# Глава 2

# ЭЛЕМЕНТЫ ТЕОРИИ ВЕРОЯТНОСТЕЙ И СТАТИСТИКИ

## 2.1. ВВЕДЕНИЕ

С приложениями анализа временных рядов тесно связаны некоторые статистические понятия. Цель этой главы состоит в том, чтобы определить такие крайне необходимые на практике ключевые параметры, как среднее, дисперсию и вероятностную плотность. Нередко приходится иметь дело с данными, испорченными нежелательным шумом, приводящим к определенным ошибкам при вычислениях различных параметров данных временного ряда. Однако мы ограничимся лишь теми понятиями, которые непосредственно относятся к анализу временных рядов. Вместе с тем эта книга — не учебник по математической статистике. Мы, как правило, только определяем термины и не доказываем теорем. Читателю, которому нужно более полное знакомство с основами статистики, советуем обратиться к книгам Большева и Смирнова (1965), Диксона и Мэсси (1969), содержащим полные и всеобъемлющие таблицы; теоретическим вопросам математической статистики посвящены книги Крамера (1975) и Фрэйзера (1958). В этих книгах даются более строгие математические определения статистических параметров. Например, корректное определение случайного процесса потребовало бы сложных математических построений, связанных с характером исследуемых величин. Но в наших приложениях, к счастью, те патологические функции, которые рассматриваются математиками, не встречаются. Поэтому можно сделать упрощающие допущения, которые не приводят к потере общности результатов.

1. Рассматриваемые функции ограниченны и до преобразования в цифровую форму были непрерывными.

2. Случайный процесс, определяющий функцию, является эргодическим и стационарным.

Строгое математическое определение случайной величины требует в первую очередь ее измеримости в том смысле, в котором это понимается в теории меры. Это гарантируется допущением 1, поскольку ограниченные непрерывные функции образуют подмножество множества измеримых функций. Оставшуюся часть введения мы отведем обсуждению на интуитивном уровне допущения 2.



**Вероятностная функция плотности (ВФП), среднее и дисперсия.** Представим себе, что большое число одинаковых генераторов шума было включено в некоторый момент в прошлом и с тех пор безостановочно работает. С выходом всех этих генераторов связывают вероятностную функцию плотности $f(x, t)$, имеющую следующие характеристики. Вероятность того, что в определенный момент, скажем $t_0$, выход $q$-го генератора сигналов $x_q(t_0)$ лежит в интервале между значениями $a$ и $b$, определяется интегралом

$$\mathrm{P}\{a \leqslant x_q(t_0) < b\} = \int_a^b f(x,\ t_0)\,dx.$$

Заметим, что интегрирование ведется по промежутку значений случайной величины. Математическое ожидание любой функции от $x$, обозначаемое $E[g(x)]$ (где $g(x)$ — та функция, ожидание которой ищется), определяется следующим образом:

$$E\{g[x(t_0)]\} = \int_{-\infty}^{\infty} g[x(t_0)]\,f(x,\ t_0)\,dx.$$

В частности, *истинные* или *множественные* среднее и дисперсия задаются формулами

$$\mu(t_0) = \int_{-\infty}^{\infty} x(t_0)\,f(x,\ t_0)\,dx \tag{2.1}$$

и

$$\sigma^2(t_0) = \int_{-\infty}^{\infty} [x(t_0) - \mu(t_0)]^2\,f(x,\ t_0)\,dx.$$

*Стандартным отклонением* называют положительное значение квадратного корня из дисперсии.

Если случайный процесс является *стационарным*, то параметры $\mu(t_0)$ и $\sigma^2(t_0)$ не зависят от времени, т. е. для произвольных $t_0$ и $t_1$

$$\mu(t_0) = \mu(t_1) = \mu, \quad \sigma^2(t_0) = \sigma^2(t_1) = \sigma^2.$$

В дальнейшем параметр $t_0$ в среднем и в дисперсии нами указываться не будет, поскольку принимается допущение о стационарности.

Допущение об *эргодичности* позволяет заменить усреднение по ансамблю усреднением по времени. В примере с генераторами шума все генераторы были совершенно одинаковыми, поэтому знания лишь одной случайной функции отдельного генератора было бы достаточно, чтобы по выходу одного из них определить статистические параметры для всех. Так, выражение для сред-

него (2.1) можно заменить выражением

$$\mu_x = \lim_{P \to \infty} \frac{1}{2P} \int_{-P}^{P} x(t)\, dt,$$

в котором временное усреднение ведется по одной траектории процесса. Подобное выражение для дисперсии имеет вид

$$\sigma_x^2 = \lim_{P \to \infty} \frac{1}{2P} \int_{-P}^{P} [x(t) - \mu]^2\, dt = E(x^2) - E^2(x). \qquad (2.2)$$

Среднеквадратичное значение определяется формулой

$$\psi_x^2 = \mu_x^2 + \sigma_x^2.$$

Как сказано выше, свойство эргодичности позволяет заменять усреднения по множеству реализаций случайного процесса усреднениями по времени.

Центральным в этой книге является понятие *плотности спектра мощности* (ПСМ), которое формально будет определено ниже. А пока будем подразумевать под ПСМ функцию, представляющую собой разбиение или разложение общей среднеквадратичной мощности как функции частоты.

Заметим, что будет «правильно» связывать со случайным процессом функцию плотности спектра мощности. Дело в том, что статистические свойства и характеристики гауссовского процесса с нулевым средним полностью определяются (и притом наиболее кратко) спектром мощности этого процесса или, что эквивалентно, его автоковариационной функцией [1]). Это делает спектр мощности исключительно важным параметром, поскольку большинство процессов, встречающихся на практике, являются гауссовскими.

[1]) Здесь мы несколько выходим за рамки основного предмета этой книги. Статистические характеристики процесса известны, если известны все многомерные вероятностные функции плотности, определяющие совместные вероятности для процесса на всех временных промежутках. Например, автоковариационная функция для промежутка времени $\tau$ определяет корреляционные характеристики между амплитудами процесса в моменты $t$ и $t+\tau$. Для гауссовских данных полностью определяют совместную (двумерную) вероятностную плотность величин амплитуд в моменты $t$ и $t+\tau$ ковариация и две дисперсии. В общем случае это нужно распространить и на более высокие порядки.

При сборе данных в практических экспериментах или тестах можно вычислять только оценки спектра мощности (и другие параметры) и, пользуясь эргодичностью, давать оценки по временным средним, полученным по отдельным интервалам временной последовательности. Поэтому оценка ПСМ определяется по конечному временному промежутку. При замене отрезка конечным набором точек мы не получим, используя выборочные данные и дискретные формулы, существенного (со статистической точки зрения) расхождения непрерывной и дискретной версий. Такая замена производится, пока не нарушаются требования теоремы о выборках, позволяющей избегать существенной подмены частот.



## 2.2. ВЫБОРОЧНЫЕ ВЕЛИЧИНЫ И ОЦЕНКИ

Для выборочного среднего и выборочной дисперсии — оценок групповых параметров — выбирают отличающиеся от принятых в теории обозначения для среднего и дисперсии, поскольку длина записи реализаций случайной величины не может быть бесконечной и, следовательно, можно располагать только конечной выборкой. В частности, выборочное среднее для цифровых данных вычисляется по формуле

$$m = \bar{x} = \frac{1}{N} \sum_{i=0}^{N-1} x(i).$$

Несмещенную выборочную дисперсию $s^2$ получают так:

$$s^2 = \frac{1}{N-1} \sum_{i=0}^{N-1} [x(i) - \bar{x}]^2.$$

Определение доверительных интервалов для выборочного среднего и выборочной дисперсии проводится с помощью критериев, точные формулы и таблицы для которых можно найти в книге Бендата и Пирсола (1971). Как правило, длинам записей, с которыми сталкиваются в анализе временных рядов, отвечает малая статистическая изменчивость этих двух параметров, и здесь трудностей не возникает.

Если делитель $N-1$ в выражении для $s^2$ заменить на $N$, то среднее значение $s^2$ станет равным $[(N-1)/N]\sigma^2$, т. е. оценка станет смещенной. Впрочем, для больших значений $N$ эта ошибка мала. Обычно в анализе временных рядов значение $N$ равно по крайней мере 100 и часто достигает величин порядка 10 000 и более. Поэтому вопрос о смещении, вызванном делением на $N$, как правило, несуществен.

Сделаем несколько замечаний о принятой в статистике системе обозначений. Групповые параметры по традиции обозначают греческими буквами, например $\mu$ и $\sigma^2$ для среднего и дисперсии. Оценки параметров обычно обозначаются соответствующими латинскими буквами, например $m$ и $s^2$. Несколько нарушая этот порядок, средние случайных величин часто помечают черточкой сверху, например, среднее величины $x$ обозначают $\bar{x}$. В анализе временных рядов черточку обычно оставляют для временных средних, а обозначение математического ожидания $E(x)$ используют для более общих ситуаций. Наконец, представляется довольно разумным использование крышечки над параметрами как стандартного обозначения для оценки. Это, в частности, полезно при истощении запаса букв. Так, оценку спектра мощности обозначаем через $\hat{S}_x(k)$. Иногда мы будем вынуждены применять все перечисленные обо-

значения сразу и полагаем, что причина этого станет ясна читателю из контекста.

Теперь остановимся более подробно на некоторых важных свойствах статистических оценок. Мы рассмотрим смещение, дисперсию и состоятельность.

*Смещение.* Оценка $\hat{\varphi}$ параметра $\varphi$ смещена, если среднее значение $\hat{\varphi}$ не совпадает с $\varphi$;

$$\text{Смещение} = \Delta_{\hat{\varphi}} = E(\hat{\varphi}) - \varphi.$$

Выборочная дисперсия с делителем $N$ служит примером смещенной оценки.

*Дисперсия.* Дисперсией оценки является величина

$$\sigma^2_{\hat{\varphi}} = E[\hat{\varphi} - E(\hat{\varphi})]^2,$$

поэтому среднеквадратичную ошибку определяют следующим образом:

$$\Psi^2_{\hat{\varphi}} = \sigma^2_{\hat{\varphi}} + \Delta^2_{\hat{\varphi}} = E[\hat{\varphi} - \varphi]^2.$$

*Состоятельность.* Оценка состоятельна, если ее дисперсия убывает с ростом объема выборки (длиной записи). Запишем это условие в виде равенства

$$\lim_{N\to\infty} \sigma^2_{\hat{\varphi}} = 0.$$

Примером несостоятельной оценки может служить несглаженный выборочный спектр мощности. Если в качестве оценки спектра мощности вычисляется квадрат абсолютной величины преобразования Фурье (часто называемый периодограммой), то получается несостоятельная оценка. В этом случае дисперсия остается постоянной при возрастании длины записи.

Помимо того, примерами смещенных оценок служат обычные оценки спектральных функций. Возникающее в этом случае смещение интерпретируется как ошибка, размывающая полосу частот вследствие недостаточной спектральной разрешающей способности. Это явление подробно обсуждается в гл. 8.

Рассмотрим важный пример вычисления дисперсии для синусоиды. Предположив, что сигнал $x$ определен формулой

$$x(t) = A\sin(2\pi f_c t + \varphi), \quad -\infty < t < \infty,$$



получим следующее выражение для дисперсии:

$$\sigma_x^2 = \lim_{P\to\infty} \frac{1}{P} \int_{-P/2}^{P/2} A^2 \sin^2(2\pi f_c t + \varphi)\, dt =$$
$$= A^2 \lim_{P\to\infty} \frac{1}{P} \int_{-P/2}^{P/2} \sin^2(2\pi f_c t + \varphi)\, dt =$$
$$= A^2 \lim_{P\to\infty} \frac{1}{P}\left[\frac{P}{2} + \{\text{sin-член} \leqslant 1\}\right] = \frac{A^2}{2}. \qquad (2.3)$$

Заметим, что для таких функций среднеквадратичная ошибка $\Psi_x^2$ и дисперсия $\sigma_x^2$ совпадают, поскольку среднее этих функций равно 0.

## 2.3. НОРМАЛЬНОЕ РАСПРЕДЕЛЕНИЕ

Предположим, что случайная величина $x$ имеет среднее $\mu$ и дисперсию $\sigma^2$. Тогда говорят, что она распределена *нормально* или имеет *гауссовское* распределение, если ее вероятностная функция плотности $\varphi(x)$ определяется выражением

$$\varphi(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]. \qquad (2.4)$$

Вероятностная функция распределения $\Phi(x)$ такой гауссовской величины дается формулой

$$\Phi(x) = \int_{-\infty}^{x} \varphi(\xi)\, d\xi,$$

т. е.

$$\Phi(x_0) = \mathbf{P}\{-\infty < x \leqslant x_0\}, \quad 1 - \Phi(x_0) = \mathbf{P}\{x_0 < x < \infty\},$$
$$\Phi(x_1) - \Phi(x_0) = \mathbf{P}\{x_0 < x \leqslant x_1\}$$

(предполагается, что $x_0 < x_1$).

Для гауссовских величин справедлива важная теорема, которую мы приведем без доказательства.

**Теорема.** *Случайная величина, полученная линейным преобразованием из гауссовской величины, является гауссовской.*

В частности, величина

$$z = ax + b \qquad (2.5)$$

— гауссовская, если $x$ имеет гауссовское распределение со средним $\mu_x$ и дисперсией $\sigma_x^2$, причем

$$\mu_z = a\mu_x + b \qquad (2.6)$$

ЭБ-169

и

$$\sigma_z^2 = a^2\sigma_x^2. \tag{2.7}$$

Применение преобразования (2.5) специального вида, задаваемого формулой

$$z = \frac{1}{\sigma_x}(x - \mu_x), \tag{2.8}$$

называют «стандартизацией». Из формул (2.6) и (2.7) следует, что $\mu_z = 0$, $\sigma_z^2 = 1$. С помощью этого преобразования можно составлять таблицы для тех гауссовских величин, которые фигурируют в приложении. Независимо от значений среднего и дисперсии исходных данных можно определять вероятности по таблице для величины с параметрами 0 и 1, а затем, обращая соотношение (2.8):

$$x = z\sigma_x + \mu_x,$$

переходить к значениям исходной величины. Приведем несколько часто используемых значений для функции распределения $\Phi$ при $\mu = 0$ и $\sigma = 1$:

$$\frac{1}{\sqrt{2\pi}}\int_{-1}^{1}\exp\left(-\frac{x^2}{2}\right)dx = 0.682,$$

$$\frac{1}{\sqrt{2\pi}}\int_{-2}^{2}\exp\left(-\frac{x^2}{2}\right)dx = 0.954,$$

$$\frac{1}{\sqrt{2\pi}}\int_{-3}^{3}\exp\left(-\frac{x^2}{2}\right)dx = 0.997.$$

Мы полагаем, что читателю известно, как пользоваться таблицами распределений, и не останавливаемся на этом.

Нормальное распределение — понятие теоретическое; на практике обычно не встречается данных с действительно нормальным распределением. В основном это связано с областью значений случайной величины. Большинство данных берется из ограниченного интервала, в то время как нормально распределенные данные должны принимать значения на бесконечной прямой.

Предположим, что группа людей измеряет стержень длины 2 см, используя одну и ту же линейку, и что полученные измерения имеют нормальное распределение. Допустим также, что истинное среднее измерений равно 2 см, а дисперсия равна 0.25 см. Вероятность того, что результат измерения окажется меньше −0.5 см, равна

$$\mathbf{P}\{x \leqslant -0.5\} = \frac{1}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{-0.5}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]dx.$$



После замены переменной $z = (x - \mu)/\sigma$ получим

$$\mathbf{P}\{x \leqslant -0.5\} = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{-(0.5+2)/0.25}\exp\left(-\frac{z^2}{2}\right)dz \approx 0.0000001.$$

Таким образом, есть конечная (хотя и малая) вероятность того, что результат измерения длины стержня окажется отрицательным. Разумеется, это абсурд. Тем не менее использование нормального распределения вполне оправданно, поскольку оно является приемлемым приближением к тем истинным распределениям, которые могут встречаться.

Предположение о нормальности (т. е. о том, что данные имеют гауссовское распределение) принимается по ряду причин. Главные из них следующие.

1. Нормальное распределение достаточно хорошо изучено. Это упрощает моделирование исследуемой ситуации и получение выводов о результатах статистических измерений.

2. Как показывает центральная предельная теорема (Крамер, 1975), распределение суммы случайных величин, вообще, с любым распределением стремится к нормальному, если суммируется «достаточное их число». Оказывается, что результаты фильтрации (см. гл. 3) представляют собой суммирование большого числа наблюдений и, следовательно, их распределение «стремится» к нормальному.

Перечислим несколько важных свойств нормального распределения:

1. Как плотность, так и распределение гауссовской случайной величины $x$ полностью определяются средним $\mu_x$ и дисперсией $\sigma_x^2$.

2. Моменты нормальной величины вычисляются по формулам

$$E[x^n] = \begin{cases} \mu, & n = 1, \\ \mu^2 + \sigma^2, & n = 2, \\ \mu(\mu^2 + 3\sigma^2), & n = 3, \\ \dots & \dots \end{cases}$$

3. Центральные моменты вычисляются по формулам

$$E[(x-\mu)^n] = \begin{cases} 0, & n = 1, 3, \dots, \\ 1\cdot 3\cdots(n-1)\,\sigma^n, & n = 2, 4, \dots. \end{cases}$$

Третий момент относительно среднего называют «асимметрией», а четвертый — «эксцессом». Асимметрия равна 0 для всякого симметричного распределения. Эксцесс, превосходящий гауссовский (равный $3\sigma^4$), указывает на значительное число данных с большими амплитудами. Распределения с таким эксцессом имеют более «толстые» хвосты функции плотности, чем нормальное. В § 2.10,

где приводятся примеры и применения вероятностных функций плотности, мы дадим различные примеры оценок вероятностной плотности.

Обратимся к последнему свойству случайных величин, связанному с ВФП,— *независимости*.

Две случайные величины $x$ и $y$ статистически независимы, если их совместная (двумерная) вероятностная функция плотности $p(x, y)$ распадается в произведение одномерных ВФП, т. е.

$$p(x, y) = p(x)\, p(y).$$

Другими словами, это свойство показывает, что со статистической точки зрения величины $x$ и $y$ не оказывают влияния друг на друга. Об этом пойдет речь в следующем параграфе, где будут рассмотрены корреляция и регрессия. А сейчас перейдем ко второму важному распределению.

**Распределение хи-квадрат.** Предположим, что $x_1, x_2, \ldots, x_n$ — независимые гауссовские величины с нулевым средним и единичной дисперсией. Определим величину $\chi_n^2$ формулой

$$\chi_n^2 = \sum_{i=1}^{n} x_i^2.$$

Об этой величине говорят, что она имеет распределение хи-квадрат с $n$ степенями свободы. Функция плотности $\chi_n^2$ дается формулой

$$p(\chi_n^2) = \left[2^{n/2}\Gamma\left(\frac{n}{2}\right)\right]^{-1} (\chi_n^2)^{(n/2)-1} \exp[-\chi_n^2/2], \quad \chi_n^2 \geqslant 0,$$

где $\Gamma(n/2)$ есть гамма-функция.

Распределение хи-квадрат используется в анализе временных рядов главным образом при изучении поведения выборочной дисперсии и плотностей спектра мощности. Если ряд $x(i)$ имеет нулевое среднее и при вычислении выборочной дисперсии $s^2$ используется $N$ независимых данных $x(i)$, то границы $B_1$ и $B_2$, в которых с вероятностью $(1-\alpha)$ лежит истинное значение дисперсии, т. е.

$$\mathbf{P}\{B_1 \leqslant \sigma^2 \leqslant B_2\} = 1-\alpha,$$

определяются по формулам

$$B_1 = \frac{ns^2}{\chi^2_{n;\,1-(\alpha/2)}}; \qquad B_2 = \frac{ns^2}{\chi^2_{n;\,(\alpha/2)}}, \qquad n = N-1.$$

Отметим, что $B_1$ и $B_2$ суть функции от $s^2$, $\alpha$ и $\chi_n^2$. Интервал $(B_1, B_2)$ называют доверительным интервалом; о нем говорят также как о $[(1-\alpha)\,100]\%$-м доверительном интервале[1]).

[1]) Как принято в статистике, мы более свободно пользуемся термином «доверительный»; строго говоря, это уже не вероятность. Такое расширение смысла этого понятия можно найти у Фрэйзера (1958).



Рассмотрим пример (взятый из книги Блэкмена и Тьюки (1958)), в котором по $N$ независимым наблюдениям нормально распределенной случайной величины ($N=31$) требуется определить 90%-й доверительный интервал для выборочной дисперсии. Предположим, что выборочное среднее $\bar{x}$ равно 58.61, а выборочная дисперсия $s^2$ равна 33.43. В этом случае для теоретического значения дисперсии $\sigma^2$ выполнено следующее равенство:

$$\mathbf{P}\left\{\frac{30s^2}{43.77} < \sigma^2 < \frac{30s^2}{18.49}\right\} = 0.90,$$

где $\chi^2_{30;\,0.05} = 18.49$ есть такое значение переменной $\chi^2_{30}$, что $\mathbf{P}\{\chi^2_{30} < 18.49\} = 0.05$, а $\chi^2_{30;\,0.95} = 43.77$ (соответственно $\mathbf{P}\{\chi^2_{30} < 43.77\} = 0.95$). После подстановки значения получаем следующие границы: $22.91 \leqslant \sigma^2 \leqslant 54.22$. Заметим, что значение истинной дисперсии попадает за пределы этого интервала примерно в десятой части случаев.

Важную роль в статистике играют и другие распределения вероятностей. Среди них можно, в частности, назвать 1) $t$-распределение Стьюдента, полезное в статистических выводах и при вычислении доверительных интервалов для истинных средних значений; 2) $F$-распределение, которое используется при проверке на равенство пар средних квадратов. Основные факты, касающиеся этих распределений, читатель может найти в книге Бендата и Пирсола (1971).

## 2.4. КОРРЕЛЯЦИЯ И РЕГРЕССИЯ

Статистические понятия корреляции и регрессии близко примыкают к понятиям когерентности и частотной функции отклика, широко используемым в анализе временных рядов (а также в технике). По этой причине мы остановимся на них подробнее.

Нам необходимо сначала определить понятие ковариации, которая представляет собой математическое ожидание произведения двух центрированных случайных величин, т. е.

$$\operatorname{cov}(x, y) = \sigma_{xy} = E[(x-\mu_x)(y-\mu_y)].$$

Для ковариации двух независимых величин (т. е. таких, что $p(x, y) = p(x)\, p(y)$) выполняется важное равенство $\sigma_{xy} = 0$. Заметим, что обратное неверно. Однако в очень важном частном случае, когда две случайные величины имеют совместное (двумерное) гауссовское распределение, последнее представляется произведением двух одномерных гауссовских распределений. Поэтому для гауссовских данных справедлив следующий важный вывод: *из равенства нулю ковариации гауссовских величин вытекает их независимость*, т. е.

$$\sigma_{yx} = 0 \Rightarrow \Phi(x, y) = \Phi(x)\,\Phi(y).$$

Этот факт играет главную роль, вероятно, в 90% задач практического анализа временных рядов.

Корреляция — это нормированная ковариация, а именно

$$\rho_{xy} = \frac{\operatorname{cov}(x, y)}{[\operatorname{var}(x) \operatorname{var}(y)]^{1/2}} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}.$$

Нетрудно показать, что коэффициент корреляции принимает значения в интервале, ограниченном $\pm 1$, т. е. $-1 \leqslant \rho_{xy} \leqslant 1$. Этот результат остается справедливым и для выборочной корреляции, если выборочная ковариация определяется формулой

$$s_{xy} = \frac{1}{N-1} \sum_{i=0}^{N-1} [x(i) - \bar{x}][y(i) - \bar{y}].$$

По аналогии с представлением о плотности спектра мощности как о разложении дисперсии по частотам под функцией *плотности кросс-спектра* (ПКС) подразумевают разложение по частотам ковариации.

С понятиями ковариации и корреляции тесно связан регрессионный анализ. Так называемую линию регрессии обычно определяют как линейное соотношение

$$y = \beta_0 + \beta_{xy} x$$

между двумя величинами. Если при сборе данных оценки $b_0$ и $b_{xy}$ для величин $\beta_0$ и $\beta_{xy}$ ищут с помощью критерия наименьших квадратов, то

$$b_{xy} = s_{xy}/s_x^2, \quad b_0 = \bar{y} - b_{xy}\bar{x}.$$

Полученная прямая

$$(y - \bar{y}) = b_{xy}(x - \bar{x})$$

называется линией регрессии[1]) $y$ по $x$. Можно говорить также и о регрессии $x$ по $y$. В этом случае

$$b_{yx} = s_{xy}/s_y^2.$$

Заметим, что

$$b_{xy} b_{yx} = \frac{s_{xy}^2}{s_x^2 s_y^2} = r_{xy}^2.$$

Для совместного гауссовского распределения линия регрессии допускает геометрическую интерпретацию. В этом важном случае коэффициенты регрессии суть тангенсы углов наклона главных осей эллипса постоянной вероятности.

[1]) Термин регрессия возник в некоторых давних исследованиях соотношения роста родителей и их детей. Было установлено, что рост «регрессирует» к среднему, т. е. высокие отцы и матери имеют более низких сыновей и дочерей, а низкие отцы и матери — более высоких.



**Множественная регрессия.** Двумерная регрессия без труда переносится на случай $p$ величин (без потери общности можно считать среднее равным нулю):

$$y = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p.$$

Если предположить, что имеется $N$ $(N > p)$ наблюдений для каждой величины, то, вводя векторные обозначения

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \ldots & x_{1p} \\ x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{N1} & x_{N2} & \ldots & x_{Np} \end{bmatrix}, \quad \mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_N \end{bmatrix},$$

решение задачи об оценке $\mathbf{B}$ параметров $\beta_i$ можно записать в виде (Андерсон (1963))

$$\mathbf{B} = (\mathbf{X}^*\mathbf{X})^{-1}\mathbf{X}^*\mathbf{Y},$$

где звездочка означает транспонирование.

В анализе частотных функций отклика важное значение приобретают некоторые другие множественные параметры, но мы пока отложим их изучение до гл. 9, где они понадобятся.

## 2.5. ФУНКЦИЯ ПЛОТНОСТИ СПЕКТРА МОЩНОСТИ

Как можно видеть из предыдущего, статистики второго порядка — дисперсия для одной случайной величины и ковариация для двух величин — имеют особое значение для величин с гауссовским распределением. Помимо этого в анализе временных рядов очень важная роль отведена разложению временных последовательностей по частотам. При оценивании таких важных параметров, как частота резонанса и декремент затухания, эта информация имеет решающее значение. В качестве другого применения можно назвать оценивание колебаний в таких экономических временных рядах, как индекс безработицы. Примерами использования плотности спектра мощности служат также обнаружение периодических колебаний температур в океане и определение цикла ритмических сокращений желудка коровы под воздействием определенных лекарств.

Среднеквадратичное значение на конечном отрезке определяется формулой

$$\psi_x^2 = \frac{1}{t_2 - t_1} \int_{t_1}^{t_2} x^2(t)\, dt.$$

Размерность $\psi_x^2$ пропорциональна среднему квадрату энергии в единицу времени, который по определению есть мощность.

Плотность спектра мощности (ПСМ) $S_x(f)$ функции $x$ — обобщение этого понятия. Ее можно понимать как функцию, интеграл от которой

$$\psi_x^2(f_1, f_2) = 2\int_{f_1}^{f_2} S_x(f)\,df, \quad 0 \leqslant f_1 \leqslant f_2,$$

равен среднеквадратичному значению в диапазоне частот от $f_1$ до $f_2$. Таким образом,

$$\psi_x^2 = \int_{-\infty}^{\infty} S_x(f)\,df.$$

Соответствующим выражением для $S_x$ служит следующее:

$$S_x(f) = \lim_{P\to\infty} \frac{1}{P}\left|\int_{-P/2}^{P/2} x(t)\,e^{-j\omega t}\,dt\right|^2. \tag{2.9}$$

При непосредственном вычислении оценок ПСМ с помощью формулы (2.9) могут возникнуть известные осложнения. Если не проводить сглаживания по частотам, то полученные оценки окажутся статистически несостоятельными. Помимо того, как отмечалось выше, нужно учитывать и математические тонкости. Наряду с тем, что в прошлом не было известно эффективных вычислительных процедур, эти две причины делают невозможным построение приемлемой вычислительной процедуры на основе формулы (2.9). Однако с помощью быстрого преобразования Фурье вычислительные трудности были преодолены, правильное сглаживание по области частот привело к состоятельным оценкам и, наконец, использование записей конечной длины устранило математические проблемы.

Теорема, которую часто называют теоремой Винера — Хинчина, дает эквивалентное выражение

$$S_x(f) = \int_{-\infty}^{\infty} s_x(\tau)\,e^{-j\omega\tau}\,d\tau = 2\int_0^{\infty} s_x(\tau)\cos(2\pi f\tau)\,d\tau,$$

где функция $s_x(\tau)$ (автоковариационная функция $x$) определяется соотношением

$$s_x(\tau) = \lim_{P\to\infty} \frac{1}{P}\int_{-P/2}^{P/2} x(t)\,x(t+\tau)\,dt, \quad -\infty < \tau < \infty,$$

причем предполагается, что процесс $x(t)$ имеет нулевое среднее.

Доказательство эквивалентности методов, основанных на этих двух формулах, не входит в нашу задачу. Его можно найти в нескольких монографиях (Бендат и Пирсол (1971), Винер (1963).



Существует и третий метод получения $S_x$, но его описание мы вынуждены отложить, поскольку он основан на понятии фильтрации, которое мы введем позже. А пока вернемся ко второму методу и поговорим о функции $s_x$. Ее можно рассматривать, учитывая принятое раньше предположение об эргодичности, как обобщение данного нами выше определения ковариации:

$$s_x(\tau) = E[x(t)\,x(t+\tau)] = \int_{-\infty}^{\infty} x(t)\,x(t+\tau)\,p(x, t, \tau)\,dt,$$

где $p(x, t, \tau)$ — соответствующая вероятностная функция плотности.

Допустим, что $n(t)$ — некоррелированный случайный шум с нулевым средним. Тогда для автокорреляционной функции справедливо равенство $s_n(\tau) = N\delta(\tau)$, и, следовательно, $S_n(f) = N$ для всех $f$. Таким образом, ПСМ некоррелированного случайного шума есть константа. Процессы такого типа называют *белым шумом*, поскольку мощность постоянна на любом промежутке частот, что очень напоминает белый свет, который в видимой части оптического спектра остается более или менее постоянным. Хотя для объяснения понятия ПСМ белый шум служит полезным математическим инструментом, следует заметить, что он физически нереализуем, поскольку имеет бесконечную дисперсию. Устройства, которые, как считается, создают белый шум, на самом деле производят шум, ПСМ которого остается постоянной лишь до определенной частоты, а затем на больших частотах убывает.

Предположим, что $x$ — синусоидальный сигнал с частотой $f_c$, амплитудой $A$ и фазой $\varphi$, т. е.

$$x(t) = A\sin(2\pi f_c t + \varphi).$$

Автоковариационная функция в этом случае вычисляется следующим образом:

$$\begin{aligned}
s_x(\tau) &= \lim_{P\to\infty}\frac{1}{P}\int_{-P/2}^{P/2} A^2\sin(2\pi f_c t+\varphi)\sin(2\pi f_c(t+\tau)+\varphi)\,dt = \\
&= \lim_{P\to\infty}\frac{1}{P}\left(\frac{A^2}{2}\right)\int_{-P/2}^{P/2}[\cos(2\pi f_c\tau) - \cos(2\pi f_c\tau + 4\pi f_c t + 2\varphi)]\,dt = \\
&= \lim_{P\to\infty}\frac{A^2}{2P}\left[t\cos(2\pi f_c\tau) + \frac{1}{4\pi f_c}\sin(2\pi f_c\tau + 4\pi f_c t + 2\varphi)\right]_{-P/2}^{P/2} = \\
&= \lim_{P\to\infty}\left\{\frac{A^2}{2}\cos(2\pi f_c\tau) + \left(\frac{A^2}{8\pi f_c}\right)\left(\frac{1}{P}\right)\times\right. \\
&\quad \left.\times[\sin(2\pi f_c\tau + 2\pi f_c P + 2\varphi) - \sin(2\pi f_c\tau - 2\pi f_c P + 2\varphi)]\right\} = \\
&= \frac{A^2}{2}\cos(2\pi f_c\tau).
\end{aligned}$$

Здесь мы учли, что выражение в квадратных скобках всегда меньше 2, поэтому второй член в предпоследнем выражении стремится к нулю, поскольку величина $P$, стоящая в знаменателе, стремится к бесконечности.

Отметим, что в окончательном результате фазовый угол исчез. Так как фазовый угол был произвольным, такой же результат получился бы и при замене синуса косинусом. Таким образом, независимо от начальной фазы автокорреляция для синусоидального сигнала всегда есть косинус с нулевой фазой. Вообще, фазовая информация в процессе получения автокорреляции теряется.

## 2.6. КАК ВЫЧИСЛЯТЬ СРЕДНЕЕ И ДИСПЕРСИЮ

Среднее и дисперсия вычисляются в принципе несложно. Однако, как иногда это случается при вычислениях на ЭВМ, возникают неожиданные и нетривиальные проблемы. Например, для вычисления выборочного среднего обычно берут формулу

$$\bar{x}_k = \left(\sum_{i=0}^{k-1} x(i)\right) + x(k), \quad k = 0, 1, \ldots, N-1, \quad \bar{x} = \frac{1}{N}\bar{x}_{N-1},$$

т. е. накапливают значения данных из последовательности. Если сумма $k$ величин будет существенно больше $x(k)$, то ошибка округления может оказаться того же порядка, что и значение $x(k)$. В этом случае применение вычислений с плавающей точкой или оперирование блоками с плавающей точкой из чисел с фиксированной точкой (что часто реализуется на мини-ЭВМ) может стать проблематичным. Другой метод, дающий более точные результаты, состоит в вычислении частных сумм следующим образом. Предположим, что число данных равно степени 2, $N = 2^m$. Тогда вычисление происходит в $m$ стадий:

$$\bar{x}_1(i) = {}^1/_2[x(2i) + x(2i+1)], \quad i = 0, \ldots, \frac{N}{2} - 1,$$

$$\bar{x}_2(i) = {}^1/_2[\bar{x}_1(2i) + \bar{x}_1(2i+1)], \quad i = 0, \ldots, \frac{N}{4} - 1, \ldots$$

$$\ldots, \bar{x} = \bar{x}_m(0) = {}^1/_2[\bar{x}_{m-1}(0) + \bar{x}_{m-1}(1)],$$

т. е. каждый раз мы последовательно складываем парами числа исходных данных. В качестве пояснения рассмотрим пример, в котором $N = 8$ и $m = 3$. Массив данных состоит из значений $x(0)$, $x(1)$, ..., $x(7)$. Вычисления проводятся в три этапа.

1. Вычисление значений

$$\bar{x}_1(0) = {}^1/_2[x(0) + x(1)], \quad \bar{x}_1(1) = {}^1/_2[x(2) + x(3)],$$
$$\bar{x}_1(2) = {}^1/_2[x(4) + x(5)], \quad \bar{x}_1(3) = {}^1/_2[x(6) + x(7)].$$



2. Вычисление с использованием предыдущих значений

$$\bar{x}_2(0) = {}^1/_2[\bar{x}_1(0) + \bar{x}_1(1)], \quad \bar{x}_2(1) = {}^1/_2[\bar{x}_1(2) + \bar{x}_1(3)].$$

3. Определение выборочного среднего

$$\bar{x} = \bar{x}_3(0) = {}^1/_2[\bar{x}_2(0) + \bar{x}_2(1)].$$

Этот метод позволяет избежать ошибки округления, если данные (как функция индекса $i$) случайны. Когда ведутся вычисления с плавающей точкой или блоками с плавающей точкой из чисел с фиксированной точкой, можно после каждой итерации заново нормировать массив, и значения данных никогда не поглотятся шумом, вызванным округлением. Эти замечания можно отнести ко всякому суммированию последовательности величин.

Естественно, сделанные замечания относятся также к вычислению суммы квадратов и, вообще, к вычислениям перекрестных произведений. Для вычисления дисперсии применяются несколько иные методы. Наименее чувствительна к ошибке округления формула

$$s_x^2 = \frac{1}{N-1}\sum_{i=0}^{N-1}[x(i) - \bar{x}]^2.$$

Вычитание $\bar{x}$ из каждого значения данных устраняет «динамическую составляющую» результирующей величины, что делает менее вероятным существенное влияние ошибок округления на общее суммирование. С другой формулой,

$$s_x^2 = \frac{1}{N-1}\left[\left(\sum_{i=0}^{N-1} x^2(i)\right) - N\bar{x}^2\right], \tag{2.10}$$

связана проблема, заключающаяся в том, что в ней как $\sum x^2(i)$, так и $N\bar{x}^2$ может оказаться большим числом, а $s_x^2$ — маленьким. Это делает весьма вероятным положение, при котором ошибка округления разности в правой части (2.10) будет того же порядка, что и значение $s_x^2$.

Сущность проблемы состоит в том, что такую величину можно сдвинуть вправо (за младшие значащие разряды) регистра-накопителя. Такое явление известно под названием *антипереполнения*.

В следующих главах будут обсуждаться и многие другие проблемы, обусловленные использованием ЭВМ при расчетах. Почти во всех случаях эти проблемы будут связаны с антипереполнением и накоплением ошибки округления.

## 2.7. ВЕРОЯТНОСТНЫЕ ГИСТОГРАММЫ

Из данных можно получать также выборочные вероятностные функции плотности или гистограммы. В отличие от $\bar{x}$ и $s^2$ они строятся неоднозначно, поскольку зависят от некоторых определяющих эти функции параметров. Вычисление гистограммы происходит следующим образом: интервал изменения величины $x$, скажем $a < x < b$, делят на $k$ подынтервалов равной длины (их называют «интервалами группировки») так, чтобы полная область изменения $x$ распалась на $k+2$ интервалов. После этого берут все данные и фиксируют число попаданий на каждый интервал. Изображая графически числа попаданий на каждый интервал, получают гистограмму.

Представим это более формально. Обозначим символом $\{N_j\}$ множество целых чисел, полученное при подсчете попаданий $\{x(i)\}$ в $j$-й интервал. Пусть $c=(b-a)/k$ и $d_j=a+jc$. Тогда $\{N_j\}$ определяются следующим образом:

| $j$ | $N_j$ |
|---|---|
| 0 | [число таких $x$, что $x < a$], |
| ... | |
| $j$ | [число таких $x$, что $d_{j-1} \leqslant x < d_j$], |
| ... | |
| $k$ | [число таких $x$, что $d_{k-1} \leqslant x < b$], |
| $(k+1)$ | [число таких $x$, что $x \geqslant b$]. |

Сказанное поясняет рис. 2.1. Довольно часто $\{N_j\}$ называют карманами. Один из методов указанной сортировки на цифровых

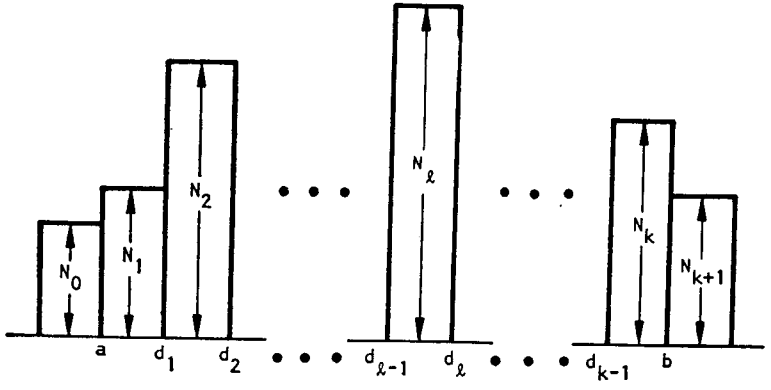


Рис. 2.1. Построение гистограммы.

ЭВМ состоит в переборе всех $x(i)$, $i=1, \ldots, N$, по очереди со следующими проверками:



1) если $x(i) < a$, то прибавить 1 к $N_0$;
2) если $x(i) \geqslant b$, то прибавить 1 к $N_{k+1}$;
3) если предыдущие условия не выполняются и, следовательно, $a \leqslant x(i) < b$, то вычислить

$$j=\left[\frac{x(i)-a}{c}\right]+1. \qquad (2.11)$$

Здесь скобки [ ] обозначают целую часть (т. е. в выражении $j=[a]$ число $j$ представляет собой наибольшее целое, не превосходящее $a$). Получив по формуле (2.11) значение $j$, остается лишь прибавить 1 к $N_j$. Такой прием нетрудно осуществить на большинстве ЭВМ.

На полученных результатах можно построить последовательности трех видов. Первая из них — гистограмма, т. е. последовательность $\{N_j\}$, взятая без всяких изменений. Вторая последовательность $\{P_j\}$, где $P_j$ — выборочная вероятность того, что $d_{j-1} \leqslant x < d_j$, определяется формулой

$$P_j=\frac{N_j}{N}, \quad j=0, 1, \ldots, k+1.$$

Третья последовательность называется выборочной вероятностной функцией плотности (ВФП), члены которой $\{p_j\}$, $j=0, 1, \ldots, k+1$, вычисляются по формуле

$$p_j=\frac{N_j k}{N(b-a)}.$$

Эти числа можно интерпретировать как значения производной функции распределения в серединах соответствующих интервалов.

Приведенной выше процедуре должен предшествовать выбор конкретных значений $a$, $b$ и $k$. Естественно, возникает вопрос: каким должен быть критерий выбора этих трех параметров? Правильного однозначного ответа на него нет. Выбор определяется главным образом предположением о распределении, которому подчиняются данные, и способом сбора данных. Если данные получаются с помощью системы, подобной гипотетической системе сбора данных, из гл. 3, то на них накладываются два ограничения: данные имеют ограниченный промежуток значений и внутри этого промежутка располагаются лишь на конечном числе уровней. Для цифрового преобразователя, имеющего 128 уровней, очевидно, бессмысленно брать $k > 128$, поскольку некоторые из уровней окажутся пустыми. Нетрудно представить также ситуацию, в которой при распределении отсчетов (или их преобразованных эквивалентов) аналого-цифрового преобразования по интервалам выборочной функции может исказиться истинная картина.

Предположим, что цифровой преобразователь имеет 16 уровней и никакого преобразования в технические единицы не произво-

дится. Далее, предположим, что все уровни 0, ..., 15 совершенно равноправны, так что вероятность попадания на каждый из них a priori одинакова и равна 1/16. Если теперь выбрать $k$ равным 12, $a$ и $b$ равными 0 и 15 соответственно, то получится следующее распределение отсчетов по уровням:

| Карман | Интервал | Содержимое уровней |
|---|---|---|
| 1 | 0–1.25 | 0,1 |
| 2 | 1.25–2.50 | 2 |
| 3 | 2.50–3.75 | 3 |
| 4 | 3.75–5.00 | 4 |
| 5 | 5.00–6.25 | 5,6 |
| 6 | 6.25–7.50 | 7 |
| 7 | 7.50–8.75 | 8 |
| 8 | 8.75–10.00 | 9 |
| 9 | 10.00–11.25 | 10,11 |
| 10 | 11.25–12.50 | 12 |
| 11 | 12.50–13.75 | 13 |
| 12 | 13.75–15.00 | 14 |

Ожидаемое число попаданий в карманы 1, 5 и 9 в два раза превосходит остальные, поэтому в гистограмме появятся значительные смещения в соответствующих интервалах, что приведет к неверному результату.

Рассмотренный сейчас пример показывает, с какой осторожностью нужно приступать к вычислениям функций плотностей, и делает очевидным необходимость тщательного изучения критериев, позволяющих избегать подобных ловушек. Один критерий для выбора параметров возникает при попытке дать ответ в другой, но связанной с рассмотренной задаче, а именно в задаче о том, как определять, подчиняются ли данные нормальному (гауссовскому) распределению.

**Критерий согласия хи-квадрат для проверки на гауссовость.** Согласно формуле (2.4), вероятностная функция плотности нормального распределения, обозначенная $\varphi$, определяется формулой

$$\varphi(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

*Вероятностная функция распределения* есть интеграл от функции плотности:

$$\mathbf{P}\{x \leqslant X\} = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{X} \exp\left(-\frac{(t-\mu)^2}{2\sigma^2}\right) dt =$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{(X-\mu)/\sigma} \exp\left(-\frac{t^2}{2}\right) dt = \Phi\left(\frac{X-\mu}{\sigma}\right).$$



Вероятность того, что значение случайной величины лежит в интервале $[\alpha, \beta)$, дает формула

$$\mathbf{P}\{\alpha \leqslant x < \beta\} = \Phi\left(\frac{\beta-\mu}{\sigma}\right) - \Phi\left(\frac{\alpha-\mu}{\sigma}\right).$$

Нормальное распределение принимается как гипотеза во многих исследованиях и возникает естественным образом во многих теоретических расчетах. Поэтому нередко желательно знать, действительно ли по отношению к собранным данным справедливо предположение о гауссовости. Одна из процедур, предназначенных для проверки такой гипотезы, известна под названием *критерия согласия хи-квадрат*. Его основу составляет использование в качестве меры расхождения наблюдаемой и теоретической функций плотности статистики хи-квадрат. При этом гипотеза об эквивалентности проверяется с помощью исследования выборочного распределения хи-квадрат. Частоты, которые следует ожидать для попаданий в $j$-й интервал группировки данных, имеющих гауссовское распределение, называют ожидаемыми частотами и обозначают $F_j$. Расхождение между наблюдаемыми и ожидаемыми частотами равно $(N_j - F_j)$. Чтобы получить полное расхождение, следует учесть каждый интервал, так как

$$\sum_{j=0}^{k+1} N_j = \sum_{i=0}^{k+1} F_j = N.$$

Сумма всех расхождений должна равняться 0. Заметим, что $F_j$, вообще говоря, не целые числа. Они вычисляются следующим образом:

$$F_0 = N\Phi\left(\frac{a-m}{s}\right),$$

$$\cdots$$

$$F_j = N\left\{\Phi\left(\frac{a+jc-m}{s}\right) - \Phi\left(\frac{a+c(j-1)-m}{s}\right)\right\},$$

$$\cdots$$

$$F_{k+1} = N\left(1 - \Phi\left(\frac{b-m}{s}\right)\right).$$

Образуем величину

$$X^2 = \sum_{j=0}^{k+1} \frac{(N_j - F_j)^2}{F_j}. \qquad (2.12)$$

Если выполняются определенные условия, то эта величина имеет распределение $\chi^2$ и может быть сопоставлена с теоретическим распределением, которое обозначается $\chi^2_{n;\,\alpha}$.

Распределение $\chi^2$, которое мы ввели в этой главе раньше, рассмотрено во многих книгах, например у Бендата и Пирсола (1971). Оно зависит от некоторого числа $n$ квадратов независимых случайных величин (называемого числом степеней свободы (ст. св.)). Значение $n$ равно разности числа $k+2$, если карманы используются все до единого, и числа разных независимых линейных ограничений, наложенных на наблюдения. Одно такое ограничение присутствует всегда, поскольку известные частоты для первых $(k+1)$ интервалов группировки определяют последнюю оставшуюся (напомним, что сумма всех частот равна $N$). Два дополнительных ограничения обусловлены подгонкой теоретической нормальной функции плотности к частотной гистограмме наблюдаемых данных. Это связано с тем, что при вычислении $\{F_j\}$ используются не истинные, а выборочные значения среднего и дисперсии. В результате следует вычитать еще двойку из числа степеней свободы. Итак, если используются все $\{N_j\}$, то

$$n=(k+2)-3=k-1.$$

На самом деле число $n$ может быть и меньше, так как карманы, для которых $F<2$, должны объединяться с другими. Детали такой операции мы обсудим позднее.

После того как число ст. св. $n$ получено, проверка гипотезы проводится следующим образом. Допустим, что величина $x$ имеет нормальное распределение. Группируя выборочные наблюдения по $(k+2)$ интервалам группировки и определяя значения частот $F_j$ для каждого интервала, основываясь на выборочных значениях среднего и дисперсии, вычисляют величину (2.12). Всякое отклонение выборочной функции плотности от нормальной приводит к возрастанию значения $X^2$. Гипотеза о том, что данные имеют нормальное распределение, принимается, если

$$X^2 \leqslant \chi^2_{n;\,\alpha}.$$

В этом случае говорят, что гипотеза принимается с уровнем значимости $\alpha$. Если же значение $X^2$ превосходит $\chi^2_{n;\,\alpha}$, то гипотеза отвергается с тем же уровнем значимости $\alpha$. Широко используются общепринятые уровни значимости 5, 10, 20% (соответствующие доверительным уровням 95, 90 и 80%). Особые значения доверительного уровня выбираются исследователями реже. Авторы, как правило, если не оговорено противное, предпочитают другим значениям $\alpha$ традиционное значение, равное 5%.

Когда для проверки на гауссовость выбирается критерий согласия хи-квадрат, можно воспользоваться выражением для числа интервалов группировки при данном значении $N$, предложенном Кендаллом и Стьартом (1973). Они, предполагая, что данные некоррелированы и $\alpha=0.05$, получили следующую формулу:

$$\text{Число интервалов группировки} = 1.87\,(N-1)^{2/5}.$$



Значения этой функции приводятся в табл. 2.1. Выше отмечалось, что смещения в гистограмме могут быть значительными, если

*Таблица 2.1.*

**Наименьшие оптимальные числа интервалов группировки $k$ для выборки объема $N$ при $\alpha=0.05$**

| $N$ | $k$ | $N$ | $k$ |
|---|---|---|---|
| 200 | 16 | 20 000 | 94 |
| 400 | 20 | 40 000 | 129 |
| 600 | 24 | 70 000 | 162 |
| 800 | 27 | 100 000 | 187 |
| 1 000 | 30 | 200 000 | 247 |
| 1 500 | 35 | 400 000 | 326 |
| 2 000 | 39 | 700 000 | 407 |
| 4 000 | 57 | 1 000 000 | 470 |
| 7 000 | 65 | 1 140 000 | 500 |
| 10 000 | 74 | | |

число интервалов группировки станет сравнимым с числом уровней отсчетов цифрового преобразователя.

Статистики, применяя критерий хи-квадрат, обычно пользуются эмпирическим правилом, согласно которому каждому интервалу должна соответствовать частота, не меньшая 2. Такое ограничение позволяет выбрать разумные значения $a$ и $b$. Наименьшего числа попаданий следует ожидать для крайних карманов. Поэтому значение $a$ определяется из уравнения

$$2=N\left\{\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{(a-m)/s} e^{-t^2/2}\,dt\right\},$$

которое можно разрешить относительно $a$. После этого значение $b$ находят по простой формуле

$$b=2m-a,$$

а параметр $k$ — по формуле

$$k=[\text{число интервалов группировки}]-2.$$

После того как эти три параметра определены, можно вычислять выборочную функцию плотности и ожидаемые частоты для нормального распределения.

На следующем шаге с помощью сравнения решается вопрос о том, принять или отвергнуть гипотезу о гауссовском распре-

делении данных. Как упоминалось выше, обычно эту процедуру сравнения $X^2$ и $\chi^2_{n;\,\alpha}$ начинают с вычисления параметра $\alpha'$, определенного равенством

$$\alpha' = \mathbf{P}\left\{X^2 > \chi^2_{n;\,\alpha'}\right\}$$

Этот параметр зависит только от $X^2$ и $n$. После вычисления его сравнивают с заранее выбранным значением $\alpha$ и проводят проверку на гауссовость, опираясь на следующее правило: гипотеза при $\alpha' \leqslant \alpha$ принимается, при $\alpha' > \alpha$ отвергается. Укажем один метод вычисления величины $\alpha'$: для нечетных $n$ применяем формулу

$$\alpha' = 2\Phi(X) - 1 - \sqrt{\frac{2}{\pi}}\, e^{-X^2/2}\left[\sum_{r=1}^{(n-1)/2} \frac{X^{2r-1}}{1\cdot 3\cdot 5\ldots(2r-1)}\right]$$

и для четных $n$ — формулу

$$\alpha' = 1 - e^{-X^2/2}\left[1 + \sum_{r=1}^{(n-2)/2} \frac{X^{2r}}{2\cdot 4\cdot 6\ldots(2r)}\right].$$

Хотя существуют более эффективные и точные алгоритмы, как правило, для задач анализа временных рядов вполне достаточно предложенного.

## 2.8. ВЕРОЯТНОСТНЫЕ ФУНКЦИИ ПЛОТНОСТИ ПИКОВ

С определением пиков, максимумов и минимумов связаны известные терминологические трудности. При изучении экстремальных значений данного ряда представляют интерес по крайней мере три различных вопроса:

распределение наибольшего (или наименьшего) значения в ряде длины $p$,

распределение наибольшего значения между двумя пересечениями нулевого уровня,

распределение значений пиков.

Рассмотрим задачу первого типа. Предположим, что имеется $N$ независимых наблюдений $\{x_i\}$, имеющих функцию плотности $f(x)$ и функцию распределения $F(x)$. Тогда функция плотности наибольшего значения имеет вид

$$f(x_{\max},\ N) = Nf(x)[F(x)]^{N-1}.$$

Функция распределения значения $x_{\max}$ определяется формулой

$$F(x_{\max},\ N) = [F(x)]^N,$$

а математическое ожидание $x_{\max}$ — формулой

$$E[x_{\max}] = \int_{-\infty}^{\infty} xNf(x)F^{N-1}(x)\,dx. \qquad (2.13)$$



Допустим, что $x_i$ $(i = 1, \ldots, N)$ — независимые величины, равномерно распределенные на $[-1/2,\ 1/2]$. Тогда

$$f(x_i) = \begin{cases} 1 & \text{при} \quad -1/2 \leqslant x_i \leqslant 1/2, \\ 0 & \text{в остальных случаях;} \end{cases}$$

$$F(x_i) = \begin{cases} 0 & \text{при} \quad x < -1/2, \\ x + 1/2 & \text{при} \quad -1/2 \leqslant x \leqslant 1/2, \\ 1 & \text{при} \quad x > 1/2; \end{cases}$$

$$F(x_{\max},\ N) = \begin{cases} 0 & \text{при} \quad x_{\max} < -1/2, \\ (x+1/2)^N & \text{при} \quad -1/2 \leqslant x_{\max} \leqslant 1/2, \\ 1 & \text{при} \quad x_{\max} > 1/2; \end{cases}$$

$$E[x_{\max}] = \int_{-1/2}^{1/2} xN(x+1/2)^{N-1}\,dx = \frac{N-1}{2(N+1)}.$$

Хотя вычисление $E[x_{\max}]$ по формуле (2.13) может в теории вызвать затруднения, нахождение $x_{\max}$ по выборке совсем просто. На практике, даже если и не интересоваться $x_{\max}$ и $x_{\min}$ самими по себе, все же разумно искать их с помощью какой-нибудь программы, поскольку это нередко оказывается полезным при проверке данных с целью удаления неправдоподобных значений.

Задача анализа пиков второго типа встречается при изучении множества наибольших значений между пересечениями нулевого уровня. Найти такие значения нетрудно. Это можно сделать с помощью следующего типичного алгоритма. Как всегда, предполагается, что ряд $\{x_i\}$ имеет $N$ точек.

1. Вычислить таблицу значений $\{I_k\}$, $k = 1, \ldots K$, в которой $I_k$ определяются парами неравенств

$$x_{I_k} \leqslant 0 \quad \text{и} \quad 0 < x_{(I_k+1)}$$

или

$$0 < x_{I_k} \quad \text{и} \quad x_{(I_k+1)} \leqslant 0.$$

2. Определить минимум значений $x_i$, $i = (I_k+1), \ldots, I_{k+1}$, если $0 < x_{I_k}$, и максимум в другом случае.

3. Далее, если $x_{I_k} < 0$, то найти максимум $x_i$, $i = (I_k+1), \ldots$ $\ldots, I_{k+1}$; в противном случае найти минимум этих же значений.

Полученные значения максимумов и минимумов можно использовать двояко. Во-первых, применить приемы предыдущего параграфа и построить выборочные функции плотности отдельно для максимумов и минимумов. Во-вторых, можно сравнить значения абсолютных значений минимумов с максимумами. Для данных многих типов статистические свойства пересечений нулевого уровня хорошо известны, поэтому максимумы (или минимумы) между

соседними пересечениями нулевого уровня, как правило, поддаются анализу. Несколько более сложные проблемы возникают в третьем случае, указанном в начале этого параграфа, а именно при изучении пиков.

Под пиком здесь подразумевается наибольшее значение между двумя относительными минимумами. Например, на рис. 2.2 есть

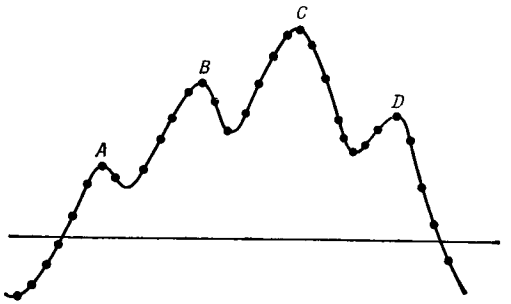


Рис. 2.2. Различные типы пиковых значений.

только один пик второго типа (точка C), а пиков третьего типа — целых три(A, B и D). Приведем процедуру для выделения таких значений.

1. Для максимумов — найти все такие $x_p$, что

$$x_p - x_{p-1} > 0, \quad x_p - x_{p+1} > 0.$$

2. Аналогично для минимумов — найти все такие $x_q$, что

$$x_q - x_{q-1} < 0, \quad x_q - x_{q+1} < 0.$$

Эти множества максимумов и минимумов также можно обработать с помощью приемов, изложенных в предыдущем параграфе.

## 2.9. МНОГОМЕРНЫЕ ФУНКЦИИ ПЛОТНОСТИ

Для двух рядов данных $\{x(i)\}$ и $\{y(i)\}$ можно вычислить их совместную выборочную ВФП, или, иначе, выборочную гистограмму. Такое вычисление производится при помощи деления обоих интервалов изменения данных на $k_1$ и $k_2$ подынтервалов соответственно, как это показано на рис. 2.3. Вообще, в двумерном случае потребуется больший по сравнению с одномерным случаем объем памяти. Например, если для самого ряда $\{x(i)\}$ потребуется $k_1+2$ ячеек, а для $\{y(i)\}$ потребуется $k_2+2$ ячеек, то необходимое число карманов для их совместного распределения может оказаться равным $(k_1+2)(k_2+2)$. Это количество может быть чрезмерно большим, особенно если обрабатывается $r$ таких рядов. В этом случае из них можно выбрать $r(r-1)/2$



различных пар рядов. Предполагая, что $k_1 = k_2 = \ldots = k_r = k$, получим такое выражение для общего необходимого объема памяти ЭВМ:

$$\frac{r(r-1)}{2}(k+2)^2.$$

При $r = 50$ и $k = 100$ это число равно 12 744 900, в то время как для одномерных гистограмм с теми же параметрами нужно всего лишь 5100 ячеек. (См. рис. 2.3.)

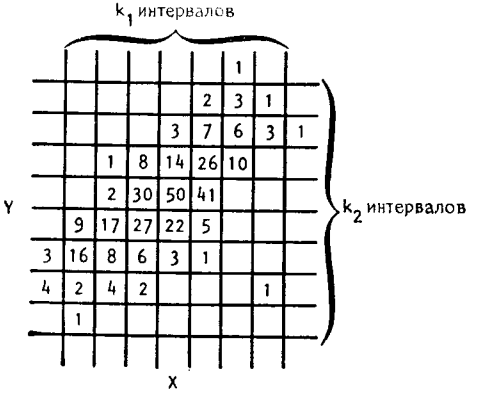


Рис. 2.3. Выборочная гистограмма для совместной функции плотности.

Добиться наглядного представления результатов на дисплее довольно трудно. Машинные распечатки выборочных совместных гистограмм неудобны из-за обилия цифр, от которых рябит в глазах. Изобразить результаты графически тоже трудно. С некоторым успехом можно пользоваться контурными графиками, но их построение не так уж просто. Несомненно, создание трехмерных дисплеев в будущем должно повысить значение совместных гистограмм. Однако пока полностью удовлетворительных методов представления результатов на дисплее нет, и это сильно усложняет построение выборочных гистограмм, мешает интерпретировать данные.

## 2.10. ПРИМЕРЫ И ПРИМЕНЕНИЯ ВЕРОЯТНОСТНЫХ ФУНКЦИЙ ПЛОТНОСТИ

На рис. 2.4 представлена выборочная гистограмма, построенная по 1001 значению равномерного белого шума, полученного на цифровой ЭВМ. Теоретическая функция плотности определяется формулой

$$f(x) = \begin{cases} 1, & \text{если } -1/2 \leqslant x < 1/2, \\ 0, & \text{в противном случае.} \end{cases}$$

В этом случае теоретические значения среднего и дисперсии равны

$$\mu = \int_{-\infty}^{\infty} xf(x)\,dx = 0,$$

$$\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\,dx =$$

$$= \int_{-1/2}^{1/2} x^2\,dx = \frac{x^3}{3}\Big|_{-1/2}^{1/2} = \frac{1}{24} + \frac{1}{24} = \frac{1}{12} = 0.08333333.$$

С этими значениями отлично согласуются величины, полученные по выборке и представленные в нижнем правом углу рисунка.

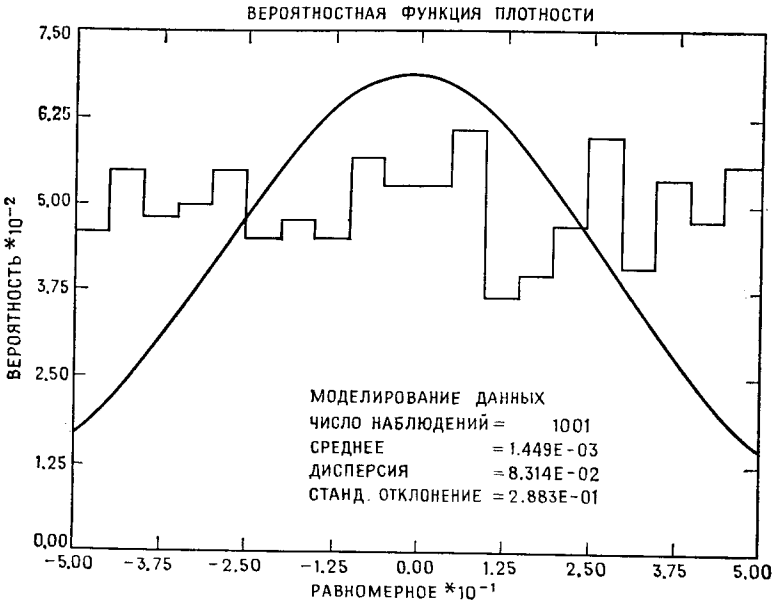


Рис. 2.4. Выборочная функция плотности последовательности значений равномерно распределенного белого шума.

Значение $P_j$ выборочной гистограммы в $j$-м интервале есть отношение числа наблюдений, попавших в $j$-й карман, ко всему числу наблюдений:

$$P_j = \frac{N_j}{N} = (\text{выборочная вероятность того, что } [d_{j-1} \leqslant x < d_j]).$$



Поэтому математическое ожидание $P_j$ равно

$$E[P_j] = \int_{-\infty}^{\infty} P_j f(x)\,dx = \int_{d_{j-1}}^{d_j} f(x)\,dx.$$

В случае рассматриваемого равномерного шума

$$E[P_j] = \int_{d_{j-1}}^{d_j} dx = d_j - d_{j-1}.$$

В частности, в нашем случае эта разность равна примерно 0.05.

Отметим, что на выборочную гистограмму наложена кривая плотности гауссовского распределения с теми же значениями выборочного среднего и дисперсии. Очевидно, что они совершенно не согласуются, и критерий $\chi^2$ здесь не нужен.

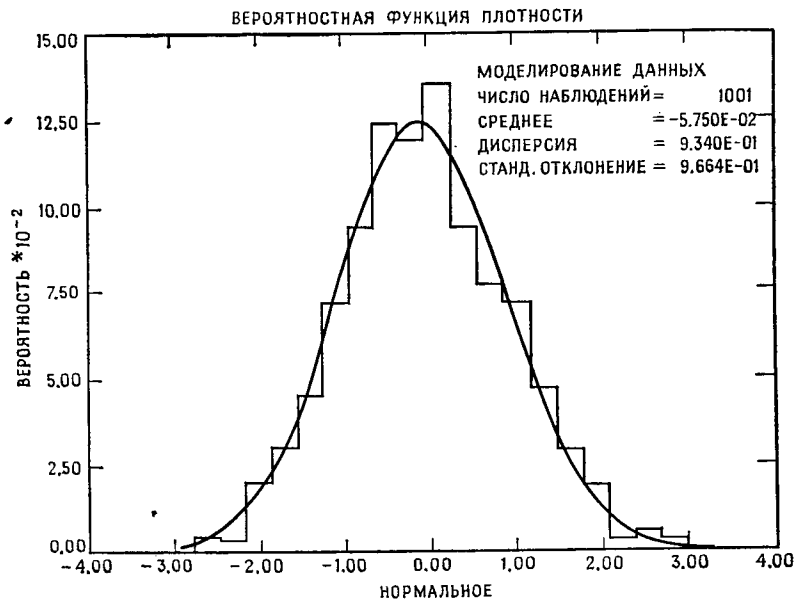


Рис. 2.5. Выборочная функция плотности последовательности значений гауссовского белого шума.

На рис. 2.5 показано, как такой критерий используется в случае гауссовского шума, полученного с помощью цифрового генератора псевдослучайных чисел. Математическое ожидание и дисперсия равны 0 и 1 соответственно. Выборочная дисперсия, рав-

ная 0.934, хотя и немного меньше теоретической, но лежит в разумных пределах ожидаемой величины. Данные согласуются с гипотезой о гауссовости на 95%-м доверительном уровне (параметры критерия: выборочная величина $X^2 = 18.39$, $\mathbf{P}\{\chi^2, 1000, 0.05\} = 0.69825$).

Теперь рассмотрим следующий случай, в котором ряд данных образован значениями синусоиды, полученными по формуле

$$x(i) = \sin(0.15i), \quad i = 0, \ldots, 1000.$$

Случайная синусоида с постоянными коэффициентами имеет такие статистические параметры:

$$\mu = 0, \quad \sigma^2 = 1/2,$$

$$f(x) = \begin{cases} \dfrac{1}{\pi\sqrt{1-x^2}}, & -1 \leqslant x \leqslant 1, \\ 0, & |x| > 1. \end{cases} \qquad (2.14)$$

Заметим, что на части интервала (—1, 1) выборочная функция плотности больше единицы. Однако интеграл от вероятностной функции плотности по всей или по какой-нибудь части интервала будет меньше или равен единице.

На рис. 2.6 изображена выборочная гистограмма. Выборочные значения среднего и дисперсии, а также гистограмма находятся в хорошем соответствии с (2.14).

Большой интерес представляет следующий пример, в котором гауссовский шум пропускается через узкополосный фильтр. На достаточно коротких отрезках данных выход такого фильтра довольно похож на простую синусоиду. Однако при вычислении выборочной функции плотности мы не увидим сходства с функцией плотности синусоиды. Это показывает рис. 2.7, на котором изображена выборочная функция плотности, построенная по 60 точкам, полученным на выходе полоснопропускающего фильтра, у которого расстояние между точками половинной мощности составляет 1% общей допустимой полосы частот (т. е. 1% частоты Найквиста), а центр частот расположен в точке, соответствующей 15% частоты Найквиста. Как видно из рисунка, эта выборочная функция плотности не похожа на функции плотностей, изображенные на рис. 2.6 и 2.5.

На рис. 2.8 приведена выборочная функция плотности для таких же данных, но взятых на более длинном промежутке. В этом случае выборочная функция плотности приближает гауссовскую функцию плотности, пожалуй, хорошо. Здесь критерий $\chi^2$ для проверки гауссовости неприменим, поскольку не выполняется предположение о независимости значений данных.



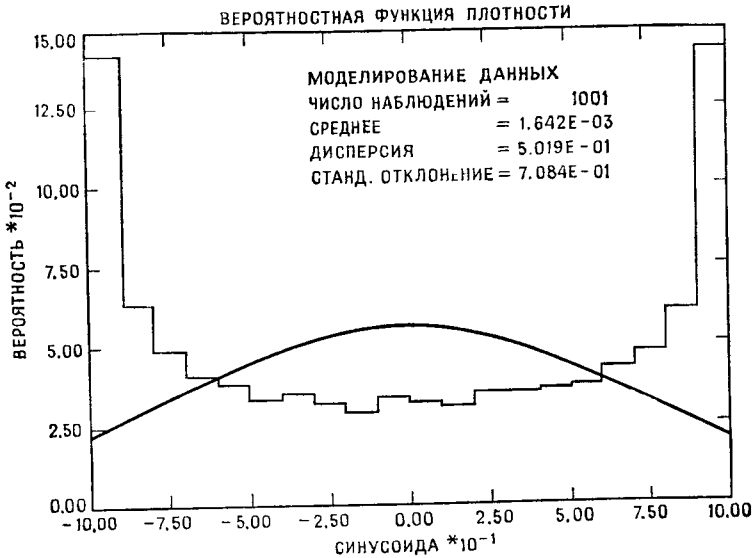


Рис. 2.6. Выборочная функция плотности синусоиды.

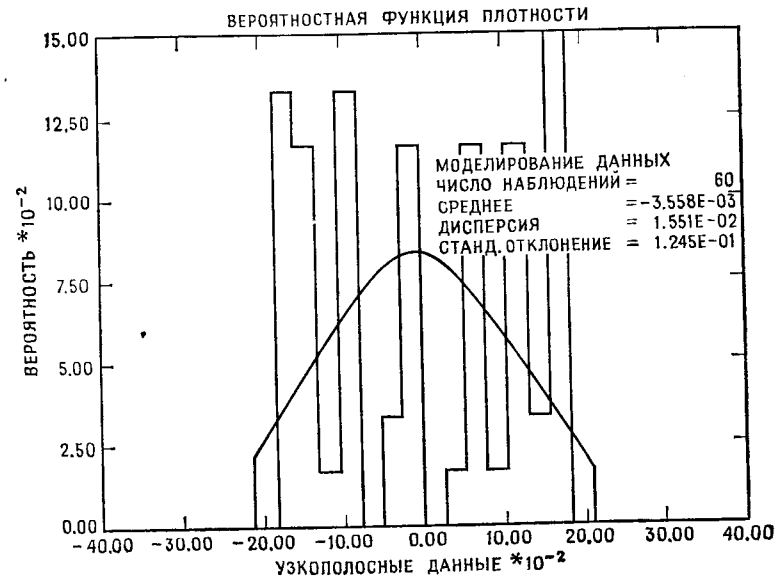


Рис. 2.7. Выборочная функция плотности для небольшого участка процесса, полученного узкополосной фильтрацией.

Этого вполне можно было бы избежать, если $N$ (которое в этом случае равно 10 001) выбирать таким образом, чтобы учитывать возможность зависимых наблюдений. В таком случае $N$ сократилось бы до 1% первоначального значения, т. е. до 100 наблюдений. В результате произошло бы соответствующее уменьшение значения выборочного $\chi^2$ разницы выборочной функции плотности и плотности гауссовского распределения, построенной по выборочному среднему и дисперсии.

На самом деле в условиях применимости центральной предельной теоремы гауссовский шум или, вообще, шум с любым распределением после линейной фильтрации сходится к гауссовскому шуму.

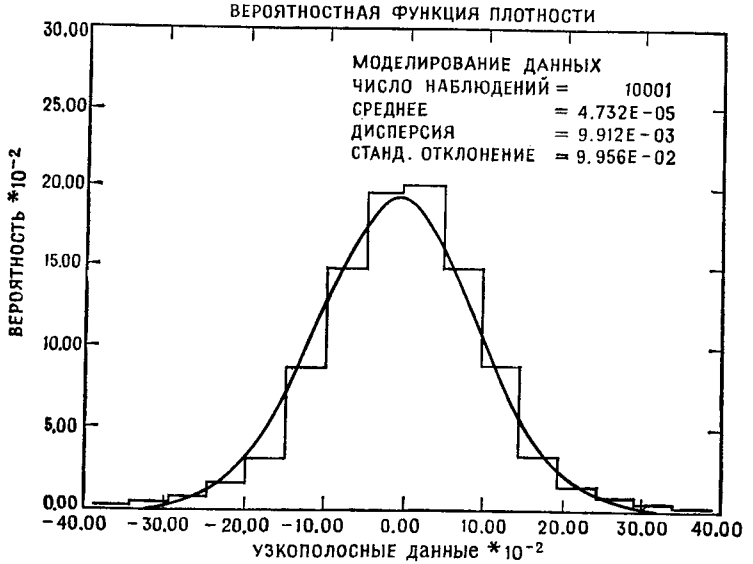


Рис. 2.8. Выборочная функция плотности для большого участка процесса, полученного узкополосной фильтрацией. Сюда включены данные, полученные для рис. 2.7.

Применение критерия $\chi^2$ в таком случае может оказаться не совсем простым делом, поскольку в выборе точно определить число независимых оценок трудно.

**Уточнение данных с помощью вероятностных гистограмм.** Один из способов уточнения данных состоит в том, чтобы изобразить все данные на графике и просмотреть их. Для больших объемов данных это утомительно, требует больших затрат времени и средств. В другом способе уточнения данных, более экономичном, используются выборочные функции плотности. Вообще говоря,



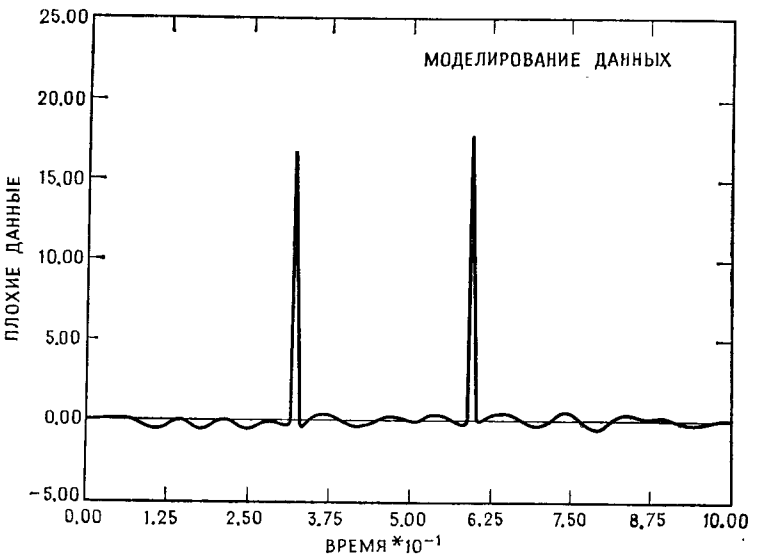


Рис. 2.9. Временная последовательность с двумя неправдоподобными значениями.

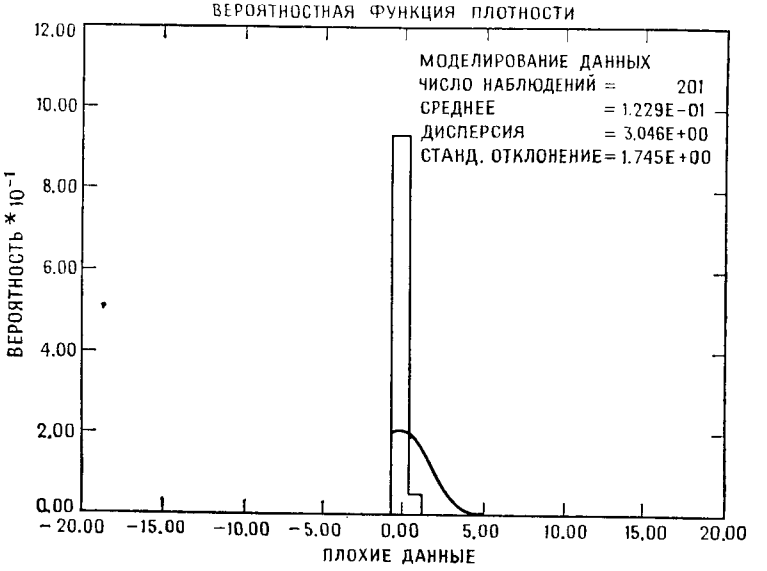


Рис. 2.10. Выборочная функция плотности данных, изображенных на рис. 2.9. Обратите внимание, как данные группируются вместе.

выборочные функции плотности несут бóльшую информацию о природе данных и выявляют проблемы разного характера. Для больших количеств данных вывод выборочной функции плотности по затратам может сравниться с полной распечаткой данных или с получением их графика.

С помощью гистограмм быстро выявляются неправдоподобные значения. На рис. 2.9 показаны данные, полученные таким способом, чтобы появлялись неправдоподобные значения. Соответствующая выборочная функция плотности представлена на рис. 2.10. Как видно, большинство данных тесно сгруппировалось вместе, а две неправдоподобные точки расположились далеко справа.

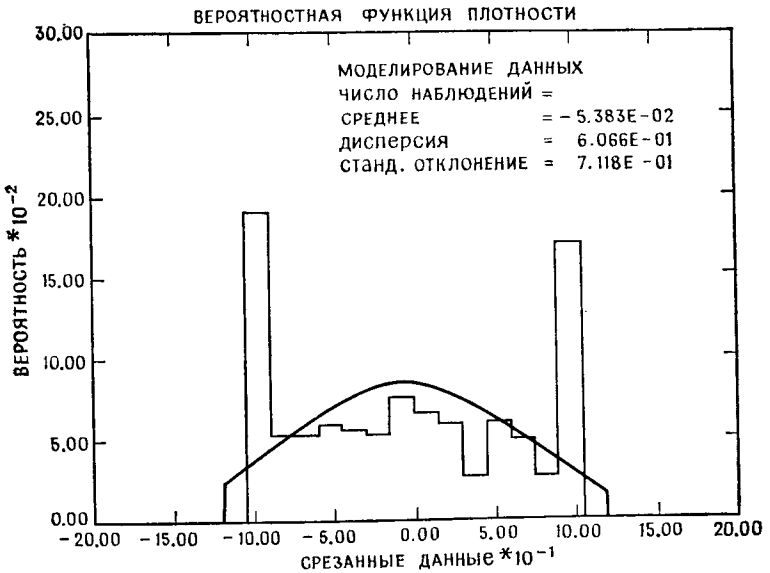


Рис. 2.11. Выборочная функция плотности срезанных данных.

Все промежуточные карманы пусты. Такая картина типична для данных с неправдоподобными значениями, поскольку наличие неправдоподобных данных загоняет все настоящие данные в один или два кармана.

Выборочная функция плотности, порождающая проблему другого характера, представлена на рис. 2.11. Здесь крайние карманы содержат значительное число наблюдений, в то время как ни один из оставшихся не содержит наблюдений. Это указывает на «срезание». Срезание может происходить по нескольким причинам. Чаще всего оно происходит из-за ухода аналогового усилителя в точки, где нарушается линейность, или из-за такой установки аналого-цифрового преобразователя, при которой не покрывается весь возможный интервал изменения данных в некотором канале.



Чтобы произвести процедуру уточнения данных, рекомендуется делать следующее. Предполагается, что для получения данных используется программа, в которой калибровка задается самим пользователем.

1. Задайте калибровку и при ней вычислите выборочную функцию плотности и основные статистические параметры для каждого файла полученных на выходе данных.

2. Прежде чем начинать дальнейшую обработку, просмотрите полученную выборочную функцию плотности. При этом нужно учесть следующее:

a) Необходимо следить за временем (или какой-то другой независимой переменной) с тем, чтобы установить, совпадает ли обработанный промежуток с истинным и согласуется ли число точек для выборочной функции плотности с предполагаемым.

b) Следует проверить статистики для каждой зависимой переменной. Необходимо установить, насколько отвечают физической реальности выборочная функция плотности с точки зрения тех проблем, которые только что были затронуты, и тех, которые появятся сами, а также среднее, дисперсия, максимум и минимум.

Предложенная процедура не гарантирует, что данные станут хорошими, но она экономно и быстро выявляет возникающие проблемы.

## Упражнения

2.1. Докажите, что $E(s^2) \neq \sigma^2$, если $s^2 = N^{-1} \sum_{i=0}^{N-1} (x_i - x)^2$.

2.2. Докажите, что дисперсия синусоиды $r(t)$ с амплитудой $A$ и произвольной фазой равна $A^2/2$.

2.3. Докажите соотношение (2.2), т. е.

$$\sigma_x^2 = E(x^2) - E^2(x)$$

Ответ:

$$\begin{aligned}\sigma^2 &= E(x-\mu)^2 = E(x^2 - 2x\mu + \mu^2) = \\ &= E(x^2) - 2E(x)\mu + \mu^2 = \\ &= E(x^2) - 2\mu^2 + \mu^2 = E(x^2) - \mu^2.\end{aligned}$$

2.4. Докажите, что корреляция ограничена следующими пределами?

$$-1 \leqslant \frac{\operatorname{cov}(x, y)}{\sigma_x \sigma_y} \leqslant 1$$

Ответ:

$$\operatorname{cov}(x, y)=E(x-\mu_x)(y-\mu_y)=E(xy-x\mu_y-y\mu_x+\mu_x\mu_y)=$$
$$=E(xy)-\mu_x\mu_y-\mu_y\mu_x+\mu_x\mu_y=$$
$$=E(xy)-\mu_x\mu_y,$$
$$\operatorname{cov}^2(x,y)=E^2(xy)-2E(xy)\mu_x\mu_y+\mu_x^2\mu_y^2\geqslant$$
$$\geqslant E(x^2)E(y^2)-E(x^2)\mu_y^2-E(y^2)\mu_x^2+\mu_x^2\sigma_y^2.$$

2.5. Вычислите ковариационную функцию синусоиды.
2.6. Приведите формулу для sin-члена в формуле (2.3).

**Задачи для решения на ЭВМ.**

2.7. Смоделируйте равномерный шум и повторите рис. 2.5.
2.8. Смоделируйте гауссовский шум и повторите рис. 2.6.
2.9. Смоделируйте синусоиду и повторите рис. 2.7.

# Глава 3

# СБОР И ПРЕДВАРИТЕЛЬНАЯ ОБРАБОТКА ДАННЫХ

## 3.1. ВВЕДЕНИЕ

В этой главе мы остановимся на некоторых основных задачах, возникающих в начальной стадии анализа временных рядов. Данные, которые будут дальше рассматриваться, распадаются по типу на два разных класса — *непрерывные* (или *аналоговые*) данные и данные, имеющие по существу *цифровой* характер.

Для работы более удобны, пожалуй, данные второго типа. К ним можно отнести сведения об уровне цен на бирже, статистику занятости, годовые уровни осадков и многие другие. Даже если поиск и регистрация таких данных могут оказаться утомительным занятием, то со статистической точки зрения или с точки зрения приложений никаких сложностей не возникает.

Исследованию же непрерывных данных сопутствует ряд проблем. Среди них измерение данных, передача и (или) их запись, преобразование из аналоговой формы в цифровую.

Вопросам измерений посвящена книга Мэгреба и Бломквиста (1971). Задача здесь сводится, по сути дела, к конструированию или выбору *датчика* — устройства, преобразующего измеряемую физическую величину в электрический потенциал. Как правило, хотя и не всегда, полученный в результате потенциал пропорционален измеряемой физической величине. К таким приборам относятся микрофоны, гидрофоны, акселерометры и устройства, измеряющие давление.

Обычно датчик выбирается так, чтобы его передаточная функция была линейной в пределах предусмотренного промежутка. Идеальное линейное устройство создать невозможно. Интервал линейности ограничен, и большинству датчиков в большей или меньшей степени присуще наличие *гистерезиса*. Последнее означает, что при изменении элемента входа от начальной точки до конечной с возвращением в начальную точку две соответствующие траектории выхода, которые должны быть одинаковыми, в точности не совпадут. Такой график называется *гистерезисной кривой* инструмента. Для ряда хорошо сконструированных аппаратов получающаяся ошибка пренебрежимо мала.

Даже и совершенно линейный датчик может иметь передаточную функцию, заглушающую вход на некоторых промежутках частот. Например, передаточные функции для некоторых микро-