статистических пакетах.

Пусть $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n, \ldots$ — независимые одинаково распределенные случайные величины (белый шум).

Определение. *Процессом скользящего среднего (первого порядка) со средним μ (сокращенно MA(1)) называют процесс $X(t)$:*

$$X(t) = \varepsilon_t + \theta\varepsilon_{t-1} + \mu, \tag{11.10}$$

где θ — некоторый числовой коэффициент, а μ — константа.

Заметим, что у процесса скользящего среднего (11.10) статистически зависимы только соседние величины $X(t-1)$ и $X(t)$. Значения процесса, разделенные промежутком времени 2 и более, статистически независимы, ибо в их формировании участвуют разные слагаемые ε_t. По этой причине процессы скользящего среднего являются непосредственным и простейшим обобщением процессов белого шума.

Описание процессов скользящего среднего второго и более высоких порядков, а также свойств этих процессов, будет дано в гл. 14.

Процессы авторегрессии. Для них часто употребляют аббревиатуру AR — от английского autoregression.

Определение. *Процессом авторегрессии (первого порядка) со средним значением μ (сокращенно AR(1)) называют случайный процесс $X(t)$, удовлетворяющий соотношению:*

$$X(t) - \mu = \phi \cdot (X(t-1) - \mu) + \varepsilon_t, \tag{11.11}$$

где ϕ и μ — некоторые числа.

Члены процесса авторегрессии, разделенные промежутком времени $h > 0$, не становятся независимыми, как бы ни было велико h. Однако зависимость между ними быстро убывает с ростом h, если $|\phi| < 1$. Именно такие процессы авторегрессии обычно встречаются в прикладных задачах.

Процессы авторегрессии порядка 2 и выше будут определены в главе 14. Там же мы обсудим их свойства и области приложений.

Марковское свойство. Поведение многих процессов в будущем определяется только их состоянием в настоящем и воздействиями на процесс, которые будут оказываться в будущем. А предыдущее развитие процесса (то есть его состояние до настоящего времени) при этом несущественно. Такие процессы называются *марковскими*. Дадим этому понятию более строгое определение.

Пусть t, $t \in T$ — произвольный момент времени, который мы назовем «настоящим». Пусть A — произвольное событие, выраженное через случайные величины $X(s)$, где $s \leqslant t-1$. Это событие, относящееся к прошлому последовательности $X(\cdot)$. Пусть B — произвольное событие, относящееся к будущему процесса $X(\cdot)$, т.е. событие B выражается через случайные величины $X(s)$, где $s \geqslant t+1$. Рассмотрим условные вероятности событий A, B и AB при фиксированном значении $X(t)$. Обозначим эти условные вероятности через $P(A|X(t))$, $P(B|X(t))$ и $P(AB|X(t))$.

Определение. *Случайная последовательность $X(t)$, $t \in T$ называется марковской, если для любых A, B и t*

$$P(AB|X(t)) = P(A|X(t))P(B|X(t)).$$

Нередко марковскому свойству последовательности $X(\cdot)$ дают несколько иное определение (впрочем, эквивалентное приведенному).

Определение. *Случайная последовательность $X(t)$, $t \in T$ называется марковской, если для любых A, B и t*

$$P(B|X(t), A) = P(B|X(t)).$$

В обычных обстоятельствах нет возможности проверить, обладает или нет наблюдаемый временной ряд этим свойством. Марковское

свойство для временного ряда обычно постулируют, когда физическая природа ряда дает для того основания.

В статистическом анализе марковское свойство процесса редко используется непосредственно. Обычно оно служит для вывода уравнений, описывающих изменения во времени каких-либо его средних характеристик (например, математического ожидания). Среди математических моделей временных рядов, которых мы далее касаемся, марковским свойством обладает процесс авторегрессии первого порядка (см. п. 14.1). Процесс авторегрессии $X(t)$ произвольного порядка $p \geqslant 1$ тоже можно представить как марковский, если его состоянием в момент t считать набор $(X(t), X(t-1), \ldots, X(t-p-1))$.

Стационарность. В теоретических исследованиях и практических задачах важную роль играют последовательности случайных величин, вероятностные свойства которых не изменяются во времени. Такие случайные последовательности называют стационарными. Их можно использовать для описания временных рядов, течение которых стабилизировалось и происходит в неизменных условиях.

Определение. *Случайный процесс $X(t)$ называется стационарным, если для любых n, $t_1, t_2, \ldots, t_n$ и τ распределения случайных величин $(X(t_1), \ldots, X(t_n))$ и $(X(t_1+\tau), \ldots, X(t_n+\tau))$ одинаковы.*

Это означает, что функции конечномерных распределений (11.8) не меняются при сдвиге времени, т.е.

$$F_{t_1+\tau, t_2+\tau \ldots, t_n+\tau}(x_1, \ldots, x_n) = F_{t_1, t_2, \ldots, t_n}(x_1, x_2, \ldots, x_n).$$

В частности, образующие стационарную случайную последовательность случайные величины $X(1), X(2), \ldots, X(t), \ldots$ распределены одинаково (но независимыми они, вообще говоря, не являются).

Этот вид стационарности называют также *стационарностью в узком смысле*. Другой вид стационарности — *стационарность в широком смысле*, — мы введем после того, как для случайных последовательностей мы определим их числовые характеристики.

Определенный ранее процесс белого шума является стационарным (в узком смысле).

11.8. Числовые характеристики временных рядов

Числовые характеристики временных рядов вводятся в полной аналогии с числовыми характеристиками случайных величин (см. п. 1.5).

Математическое ожидание *(первый момент) случайного процесса $X(t)$ — это функция $m(t)$, такая, что для каждого t значение функции $m(t)$ является математическим ожиданием случайной величины $X(t)$:*

$$m(t) = MX(t).$$

Функцию $m(t)$ часто называют *средним значением* процесса $X(t)$. Она используется для описания систематического изменения процесса. Например, для случайного процесса, допускающего запись в виде аддитивной модели (11.1), среднее значение равно $tr_t + s_t + c_t$. Заметим, что под словом «усреднение» здесь понимается усреднение случайной величины $X(t)$ при неизменном t, а не усреднение по времени, хотя такое тоже бывает. Ниже мы более подробно коснемся этого вопроса.

Ковариационная функция *случайного процесса $X(t)$ (кратко $\operatorname{cov}(X(t), X(s))$) — это величина*

$$B(s,t) = \operatorname{cov}(X(t), X(s)) = M[(X(t) - m(t))(X(s) - m(s))].$$

Она является функцией пары переменных (t, s). Иногда ее именуют функцией вторых центральных моментов.

Значение ковариационной функции при $t = s$ задает *дисперсию* случайного процесса $DX(t) = \operatorname{cov}(X(t), X(t))$. Квадратный корень из $\operatorname{cov}(X(t), X(t))$ называют *стандартным отклонением* $\sigma(t)$ случайного процесса $X(t)$:

$$\sigma(t) = \sqrt{\operatorname{cov}(X(t), X(t))}.$$

Корреляционная функция *случайного процесса $X(t)$ — это величина:*

$$\operatorname{corr}(X(t), X(s)) = \frac{\operatorname{cov}(X(t), X(s))}{\sigma(t)\sigma(s)}.$$

Как и ковариационная функция, корреляционная функция также зависит от пары переменных (t, s).

При фиксированных t и s $\operatorname{corr}(X(t), X(s))$ по определению является коэффициентом корреляции (см. п. 1.6) случайных величин $X(t)$ и $X(s)$, и для него выполняются свойства 1—4 п. 1.6. Из определения $\operatorname{cov}(X(t), X(s))$ и $\operatorname{corr}(X(t), X(s))$ следует их симметрия относительно t и s:

$$\operatorname{cov}(X(t), X(s)) = \operatorname{cov}(X(s), X(t)),$$
$$\operatorname{corr}(X(t), X(s)) = \operatorname{corr}(X(s), X(t))$$

Заметим, что функции $m(t)$, $B(s,t)$ могут и не существовать: как мы знаем, не всегда случайные величины имеют математическое ожи-

дание и дисперсию. Но в статистической практике такие временные ряды, для которых $m(t)$ и $B(s,t)$ не существуют, встречаются редко. Поэтому в дальнейшем к средним значениям временных рядов и их ковариационной или корреляционной функциям мы будем обращаться без особых оговорок.

Ковариационная и корреляционная функции играют важную роль в теоретическом и в практическом анализе случайных процессов и временных рядов. Ниже мы обсудим их свойства, а также способы оценивания этих функций по наблюдениям (см. п. 11.10). А сейчас вернемся к свойству стационарности (см. п. 11.7) и с помощью функций $m(t)$ и $B(s,t)$ дадим ему другое определение.

Из определения стационарности, данного выше, следует, что для любых s, t и любого τ:

$$m(t+\tau) = m(t), \qquad B(s+\tau, t+\tau) = B(s, t). \tag{11.12}$$

Положив $\tau = -t$, мы получаем, что

$$m(t) = m(0), \qquad B(s, t) = B(s-t, 0).$$

Отсюда следует, что у стационарного процесса функции $m(t)$ и $\sigma(t)$ постоянны, а ковариационная функция $B(s,t)$ реально зависит не от пары (s, t), как в общем случае, а от $|s-t|$. Точно так же можно убедиться, что и корреляционная функция стационарного процесса является функцией $|s-t|$.

Рассмотрим $t = s+k$, $k > 0$. Положим по определению

$$r(k) = \operatorname{corr}(X(t), X(s)) = \operatorname{corr}(X(t), X(t+k)).$$

Автокорреляционная функция. *Автокорреляционной функцией стационарного процесса $X(t)$ называют функцию $r(k) = \operatorname{corr}(X(t), X(t+k))$, где $k > 0$ — целое число.*

Величину k часто называют *задержкой*, или *лагом*. Она указывает расстояние между членами временного ряда, для которых вычисляется коэффициент корреляции.

11.9. Процессы, стационарные в широком смысле

Вообще говоря, выполнение свойства (11.12) не гарантирует того, что процесс $X(t)$ является стационарным в смысле приведенного выше определения стационарности в узком смысле. Тем не менее, свойство (11.12) определенно отражает некую неизменность во времени свойств процесса $X(t)$. Поэтому принято следующее

Определение. *Случайный процесс $X(t)$ называется стационарным в широком смысле, если его среднее значение $m(t)$ постоянно, а ковариационная функция $B(s, t)$ зависит только от расстояния между аргументами, т.е. от $|t-s|$.*

Свойство стационарности в широком смысле играет важную роль при нахождении оценок числовых характеристик временных рядов (см. п. 11.10).

Белый шум в широком смысле. Аналогичное определение можно дать для белого шума:

Определение. *Временной ряд (случайный процесс) $X(t)$ называют белым шумом (в широком смысле), если для любого t выполняется $MX(t) = 0$ и*

$$\operatorname{cov}(X(s), X(t)) = \begin{cases} \sigma^2, & \text{когда} \quad s = t, \\ 0, & \text{когда} \quad s \neq t. \end{cases}$$

Из этого определения видно, что этот белый шум является стационарным (в широком смысле) случайным процессом. На практике различие между двумя видами белого шума (в широком и в узком смысле) не всегда проводятся четко. В дальнейшем, говоря о белом шуме в связи с прикладными исследованиями, мы чаще всего будем иметь в виду белый шум именно в только что введенном широком смысле.

Хотя на практике процессы белого шума в чистом виде встречаются не часто, они играют фундаментальную роль как в теории, так и в прикладном анализе временных рядов. Типичным для такого анализа является, например, процесс «выбеливания» временного ряда, т.е. исключения из него тренда, циклической, сезонной и прочих компонент, так чтобы остаток статистически не отличался от процесса белого шума.

Гауссовские процессы. Ясно, что стационарный в узком смысле случайный процесс является одновременно и стационарным в широком смысле, если существуют функции двух первых моментов. Уже отмечалось, что обратное, вообще говоря, неверно.

Одно из исключений составляют нормальные, или гауссовские случайные процессы, то есть процессы, конечномерные распределения которых (11.8) являются гауссовскими. Для гауссовских процессов любые конечномерные распределения определяются через функции $m(t)$ и $B(s, t)$. Поэтому гауссовские процессы, стационарные в широком смысле, одновременно являются стационарными в узком смысле. Это весьма важное и полезное обстоятельство, так как на практике проверка стационарности в узком смысле не осуществима. Судить о стационарности в широком смысле значительно проще. Для этого существуют раз-

личные статистические критерии, базирующиеся на одной реализации случайного процесса. Наиболее важные из таких критериев основаны на выборочных оценках автокорреляционной функции и спектральной плотности (см. [123]).

Замечание. Для некоторых протекающих во времени процессов модель гауссовского случайного процесса дает приемлемое по качеству описание. К сожалению, в полном объеме проверить по наблюдениям, верна ли эта модель, невозможно. Поэтому гауссовский случайный процесс представляет собой, в первую очередь, удобный математический объект.

Преобразование процесса в стационарный. Наиболее распространенным случаем нарушения стационарности на практике является изменение среднего значения $m(t)$ с изменением времени t. В тех случаях, когда $m(t)$ удается тем или другим способом оценить, преобразование $Y(t) = X(t) - m(t)$ превращает процесс в стационарный. Далее $Y(t)$ изучают как стационарный, используя для этого его специфические свойства.

11.10. Оценки числовых характеристик временных рядов

В каждый фиксированный момент времени t случайный процесс $X(t)$ является случайной величиной. Следовательно, для построения оценок его моментов $m(t)$ и $B(s, t)$ теоретически можно использовать те же методы, что и для обычных случайных величин. Напомним (см. п. 1.8.1 и п. 4.3), что при этом требуется некоторая совокупность независимых реализаций этой случайной величины X, полученных при повторении опыта в неизменных условиях. Другими словами, нужна выборка $x_1, \ldots, x_k$. Применение этой методики к случайному процессу $X(t)$ требует от нас набора реализаций (траекторий) этого процесса $x_1(t), \ldots, x_k(t)$.

В технических приложениях возможности для независимых повторений опыта иногда имеются. Скажем, изучая колебания напряжения в электрических сетях в течение суток, мы можем считать временные ряды, полученные в разные сутки, независимыми реализациями одного случайного процесса. Для большей уверенности, что повторения наблюдений произведены в неизменных условиях, можно сопоставлять данные за определенные дни недели (за вторники, например), отдельно по разным сезонам и т.д.

Однако в экономических, социальных, демографических и подобных процессах мы обычно имеем дело с единственной траекторией развития, повторить которую невозможно. Поэтому при изучении статистических

свойств таких процессов приходится обходиться этой самой единственной реализацией. Зато длина ее может расти. Значительная часть математических результатов о временных рядах относится к стационарным рядам, наблюдаемым на растущем интервале времени $(0, T)$. Они формулируются в виде предельных теорем при $T \to \infty$. Многие физические, технические и естественнонаучные приложения статистической теории нуждаются именно в такой постановке проблемы.

Впрочем, для упомянутых экономических, социальных и т.п. временных рядов эти результаты дают не очень много. Во-первых, эти ряды обычно довольно коротки. Во-вторых, они не стационарны, так как условия, в которых они протекают, изменяются с течением времени. Поэтому наблюдения даже из относительно недавнего прошлого порой мало что говорят о современных тенденциях.

По единственной реализации процесса $X(t)$ мы не можем составить оценки для его среднего, дисперсии, ковариации и т.д., как мы сделали бы это, располагая выборкой. Но некоторые похожие средние величины составить можно.

Оценка среднего значения. Имея ряд $x(t_1), \ldots, x(t_n)$ последовательных наблюдений случайного процесса $X(t)$, можно составить *«среднее по реализации»*: $\overline{m} = \frac{1}{n}\sum_{i=1}^{n} x(t_i)$. В теории временных рядов для наблюденных значений $x(t_1), \ldots, x(t_n)$ используют более короткую форму записи $x_1, \ldots, x_n$. В этих обозначениях среднее по реализации есть

$$\overline{m} = \frac{1}{n}\sum_{i=1}^{n} x_i. \tag{11.13}$$

Оказывается, при некоторых условиях это среднее может служить оценкой математического ожидания процесса $X(t)$. Первым из таких условий является стационарность случайного процесса $X(t)$ (в широком смысле). Поскольку для стационарного процесса все моменты времени равноправны и его числовые характеристики неизменны во времени, в качестве оценки $m(t) = m$ естественно рассмотреть именно $\overline{m}$. Легко убедиться, что $\overline{m}$ как оценка m для стационарных процессов является несмещенной, т.е. $M\overline{m} = m$.

Рассмотрим вопрос о точности этой оценки. Естественно, хотелось бы, чтобы эта оценка $\overline{m}$ приближалась к неизвестному истинному значению с ростом числа наблюдений n, то есть была бы состоятельной (см. п. 4.5). Так как отклонение оценки от истинного значения можно описать с помощью ее дисперсии, то для состоятельности достаточно, чтобы

$$D\overline{m} \to 0 \qquad \text{при} \quad n \to \infty. \tag{11.14}$$

К сожалению, одна лишь стационарность случайного процесса не обеспечивает выполнения (11.14). Простейший и отчасти вырожденный пример стационарного процесса, для которого не выполняется свойство (11.14), устроен следующим образом. Рассмотрим стационарный процесс, для которого с вероятностью единица $X(t) = X(1)$ для любого t. Ясно, что траектории этого процесса являются константами. При этом:

$$\overline{m} = x_1 \qquad D\overline{m} = const.$$

В более общем случае типичным примером невыполнения условия (11.11) являются смеси, т.е. процессы, у которых различные участки траекторий сформированы при разных условиях. Более подробно модель таких процессов описана в [123].

Хоть мы и говорили о том, что предельные теоремы математической теории мало полезны для интересующей нас области приложений, все же приведем одно из достаточных условий для выполнения (11.14).

Теорема Слуцкого. *Для стационарного в широком смысле случайного процесса $X(t)$ оценка его среднего значения (11.13) состоятельна тогда и только тогда, когда:*

$$\frac{1}{n}\sum_{t=0}^{n-1} r_t \to 0 \qquad при \quad n \to \infty, \tag{11.15}$$

где r_t — автокорреляционная функция процесса.

Мы не будем более подробно обсуждать этот результат. Обратим внимание лишь на то, что для выполнения (11.15) достаточно, чтобы $r_t \to 0$ при $t \to \infty$. Последнее замечание позволяет на практике судить о том, можно ли использовать осреднение по одной реализации для получения состоятельных оценок его характеристик. Таким образом теорема Слуцкого подчеркивает важность анализа поведения автокорреляционной функции случайного процесса.

Еще два замечания о точности приближения оценки $\overline{m}$ к истинному значению. Первое из них касается скорости сходимости $\overline{m}$ к m. Можно показать, что стандартное отклонение $\overline{m}$ при больших n пропорционально $1/\sqrt{n}$, то есть увеличение точности оценки обратно пропорционально квадратному корню из объема наблюдений. Второе замечание относится к случаю, когда объема временного ряда недостаточно для получения достаточно точной оценки среднего значения. Определим величину T в виде:

$$T = \sum_{k=0}^{\infty} r_k$$

считая, что указанная сумма конечна. Величина T называется *временем корреляции* и дает представление о порядке величины промежутков времени τ, на которых сохраняется заметная корреляция между $X(t)$ и $X(t+\tau)$. Если объем n рассматриваемой реализации временного ряда меньше T, то оценка $\overline{m}$ считается весьма неточной. Введенная величина T позволяет также указать более точную скорость сходимости $\overline{m}$ к m. А именно, эта скорость пропорциональна $\sqrt{T/n}$.

Выборочная автокорреляционная функция. В главе 9 (п. 9.5.1) была подробно разобрана оценка коэффициента корреляции (9.17) пары случайных величин, построенная по выборкам. Методика получения оценок значений автокорреляционной функции $r(k)$ во многом напоминает случай двух выборок. Разберем ее устройство на оценке $r(1)$ — корреляции между соседними членами временного ряда X_t и X_{t+1}. (Напомним, что большие буквы X мы используем для обозначения случайного процесса, а малые буквы x — для обозначения реализации этого случайного процесса.)

Образуем из временного ряда $x_1, x_2, \ldots, x_n$ совокупность из $n-1$ пар: $(x_1, x_2), (x_2, x_3), \ldots, (x_{n-1}, x_n)$. Первый элемент каждой пары, в силу стационарности, мы можем рассматривать как реализацию случайной величины X_t, а второй — как реализацию случайной величины X_{t+1}. Тогда, согласно (9.17) оценка коэффициента корреляции между X_t и X_{t+1} может быть записана в виде:

$$\overline{r}_1 = \frac{\sum_{t=1}^{n-1}(x_t - \overline{x}_{(1)})(x_{t+1} - \overline{x}_{(2)})}{\sqrt{\left[\sum_{t=1}^{n-1}(x_t - \overline{x}_{(1)})^2 \sum_{t=1}^{n-1}(x_{t+1} - \overline{x}_{(2)})^2\right]}}, \tag{11.16}$$

где

$$\overline{x}_{(1)} = \sum_{t=1}^{n-1} x_t/(n-1), \qquad \overline{x}_{(2)} = \sum_{t=2}^{n} x_t/(n-1),$$

соответственно оценки средних значений величин X_t и X_{t+1}

При больших значениях n, учитывая что $\overline{x}_{(1)} \approx \overline{x}_{(2)} \approx \overline{x}$ и $n/(n-1) \approx 1$, выражение (11.16) часто заменяют гораздо более простым:

$$\overline{r}_1 = \frac{\sum_{t=1}^{n-1}(x_t - \overline{x})(x_{t+1} - \overline{x})}{\sum_{t=1}^{n}(x_t - \overline{x})^2}. \tag{11.17}$$

Аналогичным образом может быть определена оценка корреляции между X_t и X_{t+k} или k-го члена автокорреляционной функции r_k:

$$\overline{r}_k = \frac{\sum_{t=1}^{n-k}(x_t - \overline{x})(x_{t+k} - \overline{x})}{\sum_{t=1}^{n}(x_t - \overline{x})^2}. \tag{11.18}$$

Обратим внимание читателя, что точность приближения (11.18) заметно снижается с ростом лага k, как в силу ухудшения точности использованных выше замен, так и в силу уменьшения числа наблюдений используемых для вычисления оценки $\overline{r}_k$ Поэтому на практике обычно ограничиваются изучением небольшого числа первых членов автокорреляционной функции. Вряд ли имеет смысл рассматривать оценки r_k при $k > n/4$.

Функцию $\overline{r}_k$ аргумента k при $k = 1, 2, \ldots$ называют *выборочной автокорреляционной функцией* или, если не возникает недоразумений, просто автокорреляционной функцией. (При $k = 0$ $\overline{r}_k$ по определению равно 1 и это значение обычно исключают из рассмотрения как не несущее никакой информации.) В англоязычной литературе эту функцию также называют *сериальной корреляцией*. График выборочной автокорреляционной функции называют *коррелограммой*. На этом графике (см. рис. 11.4), кроме значений самой функции, обычно указывают доверительные пределы этой функции в предположении, что значения автокорреляционной функции равны 0 для всех $k \neq 0$. Более подробно об интерпретации графика выборочной автокорреляционной функции будет рассказано ниже при рассмотрении роли коррелограммы в практическом анализе временных рядов (см. п. 12.4.2).

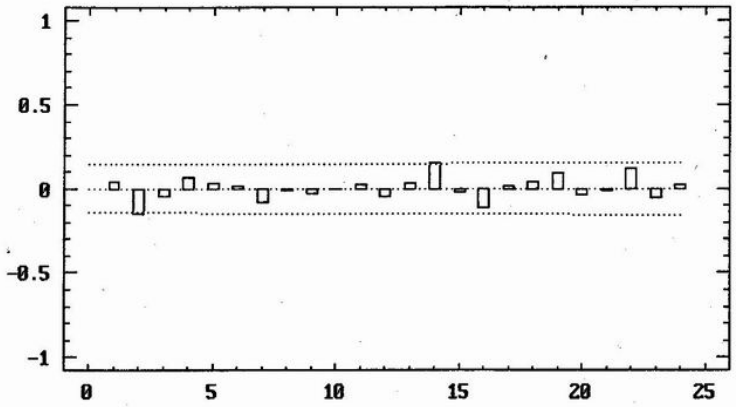


Рис. 11.4. Коррелограмма с доверительными интервалами (при равенстве нулю автокорреляционной функции для всех $k \neq 0$)

Свойства. Изучение свойств выборочных оценок автокорреляционной функции временного ряда — в общем случае довольно сложная и до конца не решенная задача.

М.Бартлетом в 1946 г. для случая бесконечного дискретного временного ряда ($-\infty < t < \infty$) было указано выражение дисперсии оценки

$\overline{r}_k$ для гауссовского процесса:

$$D\overline{r}_k = \frac{1}{n}\sum_{t=-\infty}^{t=\infty}\left[r_t^2 + r_{t-k}r_{t+k} - 4r_k r_t r_{t+k} + 2r_t^2 r_k^2\right]. \qquad (11.19)$$

Этот результат показывает, что мы не можем оценить по конечному отрезку временного ряда дисперсию оценки $\overline{r}_k$, так как она зависит от бесконечного неизвестного числа автокорреляций r_t. Поэтому на практике приходится довольствоваться лишь приближениями для выражения (11.19).

Другая проблема изучения свойств совокупности оценок $\overline{r}_k$ связана с тем, что оценки с различным лагом k коррелированы между собой. Это заметно затрудняет интерпретацию коррелограммы. Не касаясь более подробно этих и других проблем (см. [37], [51]), укажем свойства оценок $\overline{r}_k$ для наиболее простого и практически важного случая. А именно, рассмотрим свойства оценок автокорреляций для временного ряда, являющегося стационарной последовательностью независимых нормально распределенных случайных величин или, другими словами, гауссовским белым шумом (см. п. 11.7). В этом случае для любых k, не равных нулю, по определению $r_k = 0$. Таким образом, все слагаемые, стоящие под знаком суммы в выражении (11.19), равны нулю, кроме $r_0^2 = 1$. Отсюда дисперсия $\overline{r}_k$ равна:

$$D\overline{r}_k = \frac{1}{n}$$

Обратим внимание на то, что оценка $\overline{r}_k$ в форме (11.18) является смещенной. Можно показать, что $M\overline{r}_k \approx -\frac{1}{n}$, однако величина этого смещения стремится к нулю с ростом объема изучаемого ряда и не столь существенна в прикладном анализе.

Другим важным свойством оценки $\overline{r}_k$ является ее асимптотическая нормальность при $n \to \infty$.

Таким образом, для каждого отдельного значения $\overline{r}_k$ мы можем указать приблизительный 95% доверительный интервал в виде: $-1/n \pm 2/\sqrt{n}$. Границы этого доверительного интервала обычно наносятся на график коррелограммы и называются *доверительной трубкой*. Они в определенной мере позволяют судить о том, насколько изучаемый процесс напоминает белый шум. Указание 95% доверительных границ для каждого коэффициента автокорреляционной функции в отдельности не означает, что с 95% вероятностью все рассматриваемые оценки $\overline{r}_k$ одновременно попадают в доверительную трубку. Так, рассматривая 20 первых оценок $\overline{r}_k$ для гауссовского белого шума, довольно часто можно наблюдать, что одна или две из оценок выходят за границы довери-

тельной трубки. Это обстоятельство также затрудняет интерпретацию коррелограммы.

Периодограмма. Завершая рассказ об оценках основных числовых характеристик временного ряда, кратко остановимся на *периодограмме* — характеристике, особенно полезной для анализа временных рядов, допускающих представление в виде полигармонических моделей (11.5).

Многие временные ряды, возникающие в физических и технических приложениях, удобно рассматривать не во временной области значений аргумента, а в частотной. Этот переход можно совершить с помощью периодограммы. Ее назначение — обнаружение периодических составляющих в рассматриваемом ряде. Первое представление о наличии таких составляющих может дать обычный график. Если стационарный временной ряд на графике ведет себя более гладко и регулярно, чем гауссовский белый шум (см. рис. 11.1д), то можно предположить, что в нем есть периодические составляющие.

В настоящее время определение периодограммы часто использует понятие спектральной плотности, но так как это понятие не рассматривается в настоящей главе, мы дадим определение периодограммы в том виде, как оно было предложено А.Шустером в 1898 г.

Пусть x_t — временной ряд с нулевым средним, а t пробегает целые числа от 1 до n. Рассмотрим ковариацию ряда x_t с рядами $\cos(2\pi t/\lambda)$ и $\sin(2\pi t/\lambda)$, где λ — некоторая фиксированная величина, обычно именуемая периодом, или длиной волны. Пусть:

$$A = \frac{2}{n}\sum_{t=1}^{n} x_t \cos\frac{2\pi t}{\lambda}, \qquad B = \frac{2}{n}\sum_{t=1}^{n} x_t \sin\frac{2\pi t}{\lambda}.$$

Введем величину $S^2(\lambda)$:

$$S^2(\lambda) = A^2(\lambda) + B^2(\lambda)$$

Определение. *График зависимости $S^2(\lambda)$ от длины волны λ называется периодограммой.*

По замыслу, функция $S^2(\lambda)$ должна принимать большие значения (иметь локальные максимумы — пики) для тех значений λ, которые являются периодами для имеющихся у ряда x_t периодических составляющих. Практически это далеко не так, и часть максимумов $S^2(\lambda)$ к реальным периодам ряда x_t не имеет отношения. Вообще анализ периодограммы очень часто ведет к ложным выводам, и потому к нему надо подходить с осторожностью. Эти вопросы подробно освещены в литературе по спектральному анализу временных рядов. (Смотри, в частности, критический анализ в [75] и в гл. 4 книги [97].)

Глава 12

Временные ряды: практический анализ

12.1. Порядок анализа временных рядов

Кратко опишем общий порядок прикладного статистического анализа временных рядов. Обычно целью такого анализа является построение математической модели ряда, с помощью которой можно объяснить поведение ряда и осуществить прогноз его дальнейшего поведения.

Построение и изучение графика. Анализ временного ряда обычно начинается с построения и изучения его графика. Если нестационарность временного ряда очевидна, то первым делом надо выделить и удалить нестационарную составляющую ряда. Методы, используемые для этого, описаны в п. 12.3. Процесс удаления тренда и других компонент ряда, приводящих к нарушению стационарности, может проходить в несколько этапов. На каждом из них рассматривается ряд остатков, полученный в результате вычитания из исходного ряда подобранной модели тренда, или результат разностных и других преобразований ряда. Кроме графиков, признаками нестационарности временного ряда могут служить не стремящаяся к нулю автокорреляционная функция (за исключением очень больших значений лагов) и наличие ярко выраженных пиков на низких частотах в периодограмме.

Подбор модели для временного ряда. После того, как исходный процесс максимально приближен к стационарному, можно приступить к подбору различных моделей полученного процесса. Цель этого этапа — описание и учет в дальнейшем анализе корреляционной структуры рассматриваемого процесса. При этом на практике чаще всего используются два типа моделей: параметрические модели авторегрессии-скользящего среднего (ARMA-модели) и полигармонические модели (см. п. 11.6). ARMA-модели мы будем рассматривать в главе 14, а описание способов подбора полигармонических моделей можно найти в книгах [37], [75], [87], [123].

Модель может считаться подобранной, если остаточная компонента ряда является процессом типа белого шума (см. п. 11.7, 11.9). После подбора модели обычно выполняются:

- оценка дисперсии остатков, которая в дальнейшем может быть использована для построения доверительных интервалов прогноза;
- анализ остатков с целью проверки адекватности модели.

Прогнозирование или интерполяция. Последним этапом анализа временного ряда может быть прогнозирование его будущих (экстраполяция) или восстановление пропущенных (интерполяция) значений и указания точности этого прогноза на базе подобранной модели. Обратим внимание, что хорошо подобрать математическую модель удается не для всякого временного ряда. Нередко бывает и так, что для описания подходят сразу несколько моделей. Неоднозначность подбора модели может наблюдаться как на этапе выделения детерминированной компоненты ряда, так и при выборе структуры ряда остатков. Поэтому исследователи довольно часто прибегают к методу нескольких прогнозов, сделанных с помощью разных моделей.

Методы анализа. Перечислим основные группы статистических приемов, используемых для анализа временных рядов:

- графические методы представления временных рядов и их сопутствующих числовых характеристик;
- методы сведения к стационарным процессам;
- методы исследования внутренних связей между элементами временных рядов.

Ниже будет подробно рассказано о каждой из этих групп методов.

12.2. Графические методы анализа временных рядов

Зачем нужны графические методы. В выборочных исследованиях простейшие числовые характеристики описательной статистики (среднее, медиана, дисперсия, стандартное отклонение, коэффициенты асимметрии и эксцесса) обычно дают достаточно информативное представление о выборке. Графические методы представления и анализа выборок при этом играют лишь вспомогательную роль, позволяя лучше понять локализацию и концентрацию данных, их закон распределения.

Роль графических методов при анализе временных рядов совершенно иная. Дело в том, что табличное представление временного ряда и описательные статистики чаще всего не позволяют понять характер процесса, в то время как по графику временного ряда можно сделать

довольно много выводов. В дальнейшем они могут быть проверены и уточнены с помощью расчетов.

Человеческий глаз довольно уверенно определяет по графику временного ряда:

- наличие тренда и его характер;
- наличие сезонных и циклических компонент;
- степень плавности или прерывистости изменений последовательных значений ряда после устранения тренда. По этому показателю можно судить о характере и величине корреляции между соседними элементами ряда.

Так графический анализ ряда обычно задает направление его дальнейшего анализа.

Построение и изучение графика. Построение графика временного ряда — совсем не такая простая задача, как это кажется на первый взгляд. Современный уровень анализа временных рядов предполагает использование той или иной компьютерной программы для построения их графиков и всего последующего анализа. Ряд полезных рекомендаций при построении графика вручную даны в [98]. Большинство статистических пакетов и электронных таблиц снабжено теми или иными методами настройки на оптимальное представление временного ряда, но даже при их использовании могут возникать различные проблемы, например:

- из-за ограниченности разрешающей способности экранов компьютеров размеры выводимых графиков могут быть также ограничены;
- при больших объемах анализируемых рядов точки на экране, изображающие наблюдения временного ряда, могут превратиться в сплошную черную полосу.

Для борьбы с этими затруднениями используются различные способы. Наличие в графической процедуре режима «лупы» или «увеличения» позволяет изобразить более крупно выбранную часть ряда, однако при этом становится трудно судить о характере поведения ряда на всем анализируемом интервале. Приходится распечатывать графики для отдельных частей ряда и состыковывать их вместе, чтобы увидеть картину поведения ряда в целом. Иногда для улучшения воспроизведения длинных рядов используется *прореживание*, то есть выбор и отображение на графике каждой второй, пятой, десятой и т.д. точки временного ряда. Эта процедура позволяет сохранить целостное представление ряда и полезна для обнаружения трендов. На практике полезно сочетание

обеих процедур: разбиения ряда на части и прореживания, так как они позволяют подметить разные черты в поведении временного ряда.

Еще одну проблему при воспроизведении графиков создают *выбросы* — наблюдения, в несколько раз превышающие по величине большинство остальных значений ряда. Их присутствие тоже приводит к неразличимости колебаний временного ряда, так как масштаб изображения программа автоматически подбирает так, чтобы все наблюдения поместились на экране. Выбор другого масштаба на оси ординат устраняет эту проблему, но резко отличающиеся наблюдения при этом остаются за границами экрана.

Дадим еще несколько полезных советов по построению и оформлению графика временного ряда:

- внимательно следите за масштабом представления данных по каждой из осей, так как они, как правило, различаются. Многие программы автоматически используют экспоненциальную форму записи для обозначения делений осей, например 0.241E+03 вместо 241, что не всегда удобно и оправданно;
- не забывайте указывать, какие величины отображает каждая из осей и их единицы измерения. Это особенно важно при представлении экономических временных рядов, где наряду с равномерными шкалами часто используются логарифмические шкалы;
- точки на графике временного ряда обычно соединяют отрезками прямых линий. Однако в некоторых ситуациях эти линии могут вносить существенное искажение в представление о поведении ряда. Поэтому полезно построить график временного ряда с линиями между точками и без них, и внимательно изучить оба этих графика;
- наличие не слишком густой координатной сетки облегчает восприятие графиков.

Вспомогательные графики. При анализе временных рядов часто используются вспомогательные графики для числовых характеристик ряда:

- график выборочной автокорреляционной функции (коррелограммы) с доверительной зоной (трубкой) для нулевой автокорреляционной функции;
- график выборочной частной автокорреляционной функции (см. п. 14.3) с доверительной зоной для нулевой частной автокорреляционной функции;
- график периодограммы.

Первые два из этих графиков позволяют судить о связи (зависимости) соседних значений временного ряда, они используются при подборе параметрических моделей авторегрессии-скользящего среднего. График периодограммы позволяет судить о наличии гармонических составляющих во временном ряде. Эти графики и свойства соответствующих функций рассмотрены в параграфе 12.4.

Учитывая, что многие методы анализа временных рядов рассчитаны на работу с рядами с нормально распределенной случайной компонентой, в процедуры анализа временных рядов обычно включают различные графики на нормальной вероятностной бумаге. Самый распространенный из них подробно описан в главе 5.

12.3. Методы сведения к стационарности

После изучения графика временного ряда обычно пробуют выделить во временном ряде тренд, сезонные и периодические компоненты. После их исключения временной ряд должен стать стационарным. Кроме того, для облегчения дальнейшего анализа иногда используются преобразования значений временного ряда (точнее, той шкалы, в которой они измерены) — это позволяет приблизить распределение значений временного ряда к нормальному или сделать дисперсию этих значений более постоянной (иначе говоря, стабилизировать дисперсию).

В п. 12.3.1 мы рассмотрим методы оценки и удаления тренда, а в пп. 12.3.2—12.3.4 — оценки и удаления сезонных эффектов и циклических компонент временного ряда. В п. 12.3.5 рассматриваются преобразования шкалы измерений временного ряда — переход к логарифмической шкале и преобразование Бокса-Кокса.

12.3.1. Выделение тренда

Метод наименьших квадратов. Для оценки и удаления трендов из временных рядов чаще всего используется метод наименьших квадратов. Этот метод подробно обсуждался в гл. 8 при рассмотрении задач линейного регрессионного анализа.

Говоря языком регрессионного анализа, значения временного ряда x_t рассматривают как отклик (зависимую переменную), а время t — как фактор, влияющий на отклик (независимую переменную):

$$x_{t_i} = f(t_i, \theta) + \varepsilon_i, \qquad i = 1, \ldots, n$$

где f — функция тренда (она обычно предполагается гладкой), θ — неизвестные нам параметры (параметры модели временного ряда), а ε_i —

независимые и одинаково распределенные случайные величины, распределение которых мы предполагаем нормальным. Метод наименьших квадратов состоит в том, что мы выбираем функцию тренда так, чтобы

$$\sum_{i=1}^{n}[x_{t_i} - f(t_i, \theta)]^2 \to \min_{\theta}$$

Для временных рядов типично, что статистические предпосылки регрессионного анализа, как они перечислены в (8.2), (8.3), выполняются не полностью. Это особенно касается предположения о независимости случайных отклонений. Для временных рядов характерна именно взаимная зависимость его членов (по крайней мере, не далеко отстоящих по времени). Тем не менее, оценки тренда и в этих условиях обычно оказываются разумными, если выбрана адекватная модель тренда и если среди наблюдений нет больших выбросов. Упомянутые выше нарушения предпосылок регрессионного анализа сказываются не столько на значениях оценок, сколько на их статистических свойствах. Так, при наличии заметной зависимости между членами временного ряда оценки дисперсии, основанные на остаточной сумме квадратов (8.10), дают неправильные результаты. Неправильными оказываются и доверительные интервалы для коэффициентов модели, и т.д. В лучшем случае их можно рассматривать как очень приближенные.

Это положение может быть частично исправлено, если применять модифицированные алгоритмы метода наименьших квадратов, такие как взвешенный метод наименьших квадратов [50], или метод наименьших квадратов для коррелированных наблюдений [27]. Однако для этих методов требуется дополнительная информация о том, как меняется дисперсия наблюдений или их корреляция. Если же такая информация недоступна, исследователям приходится применять классический метод наименьших квадратов, несмотря на указанные недостатки.

Пример 1. Проиллюстрируем применение метода наименьших квадратов для данных об урожайности зерновых в СССР, представленных в табл. 1.2 и на рис. 11.1а. Визуальное изучения графика данных позволяет предположить, что тренд этого ряда может быть задан в виде прямой линии $tr_t = b_0 + b_1 \cdot t$. С помощью метода наименьших квадратов по формулам, аналогичным (8.7) и (8.8), находим, что

$$\hat{b}_0 = \overline{x} \qquad (\text{где } \overline{x} = \frac{1}{n}\sum_{t=1}^{n} x_t), \tag{12.1}$$

$$\hat{b}_1 = \frac{\sum_{t=1}^{n}(x_t - \overline{x})\,(t - \frac{t(t+1)}{2})}{\sum_{t=1}^{n}(t - \frac{t(t+1)}{2})^2}. \tag{12.2}$$

В формуле (12.2) в качестве независимой переменной фигурирует время t.

Для данных рассматриваемого ряда $\hat{b}_0 = 5.868$, $\hat{b}_1 = 0.275$. При этом коэффициент $\hat{b}_0$ показывает среднюю урожайность зерновых в начальный (1945 г.) момент времени рассматриваемого ряда, а коэффициент $\hat{b}_1$ дает оценку среднегодового прироста урожайности. Подробная таблица результатов процедуры выделения тренда, а также дальнейший анализ ряда, приведены в главе 13, где рассматривается решение этой задачи с помощью статистических пакетов ЭВРИСТА и SPSS.

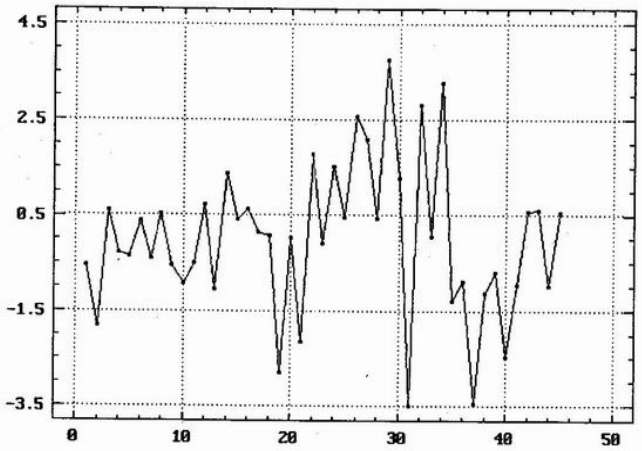


Рис. 12.1. График ряда остатков, полученный в результате удаления из ряда урожайности зерновых в СССР линейного тренда

На рис. 12.1 дан график остатков временного ряда после удаления из него подобранной модели тренда $tr_t = 5.868 + 0.275 \cdot t$. Дальнейший анализ полученного ряда (см. главу 13) показывает, что его уже можно рассматривать как последовательность независимых случайных величин. Более того, для описания остатков можно применять гауссовскую модель, согласно которой их совокупность можно рассматривать как выборку из некоторой нормальной совокупности (с нулевым средним). Последнее означает, что на базе подобранной модели тренда и модели случайной составляющей (независимые ошибки) можно осуществлять прогноз будущих значений ряда и строить доверительную зону для прогноза, используя оценку (8.11) дисперсии остатков.

Пример 2. Поведение случайной компоненты, которое мы наблюдали в примере 1 — это скорее исключение, чем правило. Чтобы убедиться в этом, рассмотрим поведение случайной компоненты у курса доллара весной и летом 1994 г. (рис. 11.1б). График этого ряда позволяет предположить, что его тренд также описывается простой линейной за-

висимостью. Найдя с помощью метода наименьших квадратов значения оценок коэффициентов $\hat{b}_0 = 1675.33$ и $\hat{b}_1 = 3.722$, и, вычтя значения тренда из рассматриваемого временного ряда, получим остаточную компоненту. Ее график приведен на рис. 12.2.

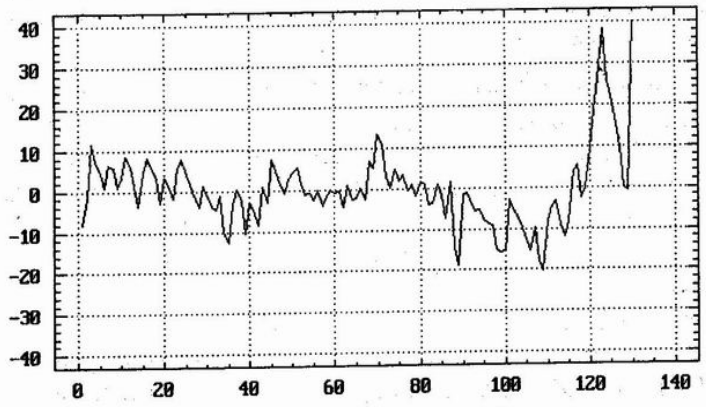


Рис. 12.2. График ряда остатков, полученный в результате удаления из ряда курса доллара линейного тренда

Даже визуальный анализ показывает, что остатки ведут себя не как последовательность независимых одинаково распределенных случайных величин (сравните рис. 12.2, например, с графиком гауссовского белого шума на рис. 11.1д). Действительно, приведенные на рис. 12.3 графики выборочной автокорреляционной функции и выборочной частной автокорреляционной функции (см. п. 14.3) показывают, что соседние значения этого ряда сильно зависимы при значениях лага от 1 до 5. При дальнейшем увеличении значения лага зависимость исчезает. Как следует интерпретировать графики указанных функций и что они означают, мы подробно расскажем ниже в п. 12.4.

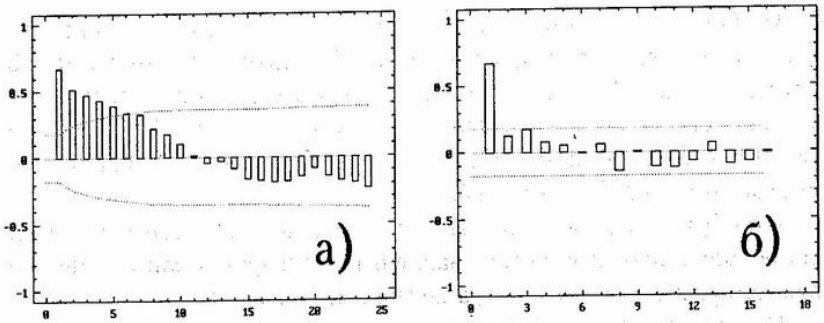


Рис. 12.3. а) — выборочная автокорреляционная функция ряда остатков, полученный в результате удаления из ряда курса доллара линейного тренда; б) — выборочная частная автокорреляционная функция того же ряда

Замечание. Как будет показано ниже в п. 12.4, такое поведение выборочной автокорреляционной и частной автокорреляционной функций характерно для процессов авторегрессии первого порядка (см. п. 11.7 и 14.1). У этих процессов значение в момент времени t формируется из их значения в предыдущий момент времени с некоторым весовым коэффициентом (в данном случае этот коэффициент равняется примерно 0.7) и независимой случайной добавки — белого шума.

Простые разностные операторы. Наряду с методом наименьших квадратов, для удаления тренда можно использовать и ряд других методов. Одним из них является метод перехода от исходного ряда к ряду разностей соседних значений ряда. В более общем виде эта идея описывается с помощью применения к ряду разностных операторов различных порядков. Эти методы сведения временного ряда к стационарному являются частным случаем общего метода, предложенного Дж.Боксом и Г.Дженкинсом в 1970 году [17]. В целом, мы относимся к разностным методам критически, но считаем нужным упомянуть о них. Они часто обсуждается в литературе и представлены во многих статистических пакетах.

Определение. *Процедура перехода от ряда x_t при $t = 1, \ldots, n$ к ряду $y_t = x_t - x_{t-1} = \nabla x_t$ при $t = 2, \ldots, n$ называется взятием первых разностей, а оператор ∇ называется простым разностным оператором первого порядка.*

Заметим, что длина ряда первых разностей y_t на единицу меньше, чем длина исходного ряда x_t. Покажем, как действует разностный оператор на временном ряде x_t, содержащем простой линейный тренд $tr_t = b_0 + b_1 \cdot t$:

$$y_t = \nabla x_t = x_t - x_{t-1} =$$
$$= b_0 + b_1 t + \varepsilon_t - b_0 - b_1(t-1) - \varepsilon_{t-1} = b_1 + \varepsilon_t - \varepsilon_{t-1} \qquad (12.3)$$

Из (12.3) видно, что в отличие от ряда x_t, преобразованный ряд y_t уже не содержит тренда, однако структура случайной компоненты в нем уже другая. Так, если ε_t была последовательностью независимых случайных величин, то последовательность $\varepsilon_t - \varepsilon_{t-1}$, $t = 2, \ldots, n$, этим свойством уже не обладает. Корреляция между соседними членами этой последовательности равна -0.5.

Итак, удалить линейный тренд из временного ряда можно разными способами: с помощью метода наименьших квадратов или с помощью простого разностного оператора первого порядка.

На рис. 12.4 приведен график ряда, полученного в результате применения разностного оператора ∇ к ряду урожайности зерновых. Дальнейшие исследования показывают, что в данном случае проще анализи-

ровать ряд остатков, полученный после удаления линейного тренда методом наименьших квадратов (рис. 12.1), чем ряд первых разностей. А в общем случае, к сожалению, нельзя сказать, какой из этих двух методов удаления тренда предпочтительней. Все зависит от заранее неизвестной структуры случайной компоненты временного ряда ε_t. Так, для временного ряда с независимыми приращениями проще анализировать ряд его первых разностей. Он будет представлять из себя просто белый шум.

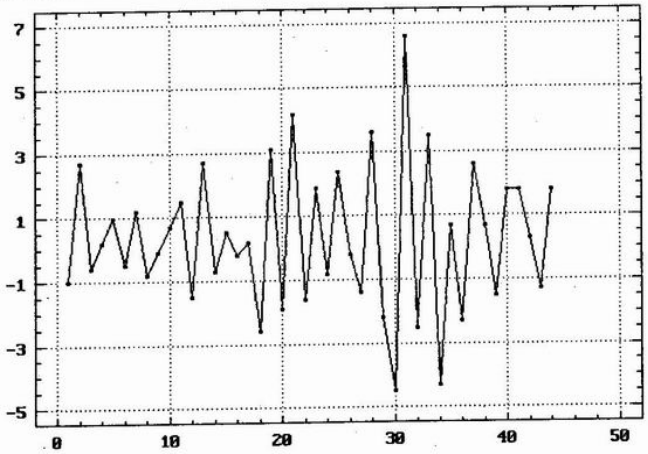


Рис. 12.4. Ряд первых разностей для урожайности зерновых

Аналогичным образом можно ввести разностный оператор второго и более высоких порядков. Так, простой разностный оператор второго порядка преобразует ряд x_t к ряду y_t, где

$$y_t = \nabla^2 x_t = \nabla(\nabla x_t) = \nabla(x_t - x_{t-1}) = \nabla x_t - \nabla x_{t-1} = \\ = x_t - 2x_{t-1} + x_{t-2}.$$

Часто для записи разностных операторов используют оператор B «сдвига назад»: $Bx_t = x_{t-1}$. При этом

$$\nabla x_t = (1-B)x_t, \quad \nabla^2 x_t = (1-B)^2 x_t, \quad \nabla^k x_t = (1-B)^k x_t.$$

Ясно, что длина ряда $\nabla^k x_t$ на k единиц меньше длины исходного ряда.

Простые разностные операторы более высоких порядков позволяют удалять из ряда полиномиальные тренды соответствующих порядков.

Возможно, разностные операторы действительно пригодны для удаления трендов, особенно если не видна подходящая аналитическая модель тренда. Беда же метода разностных операторов в том, что не всегда ясно, как приложить к исходному временному ряду результаты статистического анализа его разностей. В частности, это относится к законам распределения ошибок. К тому же эти разности могут иметь (и

часто имеют) гораздо более сложную статистическую структуру, нежели исходный ряд. Рассмотрите, например, первые разности для процесса авторегрессии первого порядка. (Об авторегрессии см. п. 14.1.)

12.3.2. Выделение сезонных эффектов

Многие временные ряды, особенно экономические, содержат *сезонные компоненты*. Сезонные компоненты ряда могут как представлять интерес сами по себе, так и выступать в роли мешающего фактора. В обоих случаях задача исследователя — выделить и устранить их из ряда.

Для этого есть несколько способов. Их выбор обычно определяется моделью подбираемого временного ряда. Ниже мы рассмотрим две наиболее распространенных модели описания экономических временных рядов. Первая из них включает в себя тренд (tr_t), сезонную (s_t) и случайную (ε_t) компоненты:

$$x_t = tr_t + s_t + \varepsilon_t \tag{12.4}$$

Вторая модель, кроме перечисленных выше компонент, включает еще и циклическую компоненту (c_t):

$$x_t = tr_t + s_t + c_t + \varepsilon_t \tag{12.5}$$

Циклическая компонента c_t в экономических временных рядах отражает периоды роста и спада экономической активности различной амплитуды и продолжительности. (Более подробно о каждой из компонент модели временного ряда рассказано в п. 11.5.)

Сезонные эффекты на фоне тренда. Предположим, что рассматриваемый временной ряд $x_1, \ldots, x_n$ может быть описан аддитивной моделью (12.4). Пусть p — период последовательности s_t, так что $s_t = s_{t+p}$ для всякого t. Наша задача — оценить значения s_t по наблюдениям x_t при том, что величина p известна.

Для этого сначала мы должны оценить тренд tr_t. Это можно сделать с помощью метода наименьших квадратов или его модификаций. Обозначим через $\hat{tr}_t$ полученную оценку тренда. Обычно она выражается в виде некоторой достаточно гладкой функции зависящей от времени t и одного или нескольких неизвестных параметров. Оценки этих параметров и дает метод наименьших квадратов. Наиболее распространенные функции тренда приведены в п. 11.6.

Затем для каждого сезона i, $1 \leqslant i \leqslant p$, рассмотрим все относящиеся к нему разности

$$x_i - \hat{tr}_i, \quad x_{i+p} - \hat{tr}_{i+p}, \quad \ldots, x_{i+mp} - \hat{tr}_{i+mp}. \tag{12.6}$$

(для простоты изложения мы предполагаем, что в рассматриваемом ряде содержится целое число периодов, т.е. $n = (m+1)p$.) Каждое из этих отклонений x_i от $\hat{tr}_i$ можно рассматривать как результат влияния сезонных изменений. Усреднение этих разностей дает нам оценку сезонной компоненты s_i. В качестве простейшей оценки можно взять простое среднее, т.е. положить

$$\hat{s}_i = \frac{1}{m+1}\sum_{l=0}^{m}(x_{i+lp} - \hat{tr}_{i+lp}) \qquad \text{для} \qquad i = 1,\ldots,p \tag{12.7}$$

В качестве других оценок $\hat{s}_i$ можно взять взвешенное среднее, цензурированное среднее, медиану и т.д. Перечисленные средние уменьшают влияние резко выделяющихся наблюдений.

Часто бывает желательно, чтобы сумма сезонных эффектов равнялась нулю. Тогда переходят к скорректированным оценкам сезонных эффектов в виде

$$s_i^* = \hat{s}_i - \frac{1}{p}\sum_{i=1}^{p}\hat{s}_i. \tag{12.8}$$

В практических задачах распространена ситуация, когда сезонные колебания пропорциональны среднему значению процесса в рассматриваемый момент времени. Для описания подобных данных можно использовать одну из следующих моделей:

$$x_t = tr_t \cdot s_t + \varepsilon_t$$
$$x_t = tr_t \cdot s_t \cdot \varepsilon_t.$$

Первая из них является смешанной мультипликативно-аддитивной моделью, вторая — мультипликативной моделью временного ряда. Для модели $x_t = tr_t \cdot s_t + \varepsilon_t$ при оценке сезонных эффектов вместо совокупности (12.6) рассматривают совокупность (12.9) частных от деления x_{i+lp} на $\hat{tr}_{i+lp}$, выраженных в процентах.

$$\frac{x_{i+lp}}{\hat{tr}_{i+lp}} \cdot 100\% \qquad \text{при} \quad l = 0,1,2,\ldots,m \tag{12.9}$$

В этом случае оценкой сезонной компоненты или *сезонным индексом* называют величину:

$$\hat{s}_i = \frac{1}{m+1}\sum_{l=0}^{m}\left(\frac{x_{i+lp}}{\hat{tr}_{i+lp}} \cdot 100\%\right) \qquad \text{где} \quad 1 \leqslant i \leqslant p \tag{12.10}$$

Так же, как и в случае аддитивной модели, вместо среднего арифметического в правой части (12.10) может фигурировать взвешенное или

цензурированное среднее, медиана или другие более устойчивые к грубым выбросам оценки. Сезонные индексы (12.10) особенно популярны при анализе экономических временных рядов. Оценка сезонного индекса для мультипликативной модели будет рассмотрена ниже в более общей ситуации.

На практике считается, что оценки сезонных эффектов недостаточно точны, если число периодов в исследуемом сезонном временном ряде меньше пяти-шести. Это означает, например, что при рассмотрении месячных данных для достаточно точной оценки сезонных эффектов необходимы, как минимум, наблюдения за пять-шесть лет.

Удаление сезонной компоненты. Получив оценки сезонных эффектов (12.7), в аддитивной модели легко провести удаление этих эффектов из рассматриваемого ряда, вычитая их из начальных значений ряда. Подобная процедура часто носит название *сезонного выравнивания ряда* или *сезонной коррекции ряда*. Еще одно название этой процедуры — *сезонная декомпозиция*. Для мультипликативно-аддитивной модели эта процедура сводится к делению значений исходного ряда на соответствующие сезонные индексы и умножению на 100%.

Проиллюстрируем оценку сезонных индексов и их использование при прогнозировании на основе данных о производстве молока в России.

Пример. В таблице 12.1 и на рис. 12.5 приведены величины месячного производства молока (в тыс. тонн) в России с января 1992 г. по октябрь 1996 г. (по данным ЦСУ Госкомстата России).

Таблица 12.1

Производство молока в России с января 1992 г. по октябрь 1996 г. (тыс. тонн в месяц)

Месяц \ год	1992	1993	1994	1995	1996
январь	2015	1759	1510	1172	1038
февраль	2123	1773	1484	1226	1104
март	2624	2361	1988	1651	1439
апрель	2891	2649	2211	1859	1521
май	3335	3203	2559	2392	1827
июнь	4071	3936	3209	2864	2446
июль	4040	3861	3204	2714	2369
август	3392	3321	2687	2420	2081
сентябрь	2467	2438	2031	1925	1577
октябрь	2092	1760	1506	1338	1081
ноябрь	1494	1299	1050	984	
декабрь	1562	1345	1054	1020	

График ряда показывает, что производство молока имеет тенденцию к сокращению, обусловленную сокращением поголовья молочного ста-

да, и подвержено сильным сезонным колебаниям с максимумом производства в летние месяцы и минимумом — в зимние. При этом величина сезонных колебаний пропорциональна среднему уровню производства.

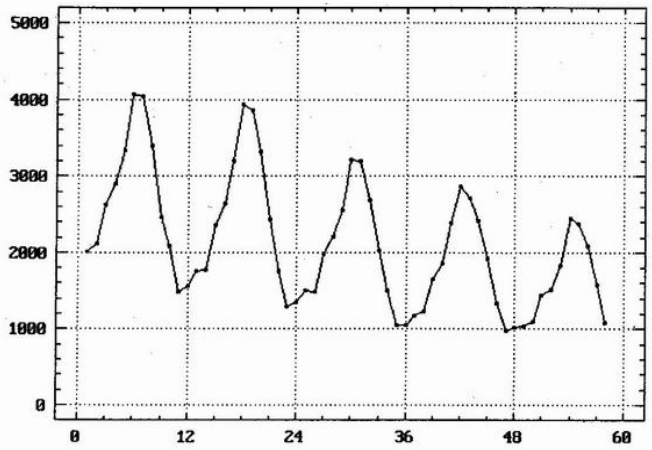


Рис. 12.5. Ежемесячное производство молока в России с 01.1992 по 10.1996 (в тыс. тонн)

Оценим сезонные индексы этого ряда и проведем выравнивание ряда с учетом сезонности. Для описания тренда используем линейную модель $tr_t = a + b \cdot t$, где $t = 1, 2, \ldots, 58$. Оценки неизвестных коэффициентов a и b методом наименьших квадратов есть: $\hat{a} = 2841.1$ и $\hat{b} = -23.63$. Таким образом, в каждой точке t можно вычислить $\hat{tr}_t = \hat{a} + \hat{b} \cdot t$. Подобранная модель тренда описывает общую тенденцию поведения ряда. Но сделать на базе этой модели достаточно точный прогноз ежемесячного производства молока в следующем году нельзя, учитывая большую сезонную изменчивость ряда. Для построения месячного прогноза необходимо оценить сезонные эффекты, или сезонные индексы.

На графике 12.5 видно, что величина сезонных колебаний пропорциональна среднему уровню производства. Поэтому для описания сезонных колебаний следует использовать мультипликативно-аддитивную или мультипликативную модель. Воспользуемся первой из этих моделей. Для получения оценок сезонных индексов используем формулы (12.9) и (12.10).

В таблице 12.2 приведены в процентах значения отношений $x_t/\hat{tr}_t$ для каждого месяца t. Обратим внимание на то, что полученные для каждого месяца индексы в таблице 12.2 довольно устойчивы. Так, производство молока в июне в среднем на 155% превышает среднегодовой уровень, а в октябре — составляет только 75% от него. Для получения сезонных индексов производства молока (12.10) для каждого

Таблица 12.2

Значения отношений $x_t/\hat{tr}_t$ для временного ряда с данными о производстве молока в России (в %)

Месяц \ год	1992	1993	1994	1995	1996
январь	71.52	69.42	67.10	59.59	61.67
февраль	75.99	70.63	66.65	63.09	66.52
март	94.72	94.95	90.23	86.01	87.96
апрель	105.26	107.55	101.45	98.05	94.34
май	122.48	131.30	118.70	127.76	115.00
июнь	150.82	162.93	150.50	154.93	156.29
июль	150.99	161.41	151.95	148.71	153.69
август	127.90	140.22	128.88	134.34	137.107
сентябрь	93.86	103.97	98.53	108.28	105.54
октябрь	80.31	75.82	73.91	76.28	73.51
ноябрь	57.88	56.54	52.13	56.86	
декабрь	61.07	59.15	52.95	59.75	

месяца следует провести усреднение данных по строкам таблицы 12.2. Полученный результат приведен в таблице 12.3.

Таблица 12.3

Сезонные индексы производство молока в России (в %)

Месяц	Индекс по всем данным	Индекс по данным 1992–1995 гг.
январь	65.86	66.96
февраль	68.58	69.23
март	90.77	91.84
апрель	101.33	103.64
май	123.05	126.01
июнь	155.09	156.18
июль	153.35	154.83
август	133.69	134.46
сентябрь	102.04	102.64
октябрь	75.97	77.73
ноябрь	55.85	56.80
декабрь	58.23	59.31

Для проведения сезонного выравнивания каждое значения исходного ряда следует разделить на соответствующий ему сезонный индекс и умножить полученный результат на 100%. Полученный результат приведен на рис. 12.6. Как видно из графика, выровненный ряд имеет ярко выраженную тенденцию линейного убывания.

Замечание. Для прогнозирования поведения рассмотренного ряда могут быть применены и другие методы. В частности, можно описывать этот ряд моделью, использующей простые и сезонные разностные операторы. Рассматривая

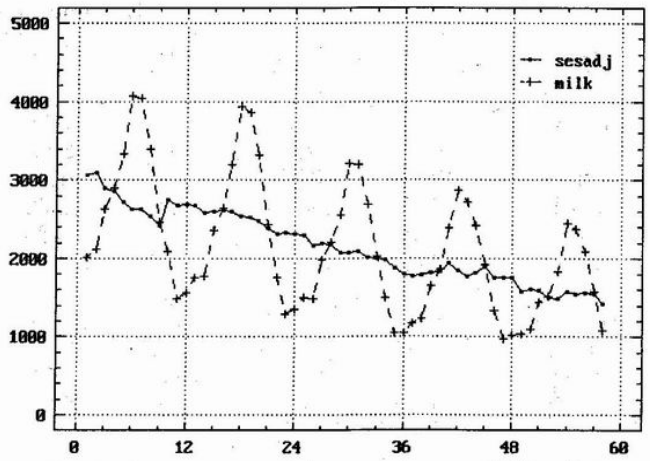


Рис. 12.6. Сезонное выравнивание ряда производства молока в России

этот пример, как иллюстрацию вычисления сезонных индексов, мы не касаемся в нем вопросов выбора наилучшей модели. Эта обширная тема, требующая определенной подготовки, выходит за рамки данной книги.

Прогнозирование. Посмотрим к каким результатам привела бы эта методика, если бы мы хотели получить прогноз на 1996 г. по данным 1992–1995 гг. Учитывая устойчивое поведение сезонного индекса в данной задаче, мы проведем его оценку по данным за 4 года. (В других задачах такой объем данных может быть недостаточным.) Повторим описанные выше действия для данных 1992–1995 гг. Подобранная модель линейного тренда $\hat{tr}_t$ имеет вид:

$$\hat{tr}_t = 2899.9 - 26.64 \cdot t \qquad (12.11)$$

Ее коэффициенты в целом не сильно отличаются от коэффициентов, полученных выше по всем данным. (Следует учитывать, что оценка $\hat{b}$ по всем данным несколько завышена. Это связано с отсутствием данных последних двух месяцев 1996 г., которые с учетом сезонности являются обычно самыми низкими в году. Здесь уместно заметить, что использование метода наименьших квадратов для подбор модели тренда сезонных рядов с незавершенными циклами, как это было сделано выше, обычно влечет за собой подобные смещения оценок. Поэтому лучше этого избегать или использовать устойчивые методы оценивания.)

Оценки сезонных индексов для данных 1992–1995 гг., рассчитанные по (12.9), приведены в таблице 12.4. Сравнение данных таблиц 12.2 и 12.4 показывает, что они хорошо согласуются между собой. Для получения сезонного индекса $\hat{s}_t$ усредним по строкам данные таблицы 12.4. Полученный результат приведен в третьем столбце табл. 12.3. Из этой

Таблица 12.4

Месяц \ год	1992	1993	1994	1995
январь	70.13	68.88	67.59	61.22
февраль	74.58	70.16	67.23	64.95
март	93.05	94.43	91.16	88.71
апрель	103.50	107.09	102.64	101.34
май	120.54	130.89	120.29	132.32
июнь	148.57	162.62	152.75	160.80
июль	148.89	161.29	154.47	154.69
август	126.25	140.30	131.23	140.06
сентябрь	92.74	104.17	100.50	113.16
октябрь	79.44	76.06	75.52	79.90
ноябрь	57.31	56.79	53.37	59.71
декабрь	60.54	59.49	54.30	62.91

таблицы видно хорошее согласие сезонных индексов, что говорит об устойчивости этого показателя.

Для осуществления прогноза ряда на 10 месяцев 1996 г. следует сначала рассчитать предварительный ежемесячный прогноз $prog_t$ по подобранной модели тренда (12.11). А именно, вычислить значения $\hat{tr}_t$ для следующих 10-ти значений t, то есть для $t = 49, 50, 51, \ldots, 58$ Результаты этого прогноза приведены во втором столбце таблицы 12.5. Для получения окончательного прогноза ряда надо скорректировать предварительный прогноз с помощью полученных сезонных индексов, вычислив $(prog_t \cdot s_t)/100$ для указанных выше значений t. Результаты этой процедуры приведены в третьем столбце табл. 12.5. В четвертом столбце таблицы приведены реальные данные за 10 месяцев 1996 г. Наглядное сравнение прогноза с реальными данными дано на рис. 12.7, где пунктирной линией обозначены реальные данные за 1995–1996 гг., а сплошной линией — построенный прогноз на 1996 г.

В табл. 12.5 и на рис. 12.7 видно хорошее согласие прогноза с реальными данными в первые 5 месяцев. Относительная погрешность прогноза здесь не превышает 3%. В последующие 4 месяца прогноз ниже реальных данных на 6–10%. Относительная ошибка прогноза в последний, 10-ый месяц не превышает 3%. Заметим, что наибольшие различия прогноза с реальными данными наблюдаются в летние месяцы, сезонный индекс которых подвержен наибольшим колебаниям (см. табл. 12.2 и 12.4). Эти колебания из года в год имеют порядок 10–20%. С учетом этого, можно признать в целом хорошее согласие прогноза с реальными данными.

Аналогичным образом можно осуществить прогноз на 1997 г. по данным 1992–1996 гг. Говорить о достоверности подобного прогноза можно лишь при сохранении общих тенденций, наблюдаемых в пре-

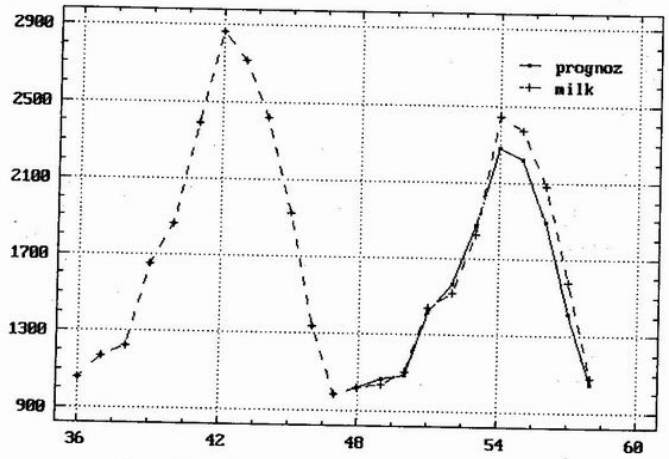


Рис. 12.7. Производство молока в России за 1995–1996 гг. и прогноз на 1996 г. (в тыс. тонн)

Таблица 12.5

Прогноз производства молока в России на 1996 г. (в тыс. тонн) и его сравнение с реальными данными

Месяц	Тренд	Прогноз на 1996 г.	Реальные данные
январь	1595	1068	1038
февраль	1568	1085	1104
март	1541	1415	1439
апрель	1515	1569	1521
май	1488	1875	1827
июнь	1461	2282	2446
июль	1435	2221	2369
август	1408	1893	2081
сентябрь	1381	1418	1577
октябрь	1355	1053	1081

дыдущие годы. Для более аргументированных прогнозов необходимо привлечение дополнительной информации (например, о тенденциях изменения поголовья молочного стада).

12.3.3. Метод скользящих средних

При наличии в ряде циклической компоненты расчет сезонных эффектов несколько отличается от описанного выше. В этом случае для выяснения сезонных вкладов в виде (12.6) или (12.9) необходимо оценить не только тренд, но и циклическую компоненту. Проще всего одновременно оценить тренд и циклическую компоненту можно с помо-

щью скользящего среднего. Этот метод полезен и тогда, когда модель тренда не ясна. Рассмотрим его подробнее.

О методе скользящих средних. Метод скользящих средних — один из самых старых и широко известных способов сглаживания временного ряда. Он основан на переходе от начальных значений ряда к их средним значениям на интервале времени, длина которого выбрана заранее. При этом сам выбранный интервал времени скользит вдоль ряда.

Получаемый таким образом ряд скользящих средних ведет себя гораздо более гладко, чем исходный ряд, за счет усреднения отклонений исходного ряда. Таким образом эта процедура дает представление об общей тенденции поведения ряда. Ее применение особенно полезно для рядов с сезонными колебаниями и неясным характером тренда. В частности, переход к ряду скользящих средних может быть использован для выявления сезонной компоненты (или сезонного индекса) временного ряда.

Вид средних. Применяя метод скользящих средних, можно использовать различные виды усреднения значений ряда: среднее арифметическое (простое или с некоторыми весами), медианы и др. К сглаживанию с помощью медианы (медианное сглаживание) прибегают тогда, когда среди наблюдений есть выбросы (резко выделяющиеся данные).

Примеры для обсуждения. Мы дадим формальные определения метода скользящих средних, используя для их иллюстрации два следующих примера. В первом из них величина интервала сглаживания равна 7, по числу дней недели. Во втором примере величина интервала сглаживания равна 12, что соответствует двенадцати месяцам года. Это типичные интервалы сглаживания в экономических временных рядах. Для ежеквартальных данных подходящим может оказаться сглаживание с интервалом 4, для почасовых данных, собираемых круглосуточно, сглаживание с интервалом 24, и т.д. Вообще говоря, величину интервала сглаживания целесообразно выбирать равным или кратным периоду сезонности. При этом каждый интервал вычисления скользящего среднего будет содержать данные, отвечающие всему периоду (периодам) сезонности.

Пример 1. На рис. 12.8а приведен среднесуточный трафик (величина загрузки) телекоммуникационного канала Париж-Москва сети Internet за четыре последовательных недели (февраль 1996 г.). Этот график характеризует интенсивность (в килобитах в секунду) получения информации западными пользователями с российских компьютерных серверов по указанному каналу. Из графика видно, что в отдельные дни недели (субботу и воскресенье) происходит уменьшение загрузки кана-

ла, в другие дни загрузка повышается. Кроме того, вероятно, имеет место плавный рост объема загрузки с начала месяца к его концу. Таким образом, можно предположить, что рассматриваемый временной ряд имеет тренд и сезонную компоненту с периодом сезонности $p = 7$ дней.

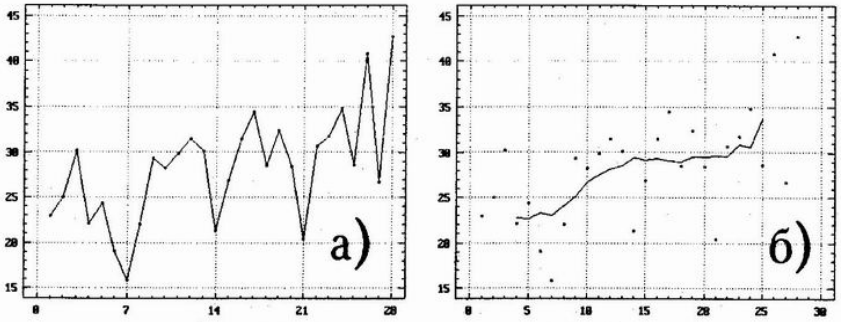


Рис. 12.8. Среднесуточный трафик в Кбит/сек телекоммуникационного канала Париж-Москва в феврале 1996 г.: а) исходный ряд; б) исходный ряд и его скользящее среднее

Пример 2. На рис. 12.9а приведен график ежемесячных продаж шампанского за ряд лет. На графике отчетливо прослеживаются сезонные колебания с пиками в декабре каждого года и спадами в летние месяцы. Период сезонности этих данных равен 12.

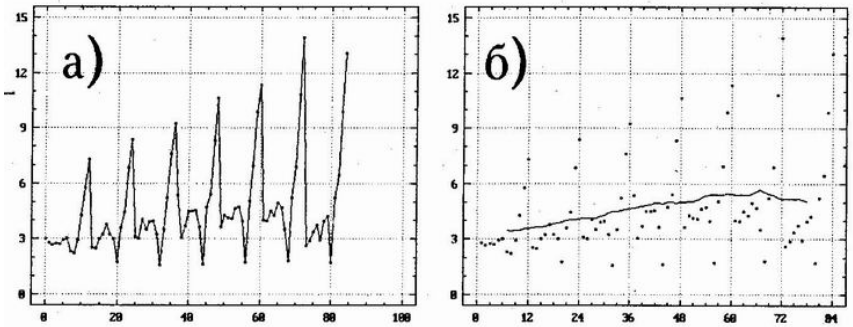


Рис. 12.9. Месячный объем продаж шампанского за ряд лет: а) исходный ряд; б) исходный ряд и его скользящее среднее

Вычисление скользящего среднего. Дадим формальное определение скользящего среднего сначала для интервалов сглаживания, длина которых выражается нечетными числами. Причина, по которой четные и нечетные длины рассматриваются порознь, выяснится чуть ниже. Пусть $p = 2m + 1$. Обозначим через $\hat{x}_t$ результат усреднения элементов ряда

$$x_{t-m}, \ldots, x_{t-1}, x_t, x_{t+1}, \ldots, x_{t+m}.$$

Если обсуждаемое среднее есть среднее арифметическое, то

$$\hat{x}_t = \frac{1}{2m+1}(x_{t-m} + \ldots + x_{t-1} + x_t + x_{t+1} + \cdots + x_{t+m}).$$

Для медианного сглаживания

$$\hat{x}_t = \mathrm{med}(x_{t-m}, \ldots, x_{t-1}, x_t, x_{t+1}, \ldots, x_{t+m}).$$

Для четных $p = 2m$ определение несколько сложнее. Причина в том, что вычисленное по аналогичным формулам (как среднее арифметическое, медиана и т.д.) усредненное значение нельзя сопоставить какому-либо определенному моменту времени t. Например, среднее арифметическое $\frac{1}{2m}\sum_{t=1}^{t=2m}(x_t)$ следовало бы сопоставить моменту времени $t = (2m+1)/2$, но такого момента во временном ряде нет. А это сильно осложняет дальнейшее выделение сезонных эффектов.

Поэтому при четном интервале сглаживания $2m$ в усреднении задействуют не $2m$, а $2m+1$ значений временного ряда, но значения на краях интервала сглаживания берут с весами $1/2$. Так, при использовании для усреднения среднего арифметического получается следующая формула:

$$\hat{x}_l = \frac{1}{2m}\left(\frac{1}{2}x_{l-m} + x_{l-m+1} + \cdots + x_{l+m-1} + \frac{1}{2}x_{l+m}\right) \qquad (12.12)$$

Выражение (12.12) задает величину простого скользящего среднего $\hat{x}_l$ для $l = m+1, m+2, \ldots, n-m$ при четной величине интервала сглаживания $p = 2m$.

Свойства скользящего среднего. Скользящее среднее, сглаживая исходный ряд, дает представление об общей тенденции поведения ряда — его тренде и циклической компоненте. Сделаем несколько замечаний о его свойствах.

1. При применении метода скользящих средних выбор величины интервала сглаживания должен делаться из содержательных соображений и привязываться к периоду сезонности для сезонных данных. Если процедура скользящего среднего используется для сглаживания несезонного ряда, то чаще всего величину интервала сглаживания выбирают равной трем, пяти или семи. Чем больше интервал усреднения, тем более гладкий вид имеет график скользящих средних.

2. Соседние члены ряда скользящих средних сильно коррелированы, так как в их формировании участвуют одни и те же члены исходного ряда. Это может приводить, к тому, что ряд скользящих средних может содержать циклические компоненты, отсутствующие в исходном ряде. Это явление носит название *эффекта Слуцкого-Юла* (см. [52], [51]).

3. В качестве метода усреднения, кроме упомянутых выше среднего арифметического и медианы, можно рассматривать *взвешенные скользящие средние*, когда значения исходного ряда суммируется с определенными весами. Подобные процедуры целесообразны, если изменение временного ряда во времени носит явно нелинейный характер. Мы не будем более касаться этого вопроса. Он подробно изложен, например, в [51].

Оценка сезонных компонент. Предположим, что наблюдаемый временной ряд имеет структуру $x_t = tr_t + c_t + s_t + \varepsilon_t$, где $tr_t + c_t$ — тренд и циклическая составляющая, s_t — сезонная составляющая, а ε_t — случайная составляющая ряда. Пусть p — период последовательности s_t, так что $s_t = s_{t+p}$ для всякого t. Пусть величина p нам известна. Мы хотим оценить значения s_t по наблюдениям x_t.

Порядок оценки сезонных компонент в этом случае, в целом, аналогичен рассмотренному в п. 12.3.2. Только вместо оценки тренда методом наименьших квадратов мы будем использовать скользящее среднее в качестве совместной оценки тренда и циклической компоненты. Обозначим через $\hat{x}_t$ скользящее среднее с периодом p, построенное по ряду x_t. Для упрощения обозначений начнем нумерацию величин $\hat{x}_t$ с единицы, так что ряд из скользящих средних есть: $\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_k$. Соответственно изменим нумерацию исходного ряда так, чтобы величине $\hat{x}_t$ соответствовал член x_t. (При этом приходится отбросить $[p/2]$ первых членов исходного ряда, для которых значения $\hat{x}_t$ не определены. Здесь через $[p/2]$ обозначена целая часть от деления p пополам.)

Ради простоты предположим, что $k = (m+1)p$, где m — положительное целое число. (Обратим внимание, что общая длина n исходного ряда при этом равна $n = (m+2)p$ при четном p и $n = (m+2)p - 1$ при нечетном p.) Для каждого сезона i, $1 \leqslant i \leqslant p$, рассмотрим все относящиеся к нему разности

$$x_i - \hat{x}_i,\ x_{i+p} - \hat{x}_{i+p},\ \ldots,\ x_{i+mp} - \hat{x}_{i+mp}. \qquad (12.13)$$

Каждое из этих отклонений x_i от $\hat{x}_i$ можно рассматривать как результат влияния сезонных изменений. Усреднение этих разностей дает нам оценку сезонной компоненты s_i. В качестве простейшей оценки можно взять простое среднее, т.е. положить

$$\hat{s}_i = \frac{1}{m+1}\sum_{l=1}^{m+1}(x_{i+lp} - \hat{x}_{i+lp}) \qquad \text{для} \qquad i = 1, \ldots, p \qquad (12.14)$$

Как и выше (см. 12.7), вместо простого среднего можно взять взвешенное среднее, цензурированное среднее, медиану и т.д., для уменьшения влияния резко выделяющихся наблюдений.

Для мультипликативной модели временного ряда, когда $x_t = tr_t \cdot c_t \cdot s_t \cdot \varepsilon_t$ целесообразно перейти к логарифмам $y_t = \log x_t$. Тогда $y_t = d_t + g_t + r_t + \delta_t$, где $d_t = \log tr_t$, $g_t = \log c_t$, $r_t = \log s_t$, $\delta_t = \log \varepsilon_t$. К ряду y_t можно применить изложенную выше методику, начиная с вычисления скользящих средних и кончая составлением оценки $\hat{r}_i$ для r_i. Оценкой для исходной величины $s_i = e^{r_i}$ будет служить $e^{\hat{r}_i}$, если

$\log x$ — натуральный логарифм x, либо $\hat{s}_i = 10^{\hat{r}_i}$, если наши логарифмы десятичные.

Удаление сезонной компоненты. Оно проводится так же, как и в разобранном выше случае. Для аддитивной модели удаление сезонной компоненты сводится к вычитанию оцененной сезонной компоненты из исходного ряда. Для мультипликативной модели эта процедура заключается в делении значений исходного ряда на соответствующие сезонные индексы.

Пример оценки и удаления сезонных компонент с помощью скользящего среднего рассмотрен ниже в главе 13 (пример 13.2к). Этот пример решается с помощью компьютерных программ SPSS и Эвриста.

12.3.4. Сезонные разностные операторы

Еще один способ удаления сезонных компонент из ряда основан на использовании специальных разностных операторов, которые называются *сезонными*. Использование этих операторов особенно распространено в линейных моделях временных рядов типа авторегрессии-скользящего среднего (см. главу 14).

Пусть $x_1, \ldots, x_n$ — реализация временного ряда, а p — период его сезонности.

Определение. *Процедура перехода от ряда x_t (при $t = 1, \ldots, n$) к ряду $y_t = x_t - x_{t-p} = \nabla_p x_t$ (при $t = p+1, \ldots, n$) называется взятием первой сезонной разности, а оператор ∇_p называется сезонным разностным оператором с периодом p.*

Преобразование $x_t - x_{t-p}$ может быть также записано с помощью оператора сдвига назад B в виде:

$$y_t = x_t - x_{t-p} = (1 - B^p)x_t$$

На рис. 12.10 изображен результат применения сезонного оператора ∇_{12} к ряду месячных продаж шампанского за 7 лет. Длина полученного ряда сократилась на 12. Разброс значений полученного ряда существенно сократился и в нем уже не просматриваются периодические колебания.

Для этого ряда теперь можно попытаться подобрать, например, линейную параметрическую модель типа авторегрессии-скользящего среднего (см. гл. 14). В случае успешного подбора модели можно осуществить прогноз для ряда разностей. Этот прогноз может быть пересчитан и для исходного ряда.

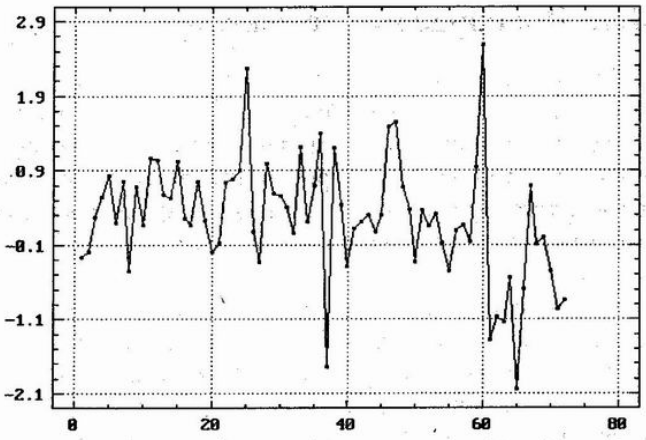


Рис. 12.10. Ряд сезонных разностей для продаж шампанского

Другим способом устранения сезонности может служить метод наименьших квадратов, в котором используется полигармоническая модель (11.5) для описания периодически повторяющихся сезонных эффектов. Сравнивая эти два подхода устранения сезонных эффектов, заметим, что метод сезонных разностей значительно проще и наглядней.

Сезонные операторы более высоких порядков. Как и в случае с простыми разностными операторами (см. п. 12.3.1), иногда бывают полезны сезонные операторы более высоких порядков. Так, сезонный оператор второго порядка с периодом p задается соотношением:

$$\nabla_p^2 x_t = \nabla_p(\nabla_p x_t) = \nabla_p(x_t - x_{t-p}) = x_t - 2x_{t-p} + x_{t-2p}$$

или, с помощью оператора сдвига назад B:

$$\nabla_p^2 x_t = (1 - B^p)^2 x_t = (1 - 2B^p + B^{2p}) x_t.$$

Смешанные разностные операторы. Выше указывалось, что простые и сезонные разностные операторы могут быть использованы соответственно для удаления тренда и сезонной компоненты из временного ряда. Если временной ряд одновременно содержит обе эти компоненты, то их удаление возможно с помощью последовательного применения простых и сезонных операторов. Нетрудно убедиться, что порядок применения этих операторов не существен:

$$\nabla\nabla_p x_t = \nabla(x_t - x_{t-p}) = (x_t - x_{t-1}) - x_{t-p} - x_{t-p-1}) = \nabla_p \nabla x_t.$$

Замечание. Существуют и другие методики оценивания и учета сезонных эффектов. Часть из них опирается на совместное использование методов однофакторного анализа и анализа временных рядов. Другие используют обобщенные сезонные модели процессов авторегрессии-скользящего среднего. Мы не будем останавливаться на этих вопросах.

12.3.5. Преобразование шкалы

К преобразованиям значений временного ряда (точнее — к преобразованиям той шкалы, в которой измерены значения временного ряда) прибегают обычно по двум причинам: либо для того, чтобы приблизить распределение к нормальному (например, избавиться от его скошенности), либо для того, чтобы сделать дисперсию временного ряда более постоянной (иными словами, стабилизировать дисперсию временного ряда).

Пусть переменная x употребляется для записи значений временного ряда. Рассмотрим преобразование x в y по правилу $y = f(x)$, где f обозначает некоторую определенную функцию. (Обычно f — монотонная функция; тогда от значений y можно однозначно вернуться к значениям x.) Применяя преобразование f к каждому члену ряда x_t, мы получим новый временной ряд $y_t = f(x_t)$.

Логарифмическое преобразование. Чаще других используемое преобразование — логарифмическое, когда

$$y = \log x, \qquad \text{либо} \qquad y = \log(x + c),$$

где c — некоторая постоянная величина, выбор которой находится в распоряжении исследователя. При логарифмическом преобразовании

$$y_t = \log(x_t + c).$$

Логарифмическое преобразование можно применять только к положительным величинам. В тех случаях, когда часть членов ряда x_t отрицательна, перед переходом к логарифмам ко всем членам ряда прибавляют постоянную c, добиваясь того, чтобы $x_t + c > 0$ при всех t.

Посмотрим, как действует логарифмическое преобразование на практике. Скошенные (асимметричные) распределения довольно часто появляются в экономической статистике. Типичным примером являются данные о душевом доходе: лиц с небольшими и средними доходами гораздо больше, чем лиц с высокими доходами. А этих последних значительно больше, чем лиц с очень высокими доходами. Примерная гистограмма распределения доходов приведена на рис. 12.11а. Прологарифмируем данные о доходах и вновь построим гистограмму. Она приведена на рис. 12.11б. Видно, что в логарифмической шкале распределение доходов близко к нормальному (гауссовскому).

Логарифмическое преобразование может оказаться полезным и при некоторых нарушениях стационарности наблюдаемого ряда. Допустим, что мы наблюдаем процесс $x_t = b_t \cdot z_t$, где z_t — стационарный ряд, а b_t — некоторая положительная неслучайная последовательность. Обозначив

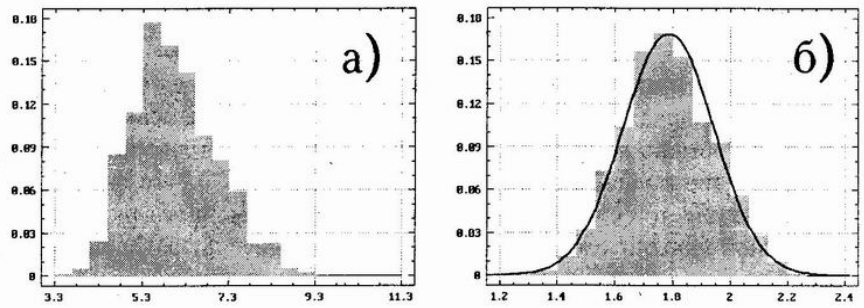


Рис. 12.11. Гистограмма данных о доходах: а) исходная шкала, б) логарифмическая шкала (для наглядности на график (б) наложена функция плотности нормального распределения)

Dz_t через σ^2, получим, что $Dx_t = \sigma^2 b_t^2$ изменяется во времени. Переход к логарифмической шкале $y_t = \log x_t$ дает

$$y_t = \log b_t + \log z_t.$$

При этом ряд $\log z_t$ — стационарный, его дисперсия во времени не изменяется. Это позволяет применить метод наименьших квадратов для выделения тренда $\log b_t$ из ряда y_t.

Примером временного ряда, дисперсия которого изменяется со временем, является ряд продаж шампанского (рис. 11.1.в). На рис. 12.12 приведены данные о продажах шампанского в логарифмической шкале. Видно, что логарифмирование устранило рост размаха сезонных колебаний значений ряда.

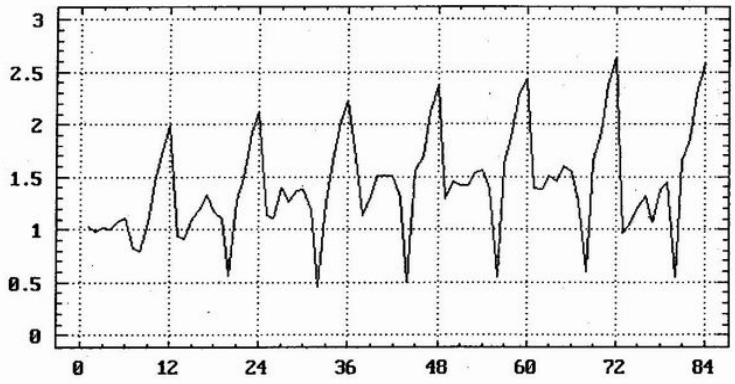


Рис. 12.12. Данные месячных продаж шампанского в логарифмической шкале

Преобразование Бокса-Кокса. Логарифмическое преобразование является частным случаем некоторого семейства преобразований, которое ввели Дж.Бокс и Д.Кокс в 1964 г. [126]. С тех пор эти преобразования приобрели популярность. Преобразования, образующие это

семейство, зависят от параметра λ, $\lambda \geqslant 0$. Если вернуться к формуле преобразований $y = f(x)$, то можно сказать, что теперь $y = f(x, \lambda)$, где значение $\lambda \geqslant 0$ исследователь может выбрать по своему усмотрению. Бокс и Кокс предложили следующую формулу

$$f(x, \lambda) = \begin{cases} (x_t^\lambda - 1)/\lambda & \text{при} \quad \lambda > 0 \\ \log x_t & \text{при} \quad \lambda = 0 \end{cases} \tag{12.15}$$

Нетрудно убедиться, что при фиксированном λ функция $f(x, \lambda)$ монотонно возрастает с ростом x, и что $f(x, \lambda)$ непрерывна не только по x, но и по λ, если $\lambda \geqslant 0$.

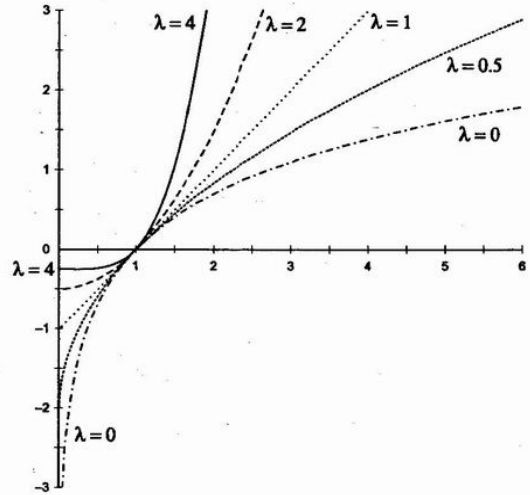


Рис. 12.13. Характер преобразования Бокса-Кокса при различных значениях параметра λ

Как видно из рис. 12.13, преобразование Бокса-Кокса при $\lambda < 1$ растягивает расстояния между малыми значениями и сжимает его между большими по величине значениями данных. При $\lambda > 1$ наблюдается обратная картина.

Следует заметить, что применение преобразования Бокса-Кокса к временным рядам может порождать определенные трудности в их дальнейшем анализе. Дело в том, что показатель степени λ существенно влияет на корреляционную функцию процесса и способен значительно усложнить дальнейший подбор модели ряда.

Ряды, имеющие отрицательные значения. Подобно логарифмическому, преобразование Бокса-Кокса можно применять только к положительным числам. Если часть членов ряда x_t отрицательна, прежде чем применить к ряду

преобразование Бокса-Кокса, ко всем членам ряда прибавляют постоянную c. Члены преобразованного ряда получают по формуле

$$y_t = \frac{(x_t + c)^\lambda - 1}{\lambda}$$

если выбранное $\lambda > 0$. Для $\lambda = 0$ преобразование Бокса-Кокса действует как уже упомянутое логарифмическое: $y_t = \log(x_t + c)$.

12.4. Методы исследования структуры стационарного временного ряда

12.4.1. Цели и методы анализа

Цели анализа. В предыдущих параграфах этой главы мы рассматривали методы выделения из временного ряда детерминированной компоненты — тренда, сезонной и циклической компонент. После удаления детерминированной компоненты временной ряд должен свестись к стационарному процессу. Так что следующим шагом после выделения детерминированной компоненты должен быть анализ остатков, то есть изучение ряда, полученного из исходного временного ряда после исключения детерминированной компоненты. При этом могут ставиться следующие цели.

1. Описание ряда с помощью той или иной модели, которая отражает зависимость между его соседними элементами. На базе построенной модели можно осуществлять прогноз будущего поведения ряда.
2. Уточнение оценки дисперсии временного ряда. Эта оценка важна для прогнозирования, так как исходя из нее вычисляется ширина доверительной трубки прогноза. Привычные оценки дисперсии, которые мы использовали в регрессионном анализе (глава 8), — например, нормированная сумма квадратов отклонений элементов реализации от их среднего, — рассчитаны на независимые случайные величины. Для статистически зависимых данных такие оценки дисперсии временного ряда могут как сильно превышать истинное значение σ^2, так и быть значительно меньше.
3. Проверка стационарности остатков (при нестационарности подбор детерминированной компоненты нуждается в уточнении).

Методы анализа. В качестве модели стационарных временных рядов чаще всего используются процессы авторегрессии, скользящего среднего и их комбинации. Этим моделям посвящена глава 14.

А для проверки стационарности ряда остатков и оценки его дисперсии на практике чаще всего используются выборочная автокорреляционная (коррелограмма, см. п. 11.10) и частная автокорреляционная функция. В пп. 12.4.2 и 12.4.3 мы рассмотрим методы интерпретации графиков этих функций.

Замечания. 1. Для выяснения статистических зависимостей между элементами временного ряда может также быть использована периодограмма (см. п. 11.10).

2. Методы исследования структуры стационарного временного ряда по одной реализации наиболее успешно и полно разработаны для нормально распределенных процессов. Это объясняется тем, что у этих процессов из стационарности в широком смысле, которая поддается определенной проверке, следует стационарность в узком смысле, которая практически не поддается проверке (см. п. 11.3.2).

12.4.2. Интерпретация графика коррелограммы

Анализ коррелограммы — это порой довольно непростая задача. О причинах возникающих при этом трудностей уже говорилось в п. 11.10. Здесь мы кратко остановимся на типичном поведении коррелограммы для некоторых классов временных рядов.

Для начала рассмотрим поведение коррелограммы для некоторых нестационарных рядов. В этом случае следует помнить, что коррелограмма практически не несет никакой информации о статистической зависимости или независимости членов временного ряда, однако она может отражать причины нарушения стационарности. Именно с этой точки зрения мы и рассматриваем два следующих примера.

Наличие тренда. Для временного ряда, содержащего тренд, коррелограмма не стремится к нулю с ростом значения лага k. Ее характерное поведение изображено на рис. 12.14, где коррелограмма построена для ряда урожайности зерновых (рис. 11.1а).

Наличие сезонных колебаний. Для ряда с сезонными колебаниями коррелограмма также будет содержать периодические всплески, соответствующие периоду сезонных колебаний. Это позволяет устанавливать предполагаемый период сезонности. Однако, как было сказано в п. 11.10, отдельные редкие выхода графика коррелограммы за границы доверительной трубки могут наблюдаться и у белого шума. Типичное поведение коррелограммы для ряда с сезонными колебаниями приведено на рис. 12.15, где она построена для данных месячных продаж шампанского в логарифмической шкале (рис. 12.12) после удаления из них линейного тренда.

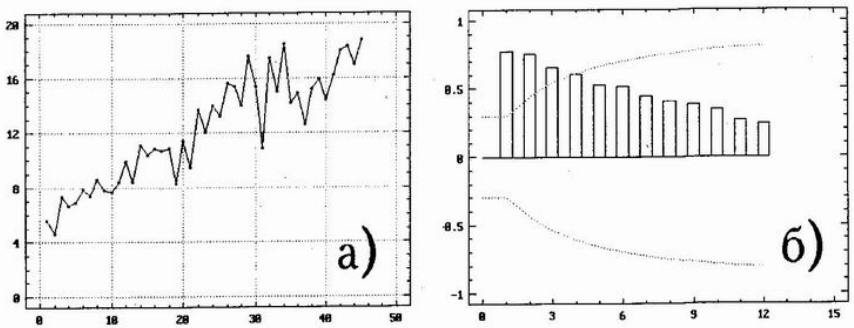


Рис. 12.14. Коррелограмма ряда урожайности зерновых:
а) исходный временной ряд; б) его коррелограмма

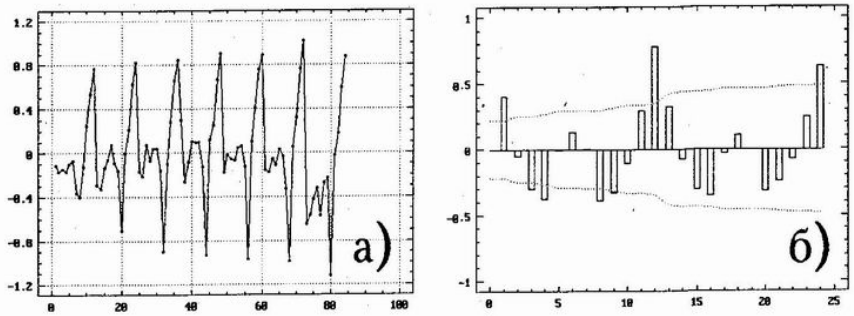


Рис. 12.15. Коррелограмма ряда месячных продаж шампанского
в логарифмической шкале (после удаления линейного тренда):
а) преобразованный ряд продаж шампанского; б) его коррелограмма

Перейдем к рассмотрению коррелограмм стационарных случайных процессов. В этом случае коррелограмма показывает коррелированность значений временного ряда при различных расстояниях между ними.

Коррелограмма белого шума. Как указывалось выше, автокорреляционная функция r_k белого шума равна нулю для всех $k \neq 0$. На рис. 12.16 изображена типичная коррелограмма белого шума. Как указывалось в п. 11.10, для гауссовского белого шума можно указать 95% доверительный интервал для каждого конкретного значения $\overline{r}_k$ в виде $-1/n \pm 2/\sqrt{n}$. Он изображен на графике коррелограммы пунктирными линиями. Если выборочные оценки корреляционной функции попадают в указанные доверительные интервалы, то можно предположить, что значения процесса являются белым шумом. Однако, как уже говорилось, довольно часто одно или несколько значений выборочной автокорреляционной функции белого шума могут выходить из указанных пределов. Особенно часто этот эффект можно наблюдать при наличии относительно небольшого числа наблюдений.

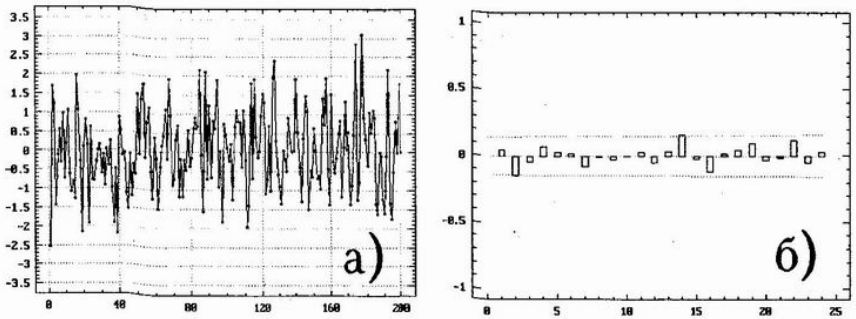


Рис. 12.16. Коррелограмма белого шума: а) исходный ряд; б) его коррелограмма

Коррелограмма процессов скользящего среднего. Траектории многих стационарных случайных процессов выглядят гораздо более гладко, чем траектории белого шума. Это связано с наличием положительной корреляции между двумя или несколькими соседними членами подобных рядов. Если же корреляция между соседними членами ряда отрицательна, то траектории подобных процессов будут более изломанными, чем траектории белого шума. Простейшим примером процессов, у которых зависимы одно или несколько соседних значений, являются процессы скользящего среднего. Определение процесса скользящего среднего первого порядка было дано в п. 11.7. Более подробно свойства этих процессов рассматриваются в п. 14.4. Здесь мы приведем вид типичных графиков этих процессов и их автокорреляционных функций.

Пусть $\varepsilon_1, \ldots, \varepsilon_n$ — гауссовский белый шум. Обозначим через $X(t)$ процесс скользящего среднего первого порядка (кратко MA(1)) с коэффициентом θ и средним равным нулю. Согласно (11.10):

$$X(t) = \varepsilon_t + \theta\varepsilon_{t-1}.$$

Нетрудно убедиться, что у этого процесса зависят между собой только соседние значения $X(t)$ и $X(t-1)$. При этом их корреляция r_1 равна:

$$r_1 = \frac{\theta}{1+\theta^2}$$

На рис. 12.17 приведены графики ста значений реализации процесса скользящего среднего с коэффициентом $\theta = 0.75$ и его коррелограммы. На рис. 12.18 приведены аналогичные графики при $\theta = -0.75$.

На графиках видно, что хотя полученные оценки значений r_k при $k = 2, 3\ldots$ не равны нулю, они значимо не отличаются от нулевых значений, так как попадают в 95% доверительный интервал, который построен в предположении равенства нулю соответствующих значений автокорреляционной функции.

Для процессов скользящего среднего второго порядка, как будет показано ниже в п. 14.4, отличаются от нуля только значения r_1 и r_2,

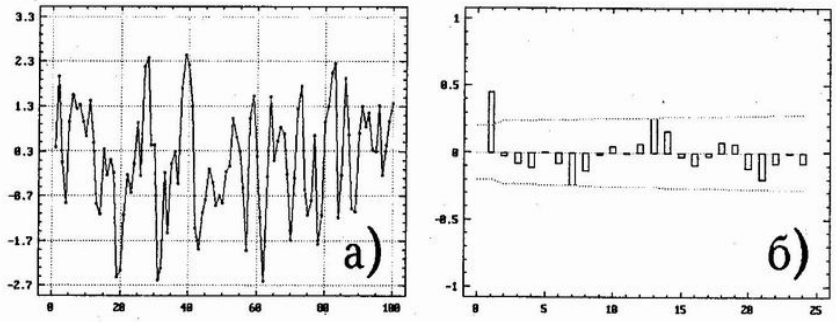


Рис. 12.17. Коррелограмма MA(1) процесса при $\theta = 0.75$: а) исходный ряд; б) его коррелограмма

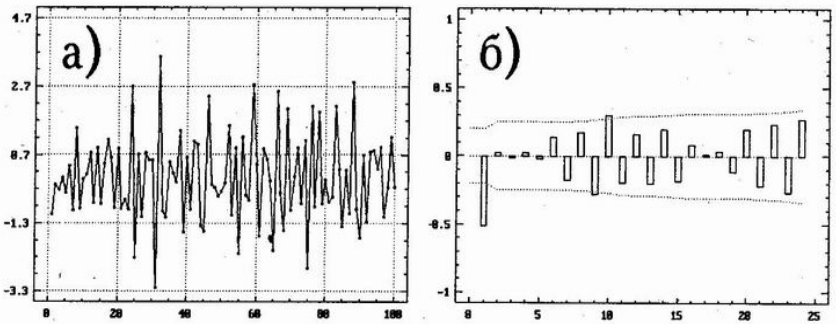


Рис. 12.18. Коррелограмма MA(1) процесса при $\theta = -0.75$: а) исходный ряд; б) его коррелограмма

а все последующие значения r_k при $k = 3, 4, \ldots$ равны нулю. Наконец, для процессов скользящего среднего порядка q отличны от нуля только первые q значений автокорреляционной функции. Строя графики коррелограмм для подобных процессов, мы можем на основании указанного свойства сделать предварительный вывод о возможном порядке процесса скользящего среднего, который может быть использован для описания наблюденного ряда.

Указанное правило хорошо, если подобранный порядок модели скользящего среднего невелик, скажем от одного до четырех-пяти. Однако на практике часто встречаются стационарные процессы с автокорреляционной функцией заметно отличной от нуля даже при больших задержках. Следуя сформулированному правилу, их можно пытаться описать процессами скользящего среднего высоких порядков. Это приводит к большому числу коэффициентов процесса скользящего среднего, которые подлежат дальнейшей оценке. При этом точность этих оценок заметно снижается. Практическая ценность таких многопараметрических моделей скользящего среднего невелика. В этой ситуации лучше попытаться описать временной ряд с помощью модели авторегрессии.

Если и эта попытка не увенчается успехом — перейти к комбинированным моделям авторегрессии-скользящего среднего. Проиллюстрируем типичное поведение автокорреляционной функции и ее оценки для процессов авторегрессии.

Коррелограмма процессов авторегрессии. Пусть, как и прежде, $\varepsilon_1, \ldots, \varepsilon_n$ — гауссовский белый шум. Напомним, что простейший процесс авторегрессии первого порядка $X(t)$ с нулевым средним задается соотношением:

$$X(t) = \phi X(t-1) + \varepsilon_t, \tag{12.16}$$

где ε_t не зависит от $X(t-1)$. Как указывалось в п. 11.7, члены даже этого простейшего процесса не становятся независимыми с ростом промежутка времени между ними. Однако при определенных условиях на коэффициенты эта зависимость быстро убывает.

В общем случае свойства этих процессов подробно разбираются нами ниже в пп. 14.1—14.3. Здесь же мы приведем два типичных графика поведения выборочных автокорреляционных функций этих процессов. Как будет показано ниже в п. 14.2 и 14.3 автокорреляционные функции этих процессов с ростом лага либо просто экспоненциально затухают либо представляют из себя экспоненциально затухающие синусоидальные волны.

На рис. 12.19 приведены графики ста значений реализации процесса авторегрессии второго порядка:

$$X(t) = \phi_1 X(t-1) + \phi_2 X(t-2) + \varepsilon_t. \tag{12.17}$$

при $\phi_1 = 0.7$ и $\phi_2 = 0.25$. Здесь автокорреляционная функция процесса и соответственно коррелограмма экспоненциально затухают с ростом лага.

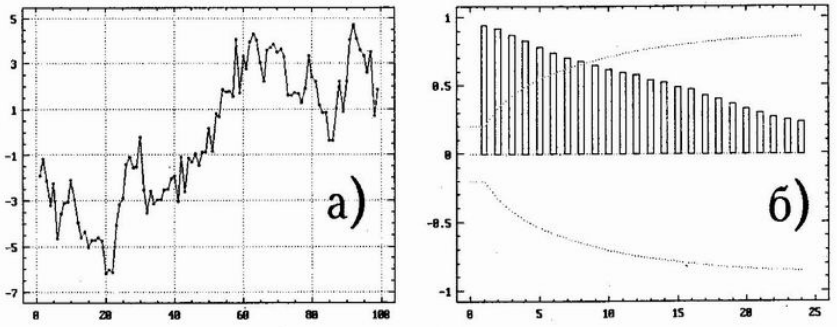


Рис. 12.19. Коррелограмма AR(2) процесса при $\phi_1 = 0.7$, $\phi_2 = 0.25$: а) исходный ряд; б) его коррелограмма

На рис. 12.20 приведены графики ста значений реализации AR(2) процесса (12.17) при $\phi_1 = 0.7$ и $\phi_2 = -0.25$. Автокорреляционная

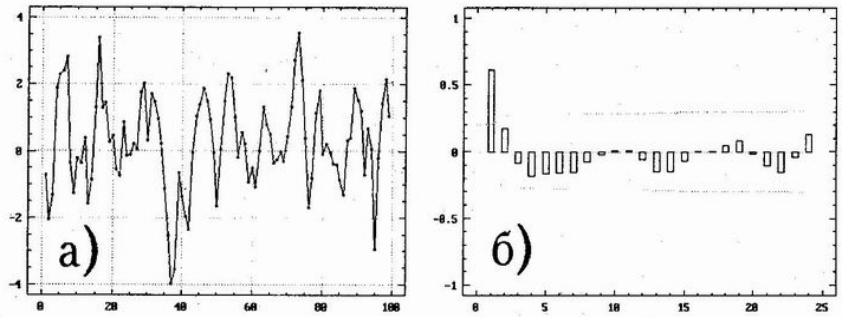


Рис. 12.20. Коррелограмма AR(2) процесса при $\phi_1 = 0.7$, $\phi_2 = -0.25$: а) исходный ряд; б) его коррелограмма

функция этого процесса и соответственно коррелограмма ведут себя с ростом лага как экспоненциально затухающая синусоида.

Замечание. Вид выборочных автокорреляционных функций процесса (12.16) для различных значений ϕ приведен на рис. 14.1 главы 14. На этом рисунке видно, что при $\phi = \pm 0.75$ значения оценок автокорреляционной функции $\overline{r}_k$ довольно сильно отличаются от нуля даже при $k = 15$.

12.4.3. Интерпретация графика частной автокорреляционной функции

Выборочная частная автокорреляционная функция. Для того, чтобы по полученной реализации процесса подобрать модель авторегрессии, необходимо предварительно указать возможный порядок этой модели. Приведенные примеры авторегрессионных процессов показывают, что непосредственно из вида выборочной автокорреляционной функции этот вывод сделать довольно трудно. Эту задачу значительно облегчает специально преобразованная автокорреляционная функция. Она называется *частной автокорреляционной функцией*. Расскажем, как следует интерпретировать поведение этой функции.

Замечание. Формальное определение частной автокорреляционной функции мы отложим до главы 14, так как это определение довольно сложно и основано на моделях авторегрессии. Однако при практическом анализе временных рядов это определение и не нужно — графики автокорреляционной функции и частной автокорреляционной функции строит статистическая программа, а человеку нужно лишь знать, как их интерпретировать.

Обозначения. Для краткости мы будем использовать сокращения: АКФ — автокорреляционная функция, ЧАКФ — частная автокорреляционная функция.

Обозначим через ϕ_{kk} значения ЧАКФ для каждого значения лага $k = 1, 2, \ldots$. Оценку этой функции по реализации временного ряда

будем обозначать через $\hat{\phi}_{kk}$ при $k = 0, 1, 2, \ldots$. Значения функций ϕ_{kk} и $\hat{\phi}_{kk}$ для каждого значения k по абсолютной величине меньше единицы.

Процессы авторегрессии первого порядка. Как показано ниже в п. 14.3, для процессов авторегрессии первого порядка отлично от нуля только значение ϕ_{11}, а все остальные значения этой функции равны нулю. Для выборочной частной автокорреляционной функции $\hat{\phi}_{kk}$ это означает, что все ее значения, начиная со второго, должны значимо не отличаться от нуля, то есть попадать в соответствующий доверительный интервал.

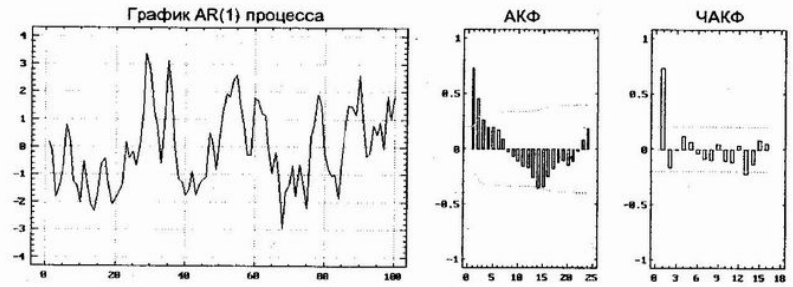


Рис. 12.21. График AR(1) процесса (значение ϕ равно 0.75, его выборочная автокорреляционная функция (АКФ) и выборочная частная автокорреляционная функция (ЧАКФ)

Это и видно на рис. 12.21. Здесь значения выборочной ЧАКФ (в отличие от выборочной АКФ), начиная со второго, малы и значимо неотличимы от нуля. Таким образом, по поведению выборочной ЧАКФ легче выяснить вид модели временного ряда.

Замечание. Стоит заметить, что при малых значениях коэффициента ϕ для того, чтобы отличить процесс от белого шума, требуется достаточно много наблюдений. Так, из рис. 12.22, на котором представлен AR(1) процесс с $\phi = -0.25$, видно, что для этого оказалось недостаточно ста наблюдений. Все значения выборочных автокорреляционной и частной автокорреляционной функций здесь попали в доверительные трубки для предположения, что процесс является белым шумом.

Процессы авторегрессии второго порядка. Для процесса авторегрессии второго порядка отличны от нуля только первые два значения ЧАКФ. На рис. 12.23а и 12.23б изображены выборочные ЧАКФ процессов, представленных на рис. 12.19 и 12.20.

Процессы скользящего среднего. Для процессов скользящего среднего (MA-процессов), в отличие от процессов авторегрессии, ЧАКФ при больших значениях лага k не обращается в ноль, а экспоненциально убывает. Мы не будем останавливаться на этом подробнее. Просто проиллюстрируем поведение выборочной ЧАКФ для MA(1) процессов,

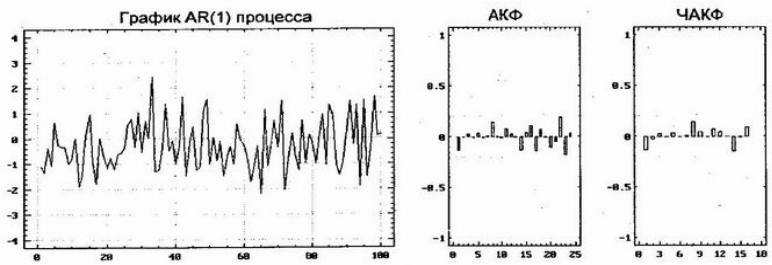


Рис. 12.22. График AR(1) процесса (значение ϕ равно -0.25), его выборочная АКФ и выборочная ЧАКФ

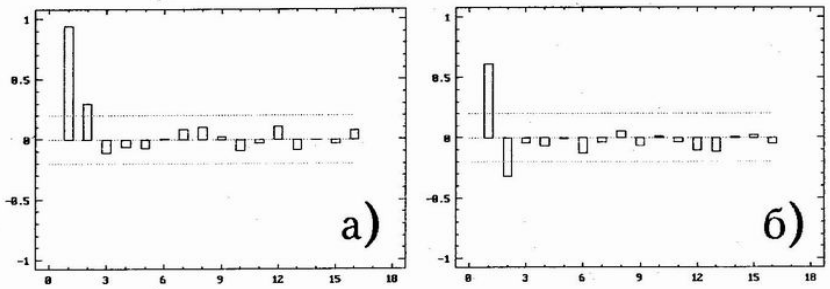


Рис. 12.23. Выборочная ЧАКФ AR(2) процессов:
а) $\phi_1 = 0.7$, $\phi_2 = 0.25$; б) $\phi_1 = 0.7$, $\phi_2 = -0.25$

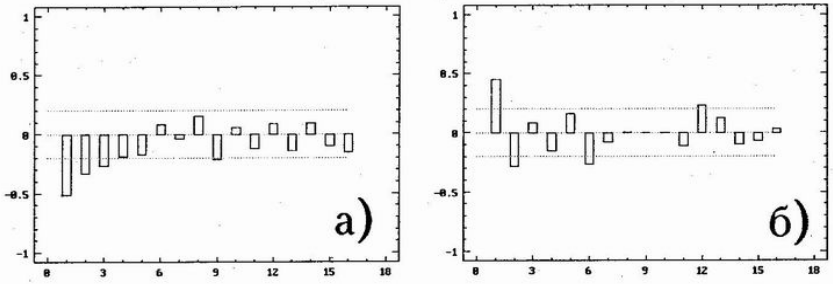


Рис. 12.24. Выборочная ЧАКФ MAR(1) процессов: а) $\theta = 0.75$; б) $\theta = -0.75$

приведенных на рис. 12.17 и 12.18. На рис. 12.24 приведены выборочные ЧАКФ MA(1) процессов при $\theta = 0.75$ и $\theta = -0.75$.

Глава 13

Анализ временных рядов на компьютере

13.1. О выборе пакетов для описания в этой книге

Методы анализа временных рядов широко представлены во многих универсальных статистических пакетах, включая разобранные в предыдущих главах STADIA и SPSS. Но анализ временных рядов — это очень специфическая область статистики, отличающаяся по кругу задач и методов их решения, а также по кругу пользователей, применяющих эти методы. Поэтому для анализа временных рядов имеются также и специализированные статистические пакеты. В этой главе мы рассмотрим способы решения рассмотренных выше задач в универсальном статистическом пакете SPSS и в специализированном статистическом пакете ЭВРИСТА. Выбор данных пакетов обусловлен следующими причинами.

Универсальный пакет SPSS занимает одно из первых мест в мире среди программ статистической обработки данных (см. приложения 1 и 2). Отечественным специалистам ранние версии SPSS в основном были известны как мощный инструмент обработки социологических и психологических данных. В связи с этим мы решили показать этот пакет с менее известной его стороны. Для нас также было важно познакомить пользователя с англоязычной терминологией в области анализа временных рядов.

Пакет ЭВРИСТА является одним из лучших специализированных отечественных пакетов для анализа временных рядов. Его функциональные возможности значительно шире стандартных процедур анализа временных рядов универсальных статистических пакетов. Пакет постоянно совершенствуется и пополняется, он хорошо зарекомендовал себя во многих организациях, в том числе активно работающих на финансовом рынке. Более подробная информация об этом пакете дана в приложениях 1 и 2, а также в [11]. Наконец, нам хотелось дать пользователям представление о более широком круге отечественных статистических пакетов.