М.С. Ермаков, А.Ф. Сизова,
Т.М. Товстик

ЭЛЕМЕНТЫ МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ

Санкт-Петербург
2001

Санкт-Петербургский государственный университет

М.С. Ермаков, А.Ф. Сизова, Т.М. Товстик

ЭЛЕМЕНТЫ МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ

Учебное пособие

Издательство С.-Петербургского университета
2001

ББК 22.172
Е72

Р е ц е н з е н т ы: проф. С.М. Ермаков (С.-Петерб. гос. ун-т),
доц. Ю.К. Кривулин (С.-Петерб. гос. ун-т)

Печатается по постановлению
Редакционно-издательского совета
С.-Петербургского государственного университета

Ермаков М.С., Сизова А.Ф., Товстик Т.М.
Е72 Элементы математической статистики: Учеб. пособие. — СПб.: Изд-во С.-Петерб. ун-та, 2001. — 148 с.

Учебное пособие содержит основную часть вводного курса по математической статистике, читаемого на факультете менеджмента. Пособие можно использовать для первого знакомства с методами математической статистики.

Предназначено для студентов факультета менеджмента, а также может быть использовано на других факультетах университета.

ББК 22.172

ВВЕДЕНИЕ

Статистика представляет собой совокупность знаний и методов, использующихся для сбора, описания, интерпретации и анализа данных с целью получения информации и аргументации в принятии решений. В настоящее время статистика стала одним из наиболее важных курсов при подготовке экономистов и бизнесменов, поскольку она является ключевым ингредиентом, необходимым для понимания бухгалтерского учета, экономики, финансов, маркетинга, организационного поведения и других курсов. В большинстве наук статистика играет столь важную роль, что выделяется даже как бы в отдельную отрасль данной науки:

биология, медицина — биометрика,
экономика — эконометрика,
техника — технометрика,
психология — психометрика.

По всем этим разделам наук выходит по несколько журналов (широко известных), где обсуждаются методы анализа данных.

Цель данного курса — не только научить обрабатывать данные, но и научить понимать насколько хорошо можно извлекать информацию в условиях неопределенности.

С точки зрения приложения математической статистики к экономике, бизнесу и финансам ее принято разделять на *описательную статистику* и *теорию статистических выводов*.

Описательная статистика занимается методами сбора статистических данных, получением наглядной информации о данных, а также первичной обработкой данных.

Теория статистических выводов (по существу это и есть математическая статистика) позволяет сделать более-менее точные оценки параметров в рассматриваемой задаче, оценить точность

оценки параметров, проверить гипотезу о соответствии предполагаемого заключения или модели реальным данным.

Если описательная статистика находит свое применение в *экономической статистике*, то методы теории статистических выводов являются основой *эконометрики*. Используя методы математической статистики, эконометрика на основе объективных данных проверяет адекватность экономической модели, оценивает параметры модели, дает прогнозы развития. Таким образом, именно эконометрика объединяет в единое целое экономическую теорию и объективные факты, соединяет реальные данные с экономикой.

Задачи теории статистических выводов принято разделять на два класса: задачи проверки гипотез и задачи статистического оценивания.

Теория проверки гипотез занимается проверкой гипотез о том, что рассматриваемая характеристика имеет данное значение, принадлежит данной совокупности или что предложенная модель адекватно описывает реальные данные.

Теория статистического оценивания занимается оцениванием значений характеристик рассматриваемой модели, а также *обязательно* указывает погрешности оценивания на основе построения *интервальных оценок*, или, что то же самое, *доверительных интервалов*.

Глава 1

ОПИСАТЕЛЬНАЯ СТАТИСТИКА

1. Выборка

Типы данных. Данные бывают *количественные* и *качественные*. Количественные данные представляют собой числовые значения. Качественные — некоторые качественные характеристики (пол, марка автомобиля, специальность человека и т. п.). Наше внимание будет обращено в основном на количественные данные.

Наиболее часто исходные данные представляют собой наблюдения, полученные при независимых повторениях опытов, происходящих примерно при одних и тех же условиях.

Последовательность n независимых одинаково распределенных наблюдений $x_1, \ldots, x_n$ случайной величины X с функцией распределения $F(x)$ назовем выборкой объема n, а сами наблюдения — выборочными значениями.

Если последовательность выборочных значений расположить в порядке возрастания $x^{(1)} \leq x^{(2)} \leq \ldots \leq x^{(n)}$, то такую последовательность принято называть *вариационным рядом.*

Ясно, что новый эксперимент приведет к новой выборке, поэтому любую выборку можно считать случайной и каждое наблюдение x_i рассматривать как случайную величину с функцией распределения $F(x)$.

Пусть выборочные значения $x_1, \ldots, x_n$ лежат в интервале $[a, b)$. Разобьем интервал $[a, b)$ на k подынтервалов

$$[e_0, e_1), \ldots, [e_{k-1}, e_k), \qquad e_0 = a, \quad e_k = b.$$

Тогда число наблюдений, попавших в интервал $[e_{i-1}, e_i), i = 1, \ldots, k$, называется *частотой* $\nu_i = \nu_i[e_{i-1}, e_i)$ наблюдений в интервале $[e_{i-1}, e_i)$.

Частота ν_i наблюдений, деленная на число наблюдений n, называется *относительной частотой* $\omega_{in} = \nu_i/n$. Ясно, что при больших объемах выборки n относительная частота ω_i является хорошей оценкой вероятности $p_i = P(e_{i-1} \leq X < e_i)$ попадания случайной величины X в интервал $[e_{i-1}, e_i)$.

Этот результат следует из закона больших чисел (а также из его простейшей формы — теоремы Бернулли), который утверждает, что

$$\omega_{in} = \frac{\nu_i}{n} \xrightarrow{P} p_i, \tag{1.1}$$

т. е. ω_{in} сходится по вероятности к p_i:

$$\lim_{n\to\infty} P(|\omega_{in} - p_i| > \epsilon) = 0 \quad \text{для любого} \quad \epsilon > 0$$

(относительная частота наблюдений в данном интервале $[e_{i-1}, e_i)$ сходится к вероятности появления наблюдений p_i в этом интервале).

Действительно, вероятность того, что каждая из случайных величин попадет в интервал $[e_{i-1}, e_i)$ (назовем это успехом), равна p_i и того, что не попадет, $1 - p_i$. Поэтому по теореме Бернулли (или по более общему утверждению — Закону Больших Чисел)

относительная частота успехов ω_{in} стремится к вероятности успеха p_i и эта сходимость имеет место по вероятности.

2. Гистограмма и эмпирическая функция распределения

Как мы знаем, случайные величины характеризуются их функцией распределения $F(x) = P(X < x)$ и плотностью распределения $f(x) = F'(x)$. Поэтому, чтобы понять, как распределена выборка, в первую очередь представляется необходимым оценить функцию распределения и ее плотность.

Для оценки функции распределения $F(x)$ строится эмпирическая (выборочная) функция распределения $F_n^*(x)$. *Эмпирическая (выборочная) функция* распределения $F_n^*(x)$ равна относительной частоте ω_x попадания наблюдений из выборки $x_1, \ldots, x_n$ в интервал $(-\infty, x)$, иначе говоря,

$$F_n^*(x) = \frac{\nu_x}{n},$$

где $\nu_x = \nu(-\infty, x)$ — общее число выборочных значений X_i, $1 \leq i \leq n$, меньших x, или, что то же самое, частота наблюдений, меньших x.

Итак, функция распределения $F(x)$ есть вероятность события $(X < x)$, а эмпирическая функция распределения $F_n^*(x)$ определяет относительную частоту этого же события. Как мы уже говорили, относительная частота события $(X < x)$ стремится к вероятности этого события, т. е. $F_n^*(x)$ стремится по вероятности к вероятности этого события, равной $F(x)$. На языке формул это записывается следующим образом: для любого $\epsilon > 0$

$$\lim_{n\to\infty} P(|F_n^*(x) - F(x)| < \epsilon) = 1. \tag{1.2}$$

Это обосновывает использование эмпирической функции распределения $F_n^*(x)$ как оценки функции распределения $F(x)$. Эмпирическая функция распределения $F_n^*(x)$ обладает всеми свойствами функции распределения.

Эмпирическая функция распределения — это ступенчатая возрастающая функция, имеющая скачок величины $1/n$ в каждом наблюдении X_i, $1 \leq i \leq n$. Таким образом, эмпирическая функция распределения приписывает каждому наблюдению $X_i, 1 \leq i \leq n$,

вероятность $1/n$. Вся информация о распределении наблюдений сосредоточена в выборке, причем каждое наблюдение несет одинаковую информацию. Поэтому мы каждому наблюдению приписываем вероятность $1/n$, и в результате получается эмпирическая функция распределения.

Часто в выборке встречаются одинаковые наблюдения. Как в этом случае выглядит эмпирическая функция распределения? В этом случае скачок эмпирической функции распределения равен числу наблюдений, попавших в данную точку, деленному на n.

Пример 2.1. Пусть у нас имеется следующая выборка объемом $n = 50$:

$$x_1,\ x_2,\ \ldots,\ x_{50},$$

и пусть эта выборка имеет 3 различных члена, причем

$$\begin{array}{ll} x_1^* = 3, & n_1 = 10, \\ x_2^* = 6, & n_2 = 15, \\ x_3^* = 9, & n_3 = 25. \end{array}$$

Рис. 1.1. Эмпирическая (выборочная) функция распределения.

Отсюда $\nu_1 = 10,\ \nu_2 = n_1 + n_2 = 25,\ \nu_3 = n_1 + n_2 + n_3 = 50$. Теперь приведем график (см. рис. 1.1) и вид выборочной функции распределения

$$\hat{F}_n(x) = \begin{cases} 0, & x \le x_1^* = 3, \\ \dfrac{10}{50} = \dfrac{1}{5}, & x_1^* = 3 < x \le x_2^* = 6, \\ \dfrac{25}{50} = \dfrac{1}{2}, & x_2^* = 6 < x \le x_3^* = 9, \\ 1, & x > x_3^* = 9. \end{cases}$$

Простейшей оценкой плотности $f(x)$ является гистограмма (обозначим ее $\hat{f}_n(x)$). Для построения гистограммы поступаем следующим образом. По выборке $x_1, \ldots, x_n$ определяем интервал $[a, b]$, в котором лежат все наблюдения $a \le X_i < b$, $i = 1, \ldots, n$, и разбиваем его на k интервалов $[e_{i-1}, e_i)$, $i = 1, \ldots, k$, $a = e_0$, $b = e_k$. Для каждого интервала $[e_{i-1}, e_i)$ находим частоту $\nu_i[e_{i-1}, e_i)$ наблюдений, попавших в этот интервал, а затем и относительную частоту $\omega_{in} = \nu_i/n$ наблюдений, попавших в этот интервал. После этого мы полагаем $\hat{f}_n(x) = \omega_{in}/(e_i - e_{i-1})$, если $x \in [e_{i-1}, e_i)$, и задаем таким образом гистограмму (рис. 1.2).

Рис. 1.2. Гистограмма.

График гистограммы выборки строится следующим образом. На оси абсцисс откладываются интервалы $[e_{i-1},\ e_i)$ и на каждом из них, как на основании, строится прямоугольник высотой $\omega_{in}/(e_i - e_{i-1})$. Площадь каждого прямоугольника равна относительной частоте попадания наблюдений $X_i, 1 \le i \le n$, в пределы данного интервала. В результате мы получим график, который принято называть *гистограммой выборки*.

Ясно, что когда число наблюдений n неограниченно растет, то

$$\frac{\omega_{in}}{e_i - e_{i-1}} \to \frac{p_i}{e_i - e_{i-1}} = \frac{1}{e_i - e_{i-1}} \int_{e_{i-1}}^{e_i} f(x)dx, \qquad (1.3)$$

и если $e_i - e_{i-1} \to 0$, то правая часть (1.3) сходится к $f(e_i)$. Это и означает, что гистограмма является оценкой плотности распределения $f(x)$.

Удобно (хотя и не обязательно) брать промежутки $e_i - e_{i-1} = h$ равными. От выбора длины промежутков (числа h) зависит большая или меньшая выразительность гистограммы. При слишком малых h гистограмма содержит слишком много случайного. При слишком большом h в гистограмме почти теряются индивидуальные черты плотности распределения. Часто рекомендуют выбирать число интервалов k не более, чем $\{3.2 \log n\}$, причем частота попадания в каждый из интервалов должна быть не меньше 5.

Пример 2.2. В таблице 1.1 записаны в порядке возрастания суммы покупок товаров в магазине 100 посетителями. На основании этих данных мы составили статистический ряд с длиной каждого интервала 5 рублей.

Таблица 1.1

$e_i - e_{i+1}$	5-10	10-15	15-20	20-25	25-30	30-35	35-40
m_i	5	14	22	23	20	10	6
w_i	0.05	0.14	0.22	0.23	0.20	0.10	0.06

3. Числовые характеристики выборки

При работе с данными естественным образом возникает вопрос об описании их положения, разброса, характере разброса. В качестве таких характеристик обычно берутся оценки соответствующих параметров для функции распределения, например математического ожидания и медианы как мер положения и дисперсии как меры разброса. Как строятся такие оценки? Мы знаем, что эмпирическая функция распределения сходится (по вероятности) к теоретической функции распределения. Поэтому и соответствующие характеристики эмпирической функции распределения будут сходиться к соответствующим характеристикам теоретической. Например, что является математическим ожиданием эмпирической функции распределения? Эмпирическая функция распределения имеет распределение с вероятностями $p_i = 1/n$ в

точках наблюдений x_i. Поэтому соответствующее математическое ожидание для F_n^* равно:

$$\bar{x} = \sum_{i=1}^{n} x_i p_i = \frac{1}{n} \sum_{i=1}^{n} x_i$$

и называется выборочным средним.

Аналогично выборочная дисперсия s^2 равна

$$s^2 = \sum_{i=1}^{n} x_i^2 p_i - \Big(\sum_{i=1}^{n} x_i p_i \Big)^2 = \frac{1}{n} \sum_{i=1}^{n} x_i^2 - \bar{x}^2,$$

а *выборочные начальные* $\hat{a}_k$ *и центральные* $\hat{m}_k$ *моменты* k*-го порядка* равны соответственно

$$\hat{a}_k = \frac{1}{n} \sum_{i=1}^{n} x_i^k, \tag{1.4}$$

$$\hat{m}_k = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^k. \tag{1.5}$$

Они являются оценками соответствующих начальных и центральных моментов

$$a_k = \mathbf{E} X^k = \int x^k f(x) dx,$$

$$m_k = E(X - EX)^k = \int (x - EX)^k f(x) dx.$$

Итак, мы будем приводить сейчас определенные характеристики положения, разброса и характера разброса для функции распределения и находить аналогичные характеристики для эмпирической функции распределения, называя их выборочными и отмечая, что они являются вполне естественными.

4. Характеристики положения

Приведем их сначала для функции распределения случайной величины X. Ими являются математическое ожидание

$$\mathbf{E}\, X = \int x f(x) dx,$$

медиана m_x, задаваемая уравнением

$$\mathbf{P}(X \le m_x) = \mathbf{P}(X > m_x) = F(m_x) = \frac{1}{2}$$

(вероятность попадания случайной величины X слева и справа от m_x одна и та же), и мода, определяемая как точка максимума плотности.

Для эмпирической функции распределения аналогами данных характеристик являются:

— выборочное среднее $\bar{x} = \dfrac{1}{n}\sum\limits_{i=1}^{n} x_i$,

— выборочная медиана $\hat{m}_x$.

Для ее построения исходные данные $x_1, \ldots, x_n$ упорядочиваются в порядке возрастания $x^{(1)} \ge x^{(2)} \ge \ldots \ge x^{(n)}$. После этого полагаем для n нечетного $\hat{m}_x = x^{(k)}$, где $k = (n+1)/2$, а для n четного $\hat{m}_x = (x^{(k)} + x^{(k+1)})/2$, где $k = n/2$.

Так, если $n = 5$, то медианой будет среднее наблюдение $x^{(3)}$, а если $n = 4$, то медиана будет лежать посередине между $x^{(2)}$ и $x^{(3)}$.

Оценка моды часто строится с помощью специальных статистических процедур, которые иногда бывают достаточно сложными с вычислительной точки зрения и проводятся на компьютерах.

5. Меры разброса

Такими характеристиками случайной величины X являются:

— *дисперсия* $\mathbf{D}X$ и *стандартное отклонение* σ

$$\mathbf{D}X = \sigma^2 = \mathbf{E}(X - \mathbf{E}X)^2 = \int x^2 f(x)dx - (\mathbf{E}X)^2;$$

— *абсолютный размах*

$$A = \int |x - m_x| f(x)dx;$$

— *p-квантиль* x_p, задаваемая уравнением

$$x_p = \inf\{x : F(x) > p\}.$$

Значение p-квантили x_p имеет следующую содержательную интерпретацию. Случайная величина X меньше x_p с вероятностью p, т. е. вероятность того, что на вещественной прямой случайная величина будет лежать левее x_p, равна p или $F(x_p) = p$;

— $1/4$-квантиль и $3/4$-квантиль называются соответственно *нижней и верхней квартилями*;

— *межквартильный размах* равен $x_{3/4} - x_{1/4}$.

Величина x_p называется также $p \times 100$-*процентилью*.

Соответствующими выборочными характеристиками являются:

— *выборочная дисперсия*

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2; \tag{1.6}$$

— *выборочное стандартное отклонение* s;

— *выборочный абсолютный размах* $\hat{A}_n$:

$$\hat{A}_n = \frac{1}{n}\sum_{i=1}^{n}|x_i - \hat{m}_x|;$$

— *выборочная* p-*квантиль*

$$\hat{x}_p = \inf\{x : \hat{F}_n(x) \geq p\}, \tag{1.7}$$

т.е. наименьшее x, такое, что pn наблюдений меньше x (pn наблюдений лежит слева от $\hat{x}_p$ на вещественной прямой);

— *нижняя и верхняя выборочные квартили* $\hat{x}_{1/4}$ и $\hat{x}_{3/4}$;

— *выборочные* p-*процентили*;

— *выборочный межквартильный размах* $x_{3/4} - x_{1/4}$.

Разброс наблюдений также характеризуют положением максимального $x^{(n)}$ и минимального $x^{(1)}$ наблюдений и разностью $\hat{x}^{(n)} - \hat{x}^{(1)}$ между ними, обычно просто называемой *размахом*.

6. Анализ характера разброса

Важным свойством распределения случайных величин является симметрия плотности распределения и его одновершинность

с вершиной в точке симметрии. В этом случае медиана, математическое ожидание и мода совпадают, и можно со всей определенностью говорить об одном параметре положения. В противном случае в оценках параметров положения может быть существенное различие.

Какие существуют меры отклонения от симметрии у плотности распределения? Это *третий момент*

$$\mu_3 = \int (x - \mathbf{E}X)^3 f(x) dx$$

и *коэффициент асимметрии Пирсона*

$$(\mathbf{E}X - m_x)/\sigma,$$

где m_x — медиана. Ясно, что в случае симметричной плотности эти коэффициенты равны нулю.

Важным свойством этих коэффициентов является то, что они не зависят от единицы измерения масштаба, т. е. случайные величины X и $cX + d$ (c и d — константы) имеют одни и те же коэффициенты асимметрии.

Соответствующими выборочными характеристиками являются *выборочный коэффициент асимметрии*

$$\frac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{ns^3} \tag{1.8}$$

и *выборочный коэффициент асимметрии Пирсона*

$$(\bar{x} - \hat{m}_x)/s.$$

Важную роль при осуществлении статистических выводов играет и приближенная нормальность распределения наблюдений. Поэтому имеет большое значение исследование вопроса, насколько закон распределения близок к нормальному закону. Одной из мер отклонения от нормального распределения является разность четвертых моментов $\mu_4 = \int (x - \mathbf{E}X)^4 f(x) dx$ распределения выборки и нормального распределения. Для нормального распределения $\mu_4 = 3\sigma^4$. Таким образом, мы приходим к определению *эксцесса*

$$\frac{\mu_4}{\sigma^4} - 3$$

и *выборочного эксцесса*

$$\frac{\sum_{i=1}^{n}(x_i-\bar{x})^4}{ns^4}-3.$$

Если эксцесс отрицательный, то плотность распределения сильнее сконцентрирована около своего математического ожидания, чем в случае нормального распределения. Если эксцесс положительный, то плотность распределения убывает медленнее, чем плотность нормального распределения. Так как плотность нормального распределения убывает достаточно быстро, то обычно эксцесс бывает положительным.

7. Моменты выборочного среднего и выборочной дисперсии

Найдем первые два момента оценок $\bar{x}$ и s^2 случайной величины X с $\mathbf{E}X=m,\ \mathbf{D}X=\sigma^2$:

$$\mathbf{E}\bar{x}=\frac{1}{n}\sum_{i=1}^{n}\mathbf{E}x_i=m. \tag{1.9}$$

Так как x_i — независимые случайные величины, то дисперсия их суммы равна сумме их дисперсий, поэтому

$$D\bar{x}=\frac{\sigma^2}{n}. \tag{1.10}$$

Нетрудно проверить, что s^2 можно представить в виде

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

Тогда

$$\mathbf{E}s^2=\mathbf{E}\frac{1}{n}\sum_{i=1}^{n}(x_i-m)^2-\mathbf{E}(\bar{x}-m)^2=\sigma^2-\frac{\sigma^2}{n}.$$

Таким образом,

$$\mathbf{E}s^2=\frac{n-1}{n}\sigma^2. \tag{1.11}$$

Используя центральные моменты $\mu_k = \mathbf{E}(X - \mathbf{E}X)^k$, приведем точное выражение для дисперсии оценки s^2

$$Ds^2 = \frac{\mu_4 - \mu_2^2}{n} - \frac{2(\mu_4 - 2\mu_2^2)}{n^2} + \frac{\mu_4 - 3\mu_2^2}{n^3}. \tag{1.12}$$

В статистике часто используется еще одна оценка дисперсии

$$\bar{s}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2. \tag{1.13}$$

Она такова, что ее математическое ожидание равно σ^2, т. е.

$$\mathbf{E}\bar{s}^2 = \sigma^2. \tag{1.14}$$

8. О качестве оценок, возникающих в описательной статистике

Все оценки (положения, разброса, асимметрии и т.п.), возникающие в описательной статистике, являются *состоятельными* и *асимптотически нормальными*.

Пусть $\hat{\Delta}_n$ — одна из таких оценок, обсуждавшихся ранее, и Δ — соответствующая характеристика.

Тогда $\hat{\Delta}_n$ — состоятельная оценка Δ, т. е. $\Delta_n \to \Delta$ при $n \to \infty$ по вероятности. Это означает следующее. Для любого $\epsilon > 0$

$$\lim_{n\to\infty} P(|\hat{\Delta}_n - \Delta| > \epsilon) = 0.$$

Оценка Δ_n также асимптотически нормальна. Это означает, что распределения $\sqrt{n}(\hat{\Delta}_n - \Delta)$ при $n \to \infty$ сходятся к нормальному распределению. Выражаясь языком формул, для любого x

$$\lim_{n\to\infty} P(\sqrt{n}(\hat{\Delta}_n - \Delta) < x) = \Phi\left(\frac{x}{\sigma}\right).$$

Здесь

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left\{-\frac{s^2}{2}\right\} ds$$

— функция стандартного нормального распределения и σ^2 — нормирующая дисперсия.

9. Роль нормального распределения в статистике

При обработке данных часто считается, что наблюдения распределены по нормальному закону. Это объясняется следующими причинами.

На точность измерений да и на саму измеряемую величину влияют несколько факторов. Приближенно можно считать, что они воздействуют независимо и их воздействие аддитивно (можно складывать). Таким образом, мы получаем сумму нескольких случайных величин, которая приближенно (в силу центральной предельной теоремы) распределена по нормальному закону.

Другим аргументом в пользу принятия предположения о нормальности является то, что именно в этом случае статистические оценки и критерии проверки гипотез имеют простой вид, являются эффективными, и параметры нормального распределения (математическое ожидание и дисперсия) имеют наглядную интерпретацию.

Английские аналоги основных терминов

Поскольку при решении задач часто используются оригинальные версии программ, приведем английские аналоги основных терминов. При этом мы будем приводить названия тех терминов, которые подлежат оценке, если названия соответствующих оценок получаются просто добавлением слова *выборочное* (sample). Например, variance — sample variance.

Выборка — sample,
среднее, $\bar{x}$ — mean,
медиана, m_x — median,
мода — moda,
дисперсия, σ^2 — variance,
средне-квадратичное отклонение, σ — standard deviation,
p-квантиль, x_p — p-quantile,
нижний квартиль, $x_{1/4}$ — lower quartile,
верхний квартиль, $x_{3/4}$ — upper quartile,
межквартильный интервал, $x_{3/4} - x_{1/4}$ — interquantile range,
процентиль — percentile,
размах — range,
асимметрия — skewness,
эксцесс — kurtosis.

Глава 2

СТАТИСТИЧЕСКОЕ ОЦЕНИВАНИЕ

При анализе своей деятельности фирме приходится оценивать целый ряд характеристик. Например:

— доля рынка, захваченного продукцией,

— средний оборот фирмы,

— время надежной работы производимых товаров (телевизоров, компьютеров и т. п.),

— процент бракованной продукции,

— средний оборот фирмы за день,

— среднее число дней, пропущенных служащими по болезни в году,

— прогноз прибыли фирмы и т. п.

Решением подобных задач занимается статистическая теория оценивания.

1. Постановка задачи статистического оценивания. Несмещенные, состоятельные и эффективные оценки

Будем считать, что мы располагаем выборкой независимых случайных величин $X_1, \ldots, X_n$, имеющих функцию распределения $F(x, \theta)$, где значение параметра θ неизвестно. Необходимо по наблюдениям $X_1, \ldots, X_n$ найти *статистическую оценку* $\hat{\theta}_n = \hat{\theta}_n(X_1, \ldots, X_n)$ параметра θ. Например, если есть выборка $X_1, \ldots, X_n$, имеющая нормальное распределение $N(\theta, \sigma^2)$ с математическим ожиданием θ и дисперсией σ^2, то выборочное среднее $\bar{X}$ является оценкой параметра θ, а выборочная дисперсия s^2 — оценкой параметра σ^2.

Качество оценок $\hat{\theta}_n$ чаще всего характеризуют математическим ожиданием их квадратичного отклонения $E_\theta(\hat{\theta}_n - \theta)^2$ от истинного значения параметра θ, которое в такой постановке называется квадратичным риском статистической оценки. Квадратичный риск показывает средние потери от принятия оценки $\hat{\theta}_n$ вместо θ при истинном значении параметра θ. Чем меньше риск статистической оценки, тем лучше статистическая оценка. Если $\mathbf{E}_\theta \hat{\theta}_n = \theta$, то

$$\mathbf{E}_\theta(\hat{\theta}_n - \theta)^2 = \mathbf{E}_\theta(\hat{\theta}_n - \mathbf{E}_\theta\hat{\theta}_n)^2 = \mathbf{D}_\theta\hat{\theta}_n$$

и квадратичный риск совпадает с дисперсией оценки.

Другой мерой качества статистических оценок является ширина их доверительных интервалов. С этим понятием мы познакомимся позднее.

Спрашивается, какими свойствами должна обладать функция $\hat{\theta}_n$, чтобы ее можно было считать хорошей оценкой для неизвестного параметра θ? Чтобы значение $\hat{\theta}_n(X_1, \ldots, X_n)$ было близко к θ, мы должны потребовать по возможности более тесной концентрации распределения вероятности $\hat{\theta}_n(X_1, \ldots, X_n)$ вблизи значения неизвестного параметра θ, другими словами, *чтобы рассеивание случайной величины $\hat{\theta}_n$ около θ было по возможности меньшим.* Это приводит нас к понятию состоятельности оценки.

а) Состоятельная оценка

Статистическая оценка $\hat{\theta}_n$, сходящаяся по вероятности к истинному значению параметра θ при неограниченном возрастании объема выборки (т. е. когда $n \to \infty$), называется состоятельной оценкой, иначе говоря,

$$\lim_{n\to\infty} P(|\hat{\theta}_n - \theta| < \epsilon) = 1$$

для любого сколь угодно малого $\epsilon > 0$.

Понятие состоятельности оценки тесно связано с понятием доверительного интервала, показывающего погрешность оценивания. С этим понятием мы познакомимся чуть позднее. Состоятельность оценки означает, что длины доверительных интервалов стремятся к нулю по вероятности, когда уровень доверия фиксирован.

Так как в силу Закона Больших Чисел

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \to E[X], \qquad \frac{1}{n}\sum_{i=1}^{n} X_i^2 \to E[X^2],$$

$$s^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2 \to E[X^2] - (E[X])^2 = D[X],$$

по вероятности при $n \to \infty$, то $\bar{X}$ и s^2 являются состоятельными оценками математического ожидания и дисперсии соответственно.

Как отмечалось выше, если у случайной величины X существует (т. е. конечен) начальный или центральный момент, то соответствующий момент выборки сходится к нему по вероятности.

Следовательно, выборочные моменты являются состоятельными оценками соответствующих моментов случайной величины X.

Итак, мы видим, что выборочные среднее и дисперсия сходятся к своим математическим ожиданиям EX и σ^2 соответственно. Это приводит нас к следующему важному понятию несмещенности оценок.

б) Несмещенные оценки

Оценка $\hat{\theta}_n$ параметра θ называется несмещенной, если ее математическое ожидание совпадает со значением оцениваемого параметра θ:

$$\mathbf{E}\hat{\theta}_n = \theta.$$

Если статистическая оценка является несмещенной, то, как мы уже показали, ее квадратичный риск $\mathbf{E}_\theta(\hat{\theta}_n - \theta)^2$ совпадает с дисперсией оценки $D_\theta(\hat{\theta}_n)$. Требование несмещенности особенно важно при малых объемах выборки, поскольку понятие состоятельности в этом случае теряет свой смысл.

Если оценка $\hat{\theta}_n$ не является несмещенной, то, чтобы охарактеризовать ее качество, задают *смещение* $\mathbf{E}\hat{\theta}_n - \theta$ оценки.

в) Асимптотическая нормальность

В силу Центральной Предельной Теоремы нормированное распределение выборочного среднего сходится к нормальному, т. е.

$$\lim_{n\to\infty} \mathbf{P}_\theta(\sqrt{n}(\bar{X} - \mathbf{E}X)/\sigma < x) = \Phi(x)$$

для любого вещественного значения x. Здесь $\Phi(x)$ — функция распределения стандартного нормального закона.

Оказывается, что нормированные распределения $\sqrt{n}(\hat{\theta}_n - \theta)/\sigma$ многих статистических оценок с ростом объема выборки ($n \to \infty$) также сходится к нормальному. Это позволяет легко анализировать погрешности оценок, заменяя при анализе погрешности распределение оценки нормальным распределением. В дальнейшем мы так поступим при построении доверительного интервала для оценки параметра биномиального распределения. Однако это только очень частный случай общей методологии.

Оценка $\hat{\theta}_n$ называется *асимптотически нормальной оценкой параметра θ*, если

$$\lim_{n\to\infty} \mathbf{P}_\theta(\sqrt{n}(\hat{\theta}_n - \theta)/\sigma < x) = \Phi(x)$$

для любого вещественного значения x. Величина $\sigma^2 = \sigma^2(\hat{\theta}_n)$ называется *асимптотической дисперсией оценки.* Ясно, что асимптотическая дисперсия является естественной мерой квадратичного риска оценок, поскольку

$$\lim_{n\to\infty} n\mathbf{E}(\hat{\theta}_n - \theta)^2 = \sigma^2.$$

Величина $\sigma/\sqrt{n}$ называется *стандартной ошибкой* статистической оценки $\hat{\theta}_n$. Она показывает средний разброс статистической оценки $\hat{\theta}_n$. Значение стандартной ошибки в пакетах программ часто выводится на экран компьютера как характеристика средней ошибки статистического оценивания.

Как уже говорилось, все оценки описательной статистики являются асимптотически нормальными, в частности, таковы выборочное среднее и дисперсия.

г) Эффективные оценки

Рассматривая в качестве меры рассеяния распределения $\hat{\theta}_n$ около значения параметра θ величину квадратичного риска $\mathbf{E}(\hat{\theta}_n - \theta)^2$, мы можем точно охарактеризовать *сравнительную эффективность* двух каких-либо оценок $\hat{\theta}_n^{(1)}$ и $\hat{\theta}_n^{(2)}$, оценивающих один и тот же параметр θ. В качестве *меры эффективности* естественно принять отношение квадратичных рисков

$$e = \frac{\mathbf{E}(\hat{\theta}_n^{(1)} - \theta)^2}{\mathbf{E}(\hat{\theta}_n^{(2)} - \theta)^2}.$$

Если $e > 1$, то оценка $\hat{\theta}_n^{(2)}$ более эффективна, чем оценка $\hat{\theta}_n^{(1)}$ (и обратно), так как ей соответствует меньшее рассеяние. Если $\hat{\theta}_n^{(1)}$ и $\hat{\theta}_n^{(2)}$ — несмещенные оценки, то e является отношением дисперсий

$$e = \frac{\sigma^2_{\hat{\theta}_n^{(1)}}}{\sigma^2_{\hat{\theta}_n^{(2)}}}, \qquad \text{так как} \qquad \mathbf{E}\hat{\theta}_n^{(1)} = \theta \quad \text{и} \quad \mathbf{E}\hat{\theta}_n^{(2)} = \theta.$$

Несмещенная оценка $\hat{\theta}_n$ называется эффективной, если при любом θ ее дисперсия не больше дисперсии любой другой несмещенной оценки θ_n^*:

$$\mathbf{D}\hat{\theta}_n \le \mathbf{D}\theta_n^*.$$

Для широкого класса распределений мы укажем точную нижнюю границу для дисперсий различных оценок одного и того же параметра.

Пусть X — случайная величина *непрерывного типа* и $f(x,\theta)$ — ее плотность распределения.

Для несмещенных оценок $\tilde{\theta}_n$ параметра θ *неравенство Рао—Крамера*

$$\mathbf{D}\tilde{\theta}_n \geq I^{-1}(\theta), \qquad I(\theta) = n\mathbf{E}\left(\frac{\partial \ln f(x,\theta)}{\partial\theta}\right)^2,$$

указывает точную нижнюю границу дисперсий оценок параметра θ. Если существует оценка, дисперсия которой в точности равна нижней границе $I^{-1}(\theta)$, то она является эффективной оценкой с минимально возможной дисперсией.

Более подробно неравенство Рао—Крамера будет рассмотрено ниже.

Помимо описанных выше понятий введем понятия *асимптотической несмещенности* и *асимптотической эффективности*.

Оценка $\hat{\theta}_n$ параметра θ называется *асимптотически несмещенной*, если

$$E\hat{\theta}_n \underset{n\to\infty}{\longrightarrow} \theta.$$

Нетрудно убедиться, что выборочная дисперсия s^2 является асимптотически несмещенной оценкой дисперсии σ^2, так как

$$\mathbf{E}s^2 = \frac{n-1}{n}\sigma^2 \underset{n\to\infty}{\longrightarrow} \sigma^2.$$

Асимптотически нормальная оценка $\hat{\theta}_n$ параметра θ называется *асимптотически эффективной*, если ее асимптотическая дисперсия $\sigma^2(\hat{\theta}_n) = I^{-1}(\theta)$. Можно доказать, что для асимптотически нормальных оценок $\hat{\theta}_n$ имеет место неравенство

$$\sigma^2(\hat{\theta}_n) \geq I^{-1}(\theta).$$

Рассмотрим *дискретный случай*. Пусть $p_i(\theta),\ i = 1, 2, \ldots,$ — вероятности, с которыми наблюдаемая случайная величина X принимает свои значения.

Несмещенная оценка $\hat{\theta}_n$ будет эффективной, если выполнено равенство

$$\mathbf{D}\hat{\theta}_n = \frac{1}{n\sum\limits_i \Big[\dfrac{\partial \ln p_i(\theta)}{\partial\theta}\Big]^2 p_i(\theta)}. \tag{2.2}$$

Введем обозначение

$$I(\theta) = \begin{cases} n\mathbf{E}\Big[\dfrac{\partial \ln f(X,\theta)}{\partial\theta}\Big]^2, & \text{если } X \text{ непрерывного типа,} \\ n\sum\limits_i \Big[\dfrac{\partial \ln p_i(\theta)}{\partial\theta}\Big]^2 p_i(\theta), & \text{если } X \text{ дискретного типа.} \end{cases}$$

Для несмещенных оценок θ_n^* параметра θ *неравенство Рао — Крамера*

$$\mathbf{D}\theta_n^* \geq \frac{1}{I(\theta)}$$

указывает точную нижнюю границу дисперсий оценок.

Пример 2.1. Пусть элементы выборки распределены по нормальному закону с плотностью распределения

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

причем $\theta = \mathbf{E}X$ неизвестно, а дисперсия σ^2 известна.

Имеем

$$\mathbf{E}\Big(\frac{\partial \ln f(X,\theta)}{\partial\theta}\Big)^2 = \mathbf{E}\Big(\frac{X-\theta}{\sigma^2}\Big)^2 = \frac{1}{\sigma^2},$$

отсюда $I(\theta) = n/\sigma^2$. С другой стороны, известно, что $D\bar{x} = \sigma^2/n$, поэтому $\mathbf{D}\bar{x} = 1/I(\theta)$, а тогда из равенства (2.1) следует, что $\bar{x}$ — эффективная оценка. Итак, если при нормальном законе дисперсия σ^2 известна, то $\bar{x}$ является несмещенной, состоятельной и эффективной оценкой математического ожидания.

Пример 2.2. Пусть X имеет распределение Пуассона с параметром λ, тогда вероятности

$$p_i(\lambda) = \frac{\lambda^i}{i!}\, e^{-\lambda}, \qquad \text{поэтому} \qquad \frac{d\,\ln p_i}{d\lambda} = \frac{i}{\lambda} - 1,$$

$$\sum_{i=0}^{\infty}\left(\frac{d\ln p_i(\lambda)}{d\lambda}\right)^2 p_i(\lambda)=\sum_{i=0}^{\infty}\left(\frac{i}{\lambda}-1\right)^2\frac{\lambda^i}{i!}\,e^{-\lambda}=\frac{1}{\lambda}.$$

Отсюда получаем, что минимальная величина дисперсии оценки λ равна $1/I(\theta)=\lambda/n$. С другой стороны, известно, что $\mathbf{E}X=\lambda$, $\mathbf{D}X=\lambda$, а выборочное среднее $\bar{x}$ является состоятельной и несмещенной оценкой математического ожидания и в нашем случае $\mathbf{E}\bar{x}=\lambda$, $\mathbf{D}\bar{x}=\lambda/n$. Из последнего равенства следует, что дисперсия $\bar{x}$ удовлетворяет равенству (2.2) и, следовательно, $\bar{x}$ является эффективной оценкой λ, а из предыдущего пункта следует ее состоятельность и несмещенность.

2. Метод моментов

В предыдущем параграфе мы сформулировали основные свойства оценок, но ничего не говорили о способах их получения. Одним из первых методов оценивания параметров был метод моментов, разработанный К. Пирсоном.

Пусть известный закон распределения $F(x;\theta_1,\theta_2,\ldots,\theta_l)$ случайной величины X определяется несколькими параметрами

$$\theta_1,\ \theta_2,\ \ldots,\ \theta_l, \tag{2.3}$$

числовое значение которых неизвестно.

Чтобы получить оценки

$$\hat{\theta}_1,\ \hat{\theta}_2,\ \ldots,\ \hat{\theta}_l \tag{2.4}$$

параметров (2.3), поступают следующим образом. Производят n независимых наблюдений x_1, x_2, ..., x_n над случайной величиной X. Следуя методу моментов, необходимо l первых моментов случайной величины X приравнять к l выборочным моментам, полученным из экспериментальных данных. Это могут быть как начальные моменты

$$\hat{a}_k=\frac{1}{n}\sum_{i=1}^{n}x_i^k,$$

так и центральные моменты

$$\hat{m}_k=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^k.$$

Ясно, что теоретические моменты случайной величины выражаются через параметры (2.3). Приравняем выборочные моменты к теоретическим моментам, в которых параметры (2.3) заменены на параметры (2.4). Решив эти уравнения, мы найдем оценки (2.4) параметров (2.3). Закон распределения будет иметь вид $F(x;\hat{\theta}_1,\hat{\theta}_2,\ldots,\hat{\theta}_l)$ и, следовательно, будет полностью определен.

Теоретическим обоснованием метода моментов служит закон больших чисел, согласно которому для рассматриваемого случая при большом объеме выборки выборочные моменты близки к моментам наблюдаемой случайной величины X.

Метод моментов позволяет получать состоятельные оценки параметров (2.3). Таким образом, надежность выводов, сделанных при его использовании, зависит от количества наблюдений. На практике этот метод часто приводит к сравнительно простым вычислениям, но при этом он иногда приводит к малоэффективным оценкам.

Пример 2.3: *Оценка параметров* Г*-распределения (гамма-распределения).* Функция распределения случайной величины X зависит от двух параметров α,β и имеет вид

$$F(x,\alpha,\beta)=\frac{\beta^\alpha}{\Gamma(\alpha)}\int_0^x x^{\alpha-1}e^{-\beta x}dx.$$

Теоретические моменты таковы:

$$\mathbf{E}X=\frac{\alpha}{\beta},\quad DX=\frac{\alpha}{\beta^2}.$$

Приравнивая первые два выборочных момента $\bar{x}$ и s^2 к теоретическим, в которых вместо α и β выступают $\hat{\alpha}$ и $\hat{\beta}$, получаем уравнения относительно $\hat{\alpha}$ и $\hat{\beta}$:

$$\frac{\hat{\alpha}}{\hat{\beta}}=\bar{x},\quad \frac{\hat{\alpha}}{\hat{\beta}^2}=s^2.$$

Отсюда находим оценки для параметров Г-распределения

$$\hat{\alpha}=\frac{(\bar{x})^2}{s^2},\quad \hat{\beta}=\frac{\bar{x}}{s^2}$$

и, подставляя их вместо α и β в закон распределения, полностью определяем этот закон $F(x,\hat{\alpha},\hat{\beta})$.

Пусть из наблюдений мы получили $\bar{x} = 3$, $s^2 = 4$, тогда $\hat{\alpha} = 9/4$, $\hat{\beta} = 3/4$.

Пример 2.4: *Оценка параметров нормального распределения.* Пусть случайная величина X имеет нормальный закон распределения $F(x, m, \sigma^2)$, который определяется двумя параметрами: математическим ожиданием $m = \mathbf{E}X$ и дисперсией $\sigma^2 = DX$. По методу моментов неизвестное математическое ожидание m оценивается выборочным средним арифметическим $\bar{x}$, а дисперсия σ^2— выборочной дисперсией s^2. Методом моментов могут быть оценены параметры, выраженные в виде определенных функций теоретических моментов.

Исследуя свойства оценок, получаемых с помощью метода моментов, английский математик Р. Фишер предложил более надежный метод оценивания параметров распределения наблюдаемой случайной величины X — метод максимального правдоподобия. Хотя метод максимального правдоподобия приводит к более сложным вычислениям, чем метод моментов, все же оценки, получаемые с его помощью как правило оказываются более надежными и особенно предпочтительными в случае малого числа наблюдений.

3. Метод максимального правдоподобия

Одним из важнейших методов для нахождения оценок параметров по данным выборки является метод максимального правдоподобия. Основная идея метода заключается в следующем.

Пусть имеется выборка объема n: x_1, x_2, ... , x_n значений случайной величины X с распределением, зависящим от параметра θ.

Функцией правдоподобия называется функция

$$L(x_1, x_2, \ldots, x_n; \theta) = f(x_1, \theta) \cdot f(x_2, \theta) \cdots f(x_n, \theta),$$

если X — случайная величина *непрерывного* типа с плотностью распределения $f(x, \theta)$, и функция

$$L(x_1, x_2, \ldots, x_n; \theta) = p(x_1, \theta) \cdot p(x_2, \theta) \cdots p(x_n, \theta),$$

если X — *дискретная* случайная величина с $p(x_i, \theta) = P_\theta(X = x_i)$.

Функцию правдоподобия в дискретном случае можно представить в другом виде, если в выборке встречаются одинаковые

величины. Пусть r — число различных элементов в выборке, $\xi_1, \xi_2, \ldots, \xi_r$ — эти элементы, а $\nu_1, \nu_2, \ldots, \nu_r$ — их частоты.

Очевидно

$$\sum_{i=1}^{r} \nu_i = n, \qquad \bar{x} = \frac{1}{n} \sum_{i=1}^{r} \nu_i \xi_i. \tag{2.5}$$

Пусть $\mathbf{P}_\theta(X = \xi_i) = f_i(\theta)$, тогда функция правдоподобия определяется следующим образом:

$$L(x_1, x_2, \ldots, x_n, \theta) = f_1^{\nu_1}(\theta) \cdot f_2^{\nu_2}(\theta) \cdots f_r^{\nu_r}(\theta). \tag{2.6}$$

Сущность метода максимального правдоподобия заключается в том, что в качестве оценки $\hat{\theta} = \hat{\theta}(x_1, \ldots, x_n)$ параметра θ берется то значение, при подстановке которого в выражение для L вместо параметра θ получаем максимальное значение функции L. Это значит, что наши наблюдения при этом значении параметра будут иметь наибольшую вероятность появления.

Ясно, что $\hat{\theta}$, при котором функция L достигает максимума, обращает в максимум и $\ln L$, поэтому для нахождения оценки следует решить относительно θ *уравнение правдоподобия*

$$\frac{\partial \ln L}{\partial \theta} = 0. \tag{2.7}$$

Решение $\hat{\theta}$ этого уравнения называется *оценкой максимального правдоподобия* параметра θ.

Если закон распределения случайной величины зависит от нескольких параметров $\theta_1, \theta_2, \ldots, \theta_s$, то ход рассуждений аналогичен предыдущему, т. е. необходимо найти такие величины $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_s$, которые максимизировали бы функцию правдоподобия. Нахождение оценок максимального правдоподобия сводится к нахождению максимума функции $\ln L$ относительно параметров $\theta_1, \theta_2, \ldots, \theta_s$ по правилу определения экстремума функции от нескольких параметров, т. е. к решению системы уравнений

$$\frac{\partial \ln L}{\partial \theta_1} = 0, \qquad \frac{\partial \ln L}{\partial \theta_2} = 0, \qquad \ldots, \qquad \frac{\partial \ln L}{\partial \theta_s} = 0. \tag{2.8}$$

Пример 2.5. Найти оценку вероятности p по данному числу m появлений события A в n независимых испытаниях.

Решение. Вероятность p можно рассматривать как параметр, входящий в распределение дискретной величины X, имеющей только два значения $\xi_1 = 1$ и $\xi_2 = 0$ в зависимости от того, появилось ли событие A в рассматриваемом испытании или не появилось. Тогда по формуле (2.6) мы имеем

$$L = p^m(1-p)^{n-m}$$

и уравнение (2.7) запишется так:

$$\frac{\partial \ln L}{\partial p} = \frac{m}{\hat{p}} - \frac{n-m}{1-\hat{p}} = 0.$$

Оно имеет единственное решение

$$\hat{p} = \frac{m}{n}.$$

Следовательно, оценкой максимального правдоподобия параметра p будет относительная частота m/n события. Мы знаем, что она является несмещенной, состоятельной. Можно также доказать, что при любом n эта оценка имеет наибольшую эффективность.

Пример 2.6. Рассмотрим случайную величину X, подчиненную закону Пуассона с неизвестным параметром λ. Нужно найти оценку этого параметра с помощью метода максимального правдоподобия.

Решение. Пусть имеется выборка объема $n : x_1, x_2, \ldots, x_n$ наблюдений X. Случайная величина X может принимать любое из значений 0, 1, 2, ... Пусть r — наибольшее из этих чисел в выборке; числа $\nu_0, \nu_1, \ldots, \nu_r$ пусть представляют частоты, с которыми встречаются в выборке числа $0, 1, 2, \ldots, r$. Тогда в формуле (2.6) $f_i(\lambda) = \lambda^i e^{-\lambda}/\,i! = \mathbf{P}(X = i)$ $(i = 0, 1, \ldots)$ и функция правдоподобия принимает вид

$$L = \prod_{i=0}^{r} \Big(\frac{\lambda^i e^{-\lambda}}{i!}\Big)^{\nu_i}.$$

Поэтому уравнение максимального правдоподобия даст

$$\frac{\partial \ln L}{\partial \lambda} = \sum_{i=0}^{r} \nu_i \Big(\frac{i}{\hat{\lambda}} - 1\Big) = \frac{1}{\hat{\lambda}} \sum_{i=0}^{r} \nu_i i - \sum_{i=0}^{r} \nu_i = 0.$$

Отсюда с учетом формулы (2.5) получаем

$$\hat{\lambda} = \frac{\sum_{i=0}^{r} \nu_i i}{\sum_{i=0}^{r} \nu_i} = \frac{1}{n} \sum_{j=1}^{n} x_j = \bar{x}.$$

В этом случае $\bar{x}$ является несмещенной состоятельной оценкой параметра λ, полученной с помощью метода максимального правдоподобия.

Пример 2.7. Найти оценки двух параметров m и σ^2 для случайной величины X, имеющей нормальное распределение $N(m, \sigma^2)$.

Решение. Так как в данном случае

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

то функция максимального правдоподобия имеет вид

$$L(x_1, x_2, \ldots, x_n; m, \sigma^2) = \frac{1}{(\sigma\sqrt{2\pi})^n} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - m)^2 \right\},$$

а ее логарифм равен

$$\ln L = n \ln \frac{1}{(\sigma\sqrt{2\pi})} - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - m)^2 =$$

$$= -\frac{n}{2} \ln 2\pi - \frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - m)^2.$$

Для нахождения оценок параметров m и σ^2 необходимо согласно формулам (2.8) решить совместно следующую систему уравнений:

$$\begin{cases} \dfrac{\partial \ln L}{\partial m} = \dfrac{1}{\sigma^2} \displaystyle\sum_{i=1}^{n} (x_i - \hat{m}) = 0, \\ \dfrac{\partial \ln L}{\partial \sigma^2} = -\dfrac{n}{2\hat{\sigma}^2} + \dfrac{\sum_{i=1}^{n} (x_i - \hat{m})^2}{2\hat{\sigma}^4} = 0. \end{cases} \tag{2.9}$$

Из первого и второго уравнений (2.9) находим соответственно

$$\hat{m} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}, \qquad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2,$$

где $\hat{\sigma}^2$ — выборочная дисперсия s^2.

Итак, если случайная величина подчинена нормальному закону распределения, то по данным выборки математическое ожидание следует оценивать с помощью средней арифметической $\bar{x}$, а дисперсию — выборочной дисперсией s^2.

Оценка $\bar{x}$ является нормальной, несмещенной, состоятельной, а если оценка математического ожидания производится при известной дисперсии σ^2, то $\bar{x}$ является эффективной. Оценка s^2 будет состоятельной, но смещенной.

Метод максимального правдоподобия обладает важными достоинствами: он всегда приводит к состоятельным (хотя иногда и смещенным) оценкам, распределенным асимптотически нормально, имеющими наименьшую возможную дисперсию по сравнению с другими оценками и наилучшим образом использующими всю информацию о неизвестном параметре, содержащуюся в выборке.

Если $\hat{\theta}_n = \hat{\theta}_n(x_1, \ldots, x_n)$ — оценки максимума правдоподобия и θ_n^* — другие эффективные оценки, то разность $\hat{\theta}_n - \theta_n^*$ пренебрежимо мала, т. е. $\sqrt{n}\left(\hat{\theta}_n - \theta_n^*\right) \to 0$ по вероятности при $n \to \infty$.

4. Некоторые распределения, связанные нормальным распределением

При определении точности оценок параметров нормального распределения, а также в задачах проверки гипотез о параметрах нормального распределения (математического ожидания и дисперсии) нам потребуется использовать распределения ряда элементарных функций от нормальных случайных величин. Эти распределения мы сейчас и изучим.

В этом разделе будет предполагаться, что

$$x_1, x_2, \cdots, x_n$$

— выборка из нормальной совокупности и

$$\mathbf{E}x_i = m, \quad Dx_i = \sigma^2, \quad \mathbf{E}(x_i - m)(x_j - m) = \delta_{ij}\sigma^2, \qquad (2.10)$$

здесь

$$\delta_{ij} = \begin{cases} 1, & i = j, \\ 0, & i \neq j, \end{cases}$$

— символ Кронекера.

Рассмотрим три распределения, которые имеют важные статистические приложения.

а) χ^2-распределение (хи-квадрат распределение)

Если при $i = 1, \ldots, n$ случайные величины $z_i \in N(0,1)$ и они независимы, то величина

$$w = \sum_{i=1}^{n} z_i^2$$

имеет χ^2-распределение с n степенями свободы, причем $\mathbf{E}w = n$, $\mathbf{D}w = 2n$. Это распределение впервые было рассмотрено Хальмертом в 1876 г. Плотность $k_n(x)$ распределения задается при $x > 0$ и имеет вид

$$k_n(x) = \frac{1}{2^{n/2}\Gamma(n/2)} x^{n/2-1} e^{-x/2}, \quad x > 0.$$

Таблицы χ^2-распределения дают возможность по значениям $\alpha = 90/100; 80/100; \ldots; 5/100; 1/100$ найти такое $\chi^2_{k,\alpha}$, что

$$\alpha = \int_{\chi^2_{n,\alpha}}^{\infty} k_n(x)dx. \tag{2.11}$$

б) t-распределение Стьюдента

Если случайная величина $z \in N(0,1)$, а величина w имеет χ^2-распределение с k степенями свободы, то величина

$$t = \frac{z\sqrt{k}}{\sqrt{w}}, \qquad w = \sum_{i=1}^{k} z_i^2,$$

распределена по закону Стьюдента с k степенями свободы. Это распределение впервые в 1908 г. было использовано английским математиком В. Госсетом, публиковавшимся под псевдонимом Стьюдент.

Обозначим плотность распределения Стьюдента с k степенями свободы через $S_k(t)$, тогда она имеет вид

$$S_k(t) = \frac{\Gamma\left(\frac{k+1}{2}\right)}{\sqrt{k\pi}\,\Gamma\left(\frac{k}{2}\right)}\left(1+\frac{t^2}{k}\right)^{-\frac{k+1}{2}}.$$

Распределение Стьюдента симметрично относительно нуля, и его таблицы связывают число степеней свободы k, вероятность α и $t_{\alpha,k}$ согласно формуле

$$\alpha = \mathbf{P}\{|t| > t_{\alpha,k}\} = 2\int_{t_{\alpha,k}}^{\infty} S_k(t)dt. \tag{2.12}$$

Ясно, что $\mathbf{P}\{-t_{\alpha,k} < t < t_{\alpha,k}\} = 1-\alpha$.

При $\alpha = 0.05,\ k = 7$ находим $t_{0.05,\,7} = 2.365$, поэтому

$$\mathbf{P}\{-2.365 < t < 2.365\} = 0.95, \quad \mathbf{P}\{|t| > 2.365\} = 0.05.$$

в) F-распределение Фишера

Предполагается, что $\xi_1, \ldots, \xi_m, \eta_1, \ldots, \eta_n$ — независимые нормальные случайные величины, принадлежащие $N(0, \sigma^2)$. Случайная величина F_{mn}, определяемая соотношением

$$F_{mn} = \frac{1}{m}\sum_{j=1}^{m}\xi_j^2\left(\frac{1}{n}\sum_{j=1}^{n}\eta_j^2\right)^{-1},$$

имеет распределение Фишера с m и n степенями свободы и плотностью распределения

$$g(x;m,n) = \frac{\Gamma((m+n)/2)}{\Gamma(m/2)\Gamma(n/2)}m^{m/2}n^{n/2}x^{m/2-1}(n+mx)^{-(m+n)/2},\ x>0.$$

В таблицах распределения Фишера для различных $\alpha = 5/100; 1/100$ и т. д. приводится величина $F_{\alpha,m,n}$, такая, что

$$\int_{F_{\alpha,m,n}}^{\infty} g(x;m,n)dx = \alpha.$$

5. Некоторые свойства среднего и дисперсии выборки из нормальной совокупности

а) Независимость среднего $\bar{x}$ и дисперсии s^2 выборки из нормальной совокупности

Пусть $x_1, x_2, \cdots, x_n$ — выборка из нормальной совокупности с параметрами (2.10) и

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

Рассмотрим линейные преобразования

$$\begin{aligned}
y_1 &= \frac{1}{\sqrt{1\cdot 2}}(x_1 - x_2),\\
y_2 &= \frac{1}{\sqrt{2\cdot 3}}(x_1 + x_2 - 2x_3),\\
\cdots &\quad \cdots\cdots\cdots\\
y_i &= \frac{1}{\sqrt{i\cdot(i+1)}}(x_1 + x_2 + \cdots + x_i - ix_{i+1}),\\
\cdots &\quad \cdots\cdots\cdots\\
y_n &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} x_i.
\end{aligned}$$

Нетрудно проверить, что

$$\mathbf{E}y_i = 0, \quad \mathbf{E}y_i^2 = \sigma^2, \qquad \mathbf{E}y_i y_k = \sigma^2\delta_{i,j}. \tag{2.13}$$

Известно, что *линейные преобразования нормальных величин распределены нормально*, поэтому

$$y_i \in N(0, \sigma^2), \quad i = 1, 2, \ldots, n,$$

и из соотношений (2.13) следует, что они независимы.

Легко убедиться, что

$$\sum_{1}^{n} y_i^2 = \sum_{1}^{n} x_i^2, \qquad y_n^2 = n\bar{x}^2,$$

поэтому

$$ns^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 = \sum_{i=1}^{n} y_i^2 - y_n^2 = \sum_{i=1}^{n-1} y_i^2.$$

Из независимости y_i следует независимость вектора

$$(y_1^2, y_2^2, \ldots, y_{n-1}^2)$$

от y_n, а следовательно, и независимость величин

$$s^2 = \frac{1}{n}\sum_{i=1}^{n-1} y_i^2 \quad \text{и} \quad \bar{x} = \frac{1}{\sqrt{n}} y_n.$$

Итак, доказано важное свойство, которое сформулировано в заголовке данного пункта.

б) Распределение случайной величины $\dfrac{ns^2}{\sigma^2}$

Если s^2 — дисперсия выборки из нормальной совокупности, то случайная величина ns^2/σ^2 имеет χ^2-распределение с $n-1$ степенями свободы.

Из выражений (2.13) следует, что случайные величины $y_i/\sigma \in N(0,1)$ и они независимы. Поэтому величина

$$\sum_{i=1}^{n-1} \frac{y_i^2}{\sigma^2} = \frac{ns^2}{\sigma^2}$$

имеет χ^2-распределение с $n-1$ степенями свободы, причем

$$\mathbf{E}\frac{ns^2}{\sigma^2} = n-1, \qquad \mathbf{D}\frac{ns^2}{\sigma^2} = 2(n-1).$$

Из последних двух равенств получаем

$$\mathbf{E}s^2 = \frac{n-1}{n}\sigma^2, \quad \mathbf{D}s^2 = \frac{2(n-1)}{n^2}\sigma^4.$$

в) Распределение случайной величины $\dfrac{(n-1)\bar{s}^2}{\sigma^2}$

Если $\bar{s}^2$ — несмещенная оценка дисперсии выборки из нормальной совокупности, то случайная величина $(n-1)\bar{s}^2/\sigma^2$ имеет χ^2-распределение с $n-1$ степенями свободы.

Для несмещенной оценки дисперсии, так как

$$\sum_{i=1}^{n-1} \frac{y_i^2}{\sigma^2} = \frac{(n-1)\bar{s}^2}{\sigma^2}, \qquad \text{где} \quad \bar{s}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2,$$

получаем, что случайная величина $(n-1)\bar{s}^2/\sigma^2$ имеет χ^2-распределение с $n-1$ степенями свободы и

$$\mathbf{E}\bar{s}^2 = \sigma^2, \quad D\bar{s}^2 = \frac{2\sigma^4}{n-1}.$$

г) Распределение случайной величины $\dfrac{(\bar{x}-m)\sqrt{n-1}}{s}$

Если $\bar{x}$ и $\bar{s}^2$ — среднее и дисперсия выборки из нормальной совокупности, то величина $(\bar{x}-m)\sqrt{n-1}/s$ имеет распределение Стьюдента с $n-1$ степенями свободы.

Так как величина $(\bar{x}-m)\sqrt{n}/\sigma \in N(0,1)$, а величина ns^2/σ^2 имеет распределение χ^2 с $n-1$ степенями свободы, то согласно п. 4б) случайная величина

$$t = \frac{(\bar{x}-m)\sqrt{n-1}}{\dfrac{\sigma}{\sqrt{n}}\sqrt{\dfrac{ns^2}{\sigma^2}}}$$

имеет распределение Стьюдента с $n-1$ степенями свободы. Простые преобразования приводят к следующему виду функции t:

$$t = \frac{(\bar{x}-m)\sqrt{n-1}}{s}.$$

Если для оценки σ^2 использовать несмещенную оценку $\bar{s}^2$, то функция t будет иметь вид

$$t = \frac{(\bar{x}-m)\sqrt{n}}{\bar{s}}.$$

6. Интервальное (доверительное) оценивание. Доверительный интервал

Построение статистических оценок бессмысленно без определения погрешности оценивания, меры разброса статистической оценки относительно истинного значения параметра. Чаще всего точность оценивания указывается построением *доверительного интервала, накрывающего истинное значение параметра с заданной вероятностью.* Построение доверительного интервала предполагает знание разброса статистической оценки, т. е. точную или приближенную информацию о ее функции распределения.

Пусть дана выборка независимых наблюдений $X_1, \ldots, X_n$ случайной величины X, имеющей функцию распределения $F(x, \theta)$, и $\hat{\theta}_n$ — оценка параметра θ. *Доверительным интервалом* $(\hat{\theta}_n - a, \hat{\theta}_n + b)$ называется интервал, которому с заданной вероятностью $1 - \alpha$ принадлежит истинное значение параметра θ, т. е.

$$P(\theta \in (\hat{\theta}_n - a, \hat{\theta}_n + b)) = 1 - \alpha.$$

Число $1 - \alpha$ называется *уровнем доверия* доверительного интервала.

Ясно, что границы доверительного интервала и его ширина зависят от уровня доверия $1 - \alpha$ и чем ближе уровень доверия к единице, тем шире доверительный интервал. С другой стороны, чем уже доверительный интервал при данном α, тем точнее оценка параметра θ.

Построение доверительного интервала возможно только на основе знания функции распределения $F_{\hat{\theta}_n}$ оценки $\hat{\theta}_n$. Действительно, пусть мы знаем

$$P_\theta(\theta - b < \hat{\theta}_n < \theta + a) = P_\theta(\hat{\theta_n} < \theta + a) - P_\theta(\hat{\theta}_n < \theta - b) = 1 - \alpha.$$

Тогда левую часть мы можем переписать так:

$$P_\theta(-b < \hat{\theta}_n - \theta < a) = P_\theta(-a < \theta - \hat{\theta}_n < b) =$$

$$P_\theta(\hat{\theta}_n - a < \theta < \hat{\theta}_n + b) = P_\theta(\theta \in (\hat{\theta}_n - a, \hat{\theta}_n + b)) = 1 - \alpha.$$

Рис. 2.3. Доверительный интервал.

Уровень доверия обычно задается заранее, причем в качестве $1 - \alpha$ берут достаточно большую вероятность, например 0.9; 0.95; 0.99 или 0.999.

Построим доверительные интервалы для параметров нормального распределения.

а) Доверительный интервал для математического ожидания нормального распределения. Требуется по уровню доверия $1-\alpha$ построить доверительный интервал для математического ожидания θ случайной величины X, распределенной по нормальному закону

$$X \in N(\theta, \sigma^2).$$

Случай 1. Рассмотрим сначала случай, когда *дисперсия* $DX = \sigma^2$ *известна.*

Возьмем в качестве оценки выборочное среднее $\bar{X}$, так как оно является эффективной оценкой математического ожидания. Для построения доверительного интервала нужно найти такое ϵ, что

$$P(\theta - \epsilon < \bar{X} < \theta + \epsilon) = P(-\epsilon < \bar{X} - \theta < \epsilon) = 1 - \alpha.$$

Мы берем $a = b = \epsilon$, поскольку плотность распределения $\bar{X}$ симметрична и для построения доверительного интервала естественно удалить наименее вероятные значения $\bar{X} - \theta$, которые расположены симметрично относительно нуля. Таким образом,

$$P(\bar{X} - \theta < -\epsilon) = P(\bar{X} - \theta > \epsilon) = \alpha/2.$$

Выборочные значения X_1, X_2, ..., X_n, по которым мы определяем $\bar{X}$, есть случайные величины одинаково распределенные, независимые, причем все $X_i \in N(\theta, \sigma^2)$. Тогда

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

также распределено по нормальному закону с математическим ожиданием $\mathbf{E}\bar{X} = \theta$ и $\mathbf{D}\bar{X} = \dfrac{\mathbf{D}X}{n} = \dfrac{\sigma^2}{n}$. Поэтому

$$\frac{(\bar{X} - \theta)\sqrt{n}}{\sigma}$$

имеет стандартное нормальное распределение и по формуле вычисления вероятностей попадания нормально распределенной вели-

чины в заданный интервал (симметричный относительно математического ожидания θ) имеем

$$P\left(\frac{|\bar{X}-\theta|\sqrt{n}}{\sigma} < x_{\alpha/2}\right) = P\left(|\bar{X}-\theta| < x_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) =$$

$$= P\left(-x_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \bar{X}-\theta < x_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = \Phi\left(x_{\alpha/2}\right) = 1-\alpha.$$

Здесь

$$\Phi(z) = \frac{1}{\sqrt{2\pi}}\int_0^z e^{-y^2/2}dy.$$

По таблице функций Лапласа $\Phi(z)$ находим такое $x_{\alpha/2}$, чтобы

$$\Phi(x_{\alpha/2}) = (1-\alpha)/2.$$

Таким образом, доверительный интервал имеет вид

$$\left(\bar{X} - \frac{x_{\alpha/2}\sigma}{\sqrt{n}},\ \bar{X} + \frac{x_{\alpha/2}\sigma}{\sqrt{n}}\right).$$

Он с доверительной вероятностью $1-\alpha$ содержит в себе истинное значение оцениваемой величины θ.

Если требуется оценить математическое ожидание с наперед заданной точностью ϵ так, чтобы вероятность такого отклонения была не меньше $1-\alpha$, то минимальный объем выборки, который обеспечит эту точность, находят по формуле $n = x_{\alpha/2}^2\sigma^2/\epsilon^2$.

Рассмотрим числовой пример. Средний вес багажа $n = 36$ пассажиров составил $\bar{X} = 14.3$ кг со стандартным отклонением $\sigma = 2.5$ кг. Найти доверительный интервал для математического ожидания, соответствующий $1-\alpha = 0.95$, предполагая, что величина веса багажа имеет распределение по нормальному закону.

Решение. $2\Phi(1.96) = 0.95$. Значит, $x_{\alpha/2} = 1.96$,

$$\epsilon = \frac{x_{\alpha/2}\sigma}{\sqrt{n}} = \frac{1.96\cdot 2.5}{\sqrt{36}} = 0.82,$$

$$\bar{X} - \epsilon = 14.3 - 0.82 = 13.48,$$

$$\bar{X} + \epsilon = 14.3 + 0.82 = 15.12.$$

Отсюда доверительный интервал для m с коэффициентом доверия 0.95 имеет вид (13.48; 15.12).

Случай 2. Рассмотрим случай, когда *математическое ожидание θ и дисперсия σ^2 нормального распределения случайной величины X нам неизвестны.*

С помощью метода максимального правдоподобия мы установили, что для выборки из нормальной совокупности наилучшей оценкой математического ожидания является $\bar{X}$, поэтому, как и в случае 1, оценкой $\hat{\theta}_n$ будет служить $\bar{X}$. Как и в случае 1, по заданной доверительной вероятности $1-\alpha$ найти доверительный интервал для математического ожидания θ — это значит найти такое ϵ, чтобы $P(|\bar{X}-\theta|<\epsilon)=1-\alpha$. Так как σ^2 нам неизвестно, то заменим его наилучшей оценкой, а именно несмещенной выборочной дисперсией $\bar{s}^2$:

$$\sigma \approx \bar{s} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2}.$$

Как было показано в предыдущей главе, случайная величина

$$T = \frac{\sqrt{n}(\bar{X}-\theta)}{\bar{s}}$$

имеет распределение Стьюдента с $n-1$ степенями свободы. Поэтому для построения доверительного интервала мы можем провести те же самые рассуждения, что и в предыдущем случае, только при вычислении $x_{\alpha/2}$ вместо нормального распределения взять распределение Стьюдента с $n-1$ степенями свободы.

Таким образом, доверительный интервал для математического ожидания θ, когда дисперсия нормально распределенной случайной величины X нам неизвестна, таков:

$$\left(\bar{X} - t_{\alpha,n-1}\frac{\bar{s}}{\sqrt{n}},\ \bar{X} + t_{\alpha,n-1}\frac{\bar{s}}{\sqrt{n}}\right),$$

где, чтобы подчеркнуть, что мы пользуемся распределением Стьюдента с $n-1$ степенями свободы, $x_{\alpha/2}$ заменено на $t_{\alpha,n-1}$.

Пример 2.8. Менеджер обнаружил, что $n=20$ клиентов внесли на депозит в среду 28.3; 28.5; 28.1; 28.2; 28.4; 28.2; 27.9;

28.4; 28.0; 28.1; 28.2; 28.0; 28.2; 28.1; 28.2; 28.3; 28.4; 27.9; 28.3; 28.1 фунтов стерлингов. Предполагая, что внесенные суммы распределены по нормальному закону, найти доверительный интервал для математического ожидания θ.

Имеем

$$\bar{X} = \frac{1}{20}\sum_{i=1}^{20} X_i = 28.19, \qquad \bar{s}^2 = \frac{1}{19}\sum_{i=1}^{20}(X_i - 28.19)^2 = 0.0283.$$

Зная $1-\alpha = 0.99$ и $n-1 = 19$, находим из таблицы $t_{\alpha,n-1} = 2.86$. Получаем

$$\epsilon = t_{\alpha,n-1}\sqrt{\frac{\bar{s}^2}{n}} = 2.86\sqrt{\frac{0.0283}{20}} = 0.108,$$

а доверительный интервал равен $(28.19 - 0.108,\ 28.19 + 0.108)$ или $(28.082,\ 28.298)$.

Доверительная вероятность $1-\alpha = 0.99$ указывает, что если произведено достаточно большое число выборок, то 99% из них определяет такие доверительные интервалы, в которых параметр θ действительно заключен; лишь в 1% случаев он может выйти за границы доверительного интервала.

Замечание. При неограниченном возрастании объема выборки n распределение Стьюдента стремится к нормальному. Поэтому при большом числе измерений (практически $n > 30$) вместо распределения Стьюдента можно пользоваться нормальным. В этом случае неизвестное значение σ_x можно приближенно заменить оценкой $\bar{s}$ и для определения ϵ воспользоваться формулой $\epsilon = \frac{z_{\alpha/2}\bar{s}}{\sqrt{n}}$, где $z_{\alpha/2}$ находится из уравнения $2\Phi(z_{\alpha/2}) = 1-\alpha$.

б) Доверительный интервал для дисперсии σ^2 нормального распределения. Как мы знаем, случайная величина

$$\chi^2_\nu = \frac{(n-1)\bar{s}^2}{\sigma^2}$$

имеет χ^2-квадрат распределение с $\nu = n-1$ степенями свободы. Поэтому доверительный интервал находится следующим образом:

$$1-\alpha = \mathbf{P}\left(\chi^2_{\nu,1-\alpha/2} < \chi^2_\nu < \chi^2_{\nu,\alpha/2}\right) =$$
$$= \mathbf{P}\left(\chi^2_{\nu,1-\alpha/2} < \frac{(n-1)\bar{s}^2}{\sigma^2} < \chi^2_{\nu,\alpha/2}\right) =$$
$$= \mathbf{P}\left(\frac{(n-1)\bar{s}^2}{\chi^2_{\nu,\alpha/2}} < \sigma^2 < \frac{(n-1)\bar{s}^2}{\chi^2_{\nu,1-\alpha/2}}\right),$$

и имеет границы

$$\frac{(n-1)\bar{s}^2}{\chi^2_{\nu,\alpha/2}};\ \frac{(n-1)\bar{s}^2}{\chi^2_{\nu,1-\alpha/2}}.$$

в) Доверительный интервал для параметра p биномиального распределения. Пусть $X_1, \ldots, X_n$ — независимые случайные величины, принимающие значение 1 с неизвестной вероятностью p и значение 0 с вероятностью $1-p$. Положим

$$\nu_n = \sum_{i=1}^{n} X_i, \quad \hat{p}_n = \frac{\nu_n}{n}.$$

Как мы знаем, $\mathbf{E}\nu_n = np$, $D\nu_n = np(1-p)$ и соответственно $\mathbf{E}\hat{p}_n = p$, $D\hat{p}_n = p(1-p)/n$.

Случайная величина

$$\frac{\nu_n - np}{\sqrt{np(1-p)}} = \frac{\sqrt{n}(\hat{p}_n - p)}{\sqrt{p(1-p)}}$$

согласно центральной предельной теореме имеет при $n \to \infty$ предельное нормальное распределение. Поэтому, заменяя стандартное отклонение $\sqrt{p(1-p)}$ ее оценкой $\sqrt{\hat{p}_n(1-\hat{p}_n)}$, получаем

$$\mathbf{P}\left(-x_{\alpha/2} < \frac{n^{1/2}(\hat{p}_n - p)}{\sqrt{p(1-p)}} < x_{\alpha/2}\right) \approx 1-\alpha.$$

Рассуждая как при выводе доверительного интервала для математического ожидания нормального распределения (заменяя θ на p и $\sigma/\sqrt{n}$ на $\sqrt{\hat{p}_n(1-\hat{p}_n)}/\sqrt{n}$), находим вид доверительного интервала

$$\hat{p}_n - x_{\alpha/2}\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}} < p < \hat{p}_n + x_{\alpha/2}\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}.$$

Глава 3

СТАТИСТИЧЕСКАЯ ПРОВЕРКА ГИПОТЕЗ

В процессе своей деятельности менеджер постоянно принимает решения на основе поступающей к нему информации. В качестве значительного аргумента для подтверждения своего решения он часто может использовать статистический аппарат, называемый теорией проверки гипотез.

1. Выбор критерия значимости

Наиболее распространенная постановка задачи проверки гипотез состоит в следующем. Дана выборка независимых наблюдений $X_1, \ldots, X_n$, имеющих функцию распределения $F(x, \theta)$ с некоторым неизвестным параметром θ. Надо проверить гипотезу $H_0 : \theta \in \Theta_0$, где Θ_0 — интересующее нас множество значений параметра θ. Например, в качестве гипотез H_0 могут быть $\theta = \theta_0$, $\theta > \theta_0$, $\theta < \theta_0$ и другие. В качестве альтернативы H_1 выступает противоположное событие.

Рассмотрим в качестве примеров следующие задачи, в которых параметр θ является математическим ожиданием.

Задача A_1. Проявление фотографии при изготовлении обычно требует 30 с. Считается, что изменение времени ухудшает качество фотографии. Менеджер корпорации CAH решил проверить это утверждение. В качестве гипотезы берем H_0: $\theta = \theta_0 = 30$ с.

Задача A_2. Необходимо проверить гипотезу о том, что рекламная кампания увеличила спрос на продукцию. Пусть θ_0 — средний спрос на продукцию до выхода рекламы. В качестве гипотезы берем $H_0 : \theta > \theta_0$, т. е. спрос увеличился, альтернатива $H_1 : \theta \leq \theta_0$. По данным о нынешнем спросе продукции проверяем гипотезу H_0.

Задача A_3. Нужно проверить гипотезу о том, что изменение технологии привело к уменьшению средних затрат θ на единицу продукции. В качестве гипотезы берем $H_0 : \theta < \theta_0$, где θ_0 — прежние средние затраты.

Как проверяется обычно гипотеза?

Задаются некоторая функция от наблюдений

$$\hat{T}_n = \hat{T}(X_1, \ldots, X_n),$$

называемая *тестовой статистикой*, и *критическое значение* C. Если $\hat{T}_n > C$, то принимается альтернатива, а в противном случае принимается гипотеза. Выбор значения C зависит от меры достоверности, с которой хотят принять гипотезу.

В задачах A_1, A_2 и A_3 в качестве тестовой статистики естественно взять нормированное выборочное среднее (времени проявления фотографий, спроса на продукцию или стоимости единицы продукции).

Математически это все облачается в следующую форму.

Определяются *область допустимых значений*

$$D_0 = \{(X_1, \ldots, X_n) : \hat{T}_n < C\},$$

при которых принимается гипотеза, и *критическая область* отклонения гипотезы

$$D_1 = \{(X_1, \ldots, X_n) : \hat{T}_n \overline{\in} C\}.$$

При проверке гипотезы мы можем совершить ошибки двоякого рода:
— отвергнуть гипотезу, когда она верна, — *ошибка первого рода*,
— принять гипотезу, когда она не верна, — *ошибка второго рода*.

Обозначим *вероятность ошибки первого рода* через α, а *вероятность ошибки второго рода* — через β.

Обычно вероятность α ошибки первого рода задается и называется уровнем значимости. По вероятности α и вычисляется критическое значение $C = x_\alpha$. Обычно берется $\alpha = 0.1;\ 0.05;\ 0.001$. Критические значения x_α находятся по распределению тестовой статистики с помощью таблиц распределения или пакетов статистических программ. Вероятности β ошибки второго рода уделяется не столь большое внимание. Она используется для анализа качества процедуры проверки гипотезы.

Таким образом, решающая процедура, называемая *критерием проверки гипотезы* имеет вид:

$$\hat{K}_n = K(X_1, \ldots, X_n) = \chi(\hat{T}_n > x_\alpha),$$

где $\chi(y > x_\alpha)$ — индикаторная функция, определяемая равенством

$$\chi(y > x_\alpha) = \begin{cases} 1, & \text{если } y > x_\alpha, \\ 0, & \text{если } y < x_\alpha. \end{cases}$$

Если $\hat{K}_n = 0$ (т. е. $\hat{T}_n < x_\alpha$), то принимается гипотеза H_0; если $\hat{K}_n = 1$ (т. е. $\hat{T}_n > x_\alpha$), то принимается альтернатива H_1 (таким образом, значение $\hat{K}_n$ совпадает с индексом величин H_0 или H_1).

Обычно тестовая статистика $\hat{T}_n$ строится по некоторой оценке $\hat{\theta}_n(X_1, \ldots, X_n)$. Обратимся к примерам.

Решение задачи A_1. Пусть в процессе проверки гипотезы $\theta_0 = 30$ получены следующие результаты. Проявление $n = 25$ качественных фотографий потребовало в среднем $\bar{X} = 31.8$ с при среднеквадратичном отклонении $\sigma = 4$ с. Считать, что время проявления фотографий распределено по нормальному закону, и уровень значимости взять $\alpha = 0.1$.

Таким образом, рассматривается гипотеза $H_0 : \theta = \theta_0 = 30$ против альтернативы $\theta \neq 30$. В качестве тестовой статистики $\hat{T}_n$ естественно взять

$$\hat{T}_n = \frac{\sqrt{n}(\bar{X} - \theta_0)}{\sigma}, \tag{3.1}$$

так как $\bar{X}$ является наилучшей оценкой параметра θ.

В данной задаче на отклонение гипотезы влияют как очень малые, так и очень большие сроки проявления фотографии. Поэтому мы удаляем как малые, так и большие значения тестовой статистики $\hat{T}_n$ с равными вероятностями $\alpha/2$. Поскольку $\hat{T}_n$ имеет стандартное нормальное распределение, то ее плотность симметрична, мы берем с обеих сторон равные критические значения (разные по знаку) и задаем критерий

$$\hat{K}_n = \frac{\sqrt{n}|\bar{X} - \theta_0|}{\sigma} > x_{\alpha/2},$$

где

$$\alpha/2 = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-x_{\alpha/2}} \exp\{-t^2/2\} dt = \frac{1}{\sqrt{2\pi}} \int_{x_{\alpha/2}}^{\infty} \exp\{-t^2/2\} dt. \tag{3.2}$$

Непосредственными вычислениями получаем

$$\hat{T}_n = \frac{\sqrt{n}|\bar{X} - \theta_0|}{\sigma} = \frac{5(31.8 - 30)}{4} = 2.25 > x_{\alpha/2} = x_{0.05} = 1.65,$$

и гипотеза отвергается при $\alpha = 0.1$.

Задачи проверки гипотез о среднем нормального распределения обычно имеют один из трех видов

$$H_0 : \theta = \theta_0; \qquad H_1 : \theta \neq \theta_0; \qquad \hat{K}_n = \chi(|\hat{T}_n| > x_{\alpha/2});$$

$$H_0 : \theta > \theta_0; \qquad H_1 : \theta \leq \theta_0; \qquad \hat{K}_n = \chi(\hat{T}_n < x_\alpha);$$

$$H_0 : \theta < \theta_0; \qquad H_1 : \theta \geq \theta_0; \qquad \hat{K}_n = \chi(\hat{T}_n > -x_\alpha).$$

Так, если предположить в задаче A_2, что после рекламной кампании ежедневный объем продаж был $X_1, \ldots, X_n$, дисперсия продаж равна примерно σ^2, и допустить, что выборка распределена по нормальному закону, то в качестве тестовой статистики естественно взять ту же статистику (3.1).

Критическая область равна

$$D_1 = \{(X_1, \ldots, X_n) : \hat{T}_n < x_\alpha\},$$

где x_α находится из уравнения

$$1 - \alpha = \frac{1}{\sqrt{2\pi}} \int_{x_\alpha}^{\infty} \exp\{-t^2/2\} dt.$$

Критерий значимости $\hat{K}_n$ имеет вид

$$\hat{K}_n = \chi(\hat{T}_n < -x_\alpha),$$

где функция $\chi(Z)$ обозначает индикатор события Z, т. е. $\chi(Z) = 1$, если имеет место событие Z, и $\chi(Z) = 0$, если Z не имеет место.

В задаче A_3 в предположении, что наблюдения имеют нормальное распределение с дисперсией σ^2, тестовая статистика $\hat{T}_n$ имеет вид (3.1), т. е. будет такой же, как и в задачах A_1 и A_2, а критерий значимости — другой:

$$\hat{K}_n = \chi(\hat{T}_n > 1 - x_\alpha\}),$$

где x_α определяется из уравнения

$$\alpha = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-x_\alpha} \exp\{-t^2/2\} dt.$$

Как уже отмечалось, при проверке гипотезы возможны ошибки двух родов рода: отвергнуть гипотезу, когда она верна (ошибка первого рода), и принять гипотезу, когда она не верна (ошибка второго рода).

Вероятности этих ошибок соответственно равны

$$\alpha_{\theta_0} = \mathbf{P}_{\theta_0}(\hat{K}_n = 1) = \mathbf{P}_{\theta_0}(\hat{T}_n > x_\alpha),$$
$$\beta_\theta = \mathbf{P}_\theta(\hat{K}_n = 0) = \mathbf{P}_\theta(\hat{T}_n < x_\alpha).$$

Таким образом, в обоих случаях мы отклоняем гипотезу, если тестовая статистика $\hat{T}_n$ оказывается в критической области, где вероятность появления наблюдений при справедливости гипотезы H_0 мала, а при справедливости альтернативы H_1 она существенно больше.

При данном критерии $\hat{K}_n$ определяется максимально допустимая ошибка первого рода, которая называется *уровнем значимости.*

После этого о качестве критерия судят по значениям вероятностей ошибок второго рода.

Гипотеза и альтернатива в приложениях играют совершенно различную роль. Вероятность ошибки первого рода — отвергнуть гипотезу при ее справедливости — мала. Гипотеза при ее справедливости отклоняется только при очень маловероятных событиях. Поэтому принятие гипотезы означает, что в принципе задача может быть подвергнута дополнительному исследованию (например, гипотеза может быть принята из-за недостаточного объема выборки). В то же время отклонение гипотезы обычно носит окончательный характер и считается, что принятая альтернатива не нуждается в дальнейшем подтверждении. Именно поэтому в задачах A_2 и A_3 мы таким образом взяли альтернативу и гипотезу. Принятие гипотезы будет означать соответственно, что рекламная кампания не увеличила спрос на товар и стоимость молока возросла меньше, чем на 30%, что традиционно для нашего государства. Аналогично в задачах проверки гипотез, что лекарство эффективно, мы берем в качестве гипотезы предположение, что лекарство неэффективно. А в судебных экспертизах на виновность человека в качестве гипотезы берется, что человек невиновен.

В заключение остановимся еще на одной классификации альтернатив и гипотез. Гипотеза и альтернатива бывают простой и

сложной. Гипотеза простая, если проверяется, что параметр имеет одно конкретное значение. Если множество возможных значений параметра больше одного, то гипотеза сложная. Аналогично — альтернатива простая, если множество ее возможных значений состоит всего из одной точки. В противном случае альтернатива сложная.

2. Проверка гипотез о параметрах распределения

Покажем, как задачу A_1 решить с помощью доверительного оценивания. Построим доверительный интервал для параметра θ:

$$(\bar{x} - x_{\alpha/2}\sigma/\sqrt{n};\ \bar{x} + x_{\alpha/2}\sigma/\sqrt{n}) =$$
$$= (31.8 - 1.64 \cdot 4/5;\ 31.8 + 1.64 \cdot 4/5) = (30.488;\ 33.112).$$

Величина $\theta_0 = 30$, соответствующая гипотезе H_0, не принадлежит данному интервалу, поэтому мы отвергаем гипотезу H_0.

Является ли непринадлежность параметра θ_0 гипотезы доверительному интервалу необходимым и достаточным условием отклонения гипотезы? Ответ утвердительный. Действительно, гипотеза при данном уровне значимости α принимается тогда и только тогда, когда значение параметра θ_0 принадлежит доверительному интервалу с уровнем доверия $1 - \alpha$.

Это доказывается довольно просто. При задании $x_{\alpha/2}$ по уровню доверия $1 - \alpha$ мы исходили из уравнения

$$\mathbf{P}\left(\frac{\sqrt{n}|\bar{X} - \theta_0|}{\sigma} < x_{\alpha/2}\right) = 1 - \alpha$$

и показывали, что соотношение

$$\hat{T_n} = \frac{\sqrt{n}|\bar{X} - \theta_0|}{\sigma} < x_{\alpha/2}$$

эквивалентно

$$\theta_0 \in (\bar{X} - x_{\alpha/2}\sigma/\sqrt{n},\ \bar{X} + x_{\alpha/2}\sigma/\sqrt{n}).$$

Но выражение $\sqrt{n}|\bar{X} - \theta_0|/\sigma$ задает тестовую статистику критерия проверки гипотезы H_0: $\theta = \theta_0$ против альтернативы H_1: $\theta \neq \theta_0$, причем $x_{\alpha/2}$ задается тем же самым уравнением (3.2).

Таким образом, мы можем сформулировать общий принцип. *Если доверительный интервал накрывает значение $\theta = \theta_0$ гипотезы, то гипотеза принимается. В противном случае гипотеза отклоняется.*

Как этот принцип действует в случае односторонних альтернатив? Оказывается, что бывают односторонние доверительные интервалы, а именно левый $(-\infty,\ \bar{X} + x_\alpha\sigma/\sqrt{n})$ с альтернативой H_1: $\theta > \theta_0$ и правый $(\bar{X} - x_\alpha\sigma/\sqrt{n},\ \infty)$ с альтернативой H_1: $\theta < \theta_0$.

Как следствие, рассуждая аналогично, мы получаем тестовые статистики и для других задач проверки гипотез.

а) Проверка гипотезы о среднем нормального распределения с неизвестной дисперсией

При проверке гипотезы H_0: $\theta = \theta_0$ против альтернативы H_1: $\theta \neq \theta_0$ критерий значимости имеет вид

$$\hat{T}_n = \frac{\sqrt{n}|\bar{X} - \theta_0|}{\bar{s}} > t_{\alpha/2,n-1},$$

где $t_{\alpha/2,n-1}$ находится из таблиц распределения Стьюдента с $n-1$ степенями свободы.

б) Проверка гипотезы о вероятности p биномиального распределения

При проверке гипотезы H_0: $p = p_0$ против альтернативы H_1: $p \neq p_0$ критерий значимости имеет вид

$$\hat{T}_n = \frac{|\hat{p}_n - p|\sqrt{n}}{\sqrt{\hat{p}(1-\hat{p})}} > x_{\alpha/2},$$

где $x_{\alpha/2}$ удовлетворяет соотношению (3.2) и находится из таблиц стандартного нормального распределения.

в) Проверка гипотезы $\sigma = \sigma_0$ против альтернативы $\sigma \neq \sigma_0$ в случае нормального распределения

Гипотеза принимается, если

$$\chi^2_{\nu,1-\alpha/2} < \frac{(n-1)\hat{s}^2}{\sigma_0^2} < \chi^2_{\nu,\alpha/2}.$$

Величины $\chi^2_{\nu,\alpha/2}$ и $\chi^2_{\nu,1-\alpha/2}$, удовлетворяющие соотношению (2.11), находим с помощью таблиц χ^2-распределения.

Для односторонних альтернатив вида $\theta > \theta_0$ критерии отличаются только тем, что в тестовой статистике $\hat{T}_n$ убирается модуль и критическое значение $x_{\alpha/2}$ заменяется на x_α.

При использовании пакетов статистических программ обычно используется понятие *p-значения (probability value)* вместо понятия уровня значимости. Именно оно обычно появляется на экране монитора.

p-значение α_p является наибольшей вероятностью ошибки первого рода, при которой принимается гипотеза.

Таким образом, правило принятия гипотезы таково.

в1) Если α_p больше уровня значимости α ($\alpha_p > \alpha$), то гипотеза H_0 принимается.

в2) Если α_p меньше уровня значимости α ($\alpha_p < \alpha$), то гипотеза H_0 отвергается.

Поясним смысл p-значения. По исходным данным вычисляется значение тестовой статистики. Допустим оно равно $t = \hat{T}_n$. Для этого t находим вероятность α_p такую, что

$$\mathbf{P}(\hat{T}_n > t) = \alpha_p. \tag{3.3}$$

Эта вероятность α_p и является p-значением.

Так, в примере A_1 мы получили $t = 2.25$, отсюда

$$\alpha_p = \mathbf{P}\left(\sqrt{n}\frac{|\bar{X} - \theta_0|}{\sigma} > 2.25\right) = \frac{2}{\sqrt{2\pi}}\int_{2.25}^{\infty} \exp\{-s^2/2\}ds = 0.025.$$

Так как в условии задачи уровень значимости $\alpha = 0.1$ и, следовательно, $\alpha > \alpha_p$, то гипотеза H_0 отвергается.

Как обосновать уравнение (3.3), задающее p-значение α_p? Для этого достаточно обосновать утверждения *в1)* и *в2)*.

Из уравнения

$$\mathbf{P}(\hat{T}_n > x_\alpha) = \alpha$$

мы видим, что чем больше x_α, тем меньше α, поэтому если $t = \hat{T}_n > x_\alpha$, то $\alpha_p < \alpha$. С другой стороны, неравенство $t = \hat{T}_n > x_\alpha$ означает, что значение t тестовой статистики $\hat{T}_n$, найденное по выборке, попало в критическую зону, поэтому мы отвергаем гипотезу и тем самым доказываем пункт *в2)*. Пункт *в1)* доказывается аналогично.

г) Проверка гипотезы о равенстве математических ожиданий двух нормально распределенных совокупностей

Довольно часто менеджеру приходится сравнивать значения средних двух выборок, чтобы прийти к определенному решению. Так, например, в задаче A_2, в принципе, мы должны были бы сравнивать числовые данные о спросе до и после рекламной кампании. Другими примерами являются перечисленные ниже задачи.

Действительно ли в компании мужчины и женщины за одинаковую работу и получают одну и ту же заработную плату?

Является ли продолжительность жизни продукции (телевизоров, компьютеров и т. п.) в двух компаниях одинаковой?

Одинакова ли в среднем стоимость продовольственных товаров в двух магазинах и т. п.?

Пусть две независимые случайные величины X и Y распределены нормально и $X \in N(m_x, \sigma_x^2)$, $Y \in N(m_y, \sigma_y^2)$. Пусть имеются две независимые выборки

$$X_1, \ldots, X_{n_1}, \quad Y_1, \ldots, Y_{n_2}$$

объемов n_1 и n_2 соответственно для X и Y и

$$\bar{X} = \frac{1}{n_1} \sum_{j=1}^{n_1} X_j, \quad \bar{Y} = \frac{1}{n_2} \sum_{j=1}^{n_2} Y_j,$$

$$s_x^2 = \frac{1}{n_1} \sum_{j=1}^{n_1} (X_j - \bar{X})^2, \quad s_y^2 = \frac{1}{n_2} \sum_{j=1}^{n_2} (Y_j - \bar{Y})^2.$$

Нужно проверить гипотезу H_0, состоящую в том, что $EX = EY$, т. е.

$$H_0 : \quad m_x = m_y.$$

Для проверки гипотезы H_0 следует воспользоваться различными критериями значимости в зависимости от того, известны или не известны дисперсии случайных величин X и Y. Рассмотрим оба случая.

г1) Пусть σ_x и σ_y известны. Так как

$$\bar{X} \in N\left(m_x, \frac{\sigma_x^2}{n_1}\right), \quad \bar{Y} \in N\left(m_y, \frac{\sigma_y^2}{n_2}\right),$$

а случайные величины $\bar{X}$ и $\bar{Y}$ независимы, то

$$\sigma^2_{(\bar{X}-\bar{Y})} = \mathbf{D}(\bar{X}-\bar{Y}) = \mathbf{D}\bar{X} + \mathbf{D}\bar{Y} = \frac{\sigma_x^2}{n_1} + \frac{\sigma_y^2}{n_2}$$

(дисперсия разности равна сумме дисперсий, так как по свойству дисперсий $\mathbf{D}(X) = \mathbf{D}(-X)$). Отсюда следует, что

$$\bar{X} - \bar{Y} \in N(m_x - m_y,\, \sigma^2(\bar{X} - \bar{Y})).$$

Для проверки гипотезы H_0 в качестве критерия рассмотрим величину

$$z = \frac{\bar{X} - \bar{Y}}{\sqrt{\dfrac{\sigma_x^2}{n_1} + \dfrac{\sigma_y^2}{n_2}}}. \tag{3.4}$$

Если гипотеза H_0 верна , то $z \in N(0,1)$ и в качестве критической области для двусторонней альтернативы следует взять область больших по модулю отклонений z, т. е. $|z| > x_{\alpha/2}$, где $x_{\alpha/2}$ определяется из (3.2), а α — уровень значимости.

Пример 3.1. Пусть $\bar{X} = 18.4,\ \bar{Y} = 19\ \sigma_X^2 = 1.2,\ \sigma_Y^2 = 3,$ $n_1 = 20,\ n_2 = 30$. Найдем значение критерия z:

$$z = \frac{18.4 - 19}{\sqrt{1.2/20 + 3/30}} = -1.5.$$

Пусть $\alpha = 0.05$, тогда $z_\alpha = 1.96$, и так как $|z| < z_\alpha(|-1.5| < 1.96)$, то принимаем гипотезу о равенстве математических ожиданий.

Пример 3.2. Пусть $\bar{X} = 18.4,\ \bar{Y} = 19, \sigma_X^2 = 24/25,\ \sigma_Y^2 = 12/5,$ $n_1 = 20,\ n_2 = 30$. В этом случае $z = 2.25$ и при $\alpha = 0.05,\ z_\alpha = 1.96$ получаем $|z| > z_\alpha$ и поэтому гипотезу H_0 отвергаем.

г2) Рассмотрим гипотезу H_0: $m_x = m_y$, когда дисперсии σ_x и σ_y не известны, но выполнено условие

$$\sigma_X^2 = \sigma_Y^2 = \sigma^2. \tag{3.5}$$

При выполнении (3.5) величина z в (3.4) принимает вид

$$\xi = \frac{\bar{X} - \bar{Y}}{\sigma\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}, \tag{3.6}$$

причем при выполнении гипотезы H_0 случайная величина $\xi \in N(0,1)$. Согласно п. 5г) гл. 2 случайные величины $\eta_1 = n_1 s_x^2/\sigma^2$ и $\eta_2 = n_2 s_y^2/\sigma^2$ имеют χ^2-распределение соответственно с $n_1 - 1$ и $n_2 - 1$ степенями свободы, а так как X и Y независимы, то $\eta_1 + \eta_2$ имеет χ^2 – распределение с $n_1 + n_2 - 2$ степенями свободы. Тогда с учетом п. 4б) гл. 2 величина

$$\frac{\xi\sqrt{n_1 + n_2 - 2}}{\sqrt{\eta_1 + \eta_1}} \tag{3.7}$$

имеет распределение Стьюдента с $n_1 + n_2 - 2$ степенями свободы и ее можно записать в виде

$$t = \frac{\bar{X} - \bar{Y}}{\sqrt{n_1 s_x^2 + n_2 s_y^2}} \sqrt{\frac{n_1 n_2 (n_1 + n_2 - 2)}{n_1 + n_2}}. \tag{3.8}$$

Далее поступаем следующим образом. Задаем уровень значимости $\alpha = 0.05$; 0.01 и т. д. и затем по таблице распределения Стьюдента находим $t_{\alpha,k}$. Если значения $|t| > t_{\alpha,k}$, то отвергаем H_0, если $|t| < t_{\alpha,k}$, то говорим, что гипотеза H_0 не противоречит данным.

Пример 3.3. Пусть $\bar{X} = 20.6$, $\bar{Y} = 21.6$, $\alpha = 0.05$, $\sigma_X^2 = 2.5$, $\sigma_Y^2 = 2, n_1 = 10$, $n_2 = 12$. Вычисляем критерий

$$t = \frac{1}{7}\sqrt{\frac{10 \cdot 20 \cdot 12}{22}} \approx 1.46,$$

находим число степеней свободы $r = 10+12-2 = 20$ и при заданном уровне значимости $\alpha = 0.05$ по таблице распределения Стьюдента находим $t_{0.05,20} = 2.074$. Так как $|t| < t_{0.05,20}$, следовательно, величина t попадает в область допустимых значений, и мы принимаем гипотезу H_0, т. е. считаем, что $m_x = m_y$.

Пример 3.4. Пусть $\bar{X} = 20.6$, $\bar{Y} = 21.6$, $q = 0.05$, $\sigma_X^2 = 0.625$, $\sigma_Y^2 = 0.4$, $n_1 = 10$, $n_2 = 12$. Вычисляем критерий $t = 2.92$, и так как $r = 20$, $t_{0.05,20} = 2.074$, то $|t| > t_{0.05,20}$, поэтому гипотезу H_0 отвергаем.

д) Проверка гипотезы о равенстве дисперсий двух нормально распределенных совокупностей

Пусть у нас есть та же самая постановка задачи, что и в предыдущем пункте. Даны две выборки независимых нормально распределенных случайных величин с неизвестными математическими ожиданиями и дисперсиями. Стоит задача проверки гипотезы о равенстве дисперсий

$$H_0: \quad \sigma_x = \sigma_y.$$

В данной задаче в качестве тестовой статистики возьмем

$$\hat{T}_n = s_x^2 / s_y^2.$$

Статистика $\hat{T}_n$ имеет распределение Фишера с $n_1 - 1$ и $n_2 - 1$ степенями свободы.

3. Проверка гипотезы о типе распределения

До сих пор мы занимались оценкой параметров распределения некой случайной величины X, имея n наблюдений этой величины X, и, кроме того, вид распределения нам был известен. На практике возникают задачи, когда бывает необходимо выявить закон распределения случайной величины X, имея ряд наблюдений этой величины X. Если закон распределения неизвестен, но есть основания предполагать, что он имеет определенный вид (например нормальный), то в качестве гипотезы выступает утверждение о виде (например нормальном) распределения случайной величины X. Какими же пользуются критериями для проверки гипотезы о виде распределения? Принцип применения *критериев согласия* состоит в следующем: по выборочным данным x_1, x_2, ... строится некоторая случайная величина v (мера расхождения), характеризующая степень расхождения теоретического и эмпирического распределений. Эта случайная величина может быть построена различными способами, причем каждый способ построения означает применение того или иного критерия. Закон распределения этой случайной величины v зависит от распределения искомой величины X и от объема выборки n.

а) Критерий χ^2 (Пирсона)

Наиболее употребительной мерой расхождения является величина χ^2, а критерий, использующий эту величину χ^2, называют соответственно критерием χ^2 (хи-квадрат).

Случайная величина χ^2 (мера расхождения) представляет собой сумму квадратов разностей между наблюденными частотами и истинными (если закон распределения задан), деленными на истинные частоты :

$$\chi^2 = \sum_{i=1}^{k} \frac{(\nu_i - np_i)^2}{np_i}. \tag{3.9}$$

Здесь все выборочные значения x_1, x_2, ..., x_n сгруппированы в k интервалов, ν_i — число наблюдений попавших в i-й интервал, а p_i — вероятность попадания X в i-й интервал в предположении, что гипотеза H_0 верна. Если значения случайной величины лежат в пределах $A \le X \le B$, то границы k интервалов запишем в виде (e_i, e_{i+1}), $i = 1, 2, \ldots, k$, причем $e_1 = A$, $e_{k+1} = B$.

Вероятность p_i попадания случайной величины X, имеющей плотность распределения $f(x)$, в i-й интервал с границами $e_i \div e_{i+1}$ вычисляется по формуле

$$p_i = \int_{e_i}^{e_{i+1}} f(x)dx \quad \text{или} \quad p_i = F(e_{i+1}) - F(e_i),$$

где $F(x)$ — функция распределения X. В этом случае величина np_i есть не что иное, как "теоретическое среднее" числа попаданий в i-й интервал при n наблюдениях.

Английским статистиком К. Пирсом доказано, что при достаточно большом n закон распределения этой суммы χ^2 практически не зависит ни от предполагаемого теоретического закона, ни от числа наблюдений n и при $n \to \infty$ приближается к известному в теории вероятностей закону распределения χ^2 с $k-1$ степенями свободы.

Число степеней свободы r этого распределения χ^2 равно разности между числом интервалов k и числом независимых условий (связей). Одной такой связью всегда является требование

$$p_1^* + p_2^* + \cdots + p_k^* = 1, \quad \text{где} \quad p_i^* = \frac{\nu_i}{n}.$$

Если проверяемый закон распределения зависит от j параметров, то их оценивают на основании той же выборки, и при этом число степеней свободы уменьшается на j и равно $k-1-j$.

Например, если проверяется принадлежность выборки к нормальной совокупности, то следует найти оценки математического ожидания и дисперсии, а для этого, как известно, используются выборочное среднее $\bar{x}$ и выборочная дисперсия s^2.

Если исходные данные заданы в виде таблицы, в которой указаны границы $e_i \div e_{i+1}$ интервалов, и $\nu_i,\ i=1,\ldots,k$, — количество наблюдений, попавших в эти интервалы, то выборочные оценки $\bar{x}$ и σ^2 вычисляются по формулам

$$\bar{x}=\sum_{i=1}^{k} x_i^* p_i^*, \quad s^2=\sum_{i=1}^{k}(x_i^*-\bar{x})^2 p_i^*=\sum_{i=1}^{k}(x_i^*)^2 p_i^*-\bar{x}^2,$$

где x_i^* — среднее значение i-го интервала $x_i^*=(e_{i+1}+e_i)/2$. В этом случае число всех связей равно 3, поэтому число степеней свободы $r=k-3$.

Ясно, что если вычисленная мера расхождения будет слишком большой, то выбранное теоретическое распределение наверняка не согласуется с экспериментальными данными. Вопрос о том, какой должна быть величина χ^2, чтобы теоретическое распределение было согласно с экспериментальными данными, зависит от выбранной доверительной вероятности β (или уровня значимости $\alpha=1-\beta$), которая выбирается обычно 0.95 (уровень значимости 0.05).

Пусть уровень значимости задан α. Тогда применение критерия Пирсона заключается в следующем. Определяем число степеней свободы r, при уровне значимости α находим по таблицам χ^2-распределения значение $\chi^2_{\alpha,r}$, такое, что

$$P(\chi^2 \geq \chi^2_{\alpha,r})=\alpha. \tag{3.10}$$

Вычисляем величину χ^2 по формуле (3.9). Если оказалось, что $\chi^2 \geq \chi^2_{\alpha,r}$, то, так как это событие имеет вероятность α (см. формулу (3.10)), мы отвергаем проверяемую гипотезу и говорим, что выбранное теоретическое распределение не согласуется с экспериментальными данными.

Если же $\chi^2 \leq \chi^2_{\alpha,r}$, то принимается гипотеза о согласованности теоретического и статистического распределений.

Возможен другой подход к понятию согласованности теоретического и выборочного распределений. Вычислим χ^2 по формуле (3.9) и обозначим $C = \chi^2$, сосчитаем число степеней свободы r и найдем по таблицам χ^2-распределения p-значение (α_p), такое, что

$$\tilde{F}(C) = 1 - \alpha_p.$$

Здесь $\tilde{F}(x)$ — χ^2-распределение. Величина α_p более точно отражает степень согласия теоретического и эмпирического законов. При $\alpha_p \geq \alpha$ мы принимаем проверяемую гипотезу, в противном случае отвергаем ее.

Для применения критерия Пирсона в общем случае необходимо, чтобы число наблюдений n было достаточно велико (практически $n \leq 50 \div 60$) и чтобы численность каждого интервала была не меньше 5. Если в каких-то интервалах окажется меньше 5 наблюдений, то следует объединить эти интервалы.

Пример 3.5. В табл. 3.1 приведены отклонения e_i от заданного размера $n = 160$ деталей, обработанных на станке. Проверить, используя критерий χ^2, гипотезу о согласии наблюдений с законом нормального распределения, приняв уровень значимости равным $\alpha = 0,05$.

Таблица 3.1

$e_i \div e_{i+1}$	ν_i	$l_i \div l_{i+1}$	$\Phi(l_i)$	p_i	np_i	χ_i^2
$-15 \div -10$	10	$-\infty \div -1.66$	-0.500	0.048	7.75	0.65
$-10 \div -5$	16	$-1.66 \div -0.99$	-0.452	0.113	18.02	0.23
$-5 \div 0$	32	$-0.99 \div -0.31$	-0.339	0.221	35.36	0.32
$0 \div 5$	48	$-0.31 \div 0.37$	0.144	0.262	41.95	0.87
$5 \div 10$	28	$0.37 \div 1.05$	0.353	0.209	33.42	0.88
$10 \div 15$	20	$1.05 \div 1.73$	0.458	0.105	16.81	0.61
$15 \div 20$	6	$1.73 \div \infty$	0.500	0.042	6.69	0.07
$\sum$	160			1.000	160	3.63

Решение. Будем считать, что математическое ожидание теоретического нормального закона равно выборочному среднему $\bar{x}$, а дисперсия — выборочной дисперсии s^2. Находим их по формулам (x_i^* — середина i-го интервала)

$$m \approx \bar{x} = \sum_{i=1}^{7} x_i^* p_i^* = 2.265, \qquad \sum_{i=1}^{7} x_i^{*2} p_i^* = 59.45, \qquad p_i^* = \frac{\nu_i}{n},$$

$$\sigma^2 \approx s^2 = \sum_{i=1}^{7} x_i^{*2} p_i^* - \bar{x}^2 = 54.32, \qquad \sigma = 7.37.$$

Зная параметры нормального закона $m = 2.265$, $\sigma = 7.37$, находим вероятности попадания в интервал $e_i \div e_{i+1}$ по формуле

$$p_i = \Phi(l_{i+1}) - \Phi(l_i), \quad l_i = \frac{e_i - m}{\sigma},$$

где $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int\limits_0^x e^{-t^2/2} dt$ — функция Лапласа, причем вместо начала первого интервала l_1 возьмем $-\infty$, а вместо конца последнего интервала l_8 возьмем ∞. Вычисления сводим в табл. 3.1.

Итак, $\chi^2 = 3.63$. Определим число степеней свободы. Число интервалов $k = 7$, а число всех связей три, так как по экспериментальным данным мы определили два параметра (m и σ) плюс еще одно обязательное условие

$$\sum_{i=1}^{7} p_i^* = 1,$$

поэтому $r = 7 - 3 = 4$. По таблице для $r = 4$ при $\alpha = 0.05$ получаем $\chi^2_{0.05,\,4} = 9.5$. Так как полученное $\chi^2 < \chi^2_{0.05,\,4}$, действительно $3.63 < 9.5$, то мы принимаем проверяемую гипотезу, т. е. считаем, что исходные данные примера подчиняются нормальному закону.

Можно найти α_p, которое отвечает полученному $\chi^2 = 3.63$, оно равно $\alpha_p = 0.46$. Так как $\alpha_p > \alpha$, то проверяемую гипотезу принимаем, и ясно, что согласование нормального закона с результатами наблюдений хорошее.

Большим преимуществом рассматриваемого метода является то, что одни и те же табличные значения используются при любом n и любых вероятностях p_i. Единственной переменной является число степеней свободы r. При этом следует отметить, что приведенные в таблице значения не являются абсолютно точными во всех случаях: это приближенные значения, справедливые лишь при достаточно больших значениях n. Достаточно большими можно считать такие значения n, для которых любое из np_i не меньше 5; однако, чтобы повысить надежность критерия, лучше брать np_i значительно большими. Если же n заранее ограничено, то нельзя

выбирать k слишком большим, так как тогда будут малыми величины np_i и неустойчивыми значения χ^2.

б) Критерий согласия Колмогорова — Смирнова

Пусть x_1, x_2, ..., x_n — выборка. Построим эмпирическую функцию распределения $F_n^*(x) = \nu_n(x)/n$, где $\nu_n(x)$ равно количеству элементов выборки, меньших чем x. Проверим гипотезу H_0 о том, что выборка представляет собой наблюдения случайной величины X с непрерывной функцией распределения $F(x)$, которая не содержит неизвестных параметров. Для проверки гипотезы воспользуемся *статистикой Колмогорова*

$$K_n = \sqrt{n} \sup_{-\infty<x<\infty} |F_n^*(x) - F(x)|. \tag{3.11}$$

Можно показать, что для любой непрерывной функции распределения $F(x)$

$$\lim_{n\to\infty} \mathbf{P}(K_n < x) = K(x), \quad \text{где} \quad K(x) = \sum_{k=-\infty}^{\infty} (-1)^k e^{-2x^2k^2}, \quad x \geq 0.$$

Имеются таблицы значений функции $K(x)$.

Правило проверки гипотезы следующее. Вычисляем K_n по формуле (3.11). Задаемся уровнем значимости α и по таблицам распределения $K(x)$ находим λ_α, такое, что $K(\lambda_\alpha) = 1 - \alpha$. Если найденная величина K_n такова, что $K_n \leq \lambda_\alpha$, то расхождение между $F_n^*(x)$ и $F(x)$ признается обусловленным случайностью наблюдений, и поэтому гипотеза H_0, состоящая в том, что закон распределения наблюдений есть $F(x)$, признается согласованным с экспериментом. В противном случае гипотеза H_0 отвергается.

Английские аналоги русских терминов в оценивании и проверке гипотез

оценка — estimate, estimator,
смещение — bias,
доверительный интервал — confidence interval,
уровень доверия — confidence level,
уровень значимости — significant level,
квадратичный риск — mean — squared error,
контрольная карта — control chart,

гипотеза — hypothesis,
проверка гипотез — hypothesis testing,
нулевая гипотеза — null hypothesis,
альтернатива — alternative hypothesis,
ошибка первого рода — type I error,
ошибка второго рода — type II error,
односторонний критерий — one-tailed test,
двусторонний критерий — two-tailed test,
критическое значение — critical value,
p-значение — probability value (p-value).

Глава 4

ДВУМЕРНОЕ НОРМАЛЬНОЕ РАСПРЕДЕЛЕНИЕ

1. Свойства двумерного нормального распределения

Статистики работают довольно часто не только с одномерными случайными величинами, но и с многомерными. Однако при работе с многомерными данными выясняется, что их можно реально описывать довольно небольшим классом вероятностных распределений таким образом, чтобы параметры распределения оценивались довольно точно.

Многомерное нормальное распределение (МНР) наиболее часто используется для описания многомерных случайных величин. Дело в том, что в многомерном случае для описания плотности распределения нужно гораздо больше параметров, а, следовательно, для оценки этих параметров и гораздо больше наблюдений. Параметры МНР являются в некотором смысле минимальным набором параметров, описывающим основные характеристики случайных величин (положение — математические ожидания, разброс — дисперсии, зависимость — корреляции).

Многомерное нормальное распределение обладает и другим удобным свойством. Любое линейное преобразование $A\bar{\xi}$ (A — матрица) нормального вектора $\bar{\xi}$ оставляет распределение вектора $A\bar{\xi}$ нормальным.

Непосредственное вхождение корреляции в параметры многомерного нормального распределения делает его незаменимым аппаратом при исследовании и описании характера зависимости случайных величин.

Мы познакомимся с двумерным нормальным распределением, поскольку многомерное принципиально не отличается от двумерного.

Приведем одну из процедур графического анализа двумерных наблюдений

$$(x_1, y_1),\ (x_2, y_2), \ldots, (x_n, y_n).$$

Их координаты наносятся на бумагу. После этого оценивается и наносится на бумагу точка параметра положения. Затем на бумаге по мере уменьшения концентрации наблюдений рисуются линии уровня, внутри которых лежит заданный процент наблюдений. Отметим, что чаще всего линии уровня похожи на эллипсы. Именно такие линии уровня у графика плотности двумерного нормального распределения.

Начнем с анализа простейших двумерных нормальных распределений с нулевым математическим ожиданием и единичной ковариационной матрицей.

Пусть ξ_1, ξ_2 — независимые случайные величины, имеющие нормальное распределение, $\mathbf{E}\xi_1 = \mathbf{E}\xi_2 = 0, \mathbf{E}\xi_1^2 = \mathbf{E}\xi_2^2 = 1$. Тогда говорят, что вектор $\bar{X} = (\xi_1, \xi_2)$ имеет двумерное нормальное распределение с нулевым математическим ожиданием и единичной ковариационной матрицей

$$R = \begin{pmatrix} \mathbf{E}\xi_1^2 & \mathbf{E}(\xi_1\xi_2) \\ \mathbf{E}(\xi_1\xi_2) & \mathbf{E}\xi_2^2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

Его функция распределения имеет вид

$$P(\xi_1 < x, \xi_2 < y) = P(\xi_1 < x)P(\xi_2 < x) = \Phi(x)\Phi(y)$$

и, следовательно, плотность распределения есть произведение плотностей

$$\phi_1(x, y) = \frac{\partial^2(\Phi(x)\Phi(y))}{\partial x \partial y} = \frac{d\Phi(x)}{dx}\frac{d\Phi(y)}{dy} = \frac{1}{2\pi}\exp\left\{-\frac{x^2}{2} - \frac{y^2}{2}\right\}.$$

Таким образом, если нарисовать график плотности $z = \phi_1(x, y)$, то мы будем иметь тело вращения графика $z = \frac{1}{2\pi}\exp\left\{-\frac{x^2}{2}\right\}$ относительно оси z. Действительно, в плоскости zx, или, что то же самое, $y = 0$, мы имеем график $z = \frac{1}{2\pi}\exp\left\{-\frac{x^2}{2}\right\}$ (плотность нормального

распределения $\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{x^2}{2}\right\}$, умноженная на постоянную $\frac{1}{\sqrt{2\pi}}$). В сечении $z = \phi_1(x,y) = \text{const}$ имеем $\frac{1}{2\pi}\exp\left\{-\frac{x^2}{2}-\frac{y^2}{2}\right\} = \text{const}$, т. е. $-\frac{x^2}{2}-\frac{y^2}{2} = \text{const}$ или $x^2+y^2 = \text{const}$, что является уравнением окружности. Отсюда и следует, что график является телом вращения относительно оси z.

Изучим теперь как выглядит график плотности двумерного нормального распределения с диагональной ковариационной матрицей.

Пусть ξ_1, ξ_2 — независимые случайные величины, имеющие нормальное распределение, но разные дисперсии:

$$\mathbf{E}\xi_1 = \mathbf{E}\xi_2 = 0, \qquad \mathbf{E}\xi_1^2 = \sigma_1^2, \ \ \mathbf{E}\xi_2^2 = \sigma_2^2.$$

Тогда говорят, что вектор $\bar{X} = (\xi_1, \xi_2)$ имеет двумерное нормальное распределение с нулевым математическим ожиданием и диагональной ковариационной матрицей

$$R = \begin{pmatrix} \mathbf{E}\xi_1^2 & \mathbf{E}(\xi_1\xi_2) \\ \mathbf{E}(\xi_1\xi_2) & \mathbf{E}\xi_2^2 \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix}.$$

Поскольку плотности ξ_1 и ξ_2 равны соответственно

$$\frac{1}{\sqrt{2\pi}\sigma_1}\exp\left\{-\frac{x^2}{2\sigma_1^2}\right\} \quad \text{и} \quad \frac{1}{\sqrt{2\pi}\sigma_2}\exp\left\{-\frac{y^2}{2\sigma_2^2}\right\},$$

то плотность распределения вектора (ξ_1, ξ_2) равна произведению их плотностей:

$$\begin{aligned} \phi_2(X) = \phi_2(x,y) &= \frac{1}{2\pi\sigma_1\sigma_2}\exp\left\{-\frac{x^2}{2\sigma_1^2}-\frac{y^2}{2\sigma_2^2}\right\} = \\ &= \frac{1}{2\pi(\det R)^{1/2}}\exp\left\{-\frac{1}{2}X^T R^{-1} X\right\}, \end{aligned}$$

где X — двумерный вектор с координатами x, y и X^T — транспонированный вектор X.

Последнее выражение есть стандартная запись плотности двумерного нормального распределения в матричной форме. Как легко видеть, $\det R = \sigma_1^2\sigma_2^2$.

В любом сечении графика плотности $z = \phi_2(x, y)$, перпендикулярном плоскости xy, мы будем опять получать плотность нормального распределения (с точностью до множителя), а в сечениях $z = \phi_2(x, y) = \text{const}$ мы будем получать эллипсы $\dfrac{x^2}{\sigma_1^2} + \dfrac{y^2}{\sigma_2^2} = \text{const}$ (эллипсы получаются из окружности $x^2 + y^2 = \text{const}$ изменением масштаба в σ_1 раз вдоль оси x и σ_2 раз вдоль оси y). Оси эллипсов направлены вдоль осей x и y системы координат.

В общем случае двумерное нормальное распределение $X = (\xi_1, \xi_2)$ задается вектором $d = (d_1, d_2)$ математических ожиданий $\mathbf{E}\xi_1 = d_1, \mathbf{E}\xi_2 = d_2$ и ковариационной матрицей

$$R = \begin{pmatrix} \mathbf{E}(\xi_1 - d_1)^2 & \mathbf{E}(\xi_1 - d_1)(\xi_2 - d_2) \\ \mathbf{E}(\xi_1 - d_1)(\xi_2 - d_2) & \mathbf{E}(\xi_2 - d_2)^2 \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & r_{12} \\ r_{12} & \sigma_2^2 \end{pmatrix},$$

или $X \in N(d, R)$. Приведем сначала вид его плотности, когда математические ожидания равны нулю:

$$\phi_3(X) = \phi_3(x, y) = \frac{1}{2\pi(\det R)^{1/2}} \exp\left\{-\frac{1}{2} X^T R^{-1} X\right\} =$$

$$= \frac{1}{2\pi(\det R)^{1/2}} \exp\left\{-\frac{1}{2} a_{11} x^2 - a_{12} xy - \frac{1}{2} a_{22} y^2\right\},$$

где a_{11}, a_{12}, a_{22} — коэффициенты матрицы R^{-1}:

$$R^{-1} = \begin{pmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{pmatrix}.$$

График плотности $z = \phi_3(x, y)$ можно получить из графика плотности $z = \phi_2(x, y)$ просто поворотом осей координат x и y. Таким образом, в сечениях графика $z = \phi_4(x, y)$ плоскостями перпендикулярными плоскости xy (что то же самое, проходящими через ось z), мы имеем графики плотности нормального распределения (с точностью до постоянного множителя), а в сечениях $z = \phi_3(x, y) = \text{const}$ – эллипсы $a_{11}x^2 + 2a_{12}xy + a_{22}y^2 = \text{const}$. Поворот осей данных эллипсов относительно осей координат x, y, на основе которого плотность $z = \phi_3(x, y)$ может приобрести вид $z = \phi_2(x, y)$, описывает, по существу, меру зависимости случайных величин. За нее отвечает коэффициент корреляции

$$\rho = \rho_{\xi_1 \xi_2} = \frac{\text{Cov}(\xi_1, \xi_2)}{\sqrt{D(\xi_1) D(\xi_2)}} = \frac{\text{Cov}(\xi_1, \xi_2)}{\sigma_1 \sigma_2}.$$

В случае ненулевых математических ожиданий плотность распределения равна

$$\phi_4(X) = \phi_4(x,y) = \frac{1}{2\pi(\det R)^{1/2}} \exp\left\{-\frac{1}{2}(X-d)^T R^{-1}(X-d)\right\} =$$

$$= \frac{1}{2\pi(\det R)^{1/2}} \times$$

$$\times \exp\left\{-\frac{1}{2}a_{11}(x-d_1)^2 + a_{12}(x-d_1)(y-d_2) + \frac{1}{2}a_{22}(y-d_2)^2]\right\},$$

где вектор d имеет координаты d_1, d_2. Ее график получается из графика $z = \phi_3(x,y)$ сдвигом в плоскости x, y на вектор d. Так как $r_{12} = \rho\sigma_1\sigma_2$, то

$$R = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix},$$

$\det R = (1-\rho^2)\sigma_1^2\sigma_2^2$, обратная матрица R^{-1} имеет элементы

$$a_{11} = \frac{1}{(1-\rho^2)\sigma_1^2}, \quad a_{12} = -\frac{\rho}{(1-\rho^2)\sigma_1\sigma_2}, \quad a_{22} = \frac{1}{(1-\rho^2)\sigma_2^2},$$

поэтому плотность $\phi_4(X)$ можно записать также в виде

$$\phi_4(X) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \times$$

$$\times \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{(x-d_1)^2}{\sigma_1^2} - \frac{2\rho(x-d_1)(y-d_2)}{\sigma_1\sigma_2} + \frac{(y-d_2)^2}{\sigma_2^2}\right]\right\}.$$

Зависимость между компонентами ξ_1 и ξ_2 двумерного нормального распределения имеет простой и естественный вид, который сформулируем в виде теоремы.

Теорема. *Пусть (ξ_1, ξ_2) — двумерный нормальный случайный вектор. Тогда случайная величина ξ_2 допускает представление*

$$\xi_2 = \rho_{\xi_1\xi_2}\frac{\sigma_2}{\sigma_1}\xi_1 + \zeta,$$

где нормальная случайная величина ζ не зависит от вектора ξ_1.

Доказательство. Как мы видим из задания плотности ϕ_2, нормальные случайные величины ξ, η независимы, если $\mathrm{Cov}(\xi, \eta) = 0$, так как тогда ковариационная матрица диагональна. Таким образом, достаточно показать, что $\mathrm{Cov}(\xi_1, \zeta) = 0$. Докажем это в предположении, что $\mathbf{E}\xi_1 = \mathbf{E}\xi_2 = 0$, а следовательно, и $\mathbf{E}\zeta = 0$. Имеем

$$\mathrm{Cov}(\xi_1, \zeta) = \mathbf{E}(\xi_1\zeta) = \mathbf{E}\xi_1 \left(\xi_2 - \rho_{\xi_1\xi_2}\frac{\sigma_2}{\sigma_1}\xi_1\right) =$$

$$= \mathbf{E}\xi_1\xi_2 - \rho_{\xi_1\xi_2}\frac{\sigma_2}{\sigma_1}\mathbf{E}\xi_1^2 == \sigma_1\sigma_2\rho_{\xi_1\xi_2} - \rho_{\xi_1\xi_2}\frac{\sigma_2}{\sigma_1}\sigma_1^2 = 0.$$

Отсюда и следует доказываемое утверждение.

Таким образом, вся мера зависимости ξ_2 от ξ_1 сосредоточена в коэффициенте корреляции $\rho_{\xi_1\xi_2}$ и эта зависимость имеет линейный вид. Аналогичный вид имеет и зависимость ξ_1 от ξ_2.

Из теоремы вытекает следующее *следствие*:

$$\mathbf{E}(\xi_2|\xi_1 = x) = \rho_{\xi_1\xi_2}\frac{\sigma_2}{\sigma_1}x + \mathrm{const}.$$

Слева стоит условное математическое ожидание ξ_2 при условии, что $\xi_1 = x$. По существу это математическое ожидание случайной величины ξ_2, если нам известно, что случайная величина $\xi_1 = x$. Эта формула допускает также следующие наглядные интерпретации: "среднее значение ξ_2 при условии, что известно, что $\xi_1 = x$" и "средний прогноз значения ξ_2 при условии, что известно, что $\xi_1 = x$".

2. Построение доверительного множества для математического ожидания

Пусть нам дана выборка $(x_1, y_1), \ldots, (x_n, y_n)$ независимых двумерных случайных величин, имеющих нормальное распределение $N(\theta, R)$ (здесь $\theta = (\theta_1, \theta_2)$ — вектор).

В качестве оценки θ естественно взять вектор $\bar{Z} = (\bar{x}, \bar{y})$, где

$$\bar{x} = \sum_{i=1}^{n} x_i, \quad \bar{y} = \sum_{i=1}^{n} y_i.$$

Возникает вопрос, как найти погрешность оценки $\bar{Z}$ и построить доверительное множество, накрывающее истинное значение параметра θ с заданным уровнем доверия $1-\alpha$. Известно, что случайный вектор $\sqrt{n}(\bar{Z}-\theta) = (\sqrt{n}(\bar{x}-\theta_1), \sqrt{n}(\bar{y}-\theta_2))$ распределен по нормальному закону $N(0,R)$ (аналогично одномерному случаю), а случайная величина $n(\bar{Z}-\theta)^T R^{-1}(\bar{Z}-\theta)$ имеет χ^2-распределение с двумя степенями свободы.

Естественно в качестве доверительного множества θ с уровнем доверия $1-\alpha$ взять множество, где появление $\bar{Z}$ в некотором смысле "наиболее вероятно", т. е. такое множество, для которого выполняется условие

$$N(\bar{Z}-\theta)^T R^{-1}(\bar{Z}-\theta) < \chi^2_{\alpha,2},$$

где $\chi^2_{\alpha,2}$ находится из таблиц хи-квадрат распределения с двумя степенями свободы (вероятность попадания $\sqrt{n}(\bar{Z}-\theta)$ в это множество максимальна, среди всех множеств, имеющих заданную площадь).

Отсюда и находим вид доверительного множества вектора θ:

$$\Omega_\alpha = \{\theta^* : n(\bar{Z}-\theta^*)^T R^{-1}(\bar{Z}-\theta^*) < \chi^2_{\alpha,2}.$$

Доверительное множество Ω_α называют доверительным эллипсом с уровнем доверия $1-\alpha$.

Если ковариационная матрица R неизвестна, то ее компоненты в задании доверительного эллипса заменяют их оценки. Оценки дисперсий имеют стандартный вид, а с методами оценивания ковариации мы познакомимся чуть позднее.

3. Проверка гипотезы для математического ожидания

Пусть для предыдущей постановки задачи нам надо проверить гипотезу

$$H_0: \ \theta = \theta_0$$

против альтернатив

$$H_1: \ \theta \neq \theta_0$$

с критерием значимости α.

Для проверки гипотезы используем тот же критерий, что и в одномерном случае: если доверительное множество накрывает

истинное значение параметра, т. е. $n(\bar{Z}-\theta)^T R^{-1}(\bar{Z}-\theta) < \chi^2_{\alpha,2}$, то гипотеза принимается. В противном случае она отвергается.

Таким образом, критерий имеет вид

$$K_n = \chi\left(n(\bar{Z}-\theta)^T R^{-1}(\bar{Z}-\theta) > \chi^2_{\alpha,2}\right),$$

где $\chi^2_{\alpha,2}$ находится из таблиц хи-квадрат распределения с двумя степенями свободы.

Глава 5
КОРРЕЛЯЦИОННЫЙ И РЕГРЕССИОННЫЙ АНАЛИЗ

Корреляционный и регрессионный анализ являются стандартными методами анализа для получения информации, необходимой для принятия решений. В бизнесе менеджер работает с большим числом переменных, и ему необходимо понимать зависимость между ними. Например:

Зависимая переменная — Независимая переменная,

Размер дивидендов — Прибыль компании,

Зарплата работника — Возраст работника,

Стоимость компании — Ее активы,

Стоимость компании — Ее прибыль,

Стоимость квартиры — Площадь квартиры,

Рост стоимости акций — Рост индекса биржи.

Таким образом, значение *зависимой переменной* может быть в определенной мере предсказано по значению другой переменной, называемой *независимой*.

Корреляционный анализ изучает меры зависимости между случайными величинами (такие, как ковариация, корреляция и т. п.).

Регрессионный анализ изучает статистические методы, позволяющие по значению независимой переменной оценить значение (сделать прогноз значения) зависимой переменной.

1. Понятие ковариации и коэффициента корреляции. Их оценивание

Если прежде мы изучали, как извлечь информацию о случайных величинах, то сейчас изучим вопрос, как извлечь информацию о зависимости между ними.

Напомним, что случайные величины X и Y независимы, если

$$\mathbf{P}(X \in A, Y \in B) - \mathbf{P}(X \in A)\mathbf{P}(Y \in B) = 0$$

для любых множеств A и B.

Есть также другое определение независимости

$$\mathbf{E}f(X)g(Y) - \mathbf{E}f(X)\mathbf{E}g(Y) = 0 \tag{5.1}$$

для любых функций f и g.

Понятно, что по конечной выборке случайных векторов

$$(x_1, y_1), \ldots, (x_n, y_n)$$

мы не можем проверить выполнение равенства (5.1) достаточно хорошо для всех функций f и g. Распространенный объем выборки $n = 50-100$ наблюдений.

Поэтому обычно (5.1) проверяют только для одного набора функций $f(X) = X$ и $g(Y) = Y$. Таким образом, мы приходим к понятиям ковариации и корреляции случайных величин.

Ковариацией пары случайных величин (X, Y) называется выражение

$$Cov(X, Y) = \mathbf{E}(X - \mathbf{E}X)(Y - \mathbf{E}Y),$$

которое можно также записать в виде

$$Cov(X, Y) = \mathbf{E}XY - \mathbf{E}X\mathbf{E}Y.$$

Коэффициентом корреляции случайных величин X и Y называется

$$\rho = \rho_{XY} = \frac{Cov(X, Y)}{\sigma_X \sigma_Y},$$

где $\sigma_X^2 = \mathbf{D}X$, $\sigma_Y^2 = \mathbf{D}Y$.

Возникает вопрос — насколько естественно оценивать зависимость случайных величин всего лишь одним набором функций f и

g. Поэтому обсудим свойства ковариации и коэффициента корреляции.

Ковариация инвариантна относительно сдвига, т. е.

$$\mathrm{Cov}(X+c, Y+d) = \mathrm{Cov}(X, Y)$$

для любых c и d, а корреляция инвариантна относительно сдвига и масштаба, т. е.

$$\rho_{XY} = \frac{\mathrm{Cov}(X,Y)}{D^{1/2}(X)D^{1/2}(Y)} = \frac{\mathrm{Cov}(aX+c, bY+d)}{D^{1/2}(aX+c)D^{1/2}(bY+d)}$$

для любых $a \neq 0$, $b \neq 0$, c, d.

Таким образом, по существу с помощью ковариации и коэффициента корреляции мы оцениваем зависимость для несколько более широких классов функций f и g ($f(x) = x + c$, $g(y) = y + d$ и $f(x) = ax + c$, $g(y) = by + d$ для ковариации и корреляции соответственно).

Можно показать, что $-1 \leq \rho_{XY} \leq 1$, причем если $\rho_{XY} = \pm 1$, то случайные величины X и Y линейно зависимы, т. е. $Y = \beta_0 + \beta_1 X$.

Понятие ковариации и корреляции тесно связано с линейной зависимостью случайных величин. Если мы рассмотрим задачу минимизации среднеквадратичных отклонений вида

$$\min_{\beta_0, \beta_1} \mathbf{E}(Y - \beta_0 - \beta_1 X)^2,$$

то получим, что выбором β_0 и β_1 являются

$$\hat{\beta}_1 = \rho_{X,Y}\frac{\sigma_Y}{\sigma_X} = \frac{Cov(X,Y)}{\sigma_X^2}, \tag{5.2}$$

$$\hat{\beta}_0 = \mathbf{E}Y - \hat{\beta}_1 \mathbf{E}X. \tag{5.3}$$

Итак, с помощью коэффициента корреляции ρ_{XY} мы получаем наилучшее линейное приближение $\beta_0 + \beta_1 X$ в смысле среднеквадратичного отклонения.

Насколько естественна такая постановка задачи? Рассмотрим более простую постановку

$$\min_{\beta_0} \mathbf{E}(Y - \beta_0)^2 = \min_{\beta_0}(\mathbf{E}Y^2 - 2\beta_0 \mathbf{E}Y + \beta_0^2).$$

Дифференцируя по β_0 и приравнивая производную к нулю, получаем

$$\frac{d}{d\beta_0}(\mathbf{E}Y^2 - 2\beta_0\mathbf{E}Y + \beta_0^2) = -2\mathbf{E}Y + 2\beta_0 = 0,$$

т. е. $\beta_0 = EY$. Следовательно,

$$\min_{\beta_0} \mathbf{E}(Y - \beta_0)^2 = \mathbf{E}(Y - \mathbf{E}Y)^2.$$

Ясно, что

$$\min_{\beta_0} \mathbf{E}(Y - \beta_0)^2 \geq \min_{\beta_0,\beta_1} \mathbf{E}(Y - \beta_0 - \beta_1 X)^2.$$

Таким образом, с помощью коэффициента корреляции мы получаем следующий более точный прогноз $\beta_1 X + \beta_0$ случайной величины Y после более грубого прогноза, равного EY.

Другая интерпретация тесно связана с зависимостью нормальных случайных величин. Если (X, Y) — двумерный нормальный случайный вектор, то Y допускает представление

$$Y = \rho_{XY}\frac{\sigma_Y}{\sigma_X}X + \xi,$$

где ξ — нормальная случайная величина, не зависящая от X, и

$$\mathbf{E}\xi = \mathbf{E}Y - \rho_{XY}\frac{\sigma_Y}{\sigma_X}\mathbf{E}X, \quad \mathbf{D}\xi = \sigma_Y^2(1 - \rho_{XY}^2).$$

Таким образом, в случае нормального случайного вектора (X, Y) коэффициент корреляции несет всю информацию о зависимости случайных величин X и Y.

Как оценить значение ковариации и коэффициента корреляции по выборке пар случайных величин

$$(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)\,?$$

Очень просто. Мы приписываем каждой паре случайных величин $(X_i, Y_i), 1 \leq i \leq n$ вероятность $1/n$ и находим соответствующие значения ковариации и коэффициента корреляции.

Итак, оценками ковариации и коэффициента корреляции являются соответственно

$$s_{XY} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) = \frac{1}{n}\sum_{i=1}^{n} X_iY_i - \bar{X}\bar{Y}, \tag{5.4}$$

$$r = r_{XY} = \frac{s_{XY}}{s_X s_Y}, \tag{5.5}$$

где

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i,$$

$$s_X^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2, \quad s_Y^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2. \tag{5.6}$$

2. Проверка гипотезы об отсутствии корреляционной связи

Если случайные величины X, Y имеют нормальное распределение и их коэффициент корреляции равен нулю, то они независимы. В общем случае произвольных случайных величин X, Y равенство коэффициента корреляции нулю является одним из признаков их независимости.

Пусть имеются совместные наблюдения двух случайных величин Y и X

$$(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, X_n), \tag{5.7}$$

о которых мы предполагаем известным, что они имеют нормальное распределение.

Нужно проверить гипотезу об отсутствии корреляционной связи, иначе говоря: H_0: $\rho_{XY} = 0$ при альтернативе H_1: $\rho_{XY} \neq 0$.

Можно показать, что если гипотеза H_0 верна, то случайная величина

$$t = r\sqrt{n-2}/\sqrt{1-r^2},$$

в которой r — выборочный коэффициент корреляции (5.5), имеет распределение Стьюдента с $n-2$ степенями свободы и может служить тестовой статистикой при проверке гипотезы H_0. Нужно выбрать уровень значимости α, найти $t_{q,n-2}$ по таблицам распределения Стьюдента с $n-2$ степенями свободы. Если

$$|t| > t_{q,n-2}, \tag{5.8}$$

значит мы попадаем в критическую область и должны отвергнуть гипотезу H_0 и, следовательно, считаем, что имеется корреляционная связь между случайными величинами Y и X. Если $|t| < t_{q,n-2}$,

то говорим, что данные не противоречат гиипотезе H_0, т. е. корреляционная связь между Y и X отсутствует.

Неравенство (5.8) эквивалентно следующему:

$$|r| > \frac{t_{q,n-2}}{\sqrt{t^2_{q,n-2} + n - 2}}, \tag{5.9}$$

и если оно выполнено, то следует отвергнуть гипотезу H_0.

3. Общая постановка задачи линейного регрессионного анализа

Пару "зависимая переменная Y и независимая переменная X" будем рассматривать как двумерный случайный вектор (Y, X). Цель регрессионного анализа по наблюдениям $(Y_1, X_1), \ldots, (Y_n, X_n)$ случайного вектора (Y, X) — оценить значение Y, если известно, что $X = x$. Случайная величина Y при фиксированном $X = x$ является одномерной случайной величиной, и мы можем определить математическое ожидание Y при условии, что известно $X = x$. Такое математическое ожидание называется в теории вероятностей условным математическим ожиданием Y при условии $X = x$ и обозначается $E(Y|X = x)$.

Ясно, что

$$\mathbf{E}(Y|X = x) = f(x)$$

является некоторой функцией, зависящей от x. Эта функция называется функцией регрессии. Именно оцениванием функции регрессии как прогноза значения Y по значению $X = x$ занимается регрессионный анализ.

Очевидно, что если $Y = f(X)$, то $\mathbf{E}(f(X)|X = x) = f(x)$, а если Y и X независимы, то $\mathbf{E}(Y|X = x) = \mathbf{E}Y$.

Если

$$Y = f(X) + \xi, \tag{5.10}$$

где ξ — случайная величина, не зависящая от X и $\mathbf{E}\xi = 0$, то

$$\mathbf{E}(Y|X = x) = f(x) + \mathbf{E}(\xi|X = x) = f(x) + \mathbf{E}\xi = f(x).$$

Понятие *регрессия* было введено английским ученым Гальтоном. В 1875 г. он посадил горошины различных размеров и обнаружил,

что из более крупных горошин выросли более мелкие, а из более мелких — более крупные. Это явление Гальтон назвал регрессией.

Как строго определить условное математическое ожидание $E(Y|X=x)$? Предположим, что у случайного вектора (Y,X) существует плотность $p(y,x)$. Тогда

$$\mathbf{E}(Y|X=x)=\int y\,p(y|X=x)\,dy,$$

где $p(y|X=x)$ — условная плотность распределения Y при условии $X=x$.

Условная функция распределения Y при условии $X=x$ определяется следующим образом:

$$F(y|X=x)=\mathbf{P}(Y<y|X=x),$$

но, так как вероятность условия $P(X=x)$ равна нулю, мы рассматриваем $F(y|X=x)$ как предел отношения вероятностей

$$\frac{\mathbf{P}(Y<y,x<X<x+\delta x)}{\mathbf{P}(x<X<x+\delta x)}$$

для малых окрестностей $(x,x+\delta x)$ точки x. Находим

$$F(y|X=x)=\lim_{\delta x\to 0}\frac{\mathbf{P}(Y<y,x<X<x+\delta x)}{\mathbf{P}(x<X<x+\delta x)}=$$

$$=\lim_{\delta x\to 0}\frac{\dfrac{1}{\delta x}\displaystyle\int_{-\infty}^{y}ds\int_{x}^{x+\delta x}p(s,t)dt}{\dfrac{1}{\delta x}\displaystyle\int_{-\infty}^{\infty}ds\int_{x}^{x+\delta x}p(s,t)dt}=\frac{\displaystyle\int_{-\infty}^{y}p(s,x)ds}{\displaystyle\int_{-\infty}^{\infty}p(s,x)ds}.$$

Условная плотность распределения равна

$$p(y|X=x)=\frac{\partial F(y|X=x)}{\partial y}=\frac{p(y,x)}{\displaystyle\int_{-\infty}^{\infty}p(s,x)ds}.$$

4. Простая линейная регрессия

Рассмотрим зависимость типа (5.10), где

$$f(x) = \beta_0 + \beta_1 x \tag{5.11}$$

и β_0, β_1 — оцениваемые постоянные. Тогда наблюдения Y_i, $1 \le i \le n$, при заданных X_i, $1 \le i \le n$, представляются в виде

$$Y_i = \beta_0 + \beta_1 X_i + \xi_i, \tag{5.12}$$

а ξ_i — ошибки измерений, предполагаются независимыми одинаково распределенными случайными величинами, причем

$$\mathbf{E}\xi_i = 0, \quad \mathbf{D}\xi_i = \sigma^2, \tag{5.13}$$

и, как правило, предполагается, что $\xi_i \in N(0, \sigma^2)$. Число β_1 называется *коэффициентом регрессии* или *наклоном.*

В реальных экономических задачах предположения о распределении случайных величин ξ_i часто нарушаются. Они бывают зависимыми, имеют разные дисперсии, не только не подчиняются нормальному закону, но и имеют даже асимметричную плотность распределения. Учет и анализ всех этих факторов представляет большую трудность в статистических исследованиях.

Величины X_i обычно являются случайными (иногда бывают детерминированными). Распространенное предположение о нормальности X_i вместе с предположением о нормальности ξ_i означает, что случайные величины Y_i тоже распределены по нормальному закону. Более того, (Y_i, X_i) имеют двумерное нормальное распределение, удовлетворяют соотношению (5.12), в котором

$$\beta_1 = \frac{r_{xy}\sigma_y}{\sigma_x}, \quad \beta_0 = EY - \beta_1 EX, \quad \mathbf{E}\xi = 0, \quad \mathbf{D}\xi = (1 - r_{xy}^2)\sigma_y^2.$$

Отметим, что модель простой линейной регрессии является простейшей. В практических приложениях часто встречается многомерная линейная регрессия (стоимость акции зависит от активов, прибыли, оборота компании и т. д.). В этом случае мы наблюдаем величины

$$Y_i, \;\; X_{1i}, \ldots, X_{ki}, \quad i = 1, \ldots, n,$$

связанные линейной зависимостью

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + \xi_i, \tag{5.14}$$

где ошибки ξ_i распределены так же, как в формулах (5.13).

Часто встречаются и модели нелинейной регрессии, некоторое представление о которых мы получим позднее.

5. Оценка параметров линейной регрессии по методу наименьших квадратов

Предположим, что наблюдения (Y_i, X_i), $1 \le i \le n$, удовлетворяют модели простой линейной регрессии (5.12) и нам надо оценить коэффициенты β_0 и β_1.

Оценками параметров β_0 и β_1, найденных по методу наименьших квадратов, являются величины $\hat{\beta}_0$ и $\hat{\beta}_1$, которые минимизируют квадратичные отклонения наблюдений Y_i от линии регрессии (5.11). Иначе говоря, наилучшие оценки $\hat{\beta}_0$ и $\hat{\beta}_1$ — те, при которых сумма квадратичных отклонений

$$Q = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

достигает минимума.

Взяв производные от Q по β_0 и β_1 и приравняв их к нулю, получим систему двух линейных уравнений относительно $\hat{\beta}_0$ и $\hat{\beta}_1$. Решив их, находим оценки

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} Y_i X_i - n\bar{Y}\bar{X}}{\sum_{i=1}^{n} (X_i - \bar{X})^2}, \tag{5.15}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}. \tag{5.16}$$

Уравнение регрессии получаем в виде

$$\hat{Y}_x = \hat{\beta}_0 + \hat{\beta}_1 x. \tag{5.17}$$

Кроме параметров β_1 и β_0, следует найти оценку $\hat{\sigma}^2$ параметра σ^2 — дисперсии ошибок ξ_i. Она имеет вид

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2. \tag{5.18}$$

Величина $\hat{\sigma}$ называется оценкой стандартной ошибки σ и рассматривается как мера отклонения наблюдений от линии регрессии. Величины $y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$ называются остатками.

6. Коэффициент простой детерминации

Часто возникает вопрос, насколько существенна информация, получаемая за счет введения модели линейной регрессии. В этой ситуации естественно сравнить сумму квадратичных отклонений (sums of squares due to errors), вызванных ошибками прогноза

$$SSE = \min_{\beta_0,\beta_1} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

в модели линейной регрессии с суммой квадратов отклонений наблюдений Y_i относительно некоторой точки. При этом в качестве центра, относительно которого измеряется квадратичный разброс Y_i, возьмем $\bar{Y}$, поскольку

$$ns_y^2 = SST = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 = \min_{\beta} \sum_{i=1}^{n} (Y_i - \beta)^2. \tag{5.19}$$

Назовем SST общей суммой квадратов (total sums of squares).

Отношение разности $SST - SSE$ сумм квадратичных разбросов этих моделей к общему разбросу SST называется *коэффициентом простой детерминации.* Он обозначается R^2 и равен

$$R^2 = \frac{SST - SSE}{SST} = 1 - \frac{\sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}. \tag{5.20}$$

Коэффициент простой детерминации показывает ту меру погрешности, какую удается отсечь за счет использования в качестве прогноза функции регрессии $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$ по сравнению с прогнозом вида $\hat{Y} = \hat{\beta}_0 = \bar{Y}$.

Используемое обозначение R^2 — не случайное совпадение, так как правая часть (5.20) равна квадрату выборочного коэффициента корреляции (см. п. 9.2)

$$R^2 = r^2 = r_{xy}^2. \tag{5.21}$$

Величину R^2 можно вычислить и по другой формуле:

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}, \tag{5.22}$$

так как

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2. \tag{5.23}$$

Доказательство равенств (5.21) и (5.23) см. в п. 9.

Таким образом, общая сумма квадратов SST равна "необъясненной" сумме квадратов ошибок SSE плюс сумма квадратов SSR ошибок (sums of squares due to regression), "объясненных" за счет введения функции регрессии:

$$SSR = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2.$$

Заметим, что тождество (5.23) можно записать в виде

$$SST = SSE + SSR.$$

"Объясненная" сумма SSR представляет собой сумму квадратов отклонений значений $\hat{y}_i$ линии регрессии в точках наблюдений x_i от выборочного среднего $\bar{y}$. Именно она показывает, какое количество квадратичного разброса s_Y^2 наблюдений y_i от $\bar{y}$ мы теряем за счет введения их аппроксимации линией регрессии. После нормировки этой информации делением на $SST = ns_Y^2$ мы получаем коэффициент детерминации R^2.

Если $R^2 = 1$, то точки (X_i, Y_i) лежат на прямой $Y = \hat{\beta}_0 + \hat{\beta}_1 X$.

Наряду с коэффициентом детерминации R^2 в статистических исследованиях используют отрегулированный (adjusted) коэффициент детерминации $\bar{R}^2$. Используя тот факт, что дробь

$$\frac{SSE/(n-2)}{SST/(n-1)}$$

имеет распределение Фишера с $n-2$ и $n-1$ степенями свободы, полагают

$$\bar{R}^2 = 1 - \frac{SSE/(n-2)}{SST/(n-1)}.$$

Таким образом, распределение $\bar{R}^2$ легко получается из стандартного распределения Фишера и может быть легко изучено.

7. Прогноз значения Y_{x_0} в точке x_0

Задачи прогноза возникают в бизнесе и финансах довольно часто. Например, известны доходы компании за пять лет, и надо дать прогноз на следующий год или необходимо дать прогноз роста акций на бирже. Естественно, что эти задачи помимо факторов, описываемых математически, содержат целый ряд и других факторов.

Линия регрессии (5.17) может служить прогнозом в любой точке x_0, и прогноз будет иметь вид

$$\hat{Y}_{x_0} = \hat{\beta}_0 + \hat{\beta}_1 x_0.$$

Доверительный интервал для будущего наблюдения Y_{x_0} в точке x_0 вычисляется по формуле

$$\hat{\beta}_0 + \hat{\beta}_1 x_0 \pm \frac{t_{\alpha,n-2}\hat{\sigma}}{\sqrt{n-2}}\sqrt{1 + \frac{n(x_0 - \bar{x})^2}{\sum_{i=1}^n x_i^2 - n\bar{x}^2}}, \tag{5.24}$$

где $t_{\alpha,n-2}$ определяется с помощью распределения Стьюдента с $n-2$ степенями свободы и уровнем значимости α.

Пример 5.1. Пусть за 5 лет фирма имела следующий доход:

$X_i = i$	$\ldots$	1	2	3	4	5
Y_i	$\ldots$	2.1	2.7	3.5	4.1	4.9

По формулам (5.15) и (5.16) получаем $\hat{\beta}_0 = 1.36$, $\hat{\beta}_1 = 0.7$, и прямая регрессии (5.17), которая может служить прогнозом в точке x, приобретает вид $\hat{Y}_x = 1.36 + 0.7 \cdot x$. При $x_0 = 6$ величина $\hat{Y}_6$, являющаяся прогнозом в следующем после наблюдений году, равна $\hat{Y}_6 = 5.56$.

Далее определяем число степеней свободы $n - 2 = 3$ и при $\alpha = 0.05$ по таблицам Стьюдента находим $t_{0.05,3} = 3.18$, затем по формуле (5.18) $\hat{\sigma} = 0.05$ и окончательно получаем доверительный интервал для будущего наблюдения в точке $x_0 = 6$ в виде 5.56 ± 0.21.

8. Проверка гипотезы о равенстве нулю коэффициента наклона β_1

Различают две постановки задачи, когда наблюдения x_i не случайны и когда они случайны.

Если X_i не случайны, то равенство нулю β_1 является одним из признаков их одинаковой распределенности. В частности, если переменная X является временной, то это является признаком независимости от времени математического ожидания Y_i.

Если X_i — случайные величины, то, поскольку $\beta_1 = \rho_{XY}\frac{\sigma_Y}{\sigma_X}$, гипотеза $\beta_1 = 0$ эквивалентна гипотезе о некоррелированности случайных величин X и Y, и проверку их некоррелированности можно осуществить с помощью критерия (5.2).

Если x_i не случайны и представляют собой время, то равенство $\beta_1 = 0$ является признаком независимости от времени математического ожидания Y. Уравнение регрессии имеет вид

$$\mathbf{E}(Y/X) = \beta_0.$$

Если в выборке (5.7) переменные X_i — не случайные величины, а Y_i распределены нормально, то полученная в формуле (5.15) оценка $\hat{\beta}_1$ имеет нормальное распределение, причем

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right).$$

Критерием для проверки гипотезы H_0: $\beta_1 = 0$ с уровнем значимости q является неравенство

$$\left|\frac{\hat{\beta}_1}{s_{\hat{\beta}_1}}\right| > t_{q/2,n-2}, \tag{5.25}$$

где выборочное среднеквадратическое отклонение $s_{\hat{\beta}_1}$ статистики $\hat{\beta}_1$ имеет вид

$$s_{\hat{\beta}_1} = \frac{\sqrt{n}\hat{\sigma}}{\sqrt{(n-2)\sum_{i=1}^n (x_i - \bar{x})^2}}, \tag{5.26}$$

а величина $\hat{\sigma}$ находится по формуле (5.18). Значение $t_{q/2,n-2}$ получаем, используя таблицы распределения Стьюдента с $n-2$ степенями свободы.

Если неравенство (5.25) для нашей выборки будет выполнено, то гипотеза отвергается. В противном случае она принимается.

При выполнении гипотезы H_0 уравнение регрессии (5.17) превращается в $\hat{Y}_x = \hat{\beta}_0$, т. е. при любом X величина $\hat{Y}_x$ является постоянной.

9. Доказательство основных формул простого регрессионного анализа

9.1. *Вывод оценок наименьших квадратов для коэффициентов простой линейной регрессии.* Обозначим

$$H(\beta_0, \beta_1) = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2.$$

Тогда $\min H(\beta_0, \beta_1)$ достигается для значений β_0, β_1, для которых

$$\begin{cases} \dfrac{\partial}{\partial \beta_0} H(\beta_0, \beta_1) = 0, \\ \dfrac{\partial}{\partial \beta_1} H(\beta_0, \beta_1) = 0. \end{cases} \tag{5.27}$$

Имеем

$$\frac{\partial}{\partial \beta_0} H(\beta_0, \beta_1) = \sum_{i=1}^{n} \frac{\partial}{\partial \beta_0} (Y_i - \beta_0 - \beta_1 X_i)^2 =$$

$$= -2 \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i) = -2n(\bar{Y} - \beta_0 - \beta_1 \bar{X}) = 0$$

и, следовательно,

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}. \tag{5.28}$$

Используя второе уравнение (5.27) и подставляя в процессе вычислений $\beta_0 = \hat{\beta}_0$, имеем

$$\frac{\partial}{\partial \beta_1} H(\beta_0, \beta_1) = \sum_{i=1}^{n} \frac{\partial}{\partial \beta_1} (Y_i - \beta_0 - \beta_1 X_i)^2 =$$

$$= -2 \sum_{i=1}^{n} X_i (Y_i - \beta_0 - \beta_1 X_i) = -2 \sum_{i=1}^{n} X_i Y_i + 2\beta_0 \sum_{i=1}^{n} X_i + 2\beta_1 \sum_{i=1}^{n} X_i^2 =$$

$$= -2\sum_{i=1}^{n} X_i Y_i + 2n(\bar{Y} - \hat{\beta}_1 \bar{X})\bar{X} + 2\beta_1 \sum_{i=1}^{n} X_i^2 =$$

$$= -2\left(\sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y}\right) + 2\beta_1 \left(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right) = 0. \qquad (5.29)$$

Отсюда получаем оценку

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y}}{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2},$$

которую можно записать также в виде

$$\hat{\beta}_1 = \frac{s_{XY}}{s_X^2} = r_{XY}\frac{s_Y}{s_X},$$

так как

$$s_{XY} = \frac{1}{n}\sum_{i=1}^{n} X_i Y_i - \bar{X}\bar{Y}, \quad s_X^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2.$$

9.2. *Доказательство тождества*

$$SST = SSE + SSR. \qquad (5.30)$$

Имеем

$$SST = \sum_{i=1}^{n}[(Y_i - \hat{Y}_i) + (\hat{Y}_i - \bar{Y})]^2 = SSE + SSR + 2\sum_{i=1}^{n}(Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}).$$

Таким образом, достаточно доказать, что

$$\sum_{i=1}^{n}(Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) = 0. \qquad (5.31)$$

Так как $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$, а $\bar{Y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{X}$, то левая часть (5.31) равна

$$\sum_{i=1}^{n}(Y_i - \hat{Y}_i)(\hat{\beta}_0 + \hat{\beta}_1 X_i - \hat{\beta}_0 - \hat{\beta}_1 \bar{X}) = \hat{\beta}_1 \sum_{i=1}^{n}(Y_i - \hat{Y}_i)(X_i - \bar{X}) =$$

$$= \hat{\beta}_1 \sum_{i=1}^{n} (Y_i - \hat{Y}_i) X_i - \hat{\beta}_1 \sum_{i=1}^{n} (Y_i - \hat{Y}_i) \bar{X}, \tag{5.32}$$

и остается показать, что каждое из слагаемых в правой части (5.32) равно нулю. В силу (5.29)

$$\sum_{i=1}^{n} (Y_i - \hat{Y}_i) X_i = \sum_{i=1}^{n} X_i (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0. \tag{5.33}$$

Осталось проверить, что

$$\bar{X} \sum_{i=1}^{n} (Y_i - \hat{Y}_i) = 0, \tag{5.34}$$

но

$$\sum_{i=1}^{n} \hat{Y}_i = \sum_{i=1}^{n} (\hat{\beta}_0 + \hat{\beta}_1 X_i) = n\hat{\beta}_0 + n\hat{\beta}_1 \bar{X}, \tag{5.35}$$

а из равенства (5.28) следует, что

$$\sum_{i=1}^{n} Y_i = n\bar{Y} = n\hat{\beta}_0 + n\hat{\beta}_1 \bar{X}, \tag{5.36}$$

поэтому из равенств (5.35) и (5.36) получаем (5.34).

Таким образом, из равенств (5.31)–(5.36) следует тождество (5.30).

9.3. *Доказательство равенства* $R^2 = r^2$. Имеем

$$SSR = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 = \sum_{i=1}^{n} (\hat{\beta}_0 + \hat{\beta}_1 X_i - \hat{\beta}_0 - \hat{\beta}_1 \bar{X})^2 =$$

$$= \hat{\beta}_1^2 \sum_{i=1}^{n} (X_i - \bar{X})^2 = n\hat{\beta}_1^2 s_X^2 = nr^2 s_Y^2.$$

Следовательно,

$$R^2 = \frac{SSR}{n s_Y^2} = r^2.$$

Глава 6
НЕПАРАМЕТРИЧЕСКИЕ МЕТОДЫ СТАТИСТИКИ

1. Ранговые критерии для проверки гипотезы однородности

Пусть у нас есть две выборки и мы хотим проверить гипотезу, что случайные величины в них имеют одну и ту же функцию распределения против альтернативы, что "случайные величины в одной выборке обычно принимают большие значения, чем в другой". Такая постановка возникает, например, при проверке гипотезы, что

одно лекарство лучше другого,
один сорт кофе лучше другого,
одна марка автомобиля лучше другой.

Именно для проверки таких гипотез и применяются ранговые критерии, поскольку они естественно приспособлены к данной постановке задачи. Одним из достоинств ранговых критериев является возможность их применения как к количественным, так и к качественным данным. В частности, ранговые критерии находят широкое применение в экспертных оценках. Ранговые критерии считаются непараметрическими, поскольку их использование не предполагает известным функции распределения наблюдений даже с точностью до параметров.

2. Ранговые критерии Уилкоксона и Манна—Уитни

Начнем рассмотрение с конкретной задачи.

Задача 6.1. Владелец компании прохладительных напитков решил улучшить качество продукции. Был изготовлен новый напиток, и пять экспертов (дегустаторов) оценивали качество нового и стандартного напитков в пятибалльной шкале. Стандартный напиток получил следующие баллы:

$$2,\ 5,\ 4,\ 1,\ 4, \tag{6.1}$$

а новый:

$$3,\ 4,\ 2,\ 2,\ 1. \tag{6.2}$$

Как на основе заключений экспертов проверить гипотезу, что новый напиток не лучше стандартного?

Решение. В нашем случае гипотеза H_0 соответствует тому, что новый и старый напитки не различаются между собой. Объединим совокупности баллов (6.1), (6.2) и запишем их в порядке возрастания:

$$1,\ 1,\ 2,\ 2,\ 2,\ 3,\ 4,\ 4,\ 4,\ 5. \tag{6.3}$$

Каждому из чисел в ряду (6.3) сопоставим его ранг, который определим следующим образом. Если в ряду (6.3) данное число, например 5, встречается один раз, то ранг этого числа равен порядковому номеру этого числа в ряду (6.3). Так, цифра 5 в ряду (6.3) стоит на десятом месте, поэтому ее ранг равен 10.

Если в ряду (6.3) какое-либо число встречается несколько раз, то для подсчета его ранга нужно просуммировать порядковые номера этого числа и полученную сумму разделить на их количество. Например, цифра 4 в ряду (6.3) присутствует трижды и стоит на местах $7, 8, 9$, следовательно, ранг четверки равен (7+8+9)/3 = 8, ранг двойки равен (3+4+5)/3 = 4, а ранг единицы равен (1+2)/2=1.5.

Ранги ряда (6.3) будут иметь вид

$$1.5,\ \ 1.5,\ \ 4,\ \ 4,\ \ 4,\ \ 6,\ \ 8,\ \ 8,\ \ 8,\ \ 10. \tag{6.4}$$

Запишем ряды рангов экспертных оценок (6.1) и (6.2):

$$4,\ \ 10,\ \ 8,\ \ 1.5,\ \ 8, \tag{6.5}$$

$$6,\ \ 8,\ \ 4,\ \ 4,\ \ 1.5. \tag{6.6}$$

Мы будем судить о качестве напитков не по баллам экспертов, а по их рангам. Достоинство такого рангового подхода состоит в том, что при изменении шкалы (например, на семибалльную) баллы экспертов для каждого из напитков существенно изменятся, в то время как ранги меняются не столь значительно.

Найдем суммы рангов стандартного и нового напитков:

$$4 + 10 + 6 + 1.5 + 8 = 31.5, \tag{6.7}$$

$$4 + 8 + 8 + 4 + 1.5 = 23.5, \tag{6.8}$$

и будем осуществлять проверку гипотезы по суммам (6.7), (6.8).

Такой подход к проверке гипотез предложил Уилкоксон. Поэтому соответствующий критерий и называется критерием Уилкоксона.

Поскольку сумма рангов нового напитка (23.5) меньше, чем у стандартного (31.5), новый напиток не лучше стандартного.

Отметим, что $31.5 + 23.5 = 55$ и это число равно сумме рангов ряда (6.4), т. е. сумме чисел от 1 до 10. Поэтому для проверки гипотезы в качестве тестовой статистики W достаточно взять только одну из сумм (6.7) или (6.8), как правило берут максимальную. В нашем случае $W = 31.5$.

Статистика W и порождает критерий Уилкоксона. При малых объемах выборки $n < 25$ и уровне значимости α $(0 < \alpha < 0.5)$ для критерия Уилкоксона верхнее критическое значение W_α, где

$$\mathbf{P}\{W \geq W_\alpha\} \leq \alpha, \tag{6.9}$$

находится из таблиц распределения Уилкоксона.

Возьмем $\alpha = 0.1$. Тогда $W_{0.1} = 37$. Поскольку $W = 31.5 < W_{0.1} = 37$, то мы принимаем гипотезу H_0. Итак, как уже отмечалось, новый напиток не лучше старого, но и не существенно хуже его.

В рассмотренной задаче мы столкнулись с новой шкалой измерений данных. В ней мы не можем дать точную числовую характеристику объекту, а можем только сказать — больше или меньше характеристика данного объекта, чем другого. Существенен только взаимный порядок значений измерений, а не их точные числовые выражения. В этом случае говорят, что наблюдения даны в *порядковой* или *ординарной шкале*.

3. Критерий Уилкоксона. Общий случай

Пусть даны две независимых выборки

$$X_1, \ldots, X_n \quad \text{и} \quad Y_1, \ldots, Y_m,$$

имеющие функции распределения F и G соответственно.

С помощью критерия Уилкоксона проверяется гипотеза, заключающаяся в том, что обе выборки извлечены из одной совокупности, т. е. имеют одну и ту же функцию распределения

$$H_0: \quad F = G,$$

против альтернативы

$$H_1:\ \mathbf{P}(X < x) = F(x) \geq G(x) = \mathbf{P}(Y < x)\ \text{ для всех } x, \text{ но } F \neq G. \tag{6.10}$$

Если имеют место неравенства (6.10), то говорят, что X стохастически меньше Y. Такому термину можно дать следующее объяснение. Если имеет место (6.10), то при любом x с вероятностью больше, чем $1/2$, в выборке относительная частота случайных величин X_i, таких, что $X_i < x$, больше относительной частоты случайных величин Y_i, таких, что $Y_i < x$.

Неравенство (6.10) имеет место, например, если случайные величины $X_1, \ldots, X_n$ и $Y_1, \ldots, Y_n$ имеют нормальные распределения $N(\theta_1, \sigma^2)$, $N(\theta_2, \sigma^2)$ соответственно, и $\theta_1 < \theta_2$ или в более общем случае они имеют распределения $F(x) = H(x - \theta_1)$, $G(x) = H(x - \theta_2)$ и $\theta_1 < \theta_2$. Здесь H — произвольная функция распределения. Действительно, при любом x

$$F(x) = \mathbf{P}(X < x) = H(x - \theta_1) \geq G(x) = \mathbf{P}(Y < x) = H(x - \theta_2),$$

так как $x - \theta_1 > x - \theta_2$.

Обозначим

$$Z_1 = X_1,\ Z_2 = X_2, \ldots,\ Z_n = X_n,\ Z_{n+1} = Y_1, \ldots,\ Z_{n+m} = Y_m \tag{6.11}$$

и расположим $Z_1, \ldots, Z_{n+m}$ в порядке возрастания:

$$Z^{(1)} \ \leq\ Z^{(2)} \ \leq\ \ldots \ \leq\ Z^{(n+m)}. \tag{6.12}$$

Найдем ранги чисел последовательности (6.12), считая, как и прежде, рангом числа $Z^{(i)}$ его порядковый номер i либо усредненный порядковый номер, если таких чисел в ряду (6.12) несколько.

Таким образом, ранг R_j наблюдения Z_j, $1 \leq j \leq n + m$, задается равенством $Z_j = Z^{(R_j)}$. Если $1 \leq j \leq n$, то $X_j = Z^{(R_j)}$, и если $n < j \leq n + m$, то $Y_{j-n} = Z^{(R_j)}$.

Гипотеза проверяется на основе тестовой статистики Уилкоксона

$$W = \sum_{i=1}^{n} R_i, \tag{6.13}$$

равной сумме рангов наблюдений X_j, $1 \leq j \leq n$. На каком принципе основана проверка гипотез с помощью критерия Уилкоксона?

Если **гипотеза H_0 верна**, то случайные величины $X_1, \ldots, X_n$ и $Y_1, \ldots, Y_m$ одинаково распределены и, следовательно, одинаково распределены и их ранги, принимающие значения $1, \ldots, m+n$ с равной вероятностью $1/(m+n)$. Значения рангов $R_1, \ldots, R_{n+m}$ представляют собой перестановку чисел $1, 2, \ldots, n+m$ и все такие перестановки имеют одинаковую вероятность $1/(n+m)!$, где $(n+m)! = 1 \cdot 2 \cdot 3 \cdot \ldots \cdot (n+m)$.

Таким образом, при справедливости гипотезы мы знаем распределение рангов и можем вычислить распределение статистики Уилкоксона, а с ним найти и критическое значение W_α. В частности, сумма рангов в ряду (6.12) равна

$$1 + 2 + \ldots + (m+n) = \frac{(m+n)(m+n+1)}{2},$$

следовательно, поскольку все ранги одинаково распределены:

$$\mathbf{E}R_i = \frac{1}{n+m} \sum_{j=1}^{n+m} \mathbf{E}R_j = \frac{1}{n+m} \sum_{j=1}^{n+m} R_j = \frac{m+n+1}{2},$$

$$\mathbf{E}W = \sum_{j=1}^{n} \mathbf{E}R_j = \frac{n(m+n+1)}{2}.$$

Прямыми вычислениями находится также, что

$$\mathbf{D}W = \frac{mn(m+n+1)}{12}. \tag{6.14}$$

При **справедливости альтернативы** H_1 ранги $R_j, 1 \le j \le n$, и $R_j, n < j \le n+m$, также одинаково распределены по отдельности для наблюдений $X_1, \ldots, X_n$ и $Y_1, \ldots, Y_m$.

Как уже говорилось, при любом x относительная частота событий $X_j < x, 1 \le j \le n$, больше относительной частоты событий $Y_i < x, 1 \le i \le m$, с вероятностью не меньше $1/2$, а при некоторых x строго больше $1/2$. Это означает, что

$$\mathbf{P}(R_j > R_{n+i}) > 1/2$$

и, следовательно,

$$\mathbf{E}R_j = \frac{1}{n} \sum_{j=1}^{n} \mathbf{E}R_j > \mathbf{E}R_{n+i} = \frac{1}{m} \sum_{i=1}^{m} \mathbf{E}R_{n+i}. \tag{6.15}$$

Поскольку имеют место сходимости по вероятности

$$\frac{1}{n}\sum_{j=1}^{n} R_j \xrightarrow{P} ER_j \qquad \text{при} \quad n \to \infty,$$

$$\frac{1}{m}\sum_{i=1}^{m} R_{n+i} \xrightarrow{P} ER_{n+i} \quad \text{при} \quad m \to \infty,$$

то именно неравенство (6.15) доказывает, что при справедливости альтернативы альтернатива и гипотеза различимы.

Заметим, что если критерий W используется для сравнения двух типов изделий и высокие оценки экспертов говорят о лучшем их качестве, то попадание W в область

$$\frac{(m+n)(m+n+1)}{2} - W_\alpha < W < W_\alpha$$

указывает на то, что изделия X и Y друг от друга существенно не отличаются.

Если для W выполнено неравенство

$$W > W_\alpha, \tag{6.16}$$

то это означает, что изделия типа X существенно лучше, чем Y, а если выполнено неравенство

$$W < \frac{(m+n)(m+n+1)}{2} - W_\alpha, \tag{6.17}$$

то наоборот, что изделия типа Y существенно лучше, чем X.

При $m, n \le 25$ критическое значение W_α, определяемое соотношением (6.9), находим по таблицам распределения Уилкоксона.

При больших m и n статистика W распределена приближенно нормально с параметрами, определенными формулами (6.14), поэтому для проверки гипотезы H_0 следует использовать критерий

$$\Big|\frac{W - \mathbf{E}W}{\sqrt{\mathbf{D}W}}\Big| > x_\alpha,$$

где x_α находим по таблицам нормального распределения.

Критерий Уилкоксона тесно связан с *критерием Манна—Уитни*. Тестовая статистика Манна—Уитни выражается непосредственно через статистику Уилкоксона и равна

$$U = nm + \frac{n(n+1)}{2} - W.$$

Если гипотеза H_0 верна, то

$$\mathbf{E}U = \frac{nm}{2}, \qquad \mathbf{D}U = \frac{mn(m+n+1)}{12}.$$

4. Знаковый критерий Уилкоксона

Рассмотрим несколько иную постановку задачи 6.1. Для владельца компании было бы более естественно попросить сравнивать качество напитков одних и тех же экспертов. Естественно, при этом оценки качества напитков одного и того же эксперта были бы зависимы, и, чтобы соблюдать всю строгость сравнения, нам надо менять постановку задачи и вид тестовой статистики.

Пусть даны пары независимых случайных величин

$$(X_1,\ Y_1), \ldots, (X_n,\ Y_n).$$

Наблюдения X_i можно рассматривать как оценки старого напитка, а Y_i — как оценки нового напитка. Определим случайные величины

$$Z_1 = X_1 - Y_1,\ \ldots,\ Z_n = X_n - Y_n$$

и будем основывать наши заключения на значениях этих случайных величин. Обозначим $H(z)$ функцию распределения случайных величин Z_i:

$$H(z) = \mathbf{P}(Z_i < z).$$

Ясно, что если новый напиток лучше, чем старый, то естественно ожидать, что

$$\mathbf{P}(Z_i < 0) = \mathbf{P}(X_i - Y_i < 0) > 1/2.$$

На этом и основан критерий знаков. Его тестовая статистика равна

$$S = \sum_{j=1}^{n} s(Z_i),$$

где

$$s(Z_i) = \begin{cases} 1, & \text{если } Z_i > 0, \\ 0, & \text{если } Z_i \le 0. \end{cases}$$

Таким образом, тестовая статистика критерия знаков равна числу положительных значений $Z_i, 1 \le i \le n$.

Ясно, что случайные величины $s(Z_1), \ldots, s(Z_n)$ независимы, имеют биномиальное распределение и в силу закона больших чисел

$$\frac{1}{n}\sum_{j=1}^{n} s(Z_i) \xrightarrow{P} \mathbf{P}(Z_i > 0)$$

при $n \to \infty$.

Таким образом, критерий знаков проверяет гипотезу

$$H_0 : \mathbf{P}(Z_i > 0) = 1/2$$

против альтернативы

$$H_1 : \mathbf{P}(Z_i > 0) > 1/2.$$

Критические значения критерия знаков находятся из таблиц биномиального распределения.

Знаковый критерий Уилкоксона проверяет более сложную гипотезу. Ясно, что если качество напитков одинаково, то естественно ожидать, что при любом $z > 0$

$$H_0 : \mathbf{P}(Z_i < -z) = \mathbf{P}(Y_i - X_i < -z) = \mathbf{P}(X_i - Y_i < -z) =$$

$$= \mathbf{P}(-Z_i < -z) = \mathbf{P}(Z_i > z),$$

т. е. $Z_i = X_i - Y_i$ имеет симметричное распределение.

Если же новый напиток лучше старого, то можно ожидать, что

$$H_1 : \mathbf{P}(Z_i < -z) = \mathbf{P}(Y_i - X_i < -z) < \mathbf{P}(X_i - Y_i < -z) =$$

$$= \mathbf{P}(-Z_i < -z) = \mathbf{P}(Z_i > z).$$

Таким образом, мы будем проверять гипотезу симметрии

$$H_0 : H(-z) = 1 - H(z) \text{ для всех } z > 0$$

против альтернативы

$$H_1 : H(-z) < 1 - H(z) \text{ для всех } z > 0.$$

Упорядочим случайные величины $|Z_1|, \ldots, |Z_n|$ в порядке возрастания и обозначим их ранги $R_1, \ldots, R_n$. Проверку гипотезы будем осуществлять на основе знаковой тестовой статистики Уилкоксона

$$W = \sum_{i=1}^{n} R_i s(Z_i),$$

где

$$s(Z_i) = \begin{cases} 1, & \text{если } Z_i > 0, \\ 0, & \text{если } Z_i \leq 0. \end{cases}$$

Таким образом, в W мы суммируем только ранги $|X_i - Y_i|$, для которых $X_i > Y_i$.

Если гипотеза H_0 верна, то это означает, что качество напитков одинаково. Тогда случайные величины $Z_i = X_i - Y_i$ имеют симметричную относительно нуля плотность распределения, так как $Z_i = X_i - Y_i$ и $-Z_i = Y_i - X_i$ одинаково распределены. Это означает, что одинаково распределены $R_i s(Z_i)$ и $R_i s(-Z_i)$, а значит, одинаково распределены статистики

$$W = \sum_{i=1}^{n} R_i s(Z_i)$$

и

$$W_1 = \sum_{i=1}^{n} R_i s(-Z_i)$$

и, следовательно, имеют одинаковые математические ожидания: $\mathbf{E}W = \mathbf{E}W_1$. Поскольку

$$\sum_{i=1}^{n} R_i = W + W_1 = \sum_{i=1}^{n} i = \frac{n(n+1)}{2},$$

то

$$\mathbf{E}W = \mathbf{E}\sum_{i=1}^{n} R_i s(Z_i) = \mathbf{E}\sum_{i=1}^{n} R_i s(-Z_i) = \frac{n(n+1)}{4}.$$

Прямыми вычислениями получаем

$$\mathbf{D}W = \frac{1}{2}\mathbf{D}R = \frac{n^3 - n}{24}.$$

Отсюда тестовую статистику можно нормировать следующим образом:

$$\frac{\left(W - \dfrac{n(n+1)}{4}\right)\sqrt{24}}{\sqrt{n^3 - n}}. \tag{6.18}$$

При $n > 30$ можно считать распределение тестовой статистики (6.18) стандартным нормальным.

Если $n \leq 30$, то выбирают меньшее из чисел W и W_1 и сравнивают его с критическим W_α, значения которого при уровнях значимости $\alpha = 0.05$ и $\alpha = 0.1$ приведены в табл. 6.1. Если

$$W_\alpha \leq \min(W, \quad W_1), \tag{6.19}$$

то принимается гипотеза H_0, в противном случае гипотеза H_0 отвергается.

Таблица 6.1

n	$\alpha = .05$	$\alpha = .1$	n	$\alpha = .05$	$\alpha = .1$
5		1	18	40	47
6	1	2	19	46	54
7	2	4	20	52	60
8	4	6	21	59	63
9	6	8	22	66	75
10	8	11	23	73	83
11	11	14	24	81	92
12	14	17	25	90	101
13	17	21	26	98	110
14	21	26	27	107	120
15	25	30	28	117	130
16	30	36	29	127	141
17	35	41	30	137	152

Задача 6.2. Пусть, как и в задаче 6.1, качество стандартного и нового напитков было оценено следующими $n = 9$ парами баллов:

$$(1,4),\ (2,3),\ (5,3),\ (2,5),\ (2,1),\ (1,3),\ (4,3),\ (3,4),\ (3,2).$$

На основании экспертных оценок с уровнем значимости $\alpha = 0.05$ проверить гипотезу H_0: качество напитков одинаковое.

Решение. Так как $n < 30$, то проверку справедливости гипотезы осуществим с помощью неравенства (6.19). Разности Z_i равны

$$-3,\ -1,\ 2,\ -3,\ 1,\ -2,\ 1,\ -1,\ 1.$$

Модули Z_i равны

$$3,\ 1,\ 2,\ 3,\ 1,\ 2,\ 1,\ 1,\ 1,$$

их ранги

$$8.5,\ 3,\ 6.5,\ 8.5,\ 3,\ 6.5,\ 3,\ 3,\ 3.$$

Находим знаковые статистики Уилкоксона

$$W = 6.5 + 3 + 3 + 3 = 15.5,$$

$$W_1 = 8.5 + 3 + 8.5 + 6.5 + 3 = 29.5.$$

Так как $W < W_1$, то проверку гипотезы H_0 осуществим, сравнивая $W = 15.5$ с $W_\alpha = 6$. Так как $W_\alpha < W$, то считаем, что экспертные данные не противоречат гипотезе H_0, и, следовательно, напитки имеют одинаковое качество.

5. Ранговые критерии независимости

Ранговые критерии для проверки независимости обычно используются при работе с данными, измерения которых проведены в порядковой шкале. Примерами таких задач являются:

есть ли зависимость между ценой автомобиля и его качеством,
насколько стаж сотрудника влияет на качество его работы,
есть ли тесная зависимость между оценками двух экспертов.

Наиболее распространенным является критерий, основанный на ранговом коэффициенте корреляции Спирмена.

Пусть даны пары независимых наблюдений

$$(X_1,\ Y_1), \ldots, (X_n,\ Y_n). \tag{6.20}$$

Пусть $S_1, \ldots, S_n$ — ранги наблюдений $X_1, \ldots, X_n$ и $R_1, \ldots, R_n$ — ранги $Y_1, \ldots, Y_n$.

Тогда ранговый коэффициент корреляции Спирмена равен выборочной корреляции пар рангов наблюдений

$$(S_1, R_1),\ (S_2, R_2),\ \ldots\ldots,\ (S_n, R_n).$$

Иначе, ранговый коэффициент корреляции Спирмена можно определить следующим образом.

Упорядочим пары (6.20) по возрастанию X:

$$X^{(1)} \le X^{(2)} \le \ldots \le X^{(n)}, \tag{6.21}$$

соответственно вторые компоненты в (6.20) запишем в виде

$$Y^{(1)},\ Y^{(2)},\ \ldots, Y^{(n)}. \tag{6.22}$$

Сначала рассмотрим вариант, когда в ряду (6.21) имеют место строгие неравенства

$$X^{(1)} < X^{(2)} < \ldots < X^{(n)}, \tag{6.23}$$

и ряду (6.23) припишем ранги $1, 2, \ldots, n$. Обозначим через $R_1,\ R_2, \ldots,\ R_n$ ранги ряда (6.22).

Вычислим для пар рангов

$$(1, R_1),\ (2, R_2),\ \ldots, (n, R_n) \tag{6.24}$$

коэффициент корреляции Спирмена:

$$r_s = \frac{\dfrac{1}{n}\displaystyle\sum_{i=1}^{n} iR_i - \left(\frac{1}{2}(n+1)\right)^2}{\dfrac{1}{12}(n^2-1)}, \tag{6.25}$$

так как среднее и дисперсия первых n натуральных чисел равны $(n+1)/2$ и $(n^2-1)/12$ соответственно. Выражение (6.25) можно записать в виде

$$r_s = 1 - \frac{6}{n(n^2-1)} \sum_{i=1}^{n} (R_i - i)^2. \tag{6.26}$$

Статистика r_s имеет выборочное среднеквадратическое отклонение

$$s_r = \sqrt{\frac{1 - r_s^2}{n-2}}. \tag{6.27}$$

Если случайные величины X_i и Y_i независимы, то при $n > 10$ распределение t-статистики вида

$$t = \frac{r_s}{s_r} \tag{6.28}$$

хорошо аппроксимируется распределением Стьюдента с $n-2$ степенями свободы. Критерием для проверки гипотезы независимости H_0 является неравенство

$$|t| > t_{\alpha,n-2}, \tag{6.29}$$

где $t_{\alpha,n-2}$ определяется с помощью распределения Стьюдента при уровне значимости α.

Если $n > 30$, то r_s распределено приближенно нормально, и если гипотеза H_0 верна, то

$$\mathbf{E}r_s = 0, \qquad \mathbf{D}r_s = \frac{1}{n-1}.$$

Критерием проверки гипотезы H_0 является неравенство

$$\left|\sqrt{n-1}\, r_s\right| > x_\alpha, \tag{6.30}$$

где x_α находим с помощью нормального распределения. Если для исходных данных коэффициент r_s удовлетворяет неравенству (6.29) или (6.30), то отвергаем гипотезу H_0 о независимости признаков X и Y. В противном случае принимаем эту гипотезу.

Замечание. Если неравенства (6.23) не выполнены, т. е. среди X есть одинаковые величины, то обычным способом следует найти ранги последовательности (6.23). Пусть $Z_1, Z_2, \ldots, Z_n$ — эти ранги, тогда r_s вычисляется по формуле

$$r_s = \frac{\displaystyle\sum_{i=1}^{n} \left(Z_i - \frac{n+1}{2}\right)\left(R_i - \frac{n+1}{2}\right)}{\displaystyle\sum_{i=1}^{n} \left(i - \frac{n+1}{2}\right)^2}.$$

Задача 6.3. В двух городах производится продажа холодильников 10 типов. Можно ли сказать, что спрос на холодильники

различных типов в этих городах независим, опираясь на следующие исходные данные:

X_i	...	50	40	30	19	15	12	10	8	7	4
Y_i	...	60	35	8	30	10	30	20	15	31	19

где приведено число холодильников разных типов, проданных в течение некоторого времени?

Решение. Гипотеза H_0: $\mathbf{E}r_s = 0$ — это независимость спроса.

Составим таблицу рангов этих рядов:

i	...	1	2	3	4	5	6	7	8	9	10
R_i	...	4	8	3	5	6.5	2	6.5	1	9	10
$i - R_i$	...	-3	-6	0	-1	-1.5	4	0.5	7	0	0

По формулам (6.26)–(6.28) находим $r_s = 0.34$, $s_r = 0.33$, $t = 1.03$. При $\alpha = 0.05$ из таблиц находим $t_\alpha = 2.3$. Так как $t < t_\alpha$, то принимаем гипотезу H_0, заключающуюся в том, что спрос на холодильники различных типов в этих городах независим.

Глава 7
ДИСПЕРСИОННЫЙ АНАЛИЗ

Дисперсионный анализ представляет собой совокупность статистических методов, развитых для задач проверки гипотез о равенстве средних, когда число генеральных совокупностей больше двух. Методы проверки гипотез основаны на сравнении оценок квадратичных отклонений средних, т. е., грубо говоря, их оценок дисперсий. Отсюда и возникло название *дисперсионный анализ.*

Техника дисперсионного анализа позволяет не только проверять гипотезу о равенстве средних значений, но и выявлять какие средние значения различны.

Задачи дисперсионного анализа довольно часто возникают на практике. Приведем их примеры.

1. Зависит ли спрос на товар от района города, национальности покупателя, вида рекламной кампании, заработной платы и т. п.?

2. Зависит ли заработная плата от стажа работы, специальности и т. п.?

3. Зависит ли урожай на участке от вида удобрения?

4. Зависит ли здоровье зубов от используемой зубной пасты?

Таким образом, в рамках дисперсионного анализа часто встает задача проверки гипотез о том, влияет ли определенный *фактор* на среднее значение.

1. Однофакторный дисперсионный анализ

Однофакторная модель дисперсионного анализа изучает влияние на среднее значение одного фактора (например, в первом примере — национальности, региона или зарплаты).

Для фактора устанавливаются определенные *уровни* $i = 1, 2, \ldots, m$, под которыми подразумевается определенная мера или состояние фактора (например, конкретная национальность, регион или уровень заработной платы). Для каждого уровня $i, 1 \leq i \leq m$, собираются наблюдения $x_{i1}, \ldots, x_{in_i}$ и на их основе проверяется гипотеза

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_m = \mu$$

о равенстве средних μ_i для всех уровней i, $1 \leq i \leq m$, фактора против альтернативы

$$H_1 : \mu_i \neq \mu_j$$

хотя бы для двух различных уровней $i \neq j$.

Будет предполагаться, что случайные величины x_{ij}, $1 \leq i \leq m$, $1 \leq n_i$, допускают следующее вероятностное представление:

$$x_{ij} = \mu + \tau_i + \xi_{ij},$$

где μ — общее среднее значение, относительно которого проверяется гипотеза, τ_i — отклонение от среднего значения μ, обусловленное влиянием i-го уровня, т. е. $\mathbf{E}x_{ij} = \mu + \tau_i$, ξ_{ij} — случайная компонента, $\mathbf{E}\xi_{ij} = 0$.

В этих обозначениях гипотеза может быть переписана в виде

$$H_0 : \ \tau_1 = \tau_2 = \ldots = \tau_m = 0.$$

Основное предположение дисперсионного анализа вызвано необходимостью действовать в рамках нормальной модели и получить распределение тестовой статистики в виде одного из стандартных распределений, а именно распределения Фишера. Оно состоит в следующем.

Случайные величины $\xi_{ij}, 1 \le i \le m, 1 \le j \le n_i$, независимы и распределены в соответствии с одним и тем же нормальным распределением $N(0, \sigma^2)$.

Таким образом, $\mathbf{E}\xi_{ij} = 0$ и все ξ_{ij} имеют одну и ту же дисперсию $\mathbf{D}\xi_{ij} = \sigma^2$.

Вычислительную процедуру однофакторного дисперсионного анализа опишем для случая равных объемов наблюдений для каждого из факторов, т. е. $n_1 = n_2 = \ldots = n_m$. Она состоит в следующем.

Вычисляются средние

$$\bar{x}_{i\cdot} = \frac{1}{n}\sum_{j=1}^{n} x_{ij}, \quad 1 \le i \le m,$$

наблюдений для каждого из факторов, а также находится общее среднее наблюдений

$$\bar{x}_{\cdot\cdot} = \frac{1}{nm}\sum_{i=1}^{m}\sum_{j=1}^{n} x_{ij}.$$

После этого сравниваются среднеквадратичное отклонение

$$SSG = n\sum_{i=1}^{m}(\bar{x}_{i\cdot} - \bar{x}_{\cdot\cdot})^2$$

групповых средних $\bar{x}_{i\cdot}$ от общего среднего $\bar{x}_{\cdot\cdot}$ со среднеквадратичным отклонением

$$SSE = \sum_{i=1}^{m}\sum_{j=1}^{n}(x_{ij} - \bar{x}_{i\cdot})^2 = \sum_{i=1}^{m}\sum_{j=1}^{n}(\xi_{ij} - \bar{\xi}_{i\cdot})^2, \qquad \bar{\xi}_{i\cdot} = \frac{1}{n}\sum_{j=1}^{n}\xi_{ij},$$

всех наблюдений x_{ij} от их средних $\bar{x}_{i\cdot}$ по уровням. Точнее говоря, вычисляется статистика

$$F = \frac{SSG/(m-1)}{SSE/(n-1)m}.$$

Можно показать, что при справедливости гипотезы случайные величины SSG, SSE независимы и имеют хи-квадрат распределения с $m-1$ и $(n-1)m$ степенями свободы соответственно. Таким образом, статистика F имеет распределение Фишера с $m-1$ и $(n-1)m$

степенями свободы. Это позволяет легко находить по таблицам критические значения для проверки гипотезы.

Насколько выбор такой статистики F хорош для проверки гипотезы?

В числителе F стоит сумма квадратов отклонений оценок $\bar{x}_{i\cdot}$ математических ожиданий средних по уровням от оценки $\bar{x}_{\cdot\cdot}$ математического ожидания общего среднего μ.

Знаменатель $SSE/(n-1)m$ можно рассматривать как некоторую нормировку. В случае справедливости гипотезы он представляет собой оценку дисперсии σ^2, которая обычно неизвестна.

Квадратичное отклонение SSG несет в себе ту же информацию об отклонении средних $\bar{x}_{i\cdot}$ по уровням от общего среднего $\bar{x}_{\cdot\cdot}$, что и общее квадратичное отклонение

$$SS = \sum_{i=1}^{m}\sum_{j=1}^{n}(x_{ij} - \bar{x}_{\cdot\cdot})^2.$$

Оказывается, что разность $SS - SSG$ не зависит от математических ожиданий μ_i и μ, а зависит только от случайных величин ξ_{ij}.

Таким образом, определяя статистику SSG, мы сохраняем в ней весь вклад $\tau_i = \mu_i - \mu$ в квадратичное отклонение SS, вызванный влиянием факторов, но в то же время удаляем из SS существенное квадратичное слагаемое, вызванное случайными величинами ξ_{ij}.

Обоснованием этого утверждения является следующее равенство:

$$SS = SSG + SSE. \tag{7.29}$$

Докажем это равенство. Имеем

$$SS = \sum_{i=1}^{m}\sum_{j=1}^{n}[(x_{ij} - \bar{x}_{i\cdot}) + (\bar{x}_{i\cdot} - \bar{x}_{\cdot\cdot})]^2 = SSE + SSG + 2\,I, \tag{7.30}$$

где

$$I = \sum_{i=1}^{m}\sum_{j=1}^{n}(x_{ij} - \bar{x}_{i\cdot})(\bar{x}_{i\cdot} - \bar{x}_{\cdot\cdot}) = \sum_{i=1}^{m}(\bar{x}_{i\cdot} - \bar{x}_{\cdot\cdot})\sum_{j=1}^{n}(x_{ij} - \bar{x}_{i\cdot}) = 0, \tag{7.31}$$

так как

$$\sum_{j=1}^{n}(x_{ij} - \bar{x}_{i\cdot}) = n\bar{x}_{i\cdot} - n\bar{x}_{i\cdot} = 0.$$

Из (7.30) и (7.31) следует (7.29).

Вычислим математическое ожидание $\mathbf{E}(SSG)$, чтобы еще раз убедиться, что тестовая статистика F позволяет различать гипотезу и альтернативу. Имеем

$$\frac{1}{n}\mathbf{E}(SSG) = \mathbf{E}\sum_{i=1}^{m}(\bar{x}_{i\cdot} - \bar{x}_{\cdot\cdot})^2 = \mathbf{E}\sum_{i=1}^{m}(\tau_i + \bar{\xi}_{i\cdot} - \bar{\xi}_{\cdot\cdot})^2 =$$

$$= \sum_{i=1}^{m}\tau_i^2 + 2\sum_{i=1}^{m}\tau_i\mathbf{E}(\bar{\xi}_{i\cdot} - \bar{\xi}_{\cdot\cdot}) + \sum_{i=1}^{m}\mathbf{E}(\bar{\xi}_{i\cdot} - \bar{\xi}_{\cdot\cdot})^2 =$$

$$= \sum_{i=1}^{m}\tau_i^2 + \sum_{i=1}^{m}\mathbf{E}(\bar{\xi}_{i\cdot} - \bar{\xi}_{\cdot\cdot})^2 = \sum_{i=1}^{m}\tau_i^2 + \frac{(m-1)\sigma^2}{n},$$

так как

$$\sum_{i=1}^{m}\mathbf{E}(\bar{\xi}_{i\cdot} - \bar{\xi}_{\cdot\cdot})^2 = m\mathbf{E}(\bar{\xi}_{i\cdot} - \bar{\xi}_{\cdot\cdot})^2 = mE\bar{\xi}_{i\cdot}^2 - 2m\mathbf{E}(\bar{\xi}_{i\cdot}\bar{\xi}_{\cdot\cdot}) + mE\bar{\xi}_{\cdot\cdot}^2 =$$

$$= m\frac{\sigma^2}{n} - 2\frac{\sigma^2}{n} + \frac{\sigma^2}{n} = \frac{(m-1)\sigma^2}{n}.$$

Заметим, что при любом i

$$\mathbf{E}\frac{1}{n-1}\sum_{j=1}^{n}(\xi_{ij} - \bar{\xi}_{i\cdot})^2 = \sigma^2,$$

поэтому $\mathbf{E}(SSE) = m(n-1)\sigma^2$ и, следовательно,

$$\mathbf{E}(SS) = \mathbf{E}(SSG) + \mathbf{E}(SSE) = \sum_{i=1}^{m}\tau_i^2 + \frac{(m-1)\sigma^2}{n} + m(n-1)\sigma^2. \tag{7.32}$$

2. Двухфакторный дисперсионный анализ

Двухфакторный дисперсионный анализ осуществляет проверку гипотезы, что сразу два фактора не влияют на среднее значение.

Приведем примеры таких гипотез:
1) зарплата выпускников не зависит от региона, в который они были распределены, а также от специальности (бухгалтерской учет,

маркетинг, менеджмент);
2) доходность в выбранных инвестиционных фондах не зависит от фонда и от года его работы;
3) страховую кампанию интересует, зависит ли уровень аварийности легковых машин от марки автомобиля и региона.

Для решения задачи мы разбиваем факторы на уровни и соотносим наблюдения уровням этих факторов:

$$X_{ijk}, \qquad 1 \le i \le I, \quad 1 \le j \le J, \quad 1 \le k \le K,$$

где i — номер уровня первого фактора — i-я группа, j — номер уровня второго фактора — j-й блок, k — номер наблюдения в i-й группе и j-м блоке.

3. Двухфакторный дисперсионный анализ без повторений

В этой модели предполагается дополнительно, что взаимодействие двух рассматриваемых факторов не влияет на среднее значение. Тогда модель можно упростить, взяв вместо случайных величин X_{ijk} их средние по k значения

$$X_{ij} = \frac{1}{K} \sum_{k=1}^{K} X_{ijk}$$

и решать задачу в предположении, что в каждой ячейке ij, соответствующей уровню i первого фактора и уровню j второго, находится ровно одно наблюдение.

Это и есть модель двухфакторного дисперсионного анализа без повторений.

Предполагается, что случайные величины X_{ij} допускают представление

$$X_{ij} = \mu + \lambda_i + \tau_j + \xi_{ij},$$

где ξ_{ij} — независимые случайные величины имеющие нормальное распределение $N(0, \sigma^2)$, μ — общее среднее, λ_i — влияние i-й группы на среднее, τ_j — влияние j-го блока на среднее.

Имеют место равенства

$$\mathbf{E}X_{ij} = \mu + \tau_j + \lambda_i,$$

$$\frac{1}{I}\sum_{i=1}^{I}\mathbf{E}X_{ij} = \mu_{\cdot j} = \mu + \tau_j, \qquad \frac{1}{J}\sum_{j=1}^{J}\mathbf{E}X_{ij} = \mu_{i\cdot} = \mu + \lambda_i,$$

где

$$\sum_{j=1}^{J}\tau_j = \sum_{j=1}^{J}(\mu - \mu_{\cdot j}) = J\mu - \sum_{j=1}^{J}\mu_{\cdot j} = 0, \tag{7.33}$$

$$\sum_{j=1}^{I}\lambda_i = \sum_{i=1}^{I}(\mu - \mu_{i\cdot}) = J\mu - \sum_{i=1}^{I}\mu_{i\cdot} = 0. \tag{7.34}$$

Таким образом, $\mu_{\cdot j}$ есть общее среднее для математических ожиданий для j-го блока, а $\mu_{i\cdot}$ — общее среднее для математических ожиданий для i-й группы.

Как показывают формулы (7.33), (7.34), общее среднее выбрано так, что влияние групп и блоков в среднем равно нулю.

Гипотеза H_0 состоит в том, что

$$\tau_1 = \tau_2 = \cdots = \tau_J = 0 \tag{7.35}$$

и

$$\lambda_1 = \lambda_2 = \cdots = \lambda_I = 0, \tag{7.36}$$

т. е. что оба фактора не влияют на среднее.

Альтернатива заключается в том, что хотя бы одно из равенств в (7.35), (7.36) не имеет места. Это и означает, что по крайней мере один из факторов влияет на среднее значение.

Вычислительная процедура заключается в следующем.

Вычисляются средние по группам и блокам

$$\bar{X}_{i\cdot} = \frac{1}{J}\sum_{j=1}^{J}X_{ij}, \quad \mathbf{E}\bar{X}_{i\cdot} = \mu_{i\cdot},$$

$$\bar{X}_{\cdot j} = \frac{1}{I}\sum_{i=1}^{I}X_{ij}, \quad \mathbf{E}\bar{X}_{\cdot j} = \mu_{\cdot j},$$

а также общее среднее

$$\bar{X}_{\cdot\cdot} = \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J}X_{ij}, \quad \mathbf{E}\bar{X}_{\cdot\cdot} = \mu.$$

После этого находятся квадратичные отклонения средних по группам

$$SSG = J\sum_{i=1}^{I}(\bar{X}_{i\cdot} - \bar{X}_{\cdot\cdot})^2$$

и средних по блокам

$$SSB = I\sum_{j=1}^{J}(\bar{X}_{\cdot j} - \bar{X}_{\cdot\cdot})^2$$

от общего среднего. Кроме того, находятся общая сумма квадратов

$$SST = \sum_{i=1}^{I}\sum_{j=1}^{J}(X_{ij} - \bar{X}_{\cdot\cdot})^2$$

и сумма квадратов остатков

$$SSE = \sum_{i=1}^{I}\sum_{j=1}^{J}(X_{ij} - \bar{X}_{i\cdot} - \bar{X}_{\cdot j} + \bar{X}_{\cdot\cdot})^2.$$

Остатки $\bar{X}_{ij} - \bar{X}_{i\cdot} - \bar{X}_{\cdot j} + \bar{X}_{\cdot\cdot}$ определяются исходя из следующих соображений. Ошибка ξ_{ij} наблюдений равна

$$\xi_{ij} = X_{ij} - \tau_i - \lambda_j - \mu = X_{ij} - (\mu_{i\cdot} - \mu) - (\mu_{\cdot j} - \mu) - \mu = X_{ij} - \mu_{i\cdot} - \mu_{\cdot j} + \mu$$

и $\bar{X}_{i\cdot}, \bar{X}_{\cdot j}, \bar{X}_{\cdot\cdot}$ являются несмещенными оценками $\mu_{i\cdot}, \mu_{\cdot j}$ и μ соответственно. Таким образом, значения остатков не зависят от значений λ_i и τ_j.

Аналогично случаю однофакторного дисперсионного анализа доказывается справедливость тождества

$$SST = SSG + SSB + SSE.$$

Можно показать, что случайные величины SSB, SSG, SSE независимы и имеют хи-квадрат распределение со следующими степенями свободы:

SSG	—	$I-1$	степеней свободы,
SSB	—	$J-1$	степеней свободы,
SSE	—	$(I-1)(J-1)$	степеней свободы.

Из этих соображений по аналогии с однофакторным дисперсионным анализом строятся две тестовые статистики

$$F_G = \frac{SSG/(I-1)}{MSE},$$

$$F_B = \frac{SSB/(J-1)}{MSE},$$

где

$$MSE = \frac{SSE}{(I-1)(J-1)}.$$

По существу знаменатель MSE является оценкой дисперсии σ^2.

Тестовая статистика F_G проверяет гипотезу о влиянии первого фактора, а F_B — второго.

При справедливости гипотезы статистики F_G и F_B имеют распределение Фишера с $I-1, (I-1)(J-1)$ и $J-1, (I-1)(J-1)$ степенями свободы. Это и используется для вычисления p-значений и нахождения критических значений критериев.

Отметим, что данная модель используется и для проверки гипотезы о влиянии только одного из факторов, когда влиянием другого фактора нельзя пренебречь, поскольку от его уровней сильно зависят значения наблюдений. Он является мешающим. Например, в задаче зависит ли зарплата выпускников от выбранной специальности в качестве мешающего фактора естественно учитывать успеваемость учеников.

4. Двухфакторный дисперсионный анализ с повторениями

Данная модель дополнительно проверяет гипотезу о том, что взаимодействие между двумя факторами не влияет на среднее значение. Это и обусловливает все различия в вычислительной процедуре, поскольку для проверки этой гипотезы вводится дополнительная тестовая статистика.

Математическая модель состоит в следующем.

Независимые наблюдения X_{ijk} допускают представление

$$X_{ijk} = \mu + \lambda_i + \tau_j + \gamma_{ij} + \xi_{ijk},$$

где ξ_{ijk} — независимые случайные величины, имеющие нормальное распределение $N(0, \sigma^2)$; μ — общее среднее; λ_i — влияние i-й

группы на среднее; τ_j — влияние j-го блока на среднее; γ_{ij} — влияние взаимодействия i-й группы и j-го фактора.

Таким образом,

$$EX_{ijk} = \mu + \tau_j + \lambda_i + \gamma_{ij},$$

где

$$\sum_{i=1}^{I} \lambda_i = 0, \quad \sum_{j=1}^{J} \tau_j = 0$$

и

$$\sum_{i=1}^{I} \gamma_{ij} = 0, \quad \sum_{j=1}^{J} \gamma_{ij} = 0,$$

причем последние два равенства имеют место для любых $1 \le i \le I$ и $1 \le j \le J$.

Необходимо проверить гипотезу

$$\lambda_i = 0, \quad \tau_j = 0, \quad \gamma_{ij} = 0$$

для любых $1 \le i \le I, 1 \le j \le J$ против альтернативы, что одно из этих равенств нарушено.

Вычислительная процедура состоит в следующем.

Вычисляются средние

$$\bar{X}_{...} = \frac{1}{IJK} \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} X_{ijk}, \quad E\bar{X}_{...} = \mu,$$

$$\bar{X}_{i..} = \frac{1}{JK} \sum_{j=1}^{J} \sum_{k=1}^{K} X_{ijk}, \quad \mathbf{E}\bar{X}_{i..} = \mu_{i\cdot} = \mu + \lambda_i,$$

$$\bar{X}_{.j.} = \frac{1}{IK} \sum_{i=1}^{I} \sum_{k=1}^{K} X_{ijk}, \quad \mathbf{E}\bar{X}_{.j.} = \mu_{\cdot j} = \mu + \tau_j,$$

$$\bar{X}_{ij.} = \frac{1}{K} \sum_{k=1}^{K} X_{ijk}, \quad \mathbf{E}\bar{X}_{ijj.} = \mu_{ij} = \mu + \lambda_i + \tau_j + \gamma_{ij},$$

являющиеся оценками своих математических ожиданий. После этого находятся квадратичные отклонения

$$SST = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(X_{ijk} - \bar{X}_{...})^2,$$

$$SSG = JK\sum_{i=1}^{I}(\bar{X}_{i..} - \bar{X}_{...})^2,$$

$$SSB = IK\sum_{j=1}^{J}(\bar{X}_{.j.} - \bar{X}_{...})^2,$$

$$SSI = \sum_{i=1}^{I}\sum_{j=1}^{J}(X_{ij.} - \bar{X}_{i..} - \bar{X}_{.j.} + \bar{X}_{...})^2,$$

$$SSE = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(X_{ijk} - \bar{X}_{ij.})^2.$$

Сумма квадратов SSI оценивает квадратичные отклонения вызванные взаимодействием (interaction) факторов, так как

$$\mathbf{E}(X_{ij.} - \bar{X}_{i..} - \bar{X}_{.j.} + \bar{X}_{...}) = \mu_{ij} - \mu_{i.} - \mu_{.j} + \mu =$$

$$(\mu + \lambda_i + \tau_j + \gamma_{ij}) - (\mu + \lambda_i) - (\mu + \tau_j) + \mu = \gamma_{ij}.$$

Квадратичное отклонение SSE не зависит от параметров μ, λ_i, τ_j, γ_{ij} и может быть использовано для оценки дисперсии σ^2.

Основным тождеством двухфакторного дисперсионного анализа с повторениями является

$$SST = SSG + SSB + SSI + SSE.$$

Статистики SSG, SSB, SSI, SSE независимы, имеют хи-квадрат распределение со следующими степенями свободы:

SSG	—	$I-1$	степеней свободы,
SSB	—	$J-1$	степеней свободы,
SSI	—	$(I-1)(J-1)$	степеней свободы,
SSE	—	$IJ(K-1)$	степеней свободы.

Таким образом, мы можем определить тестовые статистики для проверки гипотез влияния первого, второго факторов и взаимодействия факторов соответственно:

$$F_G = \frac{SSG/(I-1)}{MSE},$$

$$F_B = \frac{SSB/(J-1)}{MSE},$$

$$F_I = \frac{SSI/(I-1)(J-1)}{MSE},$$

где $MSE = \frac{SSE}{IJ(K-1)}$.

При справедливости гипотезы статистики имеют распределения Фишера со следующими степенями свободы:

F_G	—	$I-1$	и	$IJ(K-1)$	степеней свободы,
F_B	—	$J-1$	и	$IJ(K-1)$	степеней свободы,
F_I	—	$(I-1)(J-1)$	и	$IJ(K-1)$	степеней свободы.

Это используется для вычисления их p-значений и критических значений.

Глава 8
СТАТИСТИЧЕСКИЙ КОНТРОЛЬ КАЧЕСТВА

1. Введение

Высокое качество обслуживания основывается на высоком качестве продукции. Проверка качества продукции необходима при изготовлении товара и поступлении его в продажу. Чем раньше будут выявляться отклонения от нормы, тем меньше затраты на улучшение качества товара и тем меньше издержки производства. Сначала рассмотрим виды контроля за технологическими процессами, которые требуют проверки количественных и качественных признаков.

Назовем некоторые виды технологических процессов:

1) разлив фруктового сока по упаковкам;
2) нарезка металлических стержней;

3) упаковка крупы в расфасовочные пакеты;
4) изготовление стеклянной посуды;
5) сборка телевизоров.

Первые три варианта требуют проверки количественных, последние два — качественных признаков. Рассмотрим способы проверки каждого из этих видов признаков.

Наиболее простой и доступный вид проверки производимой продукции может быть осуществлен с помощью контрольных карт, которые были предложены Шухартом (*Shewhart*) в 20-е годы.

2. Контрольные карты количественных признаков при известных μ и σ

Естественно предположить, что все величины, подлежащие проверке, имеют свои нормативные значения μ, технологический процесс их изготовления вносит стандартную ошибку σ, а их рассеивание подчиняется нормальному закону. Поэтому если X — элемент генеральной совокупности, подлежащей проверке, то

$$X \in N(\mu, \sigma). \tag{8.1}$$

Пусть $X_1, X_2, \ldots, X_n$ — выборка объема n, $X_i \in N(\mu, \sigma)$. Контрольные карты будут построены на основе средних арифметических

$$\overline{X} = \frac{1}{n} \sum_{k=1}^{n} X_k$$

и размаха выборки

$$R = \max(X_1, \ldots, X_n) - \min(X_1, \ldots, X_n).$$

а) Контрольная карта средних арифметических

Известно, что если выполнено соотношение (8.1), то

$$\mathbf{P}(\mu - 2\sigma < X < \mu + 2\sigma) = 0.95, \tag{8.2}$$

$$\mathbf{P}(\mu - 3\sigma < X < \mu + 3\sigma) = 0.998. \tag{8.3}$$

Ввиду симметричности нормального распределения из соотношения (8.2) следует, что выход за каждую из границ $\mu \pm 2\sigma$ имеет

вероятность 0.025, т. е. в среднем одно из 40 будет больше $\mu + 2\sigma$ и одно меньше $\mu - 2\sigma$.

Из соотношения (8.3) следует, что каждое из событий $X > \mu + 3\sigma$ и $X < \mu - 3\sigma$ имеет вероятность 0.001, т. е. в среднем на каждую тысячу приходится по одному удовлетворяющему этим неравенствам.

Выборочное среднее $\overline{X}$ имеет $E\overline{X} = \mu$, $\mathbf{D}\overline{X} = \sigma^2/n$, кроме того, $\overline{X} \in N(\mu, \sigma^2/n)$, поэтому для него выполняются соотношения, аналогичные (8.2) и (8.3):

$$\mathbf{P}\left(\mu - 2\frac{\sigma}{\sqrt{n}} < X < \mu + 2\frac{\sigma}{\sqrt{n}}\right) = 0.95, \tag{8.4}$$

$$\mathbf{P}\left(\mu - 3\frac{\sigma}{\sqrt{n}} < X < \mu + 3\frac{\sigma}{\sqrt{n}}\right) = 0.998, \tag{8.5}$$

и именно на них основано построение контрольной карты.

Верхняя и нижняя 95% границы для выборочного среднего $\overline{X}$, определяемые соотношением (8.4), равны $\left(\mu - 2\frac{\sigma}{\sqrt{n}},\ \mu + 2\frac{\sigma}{\sqrt{n}}\right)$ и называются *предупреждающими границами*, а соотношение (8.5) дает 98% границы $\left(\mu - 3\frac{\sigma}{\sqrt{n}},\ \mu + 3\frac{\sigma}{\sqrt{n}}\right)$, которые называются *границами регулирования*.

Построение карты состоит из шагов:

1. Выбирается величина n объема выборки.

2. Через равные промежутки времени производится случайная выборка объема n.

3. Среднее $\overline{X}$ наносится на карту в соответствии с номером выборки.

4. Если очередное $\overline{X}$ выходит за пределы границ регулирования, то останавливается технологический процесс для выявления неслучайных причин вариации.

5. Если $\overline{X}$ попадают между предупреждающей границей и границей регулирования, то следующая выборка производится сразу же, не дожидаясь момента очередной выборки. Если два последовательно полученных значения $\overline{X}$ попадают между этими границами, то останавливается технологический процесс для выявления неисправностей.

6. Если точки на графике образуют возрастающий или убывающий тренд, то это может быть симптомом разладки.

Контрольная карта выборочного среднего имеет следующие линии:

Центральная линия	μ
Верхняя предупреждающая граница	$\mu + 2\sigma/\sqrt{n}$
Нижняя предупреждающая граница	$\mu - 2\sigma/\sqrt{n}$
Верхняя граница регулирования	$\mu + 3\sigma/\sqrt{n}$
Нижняя граница регулирования	$\mu - 3\sigma/\sqrt{n}$

Рис. 8.1. Контрольные карты $\overline{X}$.

Рис. 8.1 представляет собой две контрольных карты, на каждой из которых центральной линией является значение μ и при заданном объеме выборки n нанесены 95% и 98% границы для $\overline{X}$. Кружочками помечены средние $\overline{X}$ в семи выборках. На левой карте все средние $\overline{X}$ не выходят за предупреждающие границы. На карте справа среднее значение четвертой выборки вышло за верхнюю предупреждающую границу, следующие два значения лежат в пределах нормы, а среднее седьмой выборки вышло за нижнюю границу регулирования, поэтому процесс был остановлен.

Фактически контрольная карта дает возможность проверить гипотезу H_0, состоящую в том, что технологическое производство

налажено, т. е. генеральная совокупность случайных величин подчиняется нормальному закону, а именно выполняется соотношение (8.1). Если $\overline{X}$ лежит за предупреждающими границами, то гипотеза H_0 отклоняется при 5%-ном уровне значимости. Если $\overline{X}$ выходит за пределы границ регулирования, то мы отклоняем гипотезу H_0 при 2%-ном уровне значимости.

Пример 8.1. Производится расфасовка кофе в упаковки весом по 200 г. Известно, что фасовочный станок работает со стандартным отклонением $\sigma = 0.25$ г. Через каждые полчаса производится случайная выборка объемом 5 упаковок. Каждую упаковку взвешивают. В табл. 8.1 приведены результаты 6 последовательных выборок и их средние значения $\overline{X}$. Построить контрольную карту арифметического среднего.

Таблица 8.1

Номер выборки	1	2	3	4	5	6
Вес	200.2	199.8	199.7	200.4	199.4	200.1
	200.4	199.4	199.9	200.7	199.7	199.8
	199.7	200.1	200.2	199.8	200.6	199.7
	199.8	200.3	200.4	199.7	200.5	200.1
	200.1	199.5	200.6	200.6	200.7	200.4
$\overline{X}$	200.04	199.82	200.20	200.24	200.18	200.02

Решение. Центральная линия соответствует уровню $\mu = 200$ г. Предупреждающие границы:

$$\mu \pm 2\frac{\sigma}{\sqrt{n}} = 200 \pm 2\frac{0.25}{\sqrt{5}},$$

т. е. 199.776 и 200.224 г.

Границы регулирования:

$$\mu \pm 3\frac{\sigma}{\sqrt{n}} = 200 \pm 3\frac{0.25}{\sqrt{5}},$$

т. е. 199.664 и 200.336 г.

Нанесем средние значения на контрольную карту. Выборочное среднее 4-й выборки превысило верхнюю предупреждающую границу, однако средние последующих выборок лежат в области допустимых значений, что указывает на случайность выброса.

б) Контрольная карта изменчивости технологического процесса, основанная на размахе R выборки

При некотором виде разладки технологического процесса выборочное среднее $\overline{X}$ будет удовлетворительным, а сами наблюдения будут сильно отклоняться в обе стороны от среднего, что характерно при увеличении стандартного отклонения. Для проверки такого типа разладки используется контрольная карта размаха R, которая состоит в проверке превышения размахом определенных границ. Слишком малый размах свидетельствует либо об улучшении технологического процесса, либо о наличии измерительной ошибки, приводящей к занижению границ в контрольной карте.

Если для каждого из элементов X генеральной совокупности выполнено соотношение (8.1), то случайная величина R/σ не зависит от μ и σ, и можно найти ее распределение, а также величины d_n, r_w и r_Λ, такие, что

$$\mathbf{E}\left(\frac{R}{\sigma}\right) = d_n, \quad \mathbf{E}R = \sigma d_n, \tag{8.6}$$

$$\mathbf{P}\left(\frac{R}{\sigma} > r_w\right) = \mathbf{P}(R > r_w\sigma) = 0.025, \tag{8.7}$$

$$\mathbf{P}\left(\frac{R}{\sigma} > r_\Lambda\right) = \mathbf{P}(R > r_\Lambda\sigma) = 0.01. \tag{8.8}$$

Величины d_n, r_w и r_Λ в зависимости от $n = 2, 3, \ldots, 20$ приведены в табл. 8.2. Эти величины участвуют в построении контрольной карты размаха.

Таблица 8.2

n	d_n	r_w	r_Λ	n	d_n	r_w	r_Λ
2	1.128	3.17	4.65	12	3.258	4.92	6.09
3	1.169	3.68	5.06	13	3.336	4.99	6.14
4	2.059	3.98	5.31	14	3.407	5.04	6.19
5	2.326	4.20	5.48	15	3.472	5.09	6.23
6	2.534	4.36	5.62	16	3.532	5.14	6.27
7	2.704	4.49	5.73	17	3.588	5.18	6.31
8	2.847	4.61	5.82	18	3.640	5.22	6.35
9	2.970	4.70	5.90	19	3.689	5.26	6.38
10	3.077	4.79	5.97	20	3.735	5.30	6.41
11	3.173	4.86	6.04				

Контрольная карта размаха R состоит из следующих прямых:

Центральная линия	$d_n\sigma$	
Верхняя предупреждающая граница	$r_w\sigma$	2.5% значений размаха превышают границу
Верхняя граница регулирования	$r_\Lambda\sigma$	1% значений размаха превышают границу

Контрольную карту размаха R строят при объеме выборки не больше 20.

Заметим, что проверка продукции должна осуществляться с помощью обеих контрольных карт.

Пример 8.2. По данным примера 8.1 вычислить уровни линий контрольной карты размаха.

Решение. Объем выборки $n = 5$, стандартное отклонение $\sigma = 0.25$ г. Из табл. 8.2 находим:

$$d_n = d_5 = 2.326, \quad r_w = 4.20, \quad r_\Lambda = 5.45,$$

и по ним границы в контрольной карте:

Центральная линия	$d_n\,\sigma = 2.33 \times 0.25 = 0.58$ г
Верхняя предупрежд. граница	$r_w\,\sigma = 4.20 \times 0.25 = 1.05$ г
Верхняя граница регулирования	$r_\Lambda\,\sigma = 5.45 \times 0.25 = 1.36$ г

Для исходных данных примера 8.1 приведем величину размаха каждой выборки.

Таблица 8.3

Номер выборки	1	2	3	4	5	6
Размах выборки	0.7	0.9	0.9	1.0	1.3	0.7

Таблица 8.3 показывает, что в пятой выборке значение размаха превышает верхнюю предупреждающую границу, но в следующей выборке размах находится в норме, следовательно, нет оснований для беспокойства.

в) Контрольные карты количественных признаков при неизвестных μ и σ

В параграфе 8.2 указан способ построения контрольных карт в предположении, что у генеральной совокупности известны μ и σ. Если они нам не известны, то, как принято в статистике, их следует заменить соответствующими оценками.

Для оценки математического ожидания μ используется среднее $\overline{\overline{X}}$ от найденных ранее выборочных средних, т. е.

$$\mu \approx \overline{\overline{X}} = \frac{1}{M} \sum_{m=1}^{M} \overline{X}(m), \tag{8.9}$$

где $\overline{X}(m)$ — выборочное среднее m-й выборки, M — число выборок.

Если величина σ неизвестна, то ее оценивают с помощью среднего значения размахов $\overline{R}$ выборок, которое подчиняется уравнению (8.8). Обозначая $R(m)$ размах m-й выборки, $m = 1, \ldots, M$, получаем оценку для математического ожидания $\mathbf{E}R$ в виде

$$\mathbf{E}R = d_n \sigma \approx \overline{R} = \frac{1}{M} \sum_{m=1}^{M} R(m). \tag{8.10}$$

Из соотношений (8.10) следует, что для оценки σ можно использовать формулу

$$\hat{\sigma} \approx \frac{\overline{R}}{d_n}. \tag{8.11}$$

При неизвестных μ и σ можно использовать контрольные карты, полученные при известных μ и σ, заменив в них μ и σ соответственно по формулам (8.9) и (8.11). Тогда контрольная карта выборочного среднего имеет следующие линии:

Центральная линия	$\overline{\overline{X}}$
Предупреждающие границы	$\overline{\overline{X}} \pm \frac{2}{\sqrt{n}} \frac{\overline{R}}{d_n}$
Границы регулирования	$\overline{\overline{X}} \pm \frac{3}{\sqrt{n}} \frac{\overline{R}}{d_n}$

В контрольной карте размахов:

Центральная линия	$\overline{R}$
Верхняя предупреждающая граница	$r_w \frac{\overline{R}}{d_n}$
Верхняя граница регулирования	$r_\Lambda \frac{\overline{R}}{d_n}$

Пример 8.3. Используя данные примера 8.1, касающиеся расфасовки кофе, оценить μ и σ по 6 выборкам. Построить карты среднего $\overline{X}$ и размаха R.

В табл. 8.1 приведены средние значения $\overline{X}(m)$ и в табл. 8.3 — размахи $R(m)$ этих 6 выборок. Среднее значение всех выборок равно $\overline{\overline{X}} = 1/6 \times 1000.5 = 200.08$ г, а их средний размах равен $\overline{R}$=1/6× 5.5=0.91.

Контрольная карта среднего:

Центральная линия — 200.08.
Предупреждающие границы —

$$200.08 \pm 2\,\frac{0.91}{2.326\sqrt{5}} = 200.08 \pm 0.35 = (199.73;\ 200.43).$$

Границы регулирования —

$$200.08 \pm 3\,\frac{0.91}{2.326\sqrt{5}} = 200.08 \pm 0.52 = (199.56;\ 200.60).$$

Контрольная карта размаха:

Центральная линия — 0.91.
Верхняя предупреждающая граница — $4.2 \cdot 0.91/2.326 = 1.64$.
Верхняя граница регулирования — $5.48 \cdot 0.91/2.326 = 2.14$.

Оценить дисперсию σ^2 можно не только с помощью размаха, но и с помощью выборочной дисперсии

$$\overline{s}^2 = \frac{1}{n-1}\sum (X_i - \overline{X})^2.$$

Если произведено M выборок, то для оценки σ^2 можно использовать усредненную величину $\overline{s}^2 = \sum s^2(m)/M$, где $s^2(m)$ — выборочная дисперсия m-й выборки. При $n > 25$ распределение $\sqrt{n}(\overline{X} - \mu)/\overline{s}$ близко к нормальному, поэтому контрольная карта $\overline{X}$ имеет вид:

Центральная линия	$\overline{\overline{X}}$
Предупреждающие границы	$\overline{\overline{X}} \pm 2\dfrac{\overline{s}}{\sqrt{n}}$
Границы регулирования	$\overline{\overline{X}} \pm 3\dfrac{\overline{s}}{\sqrt{n}}$

3. Контрольные карты качественных признаков

Проверка качества продукции не всегда ограничивается количественными признаками, у некоторых видов продукции необходимо оценивать качественные признаки. Это могут быть дефекты при сборке телевизоров, различные дефекты стеклянной или керамической посуды. Задача контроля состоит в том, чтобы выявить превышение доли брака продукции по сравнению с нормативной. Контрольные карты качественных признаков способствуют выявлению таких отклонений.

Мы рассмотрим контрольные карты качественных признаков с помощью p-карт, в которых оценивается доля бракованных изделий к общему числу изделий. На основании нескольких выборок

следует оценить p — долю бракованных изделий в генеральной совокупности. Ясно, что наилучшая оценка p — это величина $\overline{p}$, равная отношению общего числа брака во всех выборках к общему числу изделий, подвергшихся проверке.

В каждой отдельной выборке объема n долю бракованных изделий обозначают $\hat{p}$, и $\hat{p} = r/n$, где r — число бракованных изделий в выборке.

Рассмотрим случайную величину ξ, представляющую число бракованных изделий в выборке объема n, она принимает значения $r = 0, 1, \ldots, n$ и подчиняется биномиальному закону:

$$\mathbf{P}(\xi = r) = C_n^r p^r q^{n-r}, \quad r = 0, 1, \ldots, n, \quad q = 1 - p,$$

а p — доля брака в генеральной совокупности. Известно, что $\mathbf{E}\xi = np$, $\mathbf{D}\xi = npq$.

A priori доля $\breve{p}$ бракованных изделий в выборке является случайной величиной, эквивалентной ξ/n. Случайная величина $\breve{p}$, как и ξ/n, принимает значения $\hat{p} = r/n$ и имеет параметры

$$\mathbf{E}\breve{p} = p, \quad \mathbf{D}\breve{p} = \frac{pq}{n}, \quad \sigma(\breve{p}) = \sqrt{\frac{pq}{n}}.$$

Биномиальный закон аппроксимируется законом Пуассона, если $n > 30$, $p < 0.1$, $np \leq 5$, и нормальным законом, если $n > 30$, $0.1 < p < 0.9$, $np > 5$ и $n(1-p) > 5$.

Так как вероятность p брака генеральной совокупности, как правило, неизвестна, а эта величина является центральной линией контрольной карты, то вместо p используют $\overline{p}$, полученную по большому числу выборок усреднением долей брака в них.

а) Аппроксимация нормальным законом

Так как нормальный закон симметричен, то контрольная карта качественных признаков имеет следующие линии:

Центральная линия	$\overline{p}$
Предупреждающие границы	$\overline{p} \pm 2\sqrt{\frac{pq}{n}}$
Границы регулирования	$\overline{p} \pm 3\sqrt{\frac{pq}{n}}$

Пример 8.4. Компания производит некоторые изделия. При хорошей наладке оборудования было произведено 20 выборок, каждая объемом в 100 единиц. В табл. 8.4 приведены результаты проверки количества бракованных изделий в каждой выборке. Нужно построить контрольную карту для проверки технологического процесса.

Таблица 8.4

Номер выборки	Число бракованных изделий	Номер выборки	Число бракованных изделий
1	4	11	5
2	2	12	3
3	5	13	5
4	7	14	6
5	3	15	8
6	8	16	5
7	7	17	2
8	6	18	4
9	4	19	7
10	7	20	6

Решение. Общее число бракованных изделий равно 104, следовательно,

$$\overline{p} = \frac{104}{20 \times 100} = 0.052.$$

Так как $np = 5.2$, то при решении задачи аппроксимировать можно нормальным и пуассоновским распределением.

Будем аппроксимировать долю брака нормальным распределением. Вычислим предупреждающие границы

$$0.052 \pm 2\sqrt{\frac{0.052(1-0.052)}{100}}, \quad \text{т. е.} \quad 0.008, \text{ и } 0.096,$$

и границы регулирования

$$0.052 \pm 3\sqrt{\frac{0.052(1-0.052)}{100}}, \quad \text{т. е.} \quad -0.014 \text{ и } 0.118.$$

Так как доля бракованных изделий не может быть отрицательной, то нижнюю границу регулирования полагаем равной нулю.

Контрольная карта проверки доли брака:

Центральная линия	0.052
Предупреждающие границы	0.008 и 0.096
Границы регулирования	0.000 и 0.118

б) Аппроксимация доли брака пуассоновским распределением

Пуассоновское распределение не является симметричным, поэтому построение контрольной карты отличается от нормального случая.

Пусть случайная величина η имеет пуассоновское распределение с параметром $\lambda = np$, представляющим собой среднее число бракованных изделий в выборке объема n. Вероятность того, что в выборке r бракованных изделий, равна

$$\mathbf{P}(\eta = r) = \frac{\lambda^r}{r!} e^{-\lambda} = \frac{(np)^r}{r!} e^{-np}. \tag{8.12}$$

Долю брака в выборке объемом n можно записать в виде $\breve{p} = \eta/n$, отсюда

$$\mathbf{P}(\eta = r) = \mathbf{P}\left(\frac{\eta}{n} = \frac{r}{n}\right) = \mathbf{P}\left(\breve{p} = \frac{r}{n}\right) = \frac{(np)^r}{r!} e^{-np}.$$

Для нахождения границ в контрольной карте нужно построить табл. 8.5, в которой приводятся значения r, $\hat{p}$, $P(\eta = r)$ и кумулятивные (суммарные) вероятности P_r:

$$P_r = \mathbf{P}(\eta \le r) = \sum_{k=0}^{r} \frac{(np)^k}{k!} e^{-np}. \tag{8.13}$$

Теперь вводим в рассмотрение две пары уровней вероятностей $p_{0.001} = 0.001$, $p_{0.999} = 0.999$ и $p_{0.025} = 0.025$, $p_{0.975} = 0.975$, которым отвечают нижняя и верхняя границы регулирования : $x_{0.001}$, $x_{0.999}$, и нижняя и верхняя предупреждающие границы: $x_{0.025}$, $x_{0.975}$. Уровень вероятности, например, $p_{0.975} = 0.975$ связан с соответствующей границей $x_{0.975}$ соотношением

$$\mathbf{P}(\breve{p} \le x_{0.975}) \ge p_{0.975} = 0.975.$$

Далее, чтобы определить величину границы, например, $x_{0.025}$, отвечающую вероятности $p_{0.025}$, находят r, такое, что

$$\sum_{k=0}^{r} \frac{(np)^k}{k!} e^{-np} < p_{0.025} < \sum_{k=0}^{r+1} \frac{(np)^k}{k!} e^{-np},$$

т. е. $P_r < p_{0.025} < P_{r+1}$, и полагают $x_{0.025} = (r + 1/2)/n$, т. е. $x_{0.025}$ равно середине интервала $(r/n; (r + 1)/n)$.

Вероятность того, что брака нет $(r = 0)$, равна $\mathbf{P}(\eta = 0) = P_0 = e^{-np}$. Если окажется, что уровень вероятности, например, $p_{0.001}$ меньше вероятности P_0 отсутствия брака, тогда полагают $x_{0.001} = 0$, так как регулирующая граница брака не может быть отрицательной.

Найдем значения прямых в контрольной карте для примера 8.4 при аппроксимации пуассоновским распределением с $n = 100$, $np = 5.2$. В табл. 8.5 для распределения Пуассона с параметром $\lambda = 5.2$, являющимся средним числом брака в выборке, представлены величины r, r/n, $P(r)$ и P_r.

Так как $p_{0.001} < P_0$, то $x_{0.001} = 0$. Величина $p_{0.025} = 0.025$ попадает между P_0 и P_1, следовательно, $x_{0.025} = (1/2)/n = 0.005$. Далее $P_9 < p_{0.975} = 0.975 < P_{10}$, следовательно, $x_{0.975} = (9 + 1/2)/n = 0.095$; $P_{12} < p_{0.999} < P_{13}$, следовательно, $x_{0.999} = (12 + 1/2)/n = 0.125$.

Контрольная карта имеет вид:

Центральная линия — $\hat{p} = 0.049$.

Предупреждающие границы — $x_{0.025} = 0.005$ и $x_{0.975} = 0.095$.

Границы регулирования — $x_{0.001} = 0$ и $x_{0.999} = 0.125$.

Пример 8.5. Та же компания по той же технологии выпускает новую продукцию. В табл. 8.6 представлена информация о числе бракованных изделий в 15 последовательных выборках по 100 штук.

Для проверки качества выпускаемой продукции воспользуемся контрольными картами, полученными в предыдущем примере. Доля брака в 11-й выборке превышает предупреждающую границу, однако 12-я выборка, которая производится сразу после этого выброса, и последующие лежат в нужных пределах. Это говорит о случайности большой доли брака в 11-й выборке. Так как в большей части выборок доля брака меньше $\hat{p} = 0.049$, то это свидетельство улучшения технологии.

Таблица 8.5

Число бракованных изделий r	Доля бракованных изделий $\hat{p}$	Вероятность $P(r)$	Кумулятивная вероятность P_r
0	0	0.0055	0.0055
1	0.01	0.0287	0.0342
2	0.02	0.0746	0.1088
3	0.03	0.1293	0.2381
4	0.04	0.1681	0.4061
5	0.05	0.1748	0.5809
6	0.06	0.1515	0.7324
7	0.07	0.1125	0.8449
8	0.08	0.0731	0.9181
9	0.09	0.0423	0.9603
10	0.10	0.0220	0.9823
11	0.11	0.0104	0.9927
12	0.12	0.0045	0.9972
13	0.13	0.0018	0.9990
14	0.14	0.0007	0.9997
15	0.15	0.0002	0.9999

Таблица 8.6

Номер выборки	Число брак. изделий	Доля брака	Номер выборки	Число брак. изделий	Доля брака
1	7	0.07	9	3	0.03
2	4	0.04	10	4	0.04
3	2	0.02	11	10	0.16
4	1	0.01	12	5	0.05
5	5	0.05	13	7	0.07
6	6	0.06	14	2	0.02
7	3	0.03	15	1	0.01
8	5	0.05			

4. Статистический приемочный контроль качества неколичественных признаков

До сих пор рассматривался контроль, который способствовал улучшению технологического процесса. Одноко проверке должны подвергаться готовые к отправке изделия, а также продукция, поступающая от других поставщиков. В этом случае часто используется следующий прием. Задаются размер выборки n и максимально допустимое в ней количество c бракованных изделий. Случайным образом производится выборка объема n. Если число бракованных изделий не превосходит c, то партия товара принимается, в противном случае она бракуется.

Ясно, что выбор параметров n и c должен зависеть от доли бракованных изделий, которую допускает клиент. Максимально допустимая доля бракованных изделий, которую будем обозначать p_0, называется *допустимым уровнем качества*.

При доле брака в партии равной p_0 может оказаться, что в случайной выборке доля брака выше p_0, поэтому поставщик вводит еще одну границу p_{00}, называемую *допустимым процентом бракованных изделий в партии*. Партия бракуется, если доля брака в выборке больше или равна p_{00}. Проверка должна быть такой, чтобы с большой вероятностью гарантировать наличие доли брака не более, чем p_0.

Выбор критерия должен решить по крайней мере две проблемы. Во-первых, уменьшить вероятность α отвергнуть партию, если у нее доля брака не больше p_0. Вероятность α называется *риском производителя*, так как при доле брака p_0 партия должна быть принята, а она отвергается.

Во-вторых, уменьшить вероятность β принять партию, если у нее вероятность брака равна p_{00}. Вероятность β называется *риском потребителя*, так как партия принимается в то время, как доля брака в ней значительно выше нормы p_0 и равна p_{00}.

Партия, которая подвергается проверке, как правило, содержит большое количество товара, так как иначе каждый предмет можно было бы проверить отдельно. Поэтому при заданных вероятностях брака p и размере выборки n с помощью биномиального распределения можно найти вероятности принять или забраковать партию.

По формуле биномиального распределения находим вероятность наличия r бракованных изделий в выборке объема n:

$$P(r) = C_n^r p^r q^{n-r}, \quad q = 1 - p, \quad r = 0, 1, \ldots, n.$$

Партия принимается, если число бракованных изделий не превосходит c с вероятностью β, поэтому риск потребителя слишком велик и он меньше при пользовании схемой B.

Эти два варианта показывают,что чем больше объем выборки, тем меньше риск потребителя, следовательно, для удовлетворительного решения проблемы необходимо увеличить объем выборки. При объеме выборки больше 30 можно пользоваться нормальным приближением.

Рассмотрим аппроксимацию нормальным законом.

Пусть ξ — случайная величина, представляющая число бракованных изделий в выборке объемом n и распределенная по биномиальному закону. Так как биномиальное распределение аппроксимируется нормальным с математическим ожиданием $\mathbf{E}\xi = np$, $\mathbf{D}\xi = npq$, то при больших значениях n случайная величина $\dfrac{\xi - np}{\sqrt{npq}} \epsilon N(0,1)$, и

$$\mathbf{P}\left(\frac{\xi - np}{\sqrt{npq}} \leq x_\gamma\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x_\gamma} e^{-t^2/2} dt = \gamma. \tag{8.14}$$

Величины x_γ по γ и, наоборот, γ по x_γ находятся из таблиц нормального распределения.

Зададим вероятности p_0 и p_{00}. Наша задача найти n и c, такие, чтобы выполнялись два условия :

1. Вероятность принять партию, если $p = p_0$, равна $1 - \alpha$, т. е.

$$\mathbf{P}(\xi \leq c | p = p_0) = \mathbf{P}\left(\frac{\xi - np_0}{\sqrt{np_0q_0}} \leq \frac{c - np_0}{\sqrt{np_0q_0}}\right) = 1 - \alpha. \tag{8.15}$$

2. Вероятность принять партию, если $p = p_{00}$, равна β, т. е.

$$\mathbf{P}(\xi \leq c | p = p_{00}) = \mathbf{P}\left(\frac{\xi - np_{00}}{\sqrt{np_{00}q_{00}}} \leq \frac{c - np_{00}}{\sqrt{np_{00}q_{00}}}\right) = \beta. \tag{8.16}$$

Из равенств (8.15) и (8.16) с учетом соотношения (8.14) получаем

$$\frac{c - np_0}{\sqrt{np_0q_0}} = x_{1-\alpha}, \qquad \frac{c - np_{00}}{\sqrt{np_{00}q_{00}}} = x_\beta, \tag{8.17}$$

где величины $x_{1-\alpha}$ и x_β находятся из таблиц нормального распределения с помощью равенства (8.15).

Так как в обоих уравнениях (8.17) величины n и c одни и те же, то это дает возможность найти их. Пусть n и c — решение этих уравнений, а именно

$$n = \frac{(x_{1-\alpha}\sqrt{p_0q_0} - x_\beta\sqrt{p_{00}q_{00}})^2}{(p_{00} - p_0)^2}, \tag{8.18}$$

$$c = np_0 + x_{1-\alpha}\sqrt{np_0q_0}. \tag{8.19}$$

Ясно, что n и c не являются целыми числами, а по существу должны быть целыми, поэтому n и c следует округлить до ближайших целых n_0 и c_0, причем

$$n_0 > n, \quad c_0 > c. \tag{8.20}$$

Величины n_0 и c_0 — это минимальные целые, для которых выполнены неравенства (8.20).

Итак, в эксперименте нужно брать объем выборки n_0, а максимально допустимое количество брака в ней c_0.

Какие же вероятности риска производителя α_1 и риска потребителя β_1 соответствуют величинам n_0 и c_0? Чтобы ответить на этот вопрос, найдем

$$x_{1-\alpha_1} = \frac{c_0 - n_0p_0}{\sqrt{n_0p_0q_0}}, \qquad x_{\beta_1} = \frac{c_0 - n_0p_{00}}{\sqrt{n_0p_{00}q_{00}}} \tag{8.21}$$

и, воспользовавшись таблицами нормального распределения, определим α_1 и β_1.

Пример 8.6. Пусть $p_0 = 0.05$, $p_{00} = 0.15$, $\alpha = 0.05$, $\beta = 0.05$. Найти n и c, для которых выполняются соотношения (8.15) и (8.16).

Решение. По таблицам нормального распределения находим $x_{1-\alpha} = x_{0.95} = 1.64$, $x_\beta = x_{0.05} = -1.64$. Из равенств (8.18) и (8.19) получаем $n = 88.93$, $c = 7.8$, а учитывая (8.20), $n_0 = 89$, $c_0 = 8$.

Теперь найдем вероятности риска производителя α_1 и риска потребителя β_1 для целых n_0 и c_0. Из равенств (8.21) находим $x_{1-\alpha_1} = 1.73$, $x_{\beta_1} = -1.59$, и им соответствуют (из таблиц нормального распределения) $\alpha_1 = 0.042$, $\beta_1 = 0.056$.

Итак, при выборке объемом $n_0 = 89$ и максимально допустимом в ней количестве $c_0 = 8$ бракованных изделий оба вида рисков производителя $\alpha_1 = 0.042$ и потребителя $\beta_1 = 0.056$ получились близкими к исходным.

Пример 8.7. Пусть $p_0 = 0.01$, $p_{00} = 0.05$, $\alpha = 0.01$, $\beta = 0.05$. Найти n и c, для которых выполняются соотношения (8.15) и (8.16).

Решение. Получаем: $x_{1-\alpha} = 2.33$, $x_\beta = -1.65$. По формулам (8.18) и (8.19) находим $n = 217.02$, $c = 5.58$, поэтому $n_0 = 218, c_0 = 6$, и $x_{\alpha_1} = 2.49$, $x_{\beta_1} = 1.52$, а $\alpha_1 = 0.005$, $\beta_1 = 0.066$.

Пример 8.8. Пусть $p_0 = 0.01$, $p_{00} = 0.05$, $\alpha = 0.01$, $\beta = 0.01$. Найти n и c, для которых выполняются соотношения (8.15) и (8.16).

Решение. Получаем: $x_{1-\alpha} = 2.33$, $x_\beta = -2.33$. Найдем n по формуле (8.19) $n_0 = 342, c_0 = 8$, $x_{\alpha_1} = 2.49$, $x_{\beta_1} = -2.33$ $\alpha_1 = 0.006$, $\beta_1 = 0.012$.

Глава 9

АНАЛИЗ ВРЕМЕННЫХ РЯДОВ

Каждая компания непрерывно следит за основными показателями своей деятельности, контролируя их каждый месяц, квартал, год. Примерами таких показателей являются активы, пассивы, прибыль, число служащих, средняя заработная плата, доля в рынке производимой продукции, количество проданной продукции и т. п. Они являются основными характеристиками финансового благополучия компании. По этой причине тщательное наблюдение за ними и их анализ являются ключевыми элементами менеджмента любой компании. Для статистического анализа данные показатели трактуются как временные ряды, т. е. наблюдения, упорядоченные во времени.

1. Декомпозиция временного ряда

Техника декомпозиции предназначена для выделения из временного ряда детерминированной компоненты, оценки его математического ожидания и разложения этой компоненты в виде нескольких компонент, имеющих содержательную экономическую интерпретацию.

Таким образом, при операции декомпозиции временного ряда $Y_t, t = 0, \pm 1, \pm 2, \ldots$ мы получаем его задание в виде

$$Y_t = D_t * I_t,$$

где D_t — детерминированная компонента, I_t — нерегулярная случайная компонента. Операция $*$ может быть как сложением, так и умножением. Модель умножения соответствует в экономике моделям начисления сложного процента (величина капитала, положенного в банк, из года в год растет экспоненциально).

Детерминированную компоненту D_t в свою очередь разлагают на композицию тренда T_t и периодической компоненты Π_t, т. е.

$$D_t = T_t * \Pi_t.$$

Тренд T_t представляет собой долгосрочную детерминированную компоненту, показывающую общую тенденции развития фирмы. На практике он часто бывает монотонной функцией.

Периодическая компонента Π_t обычно, в свою очередь, представляется как композиция двух компонент:

S_t — сезонной компоненты, учитывающей колебания внутри года (например, объем продаж шуб и купальников зависит от времени года),

C_t — циклической компоненты, учитывающей долгосрочные колебания экономики (экономические циклы продолжительностью 2–3 года и более).

Итак,

$$\Pi_t = C_t * S_t.$$

Таким образом, общая модель может быть записана в виде

$$Y_t = T_t * S_t * C_t * I_t. \tag{9.1}$$

Отметим, что три компоненты T, C и S в данном разложении мож-

но считать детерминированными и только одну I — случайной.

В экономике чаще используют мультипликативную модель, т. е. $Y = T \cdot S \cdot C \cdot I$, хотя рассматриваются и аддитивные модели $Y = T + S + C + I$.

2. Оценивание компонент временного ряда. Параметрическое оценивание тренда

В качестве модели аппроксимирующей функции тренда T чаще всего выбирают одну из моделей параметрической регрессии:

а) линейную регрессию (линейный тренд)

$$T_\beta(t) = \beta_0 + \beta_1\, t, \qquad \beta = (\beta_0,\, \beta_1),$$

б) полиномиальный тренд

$$T_\beta(t) = \beta_0 + \beta_1\, t + \beta_2\, t^2 + \ldots,$$

в) экспоненциальный тренд

$$T_\beta(t) = \beta_0 e^{\beta_1\, t}.$$

Оценивание параметров осуществляется чаще всего методом наименьших квадратов

$$\min_\beta \sum_i (T_\beta(t_i) - Y_i)^2 = \sum_i (T_{\hat\beta}(t_i) - Y_i)^2,$$

где $Y_i = Y(t_i)$ — значения временного ряда в точке t_i.

3. Непараметрическое оценивание тренда. Сглаживание

Иногда не удается подобрать простую функцию в качестве тренда, но в то же время необходимо оценить тенденцию данных. Тогда используют метод сглаживания временного ряда.

Тренд в точке t обозначаем Y_t^* и находим как взвешенное среднее величин

$$Y_{t-k},\ Y_{t-k-1},\ Y_{t-k-2},\ \ldots,\ Y_t,\ Y_{t+1},\ \ldots,\ Y_{t+k-1},\ Y_{t+k}$$

по формуле

$$Y_t^* = \sum_{j=-k}^{k} c_j Y_{t+j}, \quad k+1 \le t \le N-k, \quad c_{-j} = c_j, \quad \sum_{j=-k}^{k} c_j = 1. \tag{9.2}$$

Как следует из формулы (9.2), в этом случае тренд не может быть найден для k первых и k последних наблюдений.

В частном случае, когда все c_j равны между собой, получаем $c_j = 1/(2k+1)$. Тогда тренд в точке t является арифметическим средним соседних $2k+1$ наблюдений

$$Y_t^* = \frac{1}{2k+1} \sum_{j=-k}^{k} Y_{t+j}. \tag{9.3}$$

При анализе сезонных колебаний (см. п. 4а)) используются значения

$$c_{-k} = c_k = \frac{1}{4k}, \qquad c_j = \frac{1}{2k} \quad \text{при} \quad -k+1 \le j \le k-1.$$

Такое сглаживание подавляет сезонную компоненту S_t и ослабляет вклад нерегулярной компоненты I_t.

Действительно, рассмотрим случай аддитивной модели $Y = T + S + C + I$. Тогда если значения I_t слабо зависимы, то в силу закона больших чисел

$$\frac{1}{4k}(I_{t-k} + I_{t+k}) + \frac{1}{2k} \sum_{j=-k+1}^{k-1} I_j \approx 0$$

при больших k.

Если функция S_t периодическая с периодом l, таким, что k делится на l и $S_{t-l} = -S_t$, то

$$\frac{1}{4k}(S_{t-k} + S_{t+k}) + \frac{1}{2k} \sum_{j=-k+1}^{k-1} S_j = 0.$$

При анализе мультипликативной модели ее иногда сводят к аддитивной, рассматривая временной ряд $\ln Y_t, t = 0, \pm 1, \pm 2, \ldots$

Оценка тренда с помощью сглаживания, как правило, не применяется для прогноза, поскольку этот метод не находит величину

тренда в последних k точках. Сглаживание дает общую тенденцию изменения наблюдаемого ряда.

4. Сезонные колебания

Рассмотрим сначала схему, не содержащую циклической компоненты, и будем предполагать, что случайный процесс Y_t может быть представлен в виде

$$Y_t = Tr(t) \cdot S(t) \cdot I(t), \tag{9.4}$$

где $Tr(t)$ — тренд; $S(t)$ — сезонная компонента; $I(t)$ — случайная компонента, причем $Tr(t) > 0, S(t) > 0, I(t) > 0$.

Предполагается, что наблюдения

$$Y_1, Y_2, \ldots Y_N \tag{9.5}$$

велись в течение L лет m раз в году и общее число наблюдений равно $N = Lm$.

Исходные данные Y_t, как и их компоненты в формуле (9.4), можно записать в виде

$$\begin{aligned} &Y_t = Y_{ij}, \qquad S(t) = S_{ij}, \qquad t = i + (j-1)m, \\ &1 \le i \le m, \qquad 1 \le j \le L, \qquad 1 \le t \le N = Lm, \end{aligned} \tag{9.6}$$

где i — номер месяца при $m = 12$ или квартала при $m = 4$, а j — номер года.

Перейдем к оценке составляющих (9.6). Так как m — четное число, то полагаем $m = 2k$.

а) С помощью подвижных средних

$$Y_t^* = \frac{1}{2m}[Y_{t-k} + 2(Y_{t-k-1} + \ldots + Y_{t+k-1}) + Y_{t+k}], \quad k+1 \le t \le N-k, \tag{9.7}$$

получаем ряд Y_t^*, в котором проведено осреднение за год. В результате в Y_t^* сезонные колебания оказываются исключенными.

При $m = 12$ (ежемесячные данные) формулы (9.7) принимают вид

$$Y_t^* = \frac{1}{24}[Y_{t-6} + 2(Y_{t-5} + Y_{t-4} + \ldots + Y_{t+5}) + Y_{t+6}], \quad 7 \le t \le N-6, \tag{9.8}$$

а при $m = 4$ (ежеквартальные данные) — вид

$$Y_t^* = \frac{1}{8}[Y_{t-2} + 2(Y_{t-1} + Y_t + Y_{t+1}) + Y_{t+2}], \quad 3 \le t \le N-2. \quad (9.9)$$

б) Вычисляемые в процентах предварительные оценки $S^*(t)$ сезонности находим по формуле

$$S^*(t) = \frac{Y_t}{Y_t^*} \cdot 100\%, \qquad k+1 \le t \le N-k, \quad k = m/2. \qquad (9.10)$$

Пользуясь приемом (9.6), величины $S^*(t)$, полученные в формуле (9.10), можно записать в виде S_{ij}^*, где i — номер месяца или квартала, j — номер года.

В табл. 9.1 приведены величины Y_t, Y_t^*, $S^*(t)$, представляющие собой квартальный доход некоторой фирмы в миллионах долларов. Эти данные и некоторые другие взяты из книги [1].

Таблица 9.1

Год	Кв.	t	Y_t	Y_t^*	$S^*(t)$	$\tilde{Y}_t$	$\hat{T}r(t)$	$\hat{I}_t$
1985	1	1	1.441			1.389	1.366	101.64
	2	2	1.209			1.366	1.358	100.59
	3	3	1.526	1.371	111.32	1.352	1.350	100.12
	4	4	1.321	1.365	96.79	1.393	1.342	103.79
1986	1	5	1.414	1.348	104.93	1.363	1.334	102.13
	2	6	1.187	1.316	90.18	1.341	1.326	101.13
	3	7	1.411	1.283	109.98	1.250	1.318	94.82
	4	8	1.185	1.259	94.12	1.250	1.310	95.38
1987	1	9	1.284	1.262	101.78	1.237	1.302	95.01
	2	10	1.125	1.273	88.40	1.271	1.294	98.22
	3	11	1.493	1.279	116.74	1.322	1.286	102.82
	4	12	1.192	1.281	93.03	1.257	1.278	98.34
1988	1	13	1.327	1.275	104.04	1.279	1.270	100.67
	2	14	1.102	1.275	86.42	1.245	1.262	98.65
	3	15	1.469			1.301	1.254	103.75
	4	16	1.213			1.279	1.246	102.64

в) Обозначим символом $\bar{S}_i^*$ усредненные по годам сезонные индексы S_{ij}^*:

$$\bar{S}_i^* = \frac{1}{L-1} \sum_j S_{ij}^*, \quad 1 \le i \le m. \tag{9.11}$$

Коэффициент в формуле (9.11) равен $1/(L-1)$, так как общее число сезонных индексов в формуле (9.10) равно $(L-1)m$.

г) Так как сезонные компоненты (в количестве m) вычисляются в процентах, то ясно, что сумма их за год должна быть равна $m \cdot 100\%$, поэтому окончательную оценку сезонной компоненты получаем в виде

$$\hat{S}_i = \frac{\bar{S}_i^*}{\sum_{i=1}^{m} \bar{S}_i^*} \cdot m \cdot 100\%, \quad 1 \le i \le m, \tag{9.12}$$

и для них выполнено равенство

$$\sum_{i=1}^{m} \hat{S}_i = m \cdot 100\%. \tag{9.13}$$

В любой точке t номер сезонной компоненты $S(t)$ или ее оценки $\hat{S}(t)$ будем обозначать $\{t\}$, а определяется он будет следующим образом:

$$\{t\} = \begin{cases} m, & \text{если } \dfrac{t}{m} \text{ — целое число,} \\ t - \left[\dfrac{t}{m}\right] m, & \text{если } \dfrac{t}{m} \text{ — не целое число.} \end{cases} \tag{9.14}$$

Здесь выражение $[A]$ обозначает целую часть A. Ясно, что для любого t имеет место неравенство $1 \le \{t\} \le m$.

В дальнейшем для обозначения $S(t)$ будем использовать также выражение $S_{\{t\}}$, причем номер $\{t\}$ будет определяется по формуле (9.14). Согласно формуле (9.14) если $m = 12$, то $\{30\} = 6$, $\{5\} = 5$, $\{60\} = 12$, $\{45\} = 9$, а если $m = 4$, то $\{9\} = 1$, $\{12\} = 4$, $\{35\} = 3$.

В табл. 9.2 приведены промежуточные S_{ij}^*, $\bar{S}_i^*$ и окончательные $\hat{S}_i$ оценки сезонных компонент S_i.

Таблица 9.2

Кв.	S^*_{ij}				$\bar{S}^*_i$	$\hat{S}_i$
1		104.93	101.78	104.04	103.58	103.78
2		90.18	88.40	86.42	88.33	88.50
3	111.32	109.98	116.74		112.68	112.89
4	96.79	94.12	93.03		94.65	94.83

д) Для ряда (9.5) найдем по методу наименьших квадратов с использованием формул (5.8) оценку тренда в виде

$$\hat{T}r(t) = \hat{\beta}_0 + \hat{\beta}_1 t. \tag{9.15}$$

е) Из исходных данных (9.5), имеющих вид (9.4), исключаем сезонные компоненты

$$\tilde{Y}_{ij} = \frac{Y_{ij}}{\hat{S}_i} 100\%, \quad 1 \le i \le m, \quad 1 \le j \le L. \tag{9.16}$$

ж) Из полученных величин $\tilde{Y}_{ij}$, записанных в форме $\tilde{Y}_t$, исключаем тренд

$$\hat{I}_t = \frac{\tilde{Y}_t}{\hat{T}r(t)} 100\%, \qquad 1 \le t \le N. \tag{9.17}$$

В результате найдены оценки всех компонент процесса вида (9.4). Заметим, что при оценивании сезонных компонент величины $S^*(t)$, $\bar{S}^*(t)$, $\hat{S}(t)$ и $\hat{I}_t$ вычисляются в процентах.

Итак, сезонность есть процентное отношение значения процесса в какой-либо точке к среднему за год в этой точке.

5. Прогноз процесса

Обозначим прогноз процесса Y_t вида (9.4) от момента $t = N$ на t_0 шагов вперед через $\hat{Y}(N, t_0)$. Тогда

$$\hat{Y}(N, t_0) = \hat{T}r(N + t_0)\hat{S}_{\{N+t_0\}}\overline{I}/10^4, \qquad \overline{I} = \frac{1}{N}\sum_{t=1}^{N} \hat{I}_t, \tag{9.18}$$

где $\hat{S}_{\{N+t_0\}}$ — сезонность в точке $N+t_0$.

Пример 9.1. В предположении, что исходные данные Y_t, приведенные в табл. 9.2, подчиняются модели (9.4), найдем оценки соответствующих компонент.

В табл. 9.2 приведены сезонные (квартальные) оценки $\hat{S}_{\{i\}}$, $i = 1, 2, 3, 4$, вычисленные по формулам (9.12), а также их предварительные оценки $S^*_{ij}, \overline{S}^*_i$. Оценка тренда $Tr(t)$ производится с помощью формул (5.8) по исходным данным Y_t и равна

$$\hat{T}r(t) = 1.374 - 0.008t.$$

В табл. 9.2 приведены остатки $\hat{I}_t$. Их среднее равно $\overline{I} = 99.98$. Четыре оценки сезонности таковы:

$$\hat{S}_{\{1\}} = 103.8, \quad \hat{S}_{\{2\}} = 88.5, \quad \hat{S}_{\{3\}} = 112.9, \quad \hat{S}_{\{4\}} = 94.8.$$

Найдем прогноз на первые четыре квартала от момента $t = N = 16$ при $t_0 = 1, 2, 3, 4$. При $t_0 = 1$ получаем

$$\hat{Y}(16,\ 1) = \hat{T}r(16+1)\hat{S}_{\{16+1\}}\overline{I}/10^4 =$$

$$= (1.374 - 0.008(16+1)) \cdot 103.779 \cdot 99.981/10^4 = 1.285.$$

Далее, если $t_0 = 2$, то

$$\hat{Y}(16.2) = \hat{T}r(16+2)\hat{S}_{\{16+2\}}\overline{I}/10^4 =$$

$$= (1.374 - 0.008 \cdot 18) \cdot 88.502 \cdot 99.981/10^4 = 1.089,$$

при $t_0 = 3$ и $t_0 = 4$ получаем соответственно $\hat{Y}(16.3) = 1.380$ и $\hat{Y}(16.4) = 1.151$.

Пример 9.2. В предположении, что исходные данные Y_t, приведенные в табл. 9.3, имеют вид (9.4), найдем оценки соответствующих компонент. На рис. 9.1 представлен график сезонных компонент. В табл. 9.4 приведены сезонные (месячные) оценки $\hat{S}_{\{i\}}, i = 1,\ 2, \ldots, 12$, вычисленные по формулам (9.12), а также их предварительные оценки $S^*_{ij}, \overline{S}^*_i$.

Величины S^*_{ij} получены по формулам (9.10). Найдем оценку $\overline{S}^*_i$, например, при $i = 2$, $\overline{S}^*_2 = (80.6 + 95.4 + 90.7 + 89.5 + 76.7 +$

93.6)/6 = 87.7 Оценка тренда $Tr(t)$ производится с помощью формул (5.8) по исходным данным Y_t и равна

$$\hat{T}r(t) = 701.081 + 2.899 \cdot t.$$

Среднее остатков $\hat{I}_t$ равно $\overline{I} = 99.939$. Получим прогнозы на январь и июнь, т. е. при $t_0 = 1.6$. При $t_0 = 1$ индекс сезонности в точке $t = N + t_0 = 84 + 1$ равен $t = 84 + 1 = 1$, и согласно формуле (9.18) получаем прогноз

$$\hat{Y}(N,1) = \hat{T}r(N+1) \cdot S_1^* \cdot \overline{I} =$$
$$= (701.081 + 2.899 \cdot (84+1)) \cdot 91.4 \cdot 99.937 = 865.455.$$

Рис. 9.1. График сезонных компонент.

Таблица 9.3

Мес.	1982	1983	1984	1985	1986	1987	1988
1	509	595	747	781	913	800	774
2	546	569	782	790	822	671	810
3	626	725	835	927	848	829	919
4	672	728	837	936	906	895	852
5	708	773	886	912	918	830	874
6	717	869	928	923	1012	963	981
7	626	789	903	949	934	899	883
8	627	773	852	926	894	903	901
9	625	735	874	1105	1149	955	937
10	655	757	834	973	948	819	807
11	678	701	816	828	719	718	764
12	765	910	823	849	902	901	896

Таблица 9.4

Мес.	S^*_{ij}							$\bar{S}^*_i$	$\hat{S}_i$
1		85.9	92.1	90.2	99.2	91.3	89.3	91.3	91.4
2		81	95	91	90	77	94	88	88
3		101	101	105	92	96	106	100	100
4		100	100	104	98	105	98	101	101
5		106	105	101	100	98	101	102	102
6		118	110	102	111	114	113	111	111
7	96	105	107	104	103	106		104	104
8	96	101	101	101	100	106		101	101
9	95	94	103	120	129	111		109	109
10	98	96	97	106	107	95		100	100
11	101	88	94	91	81	83		90	90
12	112	113	95	93	102	104		103	103

Замечание 2. Исходный процесс может содержать циклическую компоненту $C(t)$ и иметь вид

$$Y_t = Tr(t) \cdot S(t) \cdot C(t) \cdot I(t). \tag{9.19}$$

Тогда помимо выполнения пунктов а)–ж) следует из $\hat{I}_t$ выделить циклическую компоненту $\hat{C}_t$ с периодом $m_1 \neq m$ тем же способом, что и сезонные компоненты, или методом скользящего среднего. Циклическую компоненту принято выражать в процентах. Далее следует исключить циклическую компоненту

$$I^*_t = \hat{I}_t / \hat{C}_{\{t\}},$$

где $\hat{C}_{\{t\}}$ — величина циклической переменной в точке t, а $\{t\}$ — номер циклической компоненты в момент времени t, который определяется по формуле (9.14) с заменой m на m_1.

Прогноз процесса (9.19) вычисляется по формуле

$$\hat{Y}(N, t_0) = \hat{Tr}(N + t_0)\hat{S}_{\{N+t_0\}}\hat{C}_{\{N+t_0\}}\bar{I}^*/10^6, \quad \bar{I}^* = \frac{1}{N}\sum_{t=1}^{N} I^*_t.$$

Глава 10

СТАЦИОНАРНЫЕ ВРЕМЕННЫЕ РЯДЫ. МОДЕЛЬ АВТОРЕГРЕССИИ

1. Оценка автокорреляционной функции

Рассмотрим временные ряды X_t, в которых отсутствует тренд (если в исходном ряде он был, то исключен с помощью описанных выше методов) и которые имеют достаточно устойчивый колебательный характер вокруг среднего. Это стационарные временные ряды, они могут быть двух видов.

Временной ряд X_t называется *стационарным в широком смысле*, если при всех t

$$\mathbf{E}X_t = \mu, \qquad \mathbf{E}|X_t|^2 < \infty,$$

а ковариационная функция $R(t,s)$ зависит только от разности $t-s$ аргументов, т. е.

$$R(t,s) = \mathbf{E}(X_t X_s) - \mathbf{E}X_t\,\mathbf{E}X_s = R(t-s).$$

Из стационарности процесса следует, что при любом h

$$\mathbf{E}X_{t+h} = \mathbf{E}X_t = \mu, \qquad \mathbf{E}(X_t X_s) = \mathbf{E}(X_{t+h}X_{s+h}) = R(t-s) + \mu^2.$$

Временной ряд X_t называется *стационарным в узком смысле*, если совместная функция распределения величин $X_{t+k_1}, \ldots, X_{t+k_m}$ не зависит от сдвига по времени t для любых $k_1, \ldots, k_m$.

Для построения стандартных оценок ковариации, корреляции, формул прогноза обычно достаточно предположения стационарности временного ряда в широком смысле. Однако для более тонкого анализа: построения доверительных интервалов для прогноза, для математического ожидания, для ковариации и корреляции приходится прибегать к предположению о нормальности временного ряда или его стационарности в узком смысле.

Надо отметить, что в экономических задачах стационарность далеко не всегда адекватно описывает динамику процессов, поэтому довольно часто используются другие классы процессов.

Так, например, часто используются временные ряды Y_t со стационарными приращениями, для которых приращения

$$X_t = Y_{t+1} - Y_t$$

представляют собой стационарный временной ряд. Предположение стационарности приращений естественно для временных рядов, описывающих процессы со стабильной динамикой развития.

Надо отметить, что для анализа процессов, происходящих в финансах, часто используются очень сложные и специальные модели временных рядов.

Поскольку $\mu = \text{const}$, то, не ограничивая общности, можно считать $\mathbf{E}X_t = \mu = 0$, так как иначе можно рассмотреть процесс $X_t - \mu$, у которого $\mathbf{E}(X_t - \mu) = 0$.

Пусть X_t — стационарный случайный процесс с

$$\mathbf{E}X_t = 0. \tag{10.1}$$

Обозначим $R(k)$ и ρ_k — ковариации и коэффициенты корреляций процесса X_t соответственно:

$$R(k) = \mathbf{E}X_{t+k}X_t, \quad \rho_k = \frac{R(k)}{R(0)}, \quad k = 0, 1, 2, \ldots$$

Они удовлетворяют равенствам

$$R(k) = R(-k), \quad \rho_k = \rho_{-k}, \quad k = 0, 1, 2, \ldots$$

Пусть

$$X_1,\ X_2,\ \ldots,\ X_T \tag{10.2}$$

— реализация случайного процесса X_t. Для оценки $R(k)$ и ρ_k в предположении (10.1) используются выборочные функции $\hat{R}(k)$ и r_k, определяемые равенствами

$$\hat{R}(k) = \hat{R}(-k) = \frac{1}{T-k}\sum_{j=1}^{T-k} X_{j+k}X_j, \quad r_k = r_{-k} = \frac{\hat{R}(k)}{\hat{R}(0)}, \tag{10.3}$$

где $l < T - 1, \quad k = 0, 1, \ldots, l$. Оценка $\hat{R}(k)$ является несмещенной, в чем мы убеждаемся из равенства

$$\mathbf{E}\hat{R}(k) = \frac{1}{T-k}\sum_{j=1}^{T-k} EX_{j+k}X_j = R(k).$$

Если математическое ожидание $\mathbf{E}X_t = m = \text{const}$ неизвестно, то $R(t)$ оценивают функцией

$$R^*(k) = R^*(-k) = \frac{1}{T-k}\sum_{j=1}^{T-k}(X_{j+k} - \overline{X})(X_j - \overline{X}),$$

где $\overline{X}$ — выборочное среднее наблюдений

$$\overline{X} = \frac{1}{T}\sum_{j=1}^{T} X_j. \tag{10.4}$$

Если число наблюдений T достаточно велико, то коэффициенты корреляции r_k распределены по нормальному закону с параметрами

$$\mathbf{E}r_k = \rho_k, \qquad \mathbf{D}r_k = 1/T.$$

2. Проверка независимости временного ряда

При статистической обработке временных рядов (10.3) иногда возникает необходимость проверки независимости данных. Мы знаем, что при этом первый коэффициент корреляции ρ_1 исходного процесса X_t должен быть равен нулю. С помощью выборочного коэффициента корреляции r_1, являющегося оценкой ρ_1, мы можем проверить гипотезу H_0: $\rho_1 = 0$. Альтернативная гипотеза H_1: $\rho_1 \neq 0$.

Критерием для проверки гипотезы H_0 с уровнем значимости α является неравенство

$$|r_1| > \frac{1}{\sqrt{T}} z_\alpha, \qquad \frac{1}{\sqrt{2\pi}}\int_{-z_\alpha}^{z_\alpha} e^{-\frac{\lambda^2}{2}} d\lambda = 1 - \alpha. \tag{10.5}$$

Если для выборочного коэффициента корреляций r_1 неравенство (10.5) имеет место, то с вероятностью $1-\alpha$ мы отвергаем гипотезу H_0 о независимости наблюдений, так как величина r_1 попала в критическую область.

Пример 10.1. Пусть $T = 16$, $r_1 = 0.6$, $\alpha = 0.05$. Тогда $z_\alpha = 1.96$, а правая часть неравенства (10.5) равна 0.49. Так как $r_1 > 0.49$, то мы отвергаем гипотезу H_0 о независимости данных и считаем, что данные зависимы.

3. Процесс авторегрессии порядка n ($\mathbf{AP}(n)$)

Определение. Процессом *авторегрессии* порядка n называется стационарный случайный процесс X_t, который удовлетворяет разностному уравнению

$$X_t + a_1 X_{t-1} + \ldots + a_n X_{t-n} = \beta \xi_t, \tag{10.6}$$

здесь ξ_t — независимые случайные величины, причем

$$\mathbf{E}\xi_t = 0, \quad D\xi_t = 1, \quad \mathbf{E}\xi_t \xi_s = 0 \quad \text{при} \quad t \neq s. \tag{10.7}$$

Соотношение (10.6) можно интерпретировать следующим образом: процесс X_t (в момент времени t) определяется предыдущими n значениями X_{t-1}, $X_{t-2}, \ldots, X_{t-n}$ процесса и случайной добавкой ξ_t, которая появляется в момент времени t, т. е.

$$X_t = -a_1 X_{t-1} - a_2 X_{t-2} - \ldots - a_n X_{t-n} + \beta \xi_t,$$

С процессом (10.6) связывают *характеристический полином*

$$\varphi(z) = a_0 + a_1 z + \cdots + a_n z^n = 0, \quad a_0 = 1. \tag{10.8}$$

Полином (10.8) имеет степень n и, следовательно, должен иметь n корней. Процесс (10.6) является стационарным только в том случае, когда все корни полинома $\varphi(z)$ по модулю больше единицы.

Так как случайные величины ξ_t независимы, то они не зависят от предыстории и, в частности, от X_{t-k}, $k = 1, 2, \ldots$, а отсюда следует:

$$\mathbf{E}X_{t-k}\mathbf{E}\xi_t = 0 \quad \text{при} \quad k = 1, 2, \ldots. \tag{10.9}$$

Умножим (10.6) на X_{t-k} и возьмем математическое ожидание от обеих частей равенства:

$$\mathbf{E}X_t X_{t-k} + a_1 \mathbf{E}X_{t-1} X_{t-k} + \ldots + a_n \mathbf{E}X_{t-n} X_{t-k} = \beta \mathbf{E}\xi_t X_{t-k}. \tag{10.10}$$

При $k \geq 1$ в силу равенств (10.9) математическое ожидание правой части (10.10) равно нулю, а левая часть при $k = 1, 2, \ldots, n$ с учетом того, что $R(k) = R(-k)$, приводит к системе равенств

$$\begin{array}{rl} a_0 R(1) + a_1 R(0) + a_2 R(1) + \ldots + a_n R(n-1) = & 0, \\ a_0 R(2) + a_1 R(1) + a_2 R(0) + \ldots + a_n R(n-2) = & 0, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots & \ldots \\ a_0 R(n) + a_1 R(n-1) + a_2 R(n-2) + \ldots + a_n R(0) = & 0. \end{array} \tag{10.11}$$

Соотношения (10.11) называются *уравнениями Юла—Уокера.* Они связывают коэффициенты авторегрессии

$$a_0, a_1, \ldots, a_n \tag{10.12}$$

с корреляциями процесса

$$R(0), R(1), \ldots, R(n). \tag{10.13}$$

Если известны корреляции (10.13), то можно найти коэффициенты авторегрессии (10.12) путем решения системы уравнений (10.11) относительно $a_1, \ldots, a_n$ $(a_0 = 1)$.

Замечание 2.1. Покажем, как найти коэффициент β. Умножим обе части уравнения (10.6) на ξ_t и возьмем математическое ожидание от обеих частей равенства:

$$\mathbf{E}X_t\xi_t + a_1\mathbf{E}X_{t-1}\xi_t + \cdots + a_n\mathbf{E}X_{t-n}\xi_t = \beta\mathbf{E}\xi_t^2.$$

Последнее выражение с учетом (10.9) и (10.7) и при исключении нулевых слагаемых приводится к виду

$$\mathbf{E}X_t\xi_t = \beta. \tag{10.14}$$

Теперь (10.6) умножим на X_t и возьмем математическое ожидание от обеих частей равенства:

$$\mathbf{E}X_t^2 + a_1\mathbf{E}X_{t-1}X_t + \cdots + a_n\mathbf{E}X_{t-n}X_t = \beta\mathbf{E}\xi_t X_t.$$

Это равенство с учетом (10.14) можно переписать в виде

$$R(0) + a_1R(1) + \ldots + a_nR(n) = \beta^2.$$

Отсюда

$$\beta = (R(0) + a_1R(1) + \ldots + a_nR(n))^{1/2}. \tag{10.15}$$

4. Оценка параметров процесса авторегрессии

Мы покажем, как для ряда наблюдений (10.2) подобрать модель авторегрессии и оценить ее параметры. Найдем выборочные корреляции ряда (10.2) по формулам (10.3):

$$\hat{R}(0),\ \hat{R}(1), \ldots, \hat{R}(n).$$

Как известно, процесс авторегрессии порядка n — это стационарный случайный процесс X_t, который удовлетворяет разностному уравнению (10.6).

Известно, что уравнения Юла — Уокера (10.11) связывают коэффициенты авторегрессии (10.12) с корреляциями процесса (10.13), что дает возможность оценить эти параметры, используя выборочные корреляции.

Запишем уравнения Юла — Уокера, в которых вместо корреляций $R(k)$ подставим их оценки $\hat{R}(k)$:

$$\begin{aligned} a_0\hat{R}(1) + a_1\hat{R}(0) + a_2\hat{R}(1) + \ldots + a_n\hat{R}(n-1) &= 0, \\ a_0\hat{R}(2) + a_1\hat{R}(1) + a_2\hat{R}(0) + \ldots + a_n\hat{R}(n-2) &= 0, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots &\ldots \\ a_0\hat{R}(n) + a_1\hat{R}(n-1) + a_2\hat{R}(n-2) + \ldots + a_n\hat{R}(0) &= 0. \end{aligned}$$

Решив эту систему линейных уравнений относительно коэффициентов (10.12), получим их оценки

$$\hat{a}_0, \hat{a}_1, \cdots, \hat{a}_n, \quad \hat{a}_0 = 1.$$

Далее следует оценить коэффициент b_0, используя равенство (10.15):

$$\hat{b}_0 = (\hat{R}(0) + \hat{a}_1\hat{R}(1) + \ldots + \hat{a}_n\hat{R}(n))^{1/2}.$$

Итак, к исходным данным (10.2) мы подобрали процесс авторегрессии

$$X_t + \hat{a}_1 X_{t-1} + \cdots + \hat{a}_n X_{t-n} = \hat{b}_0\xi_t.$$

Наконец, следует удостовериться, что полученный процесс является стационарным, а для этого достаточно проверить, что корни характеристического полинома

$$\hat{a}_0 + \hat{a}_1 z + \cdots + \hat{a}_n z^n = 0, \quad \hat{a}_0 = 1.$$

по модулю больше единицы.

Если оказалось, что хотя бы один корень по модулю меньше единицы, то модель авторегрессии порядка n — нестационарная. Стационарную модель можно пытаться получить, увеличив число наблюдений и/или изменив порядок n.

Пример 10.2. Пусть найдены оценки корреляций $\hat{R}(0) = 4.00$, $\hat{R}(1) = 3.05$, $\hat{R}(2) = 1.42$, и мы хотим подобрать модель авторегрессии второго порядка.

Решая систему

$$4.00\hat{a}_1 + 3.05\hat{a}_2 = -3.05,$$
$$3.05\hat{a}_1 + 4.00\hat{a}_2 = -2.03,$$

получаем оценки $\hat{a}_1 = -0.90$, $\hat{a}_2 = 0.18$. Характеристический полином имеет вид $\varphi(z) = 1 - 0.90z + 0.18z^2$. Корни полинома равны $z_1 = 1.67$, $z_2 = 3.33$, и очевидно, что они по модулю больше единицы, следовательно, модель является стационарной. Коэффициент $\hat{b}_0$, определяемый в случае уравнения второго порядка соотношением $\hat{b}_0 = \sqrt{\hat{R}_0 + \hat{a}_1\hat{R}_1 + \hat{a}_2\hat{R}_2}$, в нашем случае равен $\hat{b}_0 = \sqrt{4.00 - 0.90 \cdot 3.05 + 0.18 \cdot 1.43} = 1.23$, а само уравнение авторегрессии принимает вид

$$X_t - 0.90X_{t-1} + 0.18X_{t-2} = 1.23\xi_t.$$

5. Общая постановка прогноза процессов

Пусть имеется реализация дискретного процесса X_t объемом T

$$X_1,\ X_2, \cdots, X_T. \tag{10.2}$$

Прогноз процесса от момента T на τ шагов вперед — это предсказание величины $X_{T+\tau}$, и мы его обозначим $\hat{X}(T, \tau)$.

Мы будем искать прогноз в виде линейной комбинации величин (10.2), т. е. в виде

$$\hat{X}(T, \tau) = \sum_{k=1}^{T} \alpha_k X_k. \tag{10.16}$$

Введем понятие *ошибки прогноза* $\sigma(\tau)$, определяемой формулой

$$\sigma^2(\tau) = \mathbf{E}|X_{T+\tau} - \hat{X}(T, \tau)|^2.$$

Прогноз называется наилучшим, если его ошибка минимальна:

$$\sigma^2(\tau) = \min_{\hat{X}(T,\tau)} \mathbf{E}|X_{T+\tau} - \hat{X}(T, \tau)|^2.$$

Минимум берется по всевозможным функциям прогноза, имеющим вид (10.16). Минимальную ошибку $\sigma(\tau)$ называют *среднеквадратической ошибкой прогноза*.

Прогноз $\hat{X}(T,\tau)$ будет получен в предположении $\mathbf{E}X_t = 0$. Если же у прогнозируемого процесса $\mathbf{E}X_t \neq 0$, то в полученные формулы прогноза следует вместо X_t подставлять $X_t - \overline{X}$ (где $\overline{X}$ вычисляется по формуле (10.4)), а сам прогноз будет равен $\hat{X}(T,\tau) + \overline{X}$.

Естественным является утверждение, что наилучшим прогнозом на время τ вперед при $\tau \to \infty$ должно быть математическое ожидание процесса.

Прогноз последовательности одинаково распределенных независимых случайных величин на любое число шагов вперед равен их математическому ожиданию, а ошибка прогноза — среднеквадратичному отклонению. Для зависимых величин максимальная величина ошибки наилучшего линейного прогноза не превосходит среднеквадратического отклонения прогнозируемого процесса.

6. Прогноз процесса авторегрессии

Процесс авторегрессии в момент времени $T+1$ можно записать в виде

$$X_{T+1} = -a_1 X_T - a_2 X_{T-1} - \ldots - a_n X_{T-n+1} + b_0 \xi_{T+1}. \qquad (10.17)$$

Для прогнозирования X_{T+1} достаточно найти прогноз правой части (10.17). Но в нее входят величины $X_T, X_{T-1}, \ldots, X_{T-n+1}$, которые нам известны (следовательно, их прогнозировать не нужно), и не зависящая от них случайная величина ξ_{T+1}, для которой, как мы уже знаем, наилучший прогноз равен нулю. Следовательно, наилучший прогноз величины X_{T+1} от момента T на один шаг

$$\hat{X}(T,1) = -a_1 X_T - a_2 X_{T-1} - \ldots - a_n X_{T-n+1}. \qquad (10.18)$$

Чтобы найти ошибку прогноза на 1 шаг, мы найдем математическое ожидание квадрата разности между правыми частями равенств (10.17) и (10.18)

$$\sigma^2(1) = \mathbf{E}\left|\hat{X}(T,1) - X_{T+1}\right|^2 = \mathbf{E}\left|b_0 \xi_{T+1}\right|^2 = b_0^2.$$

Чтобы найти наилучший прогноз на 2 шага, запишем X_{T+2} в виде

$$X_{T+2} = -a_1 X_{T+1} - a_2 X_T - \ldots - a_n X_{T-n+2} + b_0 \xi_{T+2} \qquad (10.19)$$

и заметим, что в правой части (10.19) величины

$$X_T,\ X_{T-1},\ \ldots X_{T-n+2}$$

нам известны, прогноз ξ_{T+2}, как и прежде, равен нулю, а наилучший прогноз X_{T+1} равен $\hat{X}(T,1)$ и имеет вид (10.19), следовательно,

$$\hat{X}(T,2) = -a_1 \hat{X}(T,1) - a_2 X_T - \ldots - a_n X_{T-n+2}.$$

Далее последовательно находим прогнозы

$$\hat{X}(T,3) = -a_1 \hat{X}(T,2) - a_2 \hat{X}(T,1) - a_3 X_T - a_4 X_{T-1} - \ldots a_n X_{T-n+3},$$

$\hat{X}(T,4)$ и т. д., заменяя неизвестные величины $X_{T+1}, X_{T+2}, \ldots$ в правой части их прогнозами $\hat{X}(T,1), \hat{X}(T,2), \ldots$ Прогноз на число шагов $\tau > n$ будет полностью состоять их предыдущих прогнозов

$$\hat{X}(T,\tau) = -a_1 \hat{X}(T,\tau-1) - a_2 \hat{X}(T,\tau-2) - \ldots - a_n \hat{X}(T,\tau-n), \quad \tau > n.$$

Данный метод получения прогноза на k шагов требует предварительного подсчета прогнозов на 1, $2, \ldots, k-1$ шаг. Запишем формулы для ошибок прогноза на $\tau = 2, 3, 4$ шага:

$$\begin{aligned} \sigma^2(2) &= \sigma^2(1) + b_0^2 a_1^2, \\ \sigma^2(3) &= \sigma^2(2) + b_0^2 (a_1^2 - a_2)^2, \\ \sigma^2(4) &= \sigma^2(3) + b_0^2 (a_1^3 - 2a_1 a_2 + a_3)^2. \end{aligned} \qquad (10.20)$$

7. Процесс авторегрессии второго порядка

Это процесс, удовлетворяющий уравнению

$$X_t + a_1 X_{t-1} + a_2 X_{t-2} = b_0 \xi_t,$$

причем для его стационарности необходимо, чтобы корни характеристического полинома

$$\varphi(z) = 1 + a_1 z + a_2 z^2$$

были по модулю больше единицы.

Приведем формулы прогноза на 1, 2 и k $(k > 2)$ шагов:

$$\hat{X}(T,1) = -a_1 X_T - a_2 X_{T-1},$$
$$\hat{X}(T,2) = -a_1 \hat{X}(T,1) - a_2 X_T,$$
$$\hat{X}(T,k) = -a_1 \hat{X}(T,k-1) - a_2 \hat{X}(T,k-2).$$

Ошибки прогноза следует вычислять по формулам (10.20), считая $a_3 = 0$, так как коэффициент a_3 отсутствует у процесса авторегрессии второго порядка.

Если имеются наблюдения (10.2) и нужно подобрать параметры процесса авторегрессии второго порядка, то сначала следует найти оценки $\hat{R}(0)$, $\hat{R}(1)$, $\hat{R}(2)$. Затем, решив систему уравнений

$$\hat{R}(1) + \hat{a}_1 \hat{R}(0) + \hat{a}_2 \hat{R}(1) = 0,$$
$$\hat{R}(2) + \hat{a}_1 R(1) + \hat{a}_2 \hat{R}(0) = 0,$$

нужно найти $\hat{a}_1$ и $\hat{a}_2$ и коэффициент $\hat{b}_0$ по формуле

$$\hat{b}_0 = (\hat{R}(0) + \hat{a}_1 \hat{R}(1) + \hat{a}_2 \hat{R}(2))^{1/2}.$$

В результате к исходным данным подобрана модель

$$X_t + \hat{a}_1 X_{t-1} + \hat{a}_2 X_{t-2} = \hat{b}_0 \xi_t.$$

Далее следует убедиться, что модель является стационарной, т. е. что у полинома

$$\varphi(z) = 1 + \hat{a}_1 z + \hat{a}_2 z^2$$

корни по модулю больше единицы, и после этого можно приступать к прогнозу процесса.

Пример 10.3. Найти прогноз и ошибку прогноза на 1–3 шага вперед для процесса авторегрессии из примера 10.2

$$X_t - 0.90 X_{t-1} + 0.18 X_{t-2} = b_0 \xi_t, \quad b_0 = 1.23, \quad \sigma^2 = 4.00,$$

по наблюдениям $X_1 = 3$, $X_2 = -4$, $X_3 = -1$, $X_4 = 2$. В нашем случае $T = 4$, $n = 2$. Прогноз на 1, 2 и 3 шага соответственно

равен

$$\hat{X}(T,1) = -a_1 X_T - a_2 X_{T-1} = 0.90 \cdot 2 - 0.18 \cdot (-1) = 0.36,$$
$$\hat{X}(T,2) = 0.90 \cdot 0.36 - 0.18 \cdot 2 = 0.29,$$
$$\hat{X}(T,3) = 0.90 \cdot 0.29 - 0.18 \cdot 0.36 = 0.19.$$

Найдем ошибки прогноза на 1, 2 и 3 шага:

$$\sigma^2(1) = b_0^2 = 1.51,$$
$$\sigma^2(2) = b_0^2(1 + a_1^2) = 2.73,$$
$$\sigma^2(3) = 2.73 + 1.51((-0.90)^2 - 0.18)^2 = 3.34.$$

Видим, что ошибка прогноза растет вместе с τ. Напомним, что дисперсия процесса $\sigma^2 = 4.00$, следовательно, $\sigma^2(\tau) < \sigma^2$.

Рекомендуемая литература

1. *Hanke J.E., Reitsch A.G.* Understanding busness statistics. IRWIN. Homewood. IL 60430. Boston MA 021116, 1991. 878 p.
2. *Смирнов Н.В., Дунин-Барковский И.В.* Краткий курс математической статистики для технических приложений. М.: Физматгиз, 1959. 436 с.
3. *Кендалл М.Дж., Стьюарт А.* Статистические выводы и связи. М.: Наука, 1973. 900 с.
4. *Triola M.F., Franklin M.A.* Business statistics. Addison-Wesley Pub. Comp., 1995. 823 p.
5. *Кремер Н.Ш.* Теория вероятностей и математическая статистика. М.: Высшая школа. 2000. 543 с.
6. *Гмурман В.Е.* Теория вероятностей и математическая статистика. М.: Высшая школа. 1997.
7. *Калинина В.Н., Панкин В.Н.* Математическая статистика. М.: Высшая школа. 1997.

О г л а в л е н и е

ВВЕДЕНИЕ 3

Г л а в а 1. ОПИСАТЕЛЬНАЯ СТАТИСТИКА 4

1. Выборка 4
2. Гистограмма и эмпирическая функция распределения 6
3. Числовые характеристики выборки 9
4. Характеристики положения 10
5. Меры разброса 11
6. Анализ характера разброса 12
7. Моменты выборочного среднего и выборочной дисперсии 14
8. О качестве оценок в описательной статистике 15
9. Роль нормального распределения в статистике 16

Г л а в а 2. СТАТИСТИЧЕСКОЕ ОЦЕНИВАНИЕ 17

1. Постановка задачи статистического оценивания. Несмещенные, состоятельные и эффективные оценки 17
2. Метод моментов 23
3. Метод максимального правдоподобия 25
4. Некоторые распределения, связанные нормальным распределением 29
5. Некоторые свойства среднего и дисперсии выборки из нормальной совокупности 32
6. Интервальное (доверительное) оценивание. Доверительный интервал 34

Г л а в а 3. СТАТИСТИЧЕСКАЯ ПРОВЕРКА ГИПОТЕЗ 41

1. Выбор критерия значимости 41
2. Проверка гипотез о параметрах распределения 46
3. Проверка гипотез о типе распределения 52

Г л а в а 4. ДВУМЕРНОЕ НОРМАЛЬНОЕ РАСПРЕДЕЛЕНИЕ 58

1. Свойства двумерного нормального распределения 58
2. Построение доверительного множества для математического ожидания 63
3. Проверка гипотезы для математического ожидания 64

Г л а в а 5. КОРРЕЛЯЦИОННЫЙ И РЕГРЕССИОННЫЙ АНАЛИЗ 65

1. Понятие ковариации и коэффициента корреляции. Их оценивание 66
2. Проверка гипотезы об отсутствии корреляционной связи 69
3. Общая постановка задачи линейного регрессионного анализа 70
4. Простая линейная регрессия 72
5. Оценка параметров линейной регрессии по методу наименьших квадратов 73
6. Коэффициент простой детерминации 74
7. Прогноз значения Y_{x_0} в точке x_0 76

8. Проверка гипотезы о равенстве нулю коэффициента наклона β_1 77
9. Доказательство основных формул простого регрессионного анализа 78

Глава 6. НЕПАРАМЕТРИЧЕСКИЕ МЕТОДЫ СТАТИСТИКИ 81

1. Ранговые критерии для проверки гипотезы однородности 81
2. Ранговые критерии Уилкоксона и Манна — Уитни 81
3. Критерий Уилкоксона. Общий случай 83
4. Знаковый критерий Уилкоксона 87
5. Ранговые критерии независимости 91

Глава 7. ДИСПЕРСИОННЫЙ АНАЛИЗ 94

1. Однофакторный дисперсионный анализ 95
2. Двухфакторный дисперсионный анализ 98
3. Двухфакторный дисперсионный анализ без повторений 99
4. Двухфакторный дисперсионный анализ с повторениями 102

Глава 8. СТАТИСТИЧЕСКИЙ КОНТРОЛЬ КАЧЕСТВА 105

1. Введение 105
2. Контрольные карты количественных признаков при известных μ и σ ... 106
3. Контрольные карты качественных признаков 114
4. Статистический приемочный контроль качества неколичественных признаков 120

Глава 9. АНАЛИЗ ВРЕМЕННЫХ РЯДОВ 123

1. Декомпозиция временного ряда 124
2. Оценивание компонент временного ряда. Параметрическое оценивание тренда 125
3. Непараметрическое оценивание тренда. Сглаживание 125
4. Сезонные колебания 127
5. Прогноз процесса 130

Глава 10. СТАЦИОНАРНЫЕ ВРЕМЕННЫЕ РЯДЫ. МОДЕЛЬ АВТОРЕГРЕССИИ 134

1. Оценка автокорреляционной функции 134
2. Проверка независимости временного ряда 136
3. Процесс авторегрессии порядка n (AP(n)) 137
4. Оценка пареметров процесса авторегрессии 138
5. Общая постановка прогноза процессов 140
6. Прогноз процесса авторегрессии 141
7. Процесс авторегрессии второго порядка 142

Рекомендуемая литература 144

Учебное издание

Михаил Сергеевич Ермаков
Алла Филипповна Сизова
Татьяна Михайловна Товстик

ЭЛЕМЕНТЫ МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ

Учебное пособие

Зав. редакцией *Г.И. Чередниченко*

Лицензия ЛР N 040050 от 15.08.96.

Подписано в печать с оригинала-макета 21.05.2001.
Ф-т 60 × 84/16. Печать офсетная. Усл. печ. л. 8,6.
Уч.-изд. л. 8,72. Тираж 200. Заказ N

Редакция оперативной подготовки изданий
Издательства С.-Петербургского университета.
199034, С.-Петербург, Университетская наб., 7/9.

ЦОП типографии Издательства СПбГУ.
199034, С.-Петербург, наб. Макарова, 6.