Министерство науки и высшего образования Российской Федерации

САНКТ-ПЕТЕРБУРГСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ ПЕТРА ВЕЛИКОГО

Физико-механический институт
Высшая школа прикладной математики и вычислительной физики

С.А. Бабахина, А.Н. Баженов

ПРАКТИКУМ ПО МАТЕМАТИЧЕСКОЙ СТАТИСТИКЕ

Учебное пособие

Санкт-Петербург
2026

Оглавление

Введение

Математическая статистика — это раздел математики, связанный с обработкой данных. В ее задачи входят методы сбора, систематизации, анализа и использования данных наблюдений для изучения закономерностей случайных явлений, построения моделей, оценки параметров и проверки гипотез, опирающиеся в основном на теорию вероятностей для обеспечения точности выводов, сделанных по ограниченному объёму данных. Каждый исследователь, инженер, аналитик неизбежно использует понятия и методы математической статистики в своей деятельности.

При этом статистика, в отличие от теории вероятности, не является строгой теорией, а представляет набор инструментов для решения различных задач. Некоторые из таких инструментов имеют теоретическое обоснование в пределе бесконечных совокупностей данных, другие являются чисто эмпирическими рациональными приемами.

Пособие предназначено для учащихся 3-го курса Высшей школы математики и прикладной физики Физико-Механического института Санкт-Петерубргского университета Петра Великого по специальности «системное программирование» на основе опыта преподавания в 2020-2025 гг.

Методической основой материала являются книги [6] в области теоретико-вероятностной математической статистики и [8] в области интервальной статистики.

Глава 1

Описательная статистика

1.1 Распределения

Формулы *плотности* вероятности, а также основные характеристики классических *распределений* . Для непрерывных распределений указаны математические ожидания m_x, дисперсии D_x и, если они существуют, моды Mo. Для дискретных - моменты.

Нормальное распределение. Является предельным для распределения суммы независимых случайных величин, рассматриваемых в центральной предельной теореме. Описывает распределение различных выборочных средних, ошибок измерения, параметров деталей, координат точки падения снаряда, величины шума в управляющем устройстве

$$N(x, \mu, \sigma) = \frac{1}{\sigma \cdot \sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2 \cdot \sigma^2}}, \quad \mu \in \mathbb{R}, \sigma > 0 \tag{1.1}$$

m_x	D_x	Mo
μ	σ^2	μ

Таблица 1.1. Парметры нормального распределения

Распределение Коши. Является распределением отношения двух нормальных случайных величин с нулевыми математическими ожиданиями.

$$C(x, x_0, \gamma) = \frac{1}{\pi \cdot \gamma} \frac{1}{1 + \dfrac{(x - x_0)^2}{\gamma^2}}, \quad x_0 \in \mathbb{R}, \gamma > 0 \tag{1.2}$$

где x_0 - параметр сдвига, γ - параметр масштаба. Если $x_0 = 0$ и $\gamma = 1$, то такое распределение называется стандартным распределением Коши.

m_x	D_x	Mo
-	-	x_0

Таблица 1.2. Парметры распределения Коши

Распределение Лапласа. Является распределением случайной величины $X = X_1 - X_2 + m$, где X_1 И X_2 - независимые показательно распределенные случайные величины с параметром α

$$L(x, \beta, \alpha) = \tfrac{\alpha}{2} e^{-\alpha|x-\beta|}, \quad \beta \in \mathbb{R},\ \alpha > 0 \tag{1.3}$$

m_x	D_x	Mo
β	$2/\alpha^2$	β

Таблица 1.3. Парметры распределения Лапласа

Распределение Пуассона. Применяется для моделирования входящего потока заявок в системах массового обслуживания.

$$P(k, \lambda) = \tfrac{\lambda^k}{k!} e^{-\lambda}, \quad k = 0, 1, \ldots;\ \lambda > 0 \tag{1.4}$$

m_x	D_x
λ	λ

Таблица 1.4. Парметры распределения Пуассона

Равномерное распределение. Применяется для моделирования распределения ошибок округления, ошибок отсчета по приборам стрелочного типа. Равномерное распределение на отрезке $[0, 1]$ является стандартным, табулировано и по специальным программам может быть преобразовано в другие распределения. Оно же применяется в статистической практике для образования выборок.

$$U(x, a, b) = \begin{cases} \dfrac{1}{b-a} & \text{при} \quad x \in [a, b] \\ 0 \quad \text{при} \quad x \notin [a, b], \quad a, b \in \mathbb{R},\ a < b \end{cases} \tag{1.5}$$

m_x	D_x
$(a+b)/2$	$(b-a)^2/12$

Таблица 1.5. Парметры нормального распределения

1.2 Вариационный ряд

Определение. *Вариационным рядом* называется последовательность элементов выборки, расположенных в неубывающем порядке. Одинаковые элементы повторяются [1, с. 409].

Форма записи вариационного ряда: $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$.

Элементы вариационного ряда $x_{(i)} \quad (i = 1, 2, \ldots, n)$ называются *порядковыми статистиками.*

1.3 Выборочные числовые характеристики

С помощью выборки образуются её *числовые характеристики* . Это числовые характеристики дискретной случайной величины X^*, принимающей выборочные значения $x_1, x_2, \ldots, x_n$ [1, с. 411].

1.3.1 Характеристики положения

1. *Выборочное среднее*

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{1.6}$$

2. *Выборочная медиана*

$$med\, x = \begin{cases} x_{(l+1)} & \text{при} \quad n = 2l+1 \\ \dfrac{x_{(l)} + x_{(l+1)}}{2} & \text{при} \quad n = 2l \end{cases} \tag{1.7}$$

3. *Полусумма экстремальных выборочных элементов*

$$z_R = \frac{x_{(1)} + x_{(n)}}{2} \tag{1.8}$$

4. *Выборочный квартиль z_p порядка p*

$$z_p = \begin{cases} x_{([np]+1)} & \text{при } np \text{ дробном,} \\ x_{(np)} & \text{при } np \text{ целом.} \end{cases} \tag{1.9}$$

5. *Полусумма квартилей*

$$z_Q = \frac{z_{1/4} + z_{3/4}}{2} \tag{1.10}$$

6. *Усечённое среднее*

$$z_{tr} = \frac{1}{n-2r}\sum_{i=r+1}^{n-r} x_{(i)}, \quad r \approx \frac{n}{4} \tag{1.11}$$

1.3.2 Характеристики рассеяния

Выборочная дисперсия

$$D = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 \tag{1.12}$$

1.4 Эмпирическая функция распределения

Определение. *Статистическим рядом* называется последовательность различных элементов выборки $z_1, z_2, \ldots, z_k$, расположенных в возрастающем порядке с указанием частот $n_1, n_2, \ldots, n_k$, с которыми эти элементы содержатся в выборке.

Статистический ряд обычно записывается в виде таблицы

z	z_1	z_2	...	z_k
n	n_1	n_2	...	n_k

Таблица 1.6. Статистический ряд

Определение. *Эмпирической* (выборочной) *функцией распределения* (э. ф. р.) называется относительная частота события $X < x$, полученная по данной выборке:

$$F_n^*(x) = P^*(X < x). \tag{1.13}$$

Для получения относительной частоты $P^*(X < x)$ просуммируем в статистическом ряде, построенном по данной выборке, все частоты n_i, для которых элементы z_i статистического ряда меньше x. Тогда $P^*(X < x) = \frac{1}{n} \sum\limits_{z_i < x} n_i$. Получаем

$$F^*(x) = \tfrac{1}{n} \sum_{z_i < x} n_i. \tag{1.14}$$

$F^*(x)$ — функция распределения дискретной случайной величины X^*, заданной таблицей распределения

X^*	z_1	z_2	...	z_k
P	$\frac{n_1}{n}$	$\frac{n_2}{n}$	...	$\frac{n_k}{n}$

Таблица 1.7. Таблица распределения

Эмпирическая функция распределения является оценкой, т. е. приближённым значением, генеральной функции распределения

$$F_n^*(x) \approx F_X(x). \tag{1.15}$$

1.5 Боксплот Тьюки

Определение. *Боксплот* (англ. box plot) — график, использующийся в описательной статистике, компактно изображающий одномерное распределение вероятностей.

1.5.1 Описание

Боксплот содержит информацию о параметрах положения и масштаба, а также о весе «хвостов» распределения, асимметрии распределения данных и наличии выбросов. Несколько таких ящиков можно нарисовать бок о бок, чтобы визуально сравнивать одно распределение с другим; их можно располагать как горизонтально, так и вертикально. Расстояния между различными частями ящика позволяют определить степень разброса (дисперсии) и асимметрии данных и выявить выбросы [3].

Определение. *Выбросы* — экстремальные значения во входных данных, находящиеся далеко за пределами других наблюдений.

1.5.2 Построение

Стандартный одномерный боксплот Тьюки, как правило, определяется следующим набором параметров: минимумом и максимумом выборки, верхним LQ и нижним UQ квартилями, интерквартильным расстоянием IQR (оценка масштаба) и выборочной медианой. Помимо перечисленных параметров, определяемых непосредственно выборкой, на «качество» отсеивания аномальных значений влияет еще один *настраиваемый* параметр — k (как правило полагается равным 1.5 — оптимальному значению для нормально распределенной выборки). В привычном виде выражения для верхнего и нижнего усов бокслота примут вид:

$$X_1 = Q_1 - \frac{3}{2}(UQ - LQ), \;\; X_2 = Q_3 + \frac{3}{2}(UQ - LQ), \tag{1.16}$$

где X_1 — нижняя граница уса, X_2 — верхняя граница уса, k — настраиваемый параметр, LQ — нижний квартиль, UQ — верхний квартиль.

Повысить качество обнаружения аномальных значений и повлиять на чувствительность к выбросам можно с помощью параметров боксплота. В различных модификациях боксплота, как правило, варьируются оценка масштаба (O_n-оценка, медиана абсолютных отклонений MAD_n и т.д.) и значение параметра k, При необходимости могут быть также затронуты и границы «ящика» — LQ и UQ. Общая формула боксплота Тьюки имеет следующий вид:

$$f(n) = \begin{cases} X_1 = \max\{x_{(1)}, LQ - S \cdot k\} \\ X_2 = \min\{x_{(n)}, UQ + S \cdot k\} \end{cases} \tag{1.17}$$

Через S в общем виде обозначена оценка масштаба.

Данные, выходящие за границы усов (выбросы), как правило, отображаются на графике в виде маленьких кружков [3].

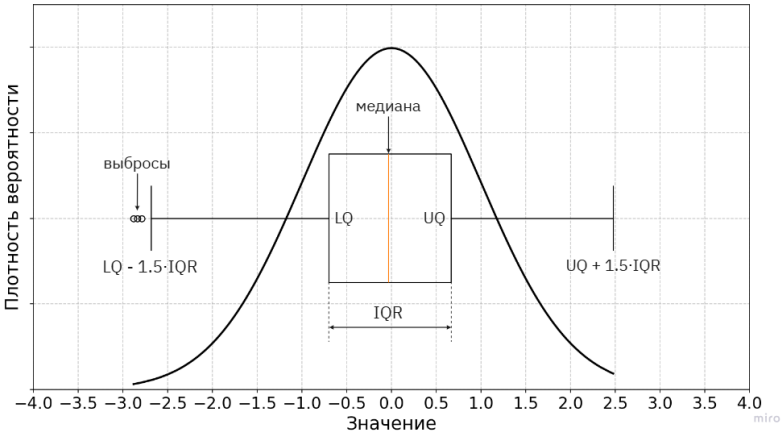

Рис. 1.1. Иллюстрация боксплота Тьюки

1.6 Гистограмма

Определение. *Гистограмма* в математической статистике — это функция, приближающая плотность вероятности некоторого распределения, построенная на основе выборки из него [2].

1.6.1 Описание

Гистограммы применяются в основном для визуализации данных на начальном этапе статистической обработки. Построение гистограмм используется для получения эмпирической *оценки плотности распределения* случайной величины. На них, в частности, построены различные способы обработки сигналов, изображений и других статистических объектов. Применение гистограмм к обработке экспериментальных данных наряду с боксплотами и прочими алгоритмами нахождения аномальных значений позволяет устранять артефакты - шумы и выбросы, мешающие работе с данными и не являющиеся содержательными

Выбросы на гистограмме будут формировать одиночные пики.

1.6.2 Построение

Для построения гистограммы наблюдаемый диапазон изменения случайной величины разбивается на несколько интервалов (разрядов, групп) и подсчитывается доля от всех измерений, попавшая в каждый из интервалов. Величина каждой доли, отнесенная к величине интервала, принимается в качестве оценки значения плотности распределения на соответствующем интервале [2]. Формализуем описанный алгоритм.

Пусть проведено n экспериментов; $\{x_i\}$, где $i = 1, \ldots, n$ - результаты, полученные в ходе экспериментов. Будем считать, что $\{x_1, \ldots, x_n\}$ - упорядоченная выборка. Таким образом, далее будем рассматривать случайную величину $X = \{x_1, \ldots, x_n\}$.

Обозначим диапазон значений X через $[a, b]$, где $a = x_1$, $b = x_n$, $x_i \in [a, b]$. Разобъем $[a, b]$ на $k \in \mathbb{N}$ интервалов:

$$[a = y_1, y_2],\ [y_2, y_3],\ \ldots,\ [y_k, y_{k+1} = b]$$

Далее для каждого разряда вычисляем число m_l, где $l = 1, \ldots, k$ — количество попаданий случайной величины x_i в этот разряд. В резуль-

тате для каждого разряда можно вычислить величину p_l — частоту попаданий в него случайной величины:

$$p_l = \frac{m_l}{n}, l = 1, \ldots, k \tag{1.18}$$

Полученные значения можно представить в табличном виде.

разряд	$[y_1, y_2]$	$[y_2, y_3]$	$\ldots$	$[y_k, y_{k+1}]$
частота	p_1	p_2	$\ldots$	p_k

Таблица 1.8. Частоты попадания x_i в разряды гистограммы

Таблицу 1.8 можно изобразить в виде графика: по оси x — разряды, по оси y — частоты. Полученный график и будет называться *гистограммой.*

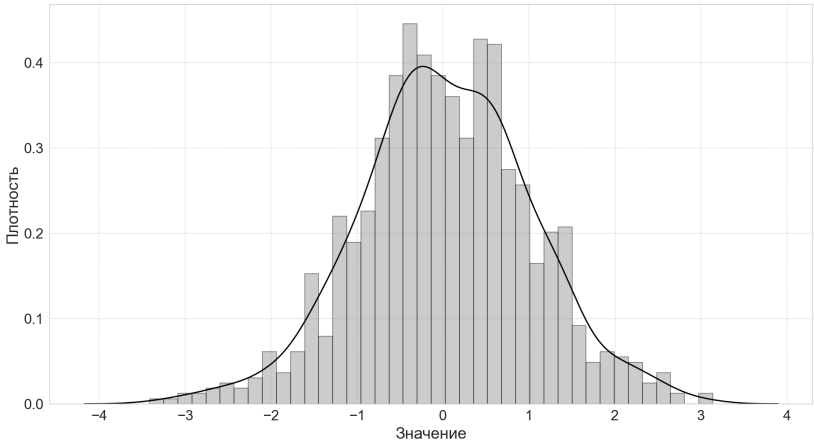

Рис. 1.2. Иллюстрация гистограммы

Нередко на гистограмме вместо частот отображают количество попаданий случайной величины в тот или иной разряд.

Аналогично таблице частот, по таблице 1.9 можно построить графическое отображение эмпирического распрделения X в виде гистограммы. Обе гистограммы изображают одно и то же и различаются только масштабом по оси y.

разряд	$[y_1, y_2]$	$[y_2, y_3]$	$\dots$	$[y_k, y_{k+1}]$
частота	m_1	m_2	$\dots$	m_k

Таблица 1.9. Количество попаданий x_i в разряды гистограммы

Замечание. При подсчете частоты (числа) попаданий x_i в разряды гистограммы может произойти ситуация, когда значение случайной величины попадает точно на граничные значения диапазона: $\exists i \in \{1, \dots, n\},\ \exists j \in \{2, \dots, k\} :\ x_i = y_j$. В таком случае правая граница всех диапазонов, кроме последнего, не включается в рассмотрение.

Глава 2

Точечное оценивание числовых характеристик

2.1 Общие теоретические сведения

Теоретические сведения раздела 2.1 заимствованы из учебного пособия по математической статистике Максимова Ю.Д [6]. Глава 2 издания 2009 года в подробностях повествует о получении точечных оценкок числовых характеристик и параметров распределения генеральной совокупности. В случае необходимости, с деталями доказательств в обширной теоретической информацией можно ознакомиться в учебном пособии. Далее будут освещены лишь некоторые определения из главы 2.

Определение. *Точечной статистической оценкой* неизвестной числовой характеристики или параметра θ распределения называется функция $\hat{\theta}_n(x_1, \ldots, x_n)$, зависящая от элементов выборки, приближенно равная θ:

$$\hat{\theta}_n(x_1, \ldots, x_n) \approx \theta \tag{2.1}$$

Для каждой конкретной выборки — это число, т. е. точка на числовой оси.

Определение. Оценка $\hat{\theta}_n = \hat{\theta}_n(x_1, \ldots, x_n)$ называется *состоятельной* оценкой θ, если она стремится по вероятности к θ с ростом n:

$$\hat{\theta}_n \xrightarrow[n\to\infty]{P} \theta. \tag{2.2}$$

Это означает, что для любого $\varepsilon > 0$ выполняется соотношение

$$\lim_{n\to\infty} \mathrm{P}(|\hat{\theta}_n - \theta| \geq \varepsilon) = 0 \tag{2.3}$$

Определение. $\hat{\theta}_n$ называется *несмещенной* оценкой θ, если математическое ожидание оценки равно θ:

$$\mathrm{M}\hat{\theta}_n = \theta. \tag{2.4}$$

В противном случае оценка называется смещенной. Разность $\mathrm{M}\hat{\theta}_n - \theta$. называется смещением оценки.

Для малых выборок требование несмещенности оценки является существенным. Для больших выборок требование несмещенности состоятельной оценки не столь важно, так как состоятельная оценка стремится к оцениваемой величине и поэтому смещение мало (стремится к нулю при $n \to \infty$).

Определение. Оценка $\hat{\theta}_n$ величины θ называется *робастной*, если она устойчива по отношению к выбросам в статистических данных.

Выбросы в выборке могут появиться вследствие сбоев регистрирующего прибора, грубых ошибок оператора. Выбросы группируются на концах вариационного ряда наблюдений. Поэтому оценки, не имеющие в своем составе элементов, близких к концам вариационного ряда, будут робастными. Это, например, выборочная медиана med и полусумма квартилей t_q .

2.2 Оценка плотности вероятности

Определение. *Оценкой плотности вероятности* $f(x)$ называется функция $\widehat{f}(x)$, построенная на основе выборки, приближённо равная $f(x)$

$$\widehat{f}(x) \approx f(x). \tag{2.5}$$

2.2.1 Ядерные оценки

Определение. Непрерывной *ядерной оценкой* [1, с. 421-423] будем называть оценку, равную сумме ядерных функций:

$$\widehat{f}_n(x) = \frac{1}{nh_n} \sum_{i=1}^{n} K\left(\frac{x-x_i}{h_n}\right). \tag{2.6}$$

Здесь функция $K(u)$, называемая ядерной (*ядром*), непрерывна и является плотностью вероятности, $x_1, \ldots, x_n$ — элементы выборки, $\{h_n\}$ — любая последовательность положительных чисел, обладающая свойствами

$$h_n \underset{n\to\infty}{\longrightarrow} 0; \qquad \frac{h_n}{n^{-1}} \underset{n\to\infty}{\longrightarrow} \infty. \tag{2.7}$$

Если плотность $f(x)$ кусочно-непрерывная, то ядерная оценка плотности является состоятельной при соблюдении условий, накладываемых на параметр сглаживания h_n, а также на ядро $K(u)$.

Гауссово (нормальное) ядро [4, с. 38]

$$K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}. \tag{2.8}$$

Правило Сильвермана [4, с. 44]

$$h_n = 1.06\hat{\sigma} n^{-1/5}, \tag{2.9}$$

где $\hat{\sigma}$ - выборочное стандартное отклонение.

2.3 Метод максимального правдоподобия

Одним из универсальных методов оценивания является *метод максимального правдоподобия*, предложенный Р.Фишером (1921).

Пусть $x_1, \ldots, x_n$ — случайная выборка из генеральной совокупности с плотностью вероятности $f(x, \theta)$; $L(x_1, \ldots, x_n, \theta)$ — функция правдоподобия (ФП), представляющая собой совместную плотность вероятности независимых с.в. $x_1, \ldots, x_n$ и рассматриваемая как функция неизвестного параметра θ:

$$L(x_1,\ \dots\ , x_n, \theta) = f(x_1, \theta)\ f(x_2, \theta)\ \dots\ f(x_n, \theta). \tag{2.10}$$

Определение. *Оценкой максимального правдоподобия* (о.м.п) будем называть такое значение $\widehat{\theta}_{\text{мп}}$ из множества допустимых значений параметра θ, для которого ФП принимает наибольшее значение при заданных $x_1,\ \dots\ , x_n$:

$$\widehat{\theta}_{\text{мп}} = \arg\max_{\theta} L(x_1,\ \dots\ , x_n, \theta). \tag{2.11}$$

Если ФП дважды дифференцируема, то её стационарные значения даются корнями уравнения

$$\frac{\partial L(x_1,\ \dots\ , x_n, \theta)}{\partial \theta} = 0. \tag{2.12}$$

Достаточным условием того, чтобы некоторое стационарное значение $\widetilde{\theta}$ было локальным максимумом, является неравенство

$$\frac{\partial^2 L}{\partial \theta^2}(x_1,\ \dots\ , x_n, \widetilde{\theta}) < 0. \tag{2.13}$$

Определив точки локальных максимумов ФП (если их несколько), находят наибольший, который и даёт решение задачи (2.10).

Часто проще искать максимум логарифма ФП, так как он имеет максимум в одной точке с ФП:

$$\frac{\partial \ln L}{\partial \theta} = \frac{1}{L}\frac{\partial L}{\partial \theta}, \quad \text{если} \quad L > 0,$$

и соответственно решать уравнение

$$\frac{\partial \ln L}{\partial \theta} = 0, \tag{2.14}$$

которое называют *уравнением правдоподобия*.

В задаче оценивания векторного параметра $\theta = (\theta_1,\ \dots\ , \theta_m)$ аналогично (2.11) находится максимум ФП нескольких аргументов:

$$\widehat{\theta}_{\text{мп}} = \arg\max_{\theta_1,\ \dots\ ,\theta_m} L(x_1,\ \dots\ , x_n, \theta_1,\ \dots\ , \theta_m), \tag{2.15}$$

и в случае дифференцируемости ФП выписывается система уравнений правдоподобия

$$\frac{\partial L}{\partial \theta_k} = 0 \quad \text{или} \quad \frac{\partial \ln L}{\partial \theta_k} = 0, \quad k = 1, \, \dots \, , m. \tag{2.16}$$

Пример 2.3.1. Оценивание м.о. m и дисперсии σ^2 нормального распределения $N(m, \sigma)$.

Составим функцию правдоподобия и получим последовательно:

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

$$= (2\pi\sigma^2)^{-n/2} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - m)^2 \right\},$$

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

Получим следующие уравнения правдоподобия:

$$\begin{cases} \frac{\partial \ln L}{\partial m} = \frac{1}{\widehat{\sigma}^2} \sum\limits_{i=1}^{n} (x_i - \widehat{m}) = \frac{n}{\widehat{\sigma}^2} (\overline{x} - \widehat{m}) = 0, \\ \frac{\partial \ln L}{\partial (\sigma^2)} = -\frac{n}{2\widehat{\sigma}^2} + \frac{1}{2(\widehat{\sigma}^2)^2} \sum\limits_{i=1}^{n} (x_i - \widehat{m})^2 = \\ = \frac{n}{2(\widehat{\sigma}^2)^2} \left[\frac{1}{n} \sum\limits_{i=1}^{n} (x_i - \widehat{m})^2 - \widehat{\sigma}^2 \right] = 0, \end{cases}$$

откуда следует, что выборочное среднее $\overline{x}$ — о.м.п. математического ожидания: $\widehat{m}_{\text{мп}} = \overline{x}$, а выборочная дисперсия $s^2 = \frac{1}{n} \sum\limits_{i=1}^{n} (x_i - \overline{x})^2$ — о.м.п. генеральной дисперсии: $\widehat{\sigma}^2_{\text{мп}} = s^2$ [1, с. 442-444].

■

Глава 3

Интервальное оценивание числовых характеристик

3.1 Доверительные интервалы для параметров нормального распределения

3.1.1 Доверительный интервал для математического ожидания m нормального распределения

Дана выборка $(x_1, x_2, \ldots, x_n)$ объёма n из нормальной генеральной совокупности. На её основе строим выборочное среднее $\overline{x}$ и выборочное среднее квадратическое отклонение s. Параметры m и σ нормального распределения неизвестны.

Доказано, что случайная величина

$$T = \sqrt{n-1} \cdot \frac{\overline{x} - m}{s}, \tag{3.1}$$

называемая *статистикой Стьюдента*, распределена по закону Стьюдента с $n-1$ степенями свободы. Пусть $f_T(x)$ — плотность вероятности этого распределения. Тогда

$$P\left(-x < \sqrt{n-1}\cdot\frac{\overline{x}-m}{s} < x\right) = P\left(-x < \sqrt{n-1}\cdot\frac{m-\overline{x}}{s} < x\right) =$$
$$= \int\limits_{-x}^{x} f_T(t)\,dt = 2\int\limits_{0}^{x} f_T(t)\,dt = 2\left(\int\limits_{-\infty}^{x} f_T(t)\,dt - \frac{1}{2}\right) = 2F_T(x) - 1.$$

Здесь $F_T(x)$ — функция распределения Стьюдента с $n-1$ степенями свободы.

Полагаем $2F_T(x) - 1 = 1 - \alpha$, где α — выбранный уровень значимости. Тогда $F_T(x) = 1 - \alpha/2$. Пусть $t_{1-\alpha/2}(n-1)$ — *квантиль распределения Стьюдента* с $n-1$ степенями свободы и порядка $1-\alpha/2$. Из предыдущих равенств мы получаем

$$\begin{aligned} &P\left(\overline{x} - \frac{sx}{\sqrt{n-1}} < m < \overline{x} + \frac{sx}{\sqrt{n-1}}\right) = 2F_T(x) - 1 = 1 - \alpha, \\ &P\left(\overline{x} - \frac{st_{1-\alpha/2}(n-1)}{\sqrt{n-1}} < m < \overline{x} + \frac{st_{1-\alpha/2}(n-1)}{\sqrt{n-1}}\right) = 1 - \alpha, \end{aligned} \tag{3.2}$$

Определение. Формула (3.2) даёт *доверительный интервал для* m с доверительной вероятностью $\gamma = 1 - \alpha$ [1, с. 457-458].

3.1.2 Доверительный интервал для среднего квадратического отклонения σ нормального распределения

Дана выборка $(x_1, x_2, \ \ldots\ , x_n)$ объёма n из нормальной генеральной совокупности. На её основе строим выборочную дисперсию s^2. Параметры m и σ нормального распределения неизвестны. Доказано, что случайная величина ns^2/σ^2 распределена по закону χ^2 с $n-1$ степенями свободы.

Задаёмся уровнем значимости α и, например, с помощью встроенных средств языка программирования R (функция qchisq) находим квантили $\chi^2_{\alpha/2}(n-1)$ и $\chi^2_{1-\alpha/2}(n-1)$. Это значит, что

$$P\left(\chi^2(n-1) < \chi^2_{\alpha/2}(n-1)\right) = \alpha/2;$$
$$P\left(\chi^2(n-1) < \chi^2_{1-\alpha/2}(n-1)\right) = 1-\alpha/2.$$

Тогда

$$P\left(\chi^2_{\alpha/2}(n-1) < \chi^2(n-1) < \chi^2_{1-\alpha/2}(n-1)\right) =$$
$$= P\left(\chi^2(n-1) < \chi^2_{1-\alpha/2}(n-1)\right) - P\left(\chi^2(n-1) < \chi^2_{\alpha/2}(n-1)\right) =$$
$$= 1-\alpha/2-\alpha/2 = 1-\alpha.$$

Отсюда

$$P\left(\chi^2_{\alpha/2}(n-1) < \frac{ns^2}{\sigma^2} < \chi^2_{1-\alpha/2}(n-1)\right) = P\left(\frac{1}{\chi^2_{1-\alpha/2}(n-1)} < \frac{\sigma^2}{ns^2} < \frac{1}{\chi^2_{\alpha/2}(n-1)}\right) =$$
$$= P\left(\frac{s\sqrt{n}}{\sqrt{\chi^2_{1-\alpha/2}(n-1)}} < \sigma < \frac{s\sqrt{n}}{\sqrt{\chi^2_{\alpha/2}(n-1)}}\right).$$

Окончательно

$$P\left(\frac{s\sqrt{n}}{\sqrt{\chi^2_{1-\alpha/2}(n-1)}} < \sigma < \frac{s\sqrt{n}}{\sqrt{\chi^2_{\alpha/2}(n-1)}}\right) = 1-\alpha, \qquad (3.3)$$

Определение. Формула (3.3) даёт *доверительный интервал для* σ с доверительной вероятностью $\gamma = 1-\alpha$ [1, с. 458-459].

3.2 Доверительные интервалы для математического ожидания m и среднего квадратического отклонения σ произвольного распределения при большом объёме выборки. Асимптотический подход

При большом объёме выборки для построения доверительных интервалов может быть использован *асимптотический метод* на основе

центральной предельной теоремы.

3.2.1 Доверительный интервал для математического ожидания m произвольной генеральной совокупности при большом объёме выборки

Выборочное среднее $\overline{x} = \sum\limits_{i=1}^{n} \frac{x_i}{n}$ при большом объёме выборки является суммой большого числа взаимно независимых одинаково распределённых случайных величин. Предполагаем, что исследуемое генеральное распределение имеет конечные математическое ожидание m и дисперсию σ^2. Тогда в силу центральной предельной теоремы центрированная и нормированная случайная величина $(\overline{x} - \mathbf{M}\overline{x})/\sqrt{\mathbf{D}\overline{x}} = \sqrt{n} \cdot (\overline{x} - m)/\sigma$ распределена приблизительно нормально с параметрами 0 и 1. Пусть

$$\Phi(x) = \tfrac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{x} e^{-t^2/2}\, dt \tag{3.4}$$

— *функция Лапласа*. Тогда

$$P\left(-x < \sqrt{n} \cdot \frac{\overline{x} - m}{\sigma} < x\right) = P\left(-x < \sqrt{n} \cdot \frac{m - \overline{x}}{\sigma} < x\right) \approx$$
$$\approx \Phi(x) - \Phi(-x) = \Phi(x) - [1 - \Phi(x)] = 2\Phi(x) - 1.$$

Отсюда

$$P\left(\overline{x} - \frac{\sigma x}{\sqrt{n}} < m < \overline{x} + \frac{\sigma x}{\sqrt{n}}\right) \approx 2\Phi(x) - 1. \tag{3.5}$$

Полагаем $2\Phi(x) - 1 = \gamma = 1 - \alpha$; тогда $\Phi(x) = 1 - \alpha/2$. Пусть $u_{1-\alpha/2}$ — квантиль нормального распределения $N(0, 1)$ порядка $1 - \alpha/2$. Заменяя в равенстве (3.5) σ на s, запишем его в виде

$$P\left(\overline{x} - \frac{s u_{1-\alpha/2}}{\sqrt{n}} < m < \overline{x} + \frac{s u_{1-\alpha/2}}{\sqrt{n}}\right) \approx \gamma, \tag{3.6}$$

Определение. Формула (3.6) даёт *доверительный интервал для* m с доверительной вероятностью $\gamma = 1 - \alpha$ [1, с. 460].

3.2.2 Доверительный интервал для среднего квадратического отклонения σ произвольной генеральной совокупности при большом объёме выборки

Выборочная дисперсия $s^2 = \sum\limits_{i=1}^{n} \frac{(x_i-\overline{x})^2}{n}$ при большом объёме выборки является суммой большого числа практически взаимно независимых случайных величин (имеется одна связь $\sum\limits_{i=1}^{n} x_i = n\overline{x}$, которой при большом n можно пренебречь). Предполагаем, что исследуемая генеральная совокупность имеет конечные первые четыре момента.

В силу центральной предельной теоремы центрированная и нормированная случайная величина $(s^2 - \mathbf{M}s^2)/\sqrt{\mathbf{D}s^2}$ при большом объёме выборки n распределена приблизительно нормально с параметрами 0 и 1. Пусть $\Phi(x)$ — функция Лапласа (3.4). Тогда

$$P\left(-x < \frac{s^2 - \mathbf{M}s^2}{\sqrt{\mathbf{D}s^2}} < x\right) \approx \Phi(x) - \Phi(-x) = \Phi(x) - [1 - \Phi(x)] = 2\Phi(x) - 1.$$

Положим $2\Phi(x) - 1 = \gamma = 1 - \alpha$. Тогда $\Phi(x) = 1 - \alpha/2$. Пусть $u_{1-\alpha/2}$ — корень этого уравнения — квантиль нормального распределения $N(0,1)$ порядка $1 - \alpha/2$. Известно, что $\mathbf{M}s^2 = \sigma^2 - \dfrac{\sigma^2}{n} \approx \sigma^2$ и $\mathbf{D}s^2 = \dfrac{\mu_4 - \mu_2^2}{n} + o\left(\dfrac{1}{n}\right) \approx \dfrac{\mu_4 - \mu_2^2}{n}$. Здесь μ_k — центральный момент k-го порядка генерального распределения; $\mu_2 = \sigma^2$; $\mu_4 = \mathbf{M}\left[(x - \mathbf{M}x)^4\right]$; $o\left(\dfrac{1}{n}\right)$ — бесконечно малая высшего порядка, чем $1/n$, при $n \to \infty$. Итак, $\mathbf{D}s^2 \approx \dfrac{\mu_4 - \sigma^4}{n}$. Отсюда

$$\mathbf{D}s^2 \approx \frac{\sigma^4}{n}\left(\frac{\mu_4}{\sigma^4} - 1\right) = \frac{\sigma^4}{n}\left(\left(\frac{\mu_4}{\sigma^4} - 3\right) + 2\right) = \frac{\sigma^4}{n}(E + 2) \approx \frac{\sigma^4}{n}(e + 2),$$

где $E = \dfrac{\mu_4}{\sigma^4} - 3$ — эксцесс генерального распределения, $e = \dfrac{m_4}{s^4} - 3$ — выборочный эксцесс; $m_4 = \frac{1}{n}\sum\limits_{i=1}^{n}(x_i - \overline{x})^4$ — четвёртый выборочный

центральный момент. Далее,

$$\sqrt{\mathbf{D}s^2} \approx \frac{\sigma^2}{\sqrt{n}}\sqrt{e+2}.$$

Преобразуем неравенства, стоящие под знаком вероятности в формуле

$$P\left(-x < \frac{s^2 - \mathbf{M}s^2}{\sqrt{\mathbf{D}s^2}} < x\right) = \gamma \ :$$
$$-\sigma^2 U < s^2 - \sigma^2 < \sigma^2 U;$$
$$\sigma^2(1-U) < s^2 < \sigma^2(1+U);$$
$$1/[\sigma^2(1+U)] < 1/s^2 < 1/[\sigma^2(1-U)];$$
$$s^2/(1+U) < \sigma^2 < s^2/(1-U);$$
$$s(1+U)^{-1/2} < \sigma < s(1-U)^{-1/2}, \tag{3.7}$$

где $U = u_{1-\alpha/2}\sqrt{(e+2)/n}$ или

$$s(1+u_{1-\alpha/2}\sqrt{(e+2)/n})^{-1/2} < \sigma < s(1-u_{1-\alpha/2}\sqrt{(e+2)/n})^{-1/2}.$$

Разлагая функции в биномиальный ряд и оставляя первые два члена, получим

$$s(1-0.5U) < \sigma < s(1+0.5U) \tag{3.8}$$

или

$$s(1-0.5u_{1-\alpha/2}\sqrt{(e+2)}/\sqrt{n}) < \sigma < s(1+0.5u_{1-\alpha/2}\sqrt{(e+2)}/\sqrt{n}).$$

Определение. Формулы (3.7) или (3.8) дают *доверительный интервал для* σ с доверительной вероятностью $\gamma = 1-\alpha$ [1, с. 461-462].

Замечание. Вычисления по формуле (3.7) дают более надёжный результат, так как в ней меньше грубых приближений.

Глава 4

Методы проверки статистических гипотез

4.1 Проверка гипотезы о законе распределения генеральной совокупности. Метод хи-квадрат

Исчерпывающей характеристикой изучаемой случайной величины является её закон распределения. Поэтому естественно стремление исследователей построить этот закон приближённо на основе статистических данных.

Сначала выдвигается гипотеза о виде закона распределения. После того как выбран вид закона, возникает задача оценивания его параметров и проверки (тестирования) закона в целом.

Для проверки гипотезы о законе распределения применяются *критерии согласия*. Таких критериев существует много. Мы рассмотрим наиболее обоснованный и наиболее часто используемый в практике — *критерий χ^2 (хи-квадрат)*, введённый К.Пирсоном (1900 г.) для случая, когда параметры распределения известны. Этот критерий был существенно уточнён Р.Фишером (1924 г.), когда параметры распределения оцениваются по выборке, используемой для проверки. Мы ограничимся рассмотрением случая одномерного распределения.

Итак, выдвинута гипотеза H_0 о генеральном законе распределения

с функцией распределения $F(x)$. Рассматриваем случай, когда гипотетическая функция распределения $F(x)$ не содержит неизвестных параметров. Разобьём генеральную совокупность, т.е. множество значений изучаемой случайной величины X на k непересекающихся подмножеств $\Delta_1, \Delta_2, \ldots, \Delta_k$.

Пусть $p_i = P(X \in \Delta_i), \quad i = 1, \ldots, k$. Если генеральная совокупность — вся вещественная ось, то подмножества $\Delta_i = (a_{i-1}, a_i]$ — полуоткрытые промежутки $(i = 2, \ldots, k-1)$. Крайние промежутки будут полубесконечными: $\Delta_1 = (-\infty, a_1]$, $\Delta_k = (a_{k-1}, +\infty)$. В этом случае $p_i = F(a_i) - F(a_{i-1})$; $a_0 = -\infty$, $a_k = +\infty$ $(i = 1, \ldots, k)$. Отметим, что $\sum\limits_{i=1}^{k} p_i = 1$. Будем предполагать, что все $p_i > 0$ $(i = 1, \ldots, k)$.

Пусть, далее, $n_1, n_2, \ldots, n_k$ — частоты попадания выборочных элементов в подмножества $\Delta_1, \Delta_2, \ldots, \Delta_k$ соответственно.

В случае справедливости гипотезы H_0 относительные частоты n_i/n при большом n должны быть близки к вероятностям p_i $(i = 1, \ldots, k)$, поэтому за меру отклонения выборочного распределения от гипотетического с функцией $F(x)$ естественно выбрать величину

$$\mathrm{Z} = \sum_{i=1}^{k} c_i \left(\frac{n_i}{n} - p_i \right)^2, \tag{4.1}$$

где c_i — какие-нибудь положительные числа (веса). К.Пирсоном в качестве весов выбраны числа $c_i = n/p_i$ $(i = 1, \ldots, k)$.

Определение. Тогда получается *статистика критерия хи-квадрат К.Пирсона*

$$\chi^2 = \sum_{i=1}^{k} \frac{n}{p_i} \left(\frac{n_i}{n} - p_i \right)^2 = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i}, \tag{4.2}$$

которая обозначена тем же символом, что и закон распределения хи-квадрат.

К.Пирсоном доказана теорема об асимптотическом поведении статистики χ^2, указывающая путь её применения.

Теорема. (К.Пирсона) Статистика критерия χ^2 асимптотически распределена по закону χ^2 с $k-1$ степенями свободы.

Это означает, что независимо от вида проверяемого распределения,

т.е. функции $F(x)$, выборочная функция распределения статистики χ^2 при $n \to \infty$ стремится к функции распределения случайной величины с плотностью вероятности

$$f_{k-1}(x) = \begin{cases} 0, & x \le 0; \\ \frac{1}{2^{\frac{k-1}{2}}\Gamma\left(\frac{k-1}{2}\right)} x^{\frac{k-3}{2}} e^{-\frac{x}{2}}, & x > 0. \end{cases} \tag{4.3}$$

Для прояснения сущности метода χ^2 сделаем ряд замечаний.

Замечание 1. Выбор подмножеств $\Delta_1, \Delta_2, \ldots, \Delta_k$ и их числа k в принципе ничем не регламентируется, так как $n \to \infty$. Но так как число n хотя и очень большое, но конечное, то k должно быть с ним согласовано. Обычно его берут таким же, как и для построения гистограммы, т.е. можно руководствоваться формулой

$$k \approx 1.72\sqrt[3]{n} \tag{4.4}$$

или формулой Старджесса

$$k \approx 1 + 3.3 \lg n. \tag{4.5}$$

При этом, если $\Delta_1, \Delta_2, \ldots, \Delta_k$ — промежутки, то их длины удобно сделать равными, за исключением крайних — полубесконечных.

Определение. *Числом степеней свободы* функции называется число её независимых аргументов.

Замечание 2. (о числе степеней свободы).
Аргументами статистики χ^2 являются частоты $n_1, n_2, \ldots, n_k$. Эти частоты связаны одним равенством $n_1 + n_2 + \ldots + n_k = n$, а в остальном независимы в силу независимости элементов выборки. Таким образом, функция χ^2 имеет $k-1$ независимых аргументов: число частот минус одна связь. В силу теоремы Пирсона число степеней свободы статистики χ^2 отражается на виде асимптотической плотности $f_{k-1}(x)$.

На основе общей схемы проверки статистических гипотез сформулируем следующее правило.

Правило проверки гипотезы о законе распределения по методу χ^2

1. Выбираем уровень значимости α.

2. По таблице [6, с. 358] находим квантиль $\chi^2_{1-\alpha}(k-1)$ распределения хи-квадрат с $k-1$ степенями свободы порядка $1-\alpha$.

3. С помощью гипотетической функции распределения $F(x)$ вычисляем вероятности $p_i = P(X \in \Delta_i)$, $i = 1, \ldots, k$.

4. Находим частоты n_i попадания элементов выборки в подмножества Δ_i, $i = 1, \ldots, k$.

5. Вычисляем выборочное значение статистики критерия χ^2:

$$\chi^2_{\text{B}} = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i}.$$

6. Сравниваем χ^2_{B} и квантиль $\chi^2_{1-\alpha}(k-1)$.

   а) Если $\chi^2_{\text{B}} < \chi^2_{1-\alpha}(k-1)$, то гипотеза H_0 на данном этапе проверки принимается.

   б) Если $\chi^2_{\text{B}} \geq \chi^2_{1-\alpha}(k-1)$, то гипотеза H_0 отвергается, выбирается одно из альтернативных распределений, и процедура проверки повторяется.

Замечание 3. Из формулы (4.2) видим, что веса $c_i = n/p_i$ пропорциональны n, т.е. с ростом n увеличиваются. Отсюда следует, что если выдвинутая гипотеза неверна, то относительные частоты n_i/n не будут близки к вероятностям p_i, и с ростом n величина χ^2_{B} будет увеличиваться. При фиксированном уровне значимости α будет фиксировано пороговое число — квантиль $\chi^2_{1-\alpha}(k-1)$, поэтому, увеличивая n, мы придём к неравенству $\chi^2_{\text{B}} > \chi^2_{1-\alpha}(k-1)$, т.е. с увеличением объёма выборки неверная гипотеза будет отвергнута.

Отсюда следует, что при сомнительной ситуации, когда $\chi^2_{\text{B}} \approx \chi^2_{1-\alpha}(k-1)$, можно попытаться увеличить объём выборки (например, в 2 раза), чтобы требуемое неравенство было более чётким.

Замечание 4. Теория и практика применения критерия χ^2 указывают, что если для каких-либо подмножеств Δ_i $(i = 1, \ldots, k)$ условие $np_i \geq 5$ не выполняется, то следует объединить соседние подмножества (промежутки).

Это условие выдвигается требованием близости величин

$$(n_i - np_i)/\sqrt{np_i},$$

квадраты которых являются слагаемыми χ^2 к нормальным $N(0,1)$. Тогда случайная величина в формуле (4.2) будет распределена по закону, близкому к хи-квадрат. Такая близость обеспечивается достаточной численностью элементов в подмножествах Δ_i [1, с. 481-485].

4.2 Проверка гипотезы об однородности выборки. Метод Фишера

Пусть заданы две выборки $x^n = (x_1, \ldots, x_n)$, $x_i \in \mathbb{R}$; $y^m = (y_1, \ldots, y_m)$, $y_i \in \mathbb{R}$.

Обозначим через σ_1^2 и σ_2^2 дисперсии выборок x^n и y^m, s_1^2 и s_2^2 — выборочные оценки дисперсий σ_1^2 и σ_2^2:

$$s_1^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2, \quad s_2^2 = \frac{1}{m-1}\sum_{i=1}^{m}(y_i - \overline{y})^2$$

где $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, $\overline{y} = \frac{1}{m}\sum_{i=1}^{m} y_i$ — выборочные средние выборок x^n и y^m.

Дополнительное предположение: выборки x^n и y^m являются нормальными. Критерий Фишера чувствителен к нарушению предположения о нормальности.

Нулевая гипотеза $H_0: \sigma_1^2 = \sigma_2^2$

При выполнении нулевой гипотезы статистика критерия Фишера $F = \frac{s_1^2}{s_2^2}$ имеет распределение Фишера с $n-1$ и $m-1$ степенями свободы.

Обычно в числителе ставится большая из двух сравниваемых дисперсий и сравнение осуществляется с «правой» квантилью распределения (т.е. критической областью критерия является правый хвост распределения Фишера), что соотвествует альтернативной гипотезе H_1'.

Критерий (при уровне значимости α):

- $H_1:\ \sigma_1^2 \neq \sigma_2^2$ если $F < F_{\alpha/2}(n-1, m-1)$ или $F > F_{1-\alpha/2}(n-1, m-1)$, то нулевая гипотеза H_0 отвергается в пользу альтернативы H_1.
- $H_1':\ \sigma_1^2 > \sigma_2^2$ если $F > F_{1-\alpha}(n-1, m-1)$, то нулевая гипотеза H_0 отвергается в пользу альтернативы H_1'

где $F_\alpha(n-1, m-1)$ есть α-квантиль распределения Фишера с $n-1$ и $m-1$ степенями свободы.

Глава 5

Взаимосвязь двух случайных величин

5.1 Двумерное нормальное распределение

Опеределение. *Двумерная* случайная величина (X, Y) называется *распределённой нормально* (или просто нормальной), если её плотность вероятности определена формулой

$$N(x, y, \overline{x}, \overline{y}, \sigma_x, \sigma_y, \rho) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \times$$
$$\times \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\overline{x})^2}{\sigma_x^2} - 2\rho\frac{(x-\overline{x})(y-\overline{y})}{\sigma_x\sigma_y} + \frac{(y-\overline{y})^2}{\sigma_y^2}\right]\right\} \quad (5.1)$$

Опеределение. *Двумерная* случайная величина (X, Y) называется *распределённой нормально* (или просто нормальной), если её плотность вероятности определена формулой

$$N(x, y, \overline{x}, \overline{y}, \sigma_x, \sigma_y, \rho) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \times$$
$$\times \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\overline{x})^2}{\sigma_x^2} - 2\rho\frac{(x-\overline{x})(y-\overline{y})}{\sigma_x\sigma_y} + \frac{(y-\overline{y})^2}{\sigma_y^2}\right]\right\}. \quad (5.2)$$

Компоненты X, Y двумерной нормальной случайной величины

также распределены нормально с математическими ожиданиями $\overline{x}, \overline{y}$ и средними квадратическими отклонениями σ_x, σ_y соответственно [1, с. 133-134].

Параметр ρ называется *коэффициентом корреляции.*

5.2 Корреляционный момент (ковариация) и коэффициент корреляции

Опеределение. *Корреляционным моментом*, иначе *ковариацией*, двух случайных величин X и Y называется математическое ожидание произведения отклонений этих случайных величин от их математических ожиданий [1, с. 141].

$$K = \operatorname{cov}(X, Y) = \mathbf{M}\left[(X - \overline{x})(Y - \overline{y})\right]. \tag{5.3}$$

Опеределение. *Коэффициентом корреляции* ρ двух случайных величин X и Y называется отношение их корреляционного момента к произведению их средних квадратических отклонений:

$$\rho = \frac{K}{\sigma_x \sigma_y}. \tag{5.4}$$

Коэффициент корреляции — это нормированная числовая характеристика, являющаяся мерой близости зависимости между случайными величинами к линейной [1, с. 150].

5.3 Выборочные коэффициенты корреляции

5.3.1 Выборочный коэффициент корреляции Пирсона

Пусть по выборке значений $\{x_i, y_i\}_1^n$ двумерной с.в. (X, Y) требуется оценить коэффициент корреляции $\rho = \frac{\operatorname{cov}(X,Y)}{\sqrt{\mathbf{D}X\mathbf{D}Y}}$. Естественной оценкой

для ρ служит его статистический аналог в виде *выборочного коэффициента корреляции*, предложенного *К.Пирсоном*

Определение. *Выборочный коэффициент корреляции Пирсона*

$$r = \frac{\frac{1}{n}\sum(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\frac{1}{n}\sum(x_i - \overline{x})^2 \frac{1}{n}\sum(y_i - \overline{y})^2}} = \frac{K}{s_X s_Y}, \tag{5.5}$$

где K, s_X^2, s_Y^2 — выборочные ковариация и дисперсии с.в. X и Y [1, с. 535].

5.3.2 Выборочный квадрантный коэффициент корреляции

Кроме выборочного коэффициента корреляции Пирсона, существуют и другие оценки степени взаимосвязи между случайными величинами. К ним относится *выборочный квадрантный коэффициент корреляции*

Определение. *Выборочный квадрантный коэффициент корреляции*

$$r_Q = \frac{(n_1 + n_3) - (n_2 + n_4)}{n}, \tag{5.6}$$

где n_1, n_2, n_3 и n_4 — количества точке с координатами (x_i, y_i), попавшими соответственно в 1, 2, 3 и 4 квадранты декартовой системы с осями $x' = x - \operatorname{med} x$, $y' = y - \operatorname{med} y$ и с центром в точке с координатами $(\operatorname{med} x, \operatorname{med} y)$ [1, с. 539].

5.3.3 Выборочный коэффициент ранговой корреляции Спирмена

На практике нередко требуется оценить степень взаимодействия между качественными признаками изучаемого объекта. Качественным называется признак, который нельзя измерить точно, но который позволяет сравнивать изучаемые объекты между собой и располагать их в порядке убывания или возрастания их качества. Для этого объекты выстраиваются в определённом порядке в соответствии с рассматриваемым признаком. Процесс упорядочения называется

ранжированием, и каждому члену упорядоченной последовательности объектов присваивается ранг, или порядковый номер. Например, объекту с наименьшим значением признака присваивается ранг 1, следующему за ним объекту — ранг 2, и т.д. Таким образом, происходит сравнение каждого объекта со всеми объектами изучаемой выборки.

Если объект обладает не одним, а двумя качественными признаками — переменными X и Y, то для исследования их взаимосвязи используют выборочный коэффициент корреляции между двумя последовательностями рангов этих признаков.

Обозначим ранги, соотвествующие значениям переменной X, через u, а ранги, соотвествующие значениям переменной Y, — через v.

Опеределение. *Выборочный коэффициент ранговой корреляции Спирмена* определяется как выборочный коэффициент корреляции Пирсона между рангами u, v переменных X, Y:

$$r_S = \frac{\frac{1}{n}\sum(u_i - \overline{u})(v_i - \overline{v})}{\sqrt{\frac{1}{n}\sum(u_i - \overline{u})^2 \frac{1}{n}\sum(v_i - \overline{v})^2}}, \tag{5.7}$$

где $\overline{u} = \overline{v} = \frac{1+2+\ldots+n}{n} = \frac{n+1}{2}$ — среднее значение рангов [1, с. 540-541].

5.4 Эллипсы рассеивания

Рассмотрим поверхность распределения, изображающую функцию (5.2). Она имеет вид холма, вершина которого находится над точкой $(\overline{x}, \overline{y})$.

В сечении поверхности распределения плоскостями, параллельными оси $N(x, y, \overline{x}, \overline{y}, \sigma_x, \sigma_y, \rho)$, получаются кривые, подобные нормальным кривым распределения. В сечении поверхности распределения плоскостями, параллельными плоскости xOy, получаются эллипсы. Напишем уравнение проекции такого эллипса на плоскость xOy:

$$\frac{(x-\overline{x})^2}{\sigma_x^2} - 2\rho\frac{(x-\overline{x})(y-\overline{y})}{\sigma_x\sigma_y} + \frac{(y-\overline{y})^2}{\sigma_y^2} = \text{const}. \tag{5.8}$$

Уравнение эллипса (5.8) можно проанализировать обычными методами аналитической геометрии. Применяя их, убеждаемся, что центр эллипса (5.8) находится в точке с координатами $(\overline{x}, \overline{y})$; что касается направления осей симметрии эллипса, то они составляют с осью Ox углы, определяемые уравнением

$$\operatorname{tg} 2\alpha = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 - \sigma_y^2}. \tag{5.9}$$

Это уравнение дает значения углов α И α_1, различающиеся на $\dfrac{\pi}{2}$.

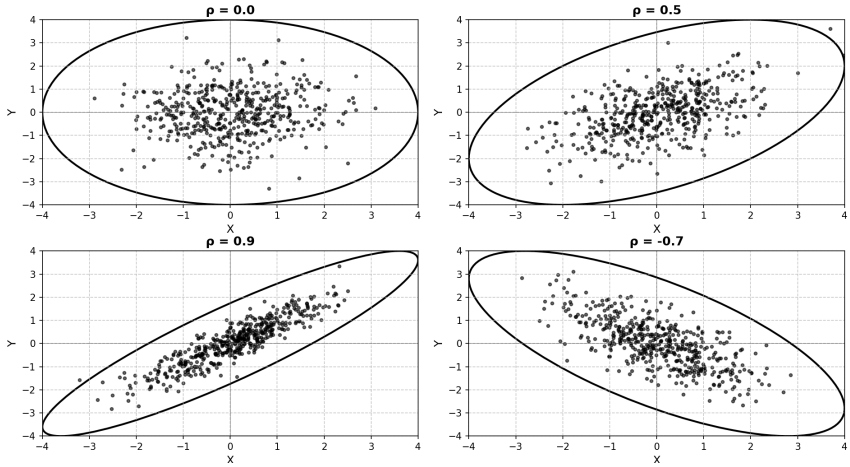

Рис. 5.1. Эллипсы рассеяния для разных значений коэффициента корреляции ρ

Таким образом, ориентация эллипса (5.8) относительно координатных осей находится в прямой зависимости от коэффициента корреляции ρ системы (X, Y); если величины не коррелированны (т.е. в данном случае и независимы), то оси симметрии эллипса параллельны координатным осям; в противном случае они составляют с координатными осями некоторый угол.

Опеределение. Пересекая поверхность распределения плоскостями, параллельными плоскости xOy, и проектируя сечения на плоскость xOy мы получим целое семейство подобных и одинаково расположенных эллипсов с общим центром $(\overline{x}, \overline{y})$. Во всех точках каждого из таких эллипсов плотность распределения $N(x, y, \overline{x}, \overline{y}, \sigma_x, \sigma_y, \rho)$ постоянна. Поэтому такие эллипсы называются эллипсами равной плотности или, короче *эллипсами рассеивания*. Общие оси всех эллипсов рассеивания называются главными осями рассеивания [5, с. 193-194].

Для построения эллипса в общем случае необходимо как минимум 4 точки. 3 точки - вырожденный случай (окружность). 2 точки - вырожденный случай (прямая).

Глава 6

Модель простой линейной регрессии

6.1 Простая линейная регрессия

6.1.1 Модель простой линейной регрессии

Опеределение. Регрессионную модель описания данных называют *простой линейной регрессией*, если

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \ldots, n, \tag{6.1}$$

где $x_1, \ldots, x_n$ — заданные числа (значения фактора); $y_1, \ldots, y_n$ — наблюдаемые значения отклика; $\varepsilon_1, \ldots, \varepsilon_n$ — независимые, нормально распределённые $N(0, \sigma)$ с нулевым математическим ожиданием и одинаковой (неизвестной) дисперсией случайные величины (ненаблюдаемые); $\beta_0,\ \beta_1$ — неизвестные параметры, подлежащие оцениванию.

В модели (6.1) отклик y зависит зависит от одного фактора x, и весь разброс экспериментальных точек объясняется только погрешностями наблюдений (результатов измерений) отклика y. Погрешности результатов измерений x в этой модели полагают существенно меньшими погрешностей результатов измерений y, так что ими можно пренебречь [1, с. 507].

6.1.2 Метод наименьших квадратов

При оценивании *параметров регрессионной модели* используют различные методы. Один из наиболее распространённых подходов заключается в следующем: вводится мера (критерий) рассогласования отклика и регрессионной функции, и оценки параметров регрессии определяются так, чтобы сделать это рассогласование наименьшим. Достаточно простые расчётные формулы для оценок получают при выборе критерия в виде суммы квадратов отклонений значений отклика от значений регрессионной функции (сумма квадратов остатков).

Опеределение. *Метод наименьших квадратов (МНК)*

$$Q(\beta_0,\ \beta_1) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \to \min_{\beta_0,\ \beta_1}. \tag{6.2}$$

Задача минимизации квадратичного критерия (6.2) носит название задачи *метода наименьших квадратов* (МНК), а оценки $\widehat{\beta}_0$, $\widehat{\beta}_1$ параметров β_0, β_1, реализующие минимум критерия (6.2), называют *МНК-оценками* [1, с. 508].

Данный метод, вообще говоря, может применяться для аппроксимации заданного набора экспериментальных данных линейной комбинацией линейно независимых функций, размер которой не превосходит мощности множества данных (в случае равенства получаем интерполяцию).

Оптимальность. Критерием оптимальности подобранной аппроксимации является l^2-норма, точнее, для простоты вычисления, её квадрат:

$$\sum_{i=1}^{n} ||\lambda_i f_i(\{x_n\}) - \{y_n\}||_{l^2}^2 \xrightarrow[\{\lambda_i\}]{} min \tag{6.3}$$

Минимум ищется по коэффициентам линейной комбинации, исходя из критерия равенства нулю градиента и положительной определённости Якобиана.

6.1.3 Расчётные формулы для МНК-оценок

МНК-оценки параметров $\widehat{\beta}_0$ и $\widehat{\beta}_1$ находятся из условия обращения функции $Q(\beta_0,\ \beta_1)$ в минимум.

Для нахождения МНК-оценок $\widehat{\beta}_0$ и $\widehat{\beta}_1$ выпишем необходимые условия экстремума:

$$\begin{cases} \frac{\partial Q}{\partial \beta_0} = -2 \sum\limits_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0, \\ \frac{\partial Q}{\partial \beta_1} = -2 \sum\limits_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) x_i = 0. \end{cases} \tag{6.4}$$

Далее для упрощения записи сумм будем опускать индекс суммирования. Из системы (6.4) получим

$$\begin{cases} n\widehat{\beta}_0 + \widehat{\beta}_1 \sum x_i = \sum y_i, \\ \widehat{\beta}_0 \sum x_i + \widehat{\beta}_1 \sum x_i^2 = \sum x_i y_i. \end{cases}$$

Разделим оба уравнения на n:

$$\begin{cases} \widehat{\beta}_0 + \left(\frac{1}{n} \sum x_i\right) \widehat{\beta}_1 = \frac{1}{n} \sum y_i, \\ \left(\frac{1}{n} \sum x_i\right) \widehat{\beta}_0 + \left(\frac{1}{n} \sum x_i^2\right) \widehat{\beta}_1 = \frac{1}{n} \sum x_i y_i \end{cases} \tag{6.5}$$

и, используя известные статистические обозначения для выборочных первых и вторых начальных моментов

$$\overline{x} = \frac{1}{n} \sum x_i, \;\; \overline{y} = \frac{1}{n} \sum y_i, \;\; \overline{x^2} = \frac{1}{n} \sum x_i^2, \;\; \overline{xy} = \frac{1}{n} \sum x_i y_i,$$

получим

$$\begin{cases} \widehat{\beta}_0 + \overline{x} \widehat{\beta}_1 = \overline{y}, \\ \overline{x} \widehat{\beta}_0 + \overline{x^2} \widehat{\beta}_1 = \overline{xy}, \end{cases} \tag{6.6}$$

откуда МНК-оценку $\widehat{\beta}_1$ наклона прямой регрессии находим по формуле Крамера

$$\widehat{\beta}_1 = \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\overline{x^2} - (\overline{x})^2}, \tag{6.7}$$

а МНК-оценку $\widehat{\beta}_0$ определяем непосредственно из первого уравнения системы (6.6):

$$\widehat{\beta}_0 = \overline{y} - \overline{x} \widehat{\beta}_1. \tag{6.8}$$

Заметим, что определитель системы (6.6)

$$\overline{x^2} - (\overline{x})^2 = n^{-1}\sum (x_i - \overline{x})^2 = s_x^2 > 0,$$

если среди значений x_1, ... , x_n есть различные, что и будем предполагать.

Доказательство минимальности функции $Q(\beta_0,\ \beta_1)$ в стационарной точке проведём с помощью известного достаточного признака экстремума функции двух переменных. Имеем:

$$\frac{\partial^2 Q}{\partial \beta_0^2} = 2n,\ \ \frac{\partial^2 Q}{\partial \beta_1^2} = 2\sum x_i^2 = 2n\overline{x^2},\ \ \frac{\partial^2 Q}{\partial \beta_0\ \partial \beta_1} = 2\sum x_i = 2n\overline{x}.$$

$$\Delta = \frac{\partial^2 Q}{\partial \beta_0^2}\cdot\frac{\partial^2 Q}{\partial \beta_1^2} - \left(\frac{\partial^2 Q}{\partial \beta_0\ \partial \beta_1}\right)^2 = 4n^2\overline{x^2} - 4n^2(\overline{x})^2 =$$
$$= 4n^2\left[\overline{x^2} - (\overline{x})^2\right] = 4n^2\left[\frac{1}{n}\sum (x_i - \overline{x})^2\right] = 4n^2 s_x^2 > 0.$$

Этот результат вместе с условием $\frac{\partial^2 Q}{\partial \beta_0^2} = 2n > 0$ означает, что в стационарной точке функция Q имеет минимум [1, с. 508-511].

6.2 Робастные оценки коэффициентов линейной регрессии

Робастность оценок коэффициентов линейной регрессии (т.е. их устойчивость по отношению к наличию в данных редких, но больших по величине выбросов) может быть обеспечена различными способами. Одним из них является использование *метода наименьших модулей* вместо метода наименьших квадратов:

Опеределение. *Метод наименьших модулей (МНМ)*

$$\sum_{i=1}^{n} |y_i - \beta_0 - \beta_1 x_i| \to \min_{\beta_0,\ \beta_1}\ . \tag{6.9}$$

Напомним, что использование метода наименьших модулей в задаче оценивания параметра сдвига распределений приводит к оценке в

виде выборочной медианы, обладающей робастными свойствами. В отличие от этого случая и от задач метода наименьших квадратов, на практике задача (6.9) решается численно. Соответствующие процедуры представлены в некоторых современных пакетах программ по статистическому анализу.

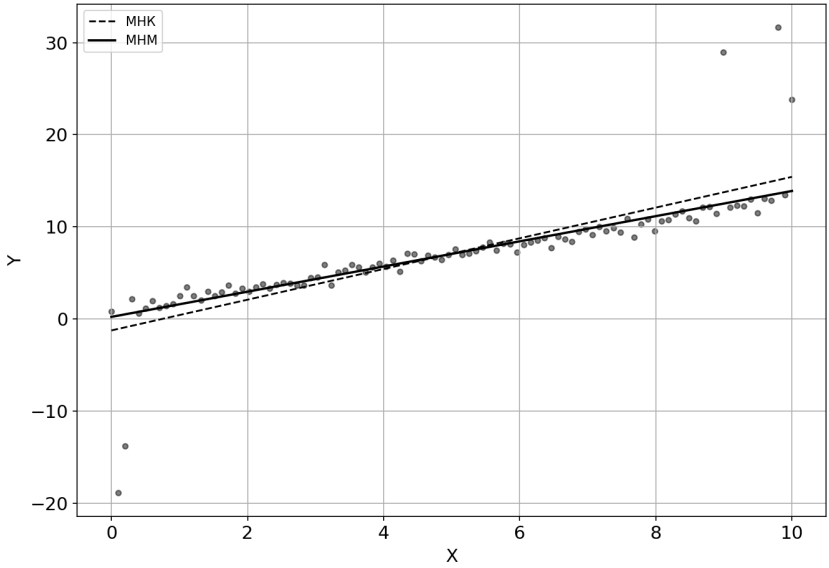

Рис. 6.1. Сравнение МНК и МНМ на данных с выбросами

Здесь мы рассмотрим простейшую в вычистлительном отношении робастную альтернативу оценкам коэффициентов линейной регрессии по МНК. Для этого сначала запишем выражения для оценок (6.7) и (6.8) в другом виде:

$$\begin{aligned} &\widehat{\beta}_1 = \frac{\overline{xy} - \overline{x}\cdot\overline{y}}{\overline{x^2} - (\overline{x})^2} = \frac{k_{xy}}{s_x^2} = \frac{k_{xy}}{s_x s_y}\cdot\frac{s_y}{s_x} = r_{xy}\frac{s_y}{s_x}, \\ &\widehat{\beta}_0 = \overline{y} - \overline{x}\widehat{\beta}_1. \end{aligned} \tag{6.10}$$

В формулах (6.10) заменим выборочные средние $\overline{x}$ и $\overline{y}$ соответственно на робастные выборочные медианы $\operatorname{med} x$ и $\operatorname{med} y$, среднеквадратические отклонения s_x и s_y на робастные нормированные интерквар-

тильные широты q_x^* и q_y^*, выборочный коэффициент корреляции r_{xy} — на знаковый коэффициент корреляции r_Q:

$$\widehat{\beta}_{1R} = r_Q \frac{q_y^*}{q_x^*}, \tag{6.11}$$

$$\widehat{\beta}_{0R} = \operatorname{med} y - \widehat{\beta}_{1R} \operatorname{med} x, \tag{6.12}$$

$$r_Q = \tfrac{1}{n} \sum_{i=1}^{n} \operatorname{sgn}(x_i - \operatorname{med} x) \operatorname{sgn}(y_i - \operatorname{med} y), \tag{6.13}$$

$$q_y^* = \frac{y_{(j)} - y_{(l)}}{k_q(n)}, \quad q_x^* = \frac{x_{(j)} - x_{(l)}}{k_q(n)}. \tag{6.14}$$

$$l = \begin{cases} [n/4] + 1 & \text{при} \quad n/4 \quad \text{дробном}, \\ n/4 & \text{при} \quad n/4 \quad \text{целом}. \end{cases}$$

$$j = n - l + 1.$$

$$\operatorname{sgn} z = \begin{cases} 1 & \text{при} \quad z > 0, \\ 0 & \text{при} \quad z = 0, \\ -1 & \text{при} \quad z < 0. \end{cases}$$

Уравнение регрессии здесь имеет вид

$$y = \widehat{\beta}_{0R} + \widehat{\beta}_{1R} x. \tag{6.15}$$

Статистики выборочной медианы и интерквартильной широты обладают робастными свойствами в силу того, что основаны на центральных порядковых статистиках, малочувствительных к большим по величине выбросам в данных. Статистика выборочного знакового коэффициента корреляции робастна, так как знаковая функция sgn z чувствительна не к величине аргумента, а только к его знаку. Отсюда оценка прямой регрессии (6.15) обладает очевидными робастными свойствами устойчивости к выбросам по координате y, но она довольно груба [1, с. 518-519].

Оптимальность. Данный метод основан на минимизации l^1-нормы разности последовательностей полученных экспериментальных данных и значений аппроксимирующей функции.

6.3 Количественная мера оценки качества регрессии

Как уже было сказано метод наименьших квадратов (МНК) минимизирует норму в l^2, а метод наименьших модулей (МНМ) норму в l^1.

Допустим, что мы будем сравнивать между собой полученные оценки для коэффициентов β_0, β_1 как модули разностей полученных значений. Тогда в случае, если для какой-то выборки окажется, что по данному критерию оба метода покажут схожие результаты, то может оказаться, что невязка по l^2-критерию все равно окажется значительно меньше для результатов, полученных с помощью МНК.

Это как раз и является следствием того, что рассмотренные методы минимизирует различные нормы. Обратная ситуация, но уже для l^1-метрики также может иметь место в некоторых случаях.

Соответственно, можно сказать, что l-метрики (в данном случае речь о близости прямых) не позволяют делать однозначных выводов о качестве линейной регрессии в смысле близости искомых коэффициентов аппроксимирующих функций.

Глава 7

Интервальный анализ и статистика

В данной главе мы рассмотрим различные способы представления экспериментальных данных и работы с ними.

Описание неопределённостей с помощью интервалов имеет ряд уникальных достоинств. В этом разделе представлена краткая информация о классической интервальной арифметике и анализе данных с интервальной неопределенностью (иначе, интервальной статистике).

Также понадобятся сведения о расширениях классической интервальной арифметики $\mathbb{IR}$, которые можно использовать для описания изотопных данных. Нам понадобится полная интервальная арифметика Каухера $\mathbb{KR}$, которая преодолевает ряд трудностей классической интервальной арифметики и позволяет естественным образом работать с непересекающимися интервалами и сиспользовать полную мощь минимаксных методов.

Для работы со структурами данных при обработке выборок понадобятся объекты, основанные на интервалах: объединения интервалов — мультиинтервалы, интервалы с интервальными вершинами — твины.

7.1 Интервальные арифметики

7.1.1 Вещественные интервалы

Первичное понятие интервального анализа данных и интервальной статистики — *интервал*. Это простое подмножество множества всех вещественных (действительных) чисел, которое задаёт целый диапазон значений интересующей нас величины. С помощью интервалов можно описывать и моделировать неопределённости и неоднозначности.

Интервалы могут определяться на вещественной оси, на комплексной плоскости, а также в многомерных пространствах. Кроме того, существуют различные определения интервалов, и некоторые их них не равносильны друг другу, задавая разные математические объекты. Далее нас будут интересовать, главным образом, вещественные интервалы, вещественные интервальные векторы и матрицы, так как именно они играют главную роль в измерениях и их обработке.

Определение. *Интервалом* $[a, b]$ вещественной оси $\mathbb{R}$ называется множество всех чисел, расположенных между заданными числами a и b, включая их самих, т. е.

$$[a, b] := \{\, x \in \mathbb{R} \mid a \le x \le b \,\}.$$

При этом a и b называются *концами* интервала $[a, b]$, *левым* (или нижним) и *правым* (или верхним) соответственно.

7.1.2 Характеристики интервала

Любой интервал полностью задаётся двумя числами — своими концами, но на практике широко используются также другие характеристики интервалов и представления интервалов на их основе.

Важнейшими характеристиками интервала являются его *середина* (центр), *радиус* и *ширина*

Определение. *Середина интервала*

$$\operatorname{mid} \boldsymbol{a} = \tfrac{1}{2}(\overline{\boldsymbol{a}} + \underline{\boldsymbol{a}}),$$

Определение. *Радиус интервала*

$$\operatorname{rad} \boldsymbol{a} = \tfrac{1}{2}(\overline{\boldsymbol{a}} - \underline{\boldsymbol{a}}).$$

Определение. *Ширина интервала*

$$\operatorname{wid} \boldsymbol{a} = \overline{\boldsymbol{a}} - \underline{\boldsymbol{a}}.$$

Таким образом, задание середины и радиуса интервала также однозначно определяет его, чем часто пользуются и в теории, и на практике.

Середина интервала — это точка, которая «представляет его» наилучшим образом, так как наименее удалена от остальных точек этого интервала.

Радиус и ширина характеризуют разброс (рассеяние) точек интервала. Интервалы нулевой ширины обычно называют *вырожденными*. Они отождествляются с вещественными числами, то есть, $[1, 1]$ — это то же самое, что и 1.

7.1.3 Отношения между интервалами

Интервалы являются множествами, и для них определены теоретико-множественные отношения и операции (объединение, пересечение и др.). Особенно важно отношение включения одного интервала в другой:

$$\boldsymbol{a} \subseteq \boldsymbol{b} \text{ равносильно тому, что } \underline{\boldsymbol{a}} \geq \underline{\boldsymbol{b}} \text{ и } \overline{\boldsymbol{a}} \leq \overline{\boldsymbol{b}}. \tag{7.1}$$

Отношение включения является частичным порядком и превращает множество интервалов в частично упорядоченное множество (см. [9]).

Важную роль играет линейное упорядочение интервалов:

Определение. Для интервалов $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{IR}$ условимся считать, что $\boldsymbol{a}$ *не превосходит* $\boldsymbol{b}$ и писать «$\boldsymbol{a} \leq \boldsymbol{b}$» тогда и только тогда, когда $\underline{\boldsymbol{a}} \leq \underline{\boldsymbol{b}}$ и $\overline{\boldsymbol{a}} \leq \overline{\boldsymbol{b}}$.
Интервал называется *неотрицательным*, т. е. «≥ 0», если неотрицательны оба его конца. Интервал называется *неположительным*, т. е. «≤ 0», если неположительны оба его конца.

7.1.4 Теоретико-множественные операции над интервалами

Если интервалы $\boldsymbol{a}$ и $\boldsymbol{b}$ имеют непустое пересечение, т. е. $\boldsymbol{a} \cap \boldsymbol{b} \neq \varnothing$, то можно дать простые выражения для результатов теоретико-множественных операций пересечения и объединения через концы этих интервалов

$$\boldsymbol{a} \cap \boldsymbol{b} = \big[\max\{\underline{\boldsymbol{a}}, \underline{\boldsymbol{b}}\}, \min\{\overline{\boldsymbol{a}}, \overline{\boldsymbol{b}}\}\big],$$

$$\boldsymbol{a} \cup \boldsymbol{b} = \big[\min\{\underline{\boldsymbol{a}}, \underline{\boldsymbol{b}}\}, \max\{\overline{\boldsymbol{a}}, \overline{\boldsymbol{b}}\}\big].$$

Если же $\boldsymbol{a} \cap \boldsymbol{b} = \varnothing$, т. е. интервалы $\boldsymbol{a}$ и $\boldsymbol{b}$ не имеют общих точек, то эти равенства уже неверны.

Обобщением операций пересечения и объединения являются операции взятия точной нижней грани и точной верхней грани относительно включения «$\subseteq$»:

$$\boldsymbol{a} \wedge \boldsymbol{b} = \big[\max\{\underline{\boldsymbol{a}}, \underline{\boldsymbol{b}}\}, \min\{\overline{\boldsymbol{a}}, \overline{\boldsymbol{b}}\}\big], \tag{7.2}$$

$$\boldsymbol{a} \vee \boldsymbol{b} = \big[\min\{\underline{\boldsymbol{a}}, \underline{\boldsymbol{b}}\}, \max\{\overline{\boldsymbol{a}}, \overline{\boldsymbol{b}}\}\big]. \tag{7.3}$$

Точная нижняя грань не обязательно присутствует во множестве, в отличие от минимума по множеству.

Операции (7.2) и (7.3) используются при обработке интервальных данных.

7.1.5 Классическая интервальная арифметика

Определение операций между интервалами производится через результаты операций между их членами, т. е. «по представителям». Именно, результат интервальной операции есть множество всевозможных результатов операции между числами из интервалов. Для двухместной операции «$\star$» имеем

$$\boldsymbol{a} \star \boldsymbol{b} \;=\; \big\{\, a \star b \mid a \in \boldsymbol{a},\, b \in \boldsymbol{b} \,\big\}. \tag{7.4}$$

Аналогично определяются интервальные одноместные операции.

Если рассматриваются арифметические операции, т. е. $\star \in \{+, -, \cdot, /\}$, то множества, задаваемые правилом (7.4), тоже являются

интервалами. Для конкретных арифметических операций имеем формулы:

$$\boldsymbol{a} + \boldsymbol{b} = \left[\, \underline{\boldsymbol{a}} + \underline{\boldsymbol{b}},\, \overline{\boldsymbol{a}} + \overline{\boldsymbol{b}} \,\right], \tag{7.5}$$

$$\boldsymbol{a} - \boldsymbol{b} = \left[\, \underline{\boldsymbol{a}} - \overline{\boldsymbol{b}},\, \overline{\boldsymbol{a}} - \underline{\boldsymbol{b}} \,\right], \tag{7.6}$$

$$\boldsymbol{a} \cdot \boldsymbol{b} = \left[\, \min\{\underline{\boldsymbol{a}}\,\underline{\boldsymbol{b}}, \underline{\boldsymbol{a}}\,\overline{\boldsymbol{b}}, \overline{\boldsymbol{a}}\,\underline{\boldsymbol{b}}, \overline{\boldsymbol{a}}\,\overline{\boldsymbol{b}}\} \,,\, \max\{\underline{\boldsymbol{a}}\,\underline{\boldsymbol{b}}, \underline{\boldsymbol{a}}\,\overline{\boldsymbol{b}}, \overline{\boldsymbol{a}}\,\underline{\boldsymbol{b}}, \overline{\boldsymbol{a}}\,\overline{\boldsymbol{b}}\} \,\right], \tag{7.7}$$

$$\boldsymbol{a}/\boldsymbol{b} = \boldsymbol{a} \cdot \left[\, 1/\overline{\boldsymbol{b}},\, 1/\underline{\boldsymbol{b}},\, \right] \qquad \text{для} \quad \boldsymbol{b} \not\ni 0. \tag{7.8}$$

Определение. Множество всех интервалов вещественной оси с операциями сложения, вычитания, умножения и деления, определёнными формулами (7.5)–(7.8), называется *классической интервальной арифметикой*, и его обозначают $\mathbb{IR}$.

7.1.6 Полная интервальная арифметика (Каухера) $\mathbb{KR}$

Определение. Элементами арифметики $\mathbb{KR}$ являются пары чисел вида $[\alpha, \beta]$. Если $\alpha \le \beta$, то $[\alpha, \beta]$ обозначает обычный интервал вещественной оси, и его называют *правильным*. Если же $\alpha > \beta$, то $[\alpha, \beta]$ — *неправильный интервал*. Таким образом, $\mathbb{IR} \subset \mathbb{KR}$.

Правильные и неправильные интервалы, две «половинки» $\mathbb{KR}$, переходят друг в друга в результате отображения *дуализации*, которое обозначается символом `dual` и меняет местами (переворачивает) концы интервала, т. е.

$$\operatorname{dual} \boldsymbol{a} := [\, \overline{\boldsymbol{a}},\, \underline{\boldsymbol{a}} \,].$$

Определение. Множество всех интервалов вещественной оси с операциями сложения, вычитания, умножения и деления, определёнными формулами (7.5)–(7.8), называется *классической интервальной арифметикой*, и его обозначают $\mathbb{IR}$.

С помощью правильной проекции из произвольного интервала получается правильный.

Арифметические операции между интервалами в $\mathbb{KR}$ продолжают операции в $\mathbb{IR}$, их подробное описание можно найти в [10]. Умножение

интервала из $\mathbb{KR}$ на число определяется совершенно так же, как и для обычных правильных интервалов.

Чрезвычайно важным в интервальной арифметике Каухера является обратимость арифметических операций. В частности, для любого интервала имеется противоположный ему, т. е. обратный по сложению. Для интервалов, не содержащих нуля, имеются обратные к ним по умножению. Для сложения (7.5) обратной операцией является не операция интервального вычитания (7.6), а операция, которую называют «алгебраическим вычитанием» и обозначают знаком «$\ominus$»:

$$\boldsymbol{a} \ominus \boldsymbol{b} = [\underline{\boldsymbol{a}} - \underline{\boldsymbol{b}}, \overline{\boldsymbol{a}} - \overline{\boldsymbol{b}}]. \tag{7.9}$$

Для любых интервалов $\boldsymbol{a}$, $\boldsymbol{b}$ из $\mathbb{KR}$ справедливы равенства

$$\boldsymbol{a} \ominus \boldsymbol{a} = 0, \qquad (\boldsymbol{a} + \boldsymbol{b}) \ominus \boldsymbol{b} = \boldsymbol{a}, \qquad (\boldsymbol{a} \ominus \boldsymbol{b}) + \boldsymbol{b} = \boldsymbol{a}.$$

Пример 7.1.1.

$$[1,2] \ominus [1,2] = [1,2] + [-1,-2] = [1-1, 2-2] = [0,0] = 0.$$

■

Полная интервальная арифметика Каухера $\mathbb{KR}$ пополняет классическую интервальную арифметику $\mathbb{IR}$ не только в алгебраическом смысле, но также и относительно естественного порядка по включению «$\subseteq$».

Утверждение. Для интервалов $\boldsymbol{a}$, $\boldsymbol{b} \in \mathbb{KR}$ выполняется включение $\boldsymbol{a} \subseteq \boldsymbol{b}$, если

$$\underline{\boldsymbol{a}} \geq \underline{\boldsymbol{b}} \quad \text{и} \quad \overline{\boldsymbol{a}} \leq \overline{\boldsymbol{b}}.$$

Относительно введённого таким образом отношения включения в $\mathbb{KR}$ для любых двух интервалов существуют интервалы точной нижней грани и точной верхней грани по включению, т. е. результаты операций $\boldsymbol{a} \wedge \boldsymbol{b}$ и $\boldsymbol{a} \vee \boldsymbol{b}$ всегда определены.

Пример 7.1.2.

$$[0,1] \wedge [4,5] = [4,1], \quad [0,1] \vee [4,5] = [0,5].$$

■

7.2 Составные интервальные объекты

Данные и результаты операций с ними можно описывать не только интервалами, но и более сложными объектами.

7.2.1 Твины

В данных мы можем использовать интервал с интервальными концами —*твин*. Слово «твин» является акронимом английского выражения «twice interval», т. е. «двойной интервал». Впервые такие объекты были рассмотрены Э. Гарденьесом с коллегами в 80-х годах XX века [11]. *Твин* можно представить в виде

$$\boldsymbol{X} = [\boldsymbol{a}, \boldsymbol{b}] = [\ [\underline{\boldsymbol{a}}, \overline{\boldsymbol{a}}], [\underline{\boldsymbol{b}}, \overline{\boldsymbol{b}}]\ ],$$

В зависимости от того, как мы определяем понятия «больше или равно» и «меньше или равно», подразумевать под ним множество всех интервалов, больших или равных $[\underline{\boldsymbol{a}}, \overline{\boldsymbol{a}}]$ и меньших или равных $[\underline{\boldsymbol{b}}, \overline{\boldsymbol{b}}]$. Так как на множествах интервалов из $\mathbb{IR}$ и $\mathbb{KR}$ существуют частичные упорядочения «$\subseteq$» и «$\leq$», то, соответственно, возможны два типа твинов: «$\subseteq$»-твины и «$\leq$»-твины.

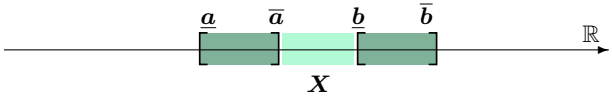

Рис. 7.1. Твины на вещественной оси

На рис. 7.1 твин $\boldsymbol{X}$ представлен в графической форме. Концы твина, т. е. интервалы $\boldsymbol{a}$ и $\boldsymbol{b}$, представлены более темной заливкой, чем остальная часть твина.

В.М.Нестеров развил идеи твинов [12]. Особенно значимы его идеи о твинах, как способе одновременного вычисления внутренних и внешних оценок.

Как представление твинов, так и работа с ними сложнее, чем с обычными интервалами. В настоящее время развиты программные средства, реализующие арифметику твинов [13].

7.2.2 Мультиинтервалы

В ряде разделов науки и техники встречаются ситуации, когда исследуемая величина содержится в неодносвязной области.

Согласно определению, приведённому в книге [10], *мультиинтервал* — это объединение конечного числа несвязных интервалов числовой оси (Рис. 7.2).

Рис. 7.2. Мультиинтервал в $\mathbb{R}$.

Между мультиинтервалами также могут быть определены арифметические операции «по представителям», аналогично тому, как это делается на множестве интервалов.

Глава 8

Интервальная статистика

Интервальная статистика — молодая ветвь анализа данных, в которой результаты измерений и их обработки име. двусторонние ограничения. В книге [8] дана система понятий интервальной статистики и подробное изложение её методов.

В изложении мы будем по необходимости использовать понятия из [8]. Важнейшим в контексте изотопных данных являются приёмы обработки постоянных величин.

8.1 Обработка постоянной величины

Постоянная величина — это величина, которая в рассматриваемом процессе сохраняет свое значение неизменным. Например, температура воды реки не меняется заметно в процессе её измерения в конкретном месте в течение непродолжительного времени, поэтому может считаться постоянной величиной.

Пусть имеется выборка измерений некоторой величины

$$\boldsymbol{x}_1, \boldsymbol{x}_2, \ \ldots, \ \boldsymbol{x}_n, \tag{8.1}$$

или кратко $\{\, \boldsymbol{x}_k \}_{k\,=\,1}^{n}$, где k — номер измерения; $\boldsymbol{x}_k$ — интервальный результат измерения. Таким образом, согласно терминологии интервального анализа рассматриваемая выборка — это вектор интервалов $\boldsymbol{x} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ \ldots, \ \boldsymbol{x}_n)$. Число n — размерность вектора данных — будем называть *длиной выборки*. По интервальным результатам измерений

или наблюдений требуется найти оценку интересующей нас величины.

Для наглядного представления выборки её интервалы преставляют в виде графика, изображенного на рис. 8.1, который называют *диаграммой рассеяния*.

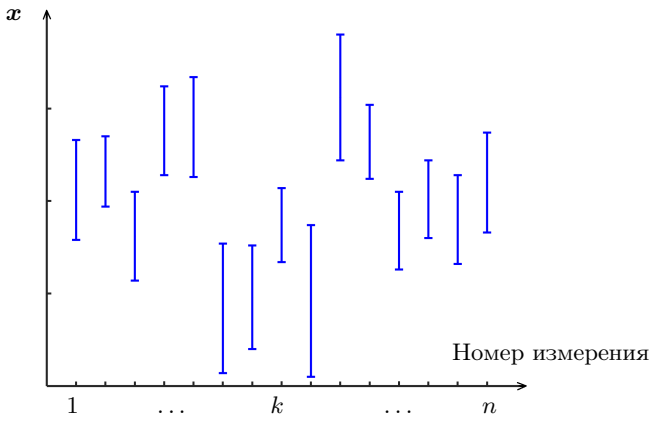

Рис. 8.1. Диаграмма рассеяния интервальных измерений постоянной величины

Значения rad $\boldsymbol{x}_k$, $k = 1, 2, \ldots, n$ показывают величины интервальной неопределенности отдельных измерений выборки.

Информационным множеством для оценивания постоянной величины по выборке интервальных данных будет интервал, который называют *информационным интервалом* [8]. Другими словами, это интервал, содержащий значения оцениваемой величины, которые совместны с измерениями выборки (согласуются с данными этих измерений).

8.1.1 Совместность выборки.

Для описания внутреннего свойства интервальной выборки, которое характеризует согласование её данных между собой вводят понятие совместности:

Определение. Выборка $\{\boldsymbol{x}_k\}_{k=1}^n$ называется *совместной*, если пересечение всех интервалов составляющих её измерений непусто, т. е.

$$\bigcap_{1\le k\le n} \boldsymbol{x}_k \neq \varnothing.$$

В противном случае, если пересечение всех интервалов $\boldsymbol{x}_k$, $k = 1,\dots,n$, является пустым, то выборка называется *несовместной*.

8.1.2 Индекс Жаккара.

Для описания выборок, помимо оценок их размеров, желательно иметь дополнительную информацию о мере сходства элементов выборки. В различных областях анализа данных, биологии, информатике, в науках о Земле используют различные меры сходства множеств. Для подобных конструкций часто используется термин *индекс Жаккара*, по имени математика, предложившего эту меру в начале XX века.

Обобщение меры Жаккара на выборки интервалов дано в [15]. В качестве числовой характеристики степени совпадения двух интервалов $\boldsymbol{x}, \boldsymbol{y}$ рассмотрим величину

$$Ji(\boldsymbol{x},\boldsymbol{y}) := \frac{\mathrm{wid}\,(\boldsymbol{x}\wedge\boldsymbol{y})}{\mathrm{wid}\,(\boldsymbol{x}\vee\boldsymbol{y})}. \tag{8.2}$$

В общем случае нижняя грань по включению в числителе выражения (8.2) может быть неправильным интервалом, и её ширина тогда отрицательна.

Рассмотренная мера обобщает обычное понятие меры совместности на различные типы взаимной совместности интервалов. Если пересечение интервалов $\boldsymbol{x}, \boldsymbol{y}$ пусто, т. е. $\boldsymbol{x}\cap\boldsymbol{y} = \varnothing$, то $\boldsymbol{x}\wedge\boldsymbol{y}$ — неправильный интервал и числитель формулы (8.2) имеет отрицательное значение. В предельном случае несовпадающих вещественных вырожденных интервалов $\boldsymbol{x} = x$ и $\boldsymbol{y} = y$, $x \neq y$, имеем

$$Ji(x,y) = -1.$$

В целом получаем

$$-1 \le Ji(\boldsymbol{x},\boldsymbol{y}) \le 1. \tag{8.3}$$

Таким образом, величина Ji непрерывно описывает ситуации от полной несовместности вещественных значений $x \neq y$ до полного перекрытия интервалов $\boldsymbol{x} = \boldsymbol{y}$.

Мера совместности, введённая для двух интервалов в форме (8.2), допускает естественное обобщение на случай интервальной выборки $\boldsymbol{X} = \{\boldsymbol{x}_i\}$, $i = 1, 2, \ldots, n$. Определим меру $Ji(\boldsymbol{X})$ для этой выборки как

$$Ji(\boldsymbol{X}) = \frac{\text{wid}\left(\bigwedge_i \boldsymbol{x}_i\right)}{\text{wid}\left(\bigvee_i \boldsymbol{x}_i\right)}. \tag{8.4}$$

Видно, что выражение (8.4) переходит в случае интервальной выборки из двух элементов в выражение (8.2).

Пример 8.1.1. Пример вычисления меры совместности для накрывающей выборки. Пусть имеется интервальная выборка из четырех элементов

$$\boldsymbol{X} = \{[1, 4], [5, 9], [1, 5, 4, 5], [6, 9]\}.$$

Выберем из неё накрывающую подвыборку

$$\boldsymbol{X}_c = \{[5, 9], [6, 9]\}.$$

Для выборки $\boldsymbol{X}_c$ имеем, по формуле (8.4),

$$\text{Ji}(\boldsymbol{X}_c) = \frac{9 - 6}{9 - 5} = 0.75.$$

Значение $\text{Ji}(\boldsymbol{X}_c)$, близкое единице, демонстрирует высокую меру сходства элементов выборки $\boldsymbol{X}_c$. ■

8.2 Арифметика твинов

8.2.1 Определение и операции твинной арифметики в нотации В. М. Нестерова

Достаточно подробно твинная арифметика была рассмотрена в диссертации В. М. Нестерова [12].

Определение. Твин – это пара интервалов $T = (X_l, X)$, где $X_l \in I(\mathbb{R}) \cup \{\emptyset\}$ и $X \in I(\mathbb{R})$.

Оценить неизвестный интервал I твином – это значит найти такой твин $T = (X_l, X)$, что $X_l \subseteq I \subseteq X$. Обозначим это как $I \sqsubseteq T$.

Определим «внутреннюю длину твина» как $|T|_l = |X_l|$. Аналогично определим «внешнюю длину твина» как $|T| = |X|$. Запись $I \sqsubseteq (\emptyset, X)$

обозначает, что существует только внешняя оценка I, равная X. Будем формально полагать, что в этом случает $|T|_l = -1$.

Твин $(\emptyset, [A, A])$ будем отождествлять с твином $([A, A], [A, A])$. Для твина $T = ([A, A], [A, A])$ будем считать, что $|T|_l = 0$.

Пусть $\diamond$ и $\circ$ – любая унарная и любая бинарная операция соответственно. Тогда все операции вводимой арифметики твинов должны удовлетворять следующим свойствам:

$$X \sqsubseteq T \Rightarrow \diamond X \sqsubseteq \diamond T; \tag{8.5}$$

$$X \sqsubseteq T_1 \ \ \& \ \ Y \sqsubseteq T_2 \Rightarrow X \circ Y \sqsubseteq T_1 \circ T_2. \tag{8.6}$$

Определение. Пусть внутренние интервалы твинов T_1 и T_2 непусты. Тогда обозначим

$$\begin{aligned} T_1 &= ([a^-, a^+], [A^-, A^+]), \\ T_2 &= ([b^-, b^+], [B^-, B^+]). \end{aligned} \tag{8.7}$$

Определение. Далее нам понадобится определение чисел p и q:

- Если внутренние интервалы твинов T_1 и T_2 непусты, то

$$p = \min(a^- + B^+, b^- + A^+); \tag{8.8}$$

$$q = \max(a^+ + B^-, b^+ + A^-). \tag{8.9}$$

- Если внутренний интервал твина T_1 является пустым множеством (но внутренний интервал твина T_2 непустой), то

$$\begin{aligned} p &= b^- + A^+ \\ q &= b^+ + A^-. \end{aligned} \tag{8.10}$$

- Если внутренний интервал твина T_2 является пустым множеством (но внутренний интервал твина T_1 непустой), то

$$\begin{aligned} p &= a^- + B^+ \\ q &= a^+ + B^-. \end{aligned} \tag{8.11}$$

- Если оба твина T_1 и T_2 одновременно являются вырожденными, то p и q являются неопределёнными.

Определение. Пусть Z – пустое или одноэлементное множество, $I_1, I_2 \in I(\mathbb{R})$. Тогда

$$\phi(I_1, I_2) = \begin{cases} \min\limits_{\subseteq}\{[c^-, c^+] \mid (c^- \in I_1 \ \ \& \ \ c^+ \in I_2) \vee \\ \qquad\qquad (c^- \in I_2 \ \ \& \ \ c^+ \in I_1)\}, & \text{если } I_1 \cap I_2 = Z; \\ \emptyset, & \text{в остальных случаях.} \end{cases} \tag{8.12}$$

Определение.

$$\psi(I_1, I_2) = \max_{\subseteq}\{[c^-, c^+] \mid c^-, c^+ \in I_1 \cup I_2\} \tag{8.13}$$

8.2.2 Формулы твинной арифметики

Теперь введём правила основных арифметических операций, которые используются в твинной арифметике.

Сложение Сложение в твинной арифметике задаётся следующей формулой:

$$T_1 + T_2 = \begin{cases} ([p, q], [A^- + B^-, A^+ + B^+]), & \text{если } |T_1| \leq |T_2|_l \vee |T_2| \leq |T_1|_l; \\ (\emptyset, \quad [A^- + B^-, A^+ + B^+]), & \text{в остальных случаях.} \end{cases}, \tag{8.14}$$

где p и q – это числа, определённые по определению 8.2.1.

Сложение твинов обладает свойствами коммутативности и ассоциативности.

Умножение Формула умножения твинов зависит от пустоты внутренних интервалов твинов-операндов:

- Если $|T_1|_l \neq -1$ и $|T_2|_l \neq -1$, то

$$T_1 \cdot T_2 = \Big(\psi\big(\phi(a^-[B^-, B^+], a^+[B^-, B^+]), \phi(b^-[A^-, A^+], b^+[A^-, A^+])\big), \tag{8.15}$$

$$[A^-, A^+] \cdot [B^-, B^+]\Big) \tag{8.16}$$

- Если $|T_1|_l = -1$ и $|T_2|_l \neq -1$, то

$$T_1 \cdot T_2 = \Big(\phi\big(b^-[A^-, A^+], b^+[A^-, A^+]\big), [A^-, A^+] \cdot [B^-, B^+]\Big) \tag{8.17}$$

- Если $|T_1|_l \neq -1$ и $|T_2|_l = -1$, то

$$T_1 \cdot T_2 = \Big(\phi\big(a^-[B^-, B^+], a^+[B^-, B^+]\big), [A^-, A^+] \cdot [B^-, B^+]\Big) \tag{8.18}$$

- Если $|T_1|_l = -1$ и $|T_2|_l = -1$, то

$$T_1 \cdot T_2 = (\emptyset, [A^-, A^+] \cdot [B^-, B^+]) \tag{8.19}$$

Умножение твинов обладает свойствами коммутативности и ассоциативности.

Другие операции Также вводят операции унарного минуса и взятие обратного твина:

$$-T_1 = (-[a^-, a^+], -[A^-, A^+]) \tag{8.20}$$

$$\frac{1}{T_1} = \left(\frac{1}{[a^-, a^+]}, \frac{1}{[A^-, A^+]}\right), \quad 0 \notin [A^-, A^+]. \tag{8.21}$$

Необходимо отметить, что введённые операции твинной арифметики удовлетворяют свойствам 8.5 и 8.6.

Глава 9

Практика

Блок лабораторных работ №1 (работы №1-4)

Целью первого блока лабораторных работ является знакомство с основными методами описательной статистики: визуализацией распределений, вычислением характеристик положения и рассеяния, анализом выбросов и построением эмпирических оценок функций распределения и плотности.

Исследуются следующие **распределения**:

1. Нормальное $N(x; 0, 1)$
2. Коши $C(x; 0, 1)$
3. Лапласа $L(x; 0, \frac{1}{\sqrt{2}})$
4. Пуассона $P(k; 5)$
5. Равномерное $U(x; -\sqrt{3}, \sqrt{3})$

Лабораторная работа №1: Гистограммы и плотности распределений

Задание: Сгенерировать выборки объёмом $n = 10, 50, 1000$ для каждого из 5 распределений. Построить на одном графике гистограмму выборки и теоретическую кривую плотности распределения.

Цели и задачи:

- Освоить принцип группировки данных и построения гистограмм.
- Сделать вывод о том, как влияет размер выборки на определение характера распределения величины с помощью гистограммы.

Ключевые моменты:

- **Количество интервалов (бинов)**

  Гистограммы должны быть представительны и содержать оптимальное количество бинов. Если взять слишком мало разрядов, получим неинформативный график. Если взять слишком много разрядов, то данных для построения в таблице может оказаться недостаточно. Слишком мало интервалов → потеря информации; слишком много → шум.

Лабораторная работа №2: Характеристики положения и рассеяния

Задание: Сгенерировать выборки объёмом $n = 10, 100, 1000$. Для каждой выборки вычислить:

1. Выборочное среднее $\bar{x}$
2. Медиану med
3. Полусумму экстремальных элементов $z_R = \frac{x_{\min} + x_{\max}}{2}$
4. Полусумму квартилей $z_Q = \frac{z_{1/4} + z_{3/4}}{2}$
5. Усечённое среднее z_{tr} (с отбрасыванием 10% наименьших и наибольших значений).

Повторить генерацию и вычисления 1000 раз. Найти среднее и дисперсию каждой характеристики:

$$E(z) = \overline{z}, \quad D(z) = \overline{z^2} - \overline{z}^2.$$

Цели и задачи:

- Исследовать сходимость выборочных характеристик к теоретическим при росте n.

- Исследовать оценки характеристик положения на устойчивость к выбросам.

Ключевые моменты:

- **Робастность**

  Среднее арифметическое далеко не всегда является лучшей относительной характеристикой ввиду своей неустойчивости к наличию в выборке аномальных значений. Необходимо наглядно продемонстрировать этот и альтернаивные подходы к оценке истинного центра распределения.

- **Округление**

  В оценке $x = E \pm \sqrt{D}$ вариации подлежит первая цифра после точки. В данном случае $x = 0.0 \pm 0.1k$, где k $-$ зависит от доверительной вероятности и вида распределения (рассматривается в дальнейшем цикле лабораторных работ).

  Округление проводить для $k = 1$

Лабораторная работа №3: Боксплот Тьюки и анализ выбросов

Задание:

1. Сгенерировать выборки объёмом $n = 20, 100$. Построить боксплот Тьюки.

2. Определить долю выбросов экспериментально: для каждого распределения сгенерировать 1000 выборок заданного объёма, вычислить среднюю долю выбросов.

Цели и задачи:

- Научиться применять боксплот Тьюки в обработке и анализе одномерных распределений.
- Проанализировать, как размер выборки влияет на долю отсеиваемых аномальных значений.

Ключевые моменты:

- **Доля аномальных значений**

  Практическая доля выбросов в выборке ситается как отношение количества выбросов к количеству всех элементов в выборке. В данной работе необходимо посчитать среднюю долю выбросов, сгнерировав выборку 1000 раз.

Лабораторная работа №4: Эмпирическая функция распределения и ядерные оценки плотности

Задание: Сгенерировать выборки объёмом $n = 20, 60, 100$. Построить на них:

1. Эмпирическую функцию распределения (э. ф. р.) и теоретическую функцию распределения.
2. Ядерную оценку плотности (ядро — гауссово) и теоретическую плотность.

Пояснение: 1–2 выполнить на отрезке $[-4, 4]$ для непрерывных распределений и на отрезке $[6, 14]$ для распределения Пуассона.

Цели и задачи:

- Сделать вывод о том, как мощность выборки влияет на приближение эмпирической функции к истинной функции распределения.
- Сравнить два способа оценки плотности распределения: ядерные оценки и гистограммы.

Ключевые моменты:

- **Эмпирическая функция распределения**

  Ступенчатая функция, сходится по вероятности к истинной функции распределения (теорема Гливенко–Кантелли).

- **Ядерное сглаживание**

  График плотности распределения, полученный с помощью гистограммы, является ступенчатой функцией. Реальные функции распределения непрерывны. Следовательно, гистограммные оценки хорошо аппроксимируют функции распределения в случаях больших объемов наблюдений или дискретных распределений. В общем случае мы стремимся получить более гладкие оценки плотности вероятности. Для этого используют модель построения называемую ядерной оценки плотности. Необходимо наглядно продемонстрировать описанную ранее разницу в двух подходах к оценке плотности вероятности.

Блок лабораторных работ №2 (работы №5-8)

Цель второго блока лабораторных работ — освоить методы статистического вывода: оценку параметров распределений, построение доверительных интервалов, проверку статистических гипотез и анализ связей между переменными. Основное внимание уделяется переходу от описания данных к вероятностным выводам о генеральной совокупности.

Лабораторная работа №5: Корреляционный анализ и эллипсы рассеивания

Задание:

1. Сгенерировать двумерные выборки (x, y) объёмом $n = 20, 60, 100$ из:

   - Двумерного нормального распределения $N(0, 0, 1, 1, \rho)$ с $\rho = 0, 0.5, 0.9$.

- Смеси: $0.9N(0, 0, 1, 1, 0.9) + 0.1N(0, 0, 10, 10, -0.9)$.

2. Для каждой выборки вычислить коэффициенты корреляции Пирсона, Спирмена и квадрантный. Повторить 1000 раз, найти средние и дисперсии.

3. Построить диаграммы рассеяния и эллипсы равновероятности.

Цели и задачи:

- Сравнить исходное заначение коэффициента корреляции с вычисленными

- Изучить способы визуализации ковариационной структуры данных.

Ключевые моменты:

- **Коэффициент корреляции**

  Коэффициент корреляции показывает тесноту линейной взаимосвязи и изменяется в диапазоне $[-1, 1]$. -1 означает полную (функциональную) линейную обратную взаимосвязь. 1 — полную (функциональную) линейную положительную взаимосвязь. 0 — отсутствие линейной корреляции (но не обязательно взаимосвязи). Необходимо содержательно охарактеризовать полученные результаты.

Лабораторная работа №6: Линейная регрессия. Метод наименьших квадратов и наименьших модулей

Задание:

1. Задать равномерные значения x_i на отрезке $[-1.8, 2]$ с шагом 0.2 (20 точек).

2. Сгенерировать $y_i = 2 + 2x_i + \varepsilon_i$, где $\varepsilon_i \sim N(0, 1)$.

3. Оценить параметры a, b двумя методами:

   - Метод наименьших квадратов (МНК).

- Метод наименьших модулей (МНМ).

4. Повторить пункты 1–3, добавив выбросы: $y_1 \leftarrow y_1 + 10,\ y_{20} \leftarrow y_{20} - 10$.

Цели и задачи:

- Сравнить устойчивость МНК и МНМ к выбросам.
- Построить графики регрессионных прямых и исходных данных.

Ключевые моменты:

- Относительная погрешность вычисляется по формуле:

$$\delta\, z = \frac{|z - \hat{z}|}{z} \cdot 100\%$$

  где z - точное значение, $\hat{z}$ - приближённое

- Оформить результаты в таблицу:

	a	$\Delta\, a$	$\delta\, a$, %	b	$\Delta\, b$	$\delta\, b$, %
МНК						
МНМ						

Лабораторная работа №7: Проверка гипотез о законе распределения. Критерий χ^2

Задание:

1. Сгенерировать выборку объёмом $n = 100$ из $N(0, 1)$.
2. Оценить параметры μ, σ методом максимального правдоподобия (ММП).
3. Проверить гипотезу H_0: выборка из $N(\hat{\mu}, \hat{\sigma})$ с помощью критерия χ^2 при $\alpha = 0.05$.
4. Исследовать чувствительность критерия: сгенерировать выборки из равномерного распределения и распределения Лапласа ($n = 20$) и проверить их на нормальность.

Цели и задачи:

- Исследовать принцип работы критерия χ^2 и его ограничения.
- Оценить мощность критерия при альтернативных распределениях.

Ключевые моменты:

- **Построение интервалов:** разбить область значений на k интервалов, чтобы ожидаемые частоты $np_i \geq 5$.

Лабораторная работа №8: Доверительные интервалы. Критерий Стьюдента и F-тест

Задание:

1. Сгенерировать две выборки объёмом $n_1 = 20,\ n_2 = 100$ из $N(0, 1)$.
2. Построить доверительные интервалы для:
   - Математического ожидания (на основе $\bar{x}$ и t-распределения Стьюдента).
   - Дисперсии (на основе s^2 и χ^2-распределения).
3. Проверить гипотезу о равенстве дисперсий двух независимых выборок с помощью **F-теста** (критерий Фишера).

Цели и задачи:

- Освоить построение интервальных оценок параметров.
- Исследовать связь между точечными и интервальными оценками.

Требования к отчету

Отчет о выполнении лабораторных работ должен быть оформлен в виде документа в формате PDF и размещен в репозитории на платформе GitHub вместе с исходным кодом программ.

Структура отчета должна включать следующие обязательные разделы в указанном порядке:

1. **Титульный лист**.
2. **Постановка задачи**. Формулировка цели и задач лабораторной работы.
3. **Теоретическая часть**. Изложение необходимого теоретического материала, используемого в работе. Нумерация присваивается только тем формулам, на которые имеются ссылки в последующих разделах. Каждая такая формула сопровождается своим порядковым номером.
4. **Реализация**. Описание практической части работы с указанием примененного языка программирования, среды разработки, использованных программных библиотек, а также краткое изложение сути реализованных методов, пояснение связи с теоретической частью.
5. **Результаты**. Представление полученных результатов в виде графиков, таблиц и текстового описания. Все иллюстрации и таблицы должны быть снабжены содержательными подписями.
6. **Обсуждение**. Анализ полученных результатов, выводы и оценка соответствия результатов поставленным задачам.
7. **Выводы**. Выводы, полученные в ходе работы.
8. **Список литературы**. Перечень использованных литературных источников, оформленный в соответствии с принятыми стандартами.
9. **Приложение**. Раздел, содержащий ссылку на репозиторий GitHub с исходным кодом программ, разработанных в ходе выполнения работ.

Литература

[1] Вероятностные разделы математики. Учебник для бакалавров технических направлений.//Под ред. Максимова Ю.Д. — Спб.: «Иван Федоров», 2001. — 592 с., илл.

[2] Histogram [Электронный ресурс] // Wikipedia. – URL: `https://en.wikipedia.org/wiki/Histogram` (дата обращения: 01.02.2026).

[3] Box plot [Электронный ресурс] // Wikipedia. – URL: `https://en.wikipedia.org/wiki/Box_plot` (дата обращения: 01.02.2026).

[4] Анатольев, Станислав. Непараметрическая регрессия // Квантиль. — 2009. — № 7. — С. 37-52.

[5] Вентцель Е.С. Теория вероятностей: учеб. для вузов. — 6-е изд., стер. — М.: Высш. шк., 1999. — 576 с.

[6] Максимов Ю.Д. Математика. Теория и практика по математической статистике. Конспект-справочник по теории вероятностей: учеб. пособие / Ю.Д. Максимов; под ред. В.И. Антонова. — СПб.: Изд-во Политехн. ун-та, 2009. — 395 с. — (Математика в политехническом университете).

[7] Кобзарь А.И. Прикладная математическая статистика. — М.: Физматлит, 2006. — 816 с.

[8] Баженов А.Н., Жилин С.И., Кумков С.И., Шарый С.П. Обработка и анализ интервальных данных / Сер. Интервальный анализ и его приложения. — М.; Ижевск: ИКИ, 2024. — 356 с.

[9] Шрейдер Ю.А. Равенство, сходство, порядок. — М.: Наука, 1971. — 256 с.

[10] Шарый С.П. Конечномерный интервальный анализ. — Новосибирск: ФИЦ ИВТ, 2024. — URL: http://www.nsc.ru/interval/Library/InteBooks/SharyBook.pdf (дата обращения: 01.02.2026).

[11] Gardeñes E., Trepat A., Janer J.M. Approaches to simulation and to the linear problem in the SIGLA system // Freiburger Intervall-Berichte. — 1981. — No. 8. — S. 1–28.

[12] Нестеров В.М. Твинные арифметики и их применение в методах и алгоритмах двустороннего интервального оценивания: дис. ... д-ра физ.-мат. наук. — СПб., 1999. — 234 с.

[13] Яворук Т. Twin: библиотека для работы с твинной арифметикой Нестерова [Электронный ресурс]. — URL: https://github.com/Tatiana655/Twins (дата обращения: 12.02.2026).

[14] Мордовин Н. FuzzyNumbers: библиотека для работы с нечеткими числами [Электронный ресурс]. — URL: https://github.com/MordovinNik/FuzzyNumbers (дата обращения: 01.02.2026).

[15] Баженов А.Н., Тельнова А.Ю. Обобщение коэффициента Жаккара для анализа данных с интервальной неопределённостью // Измерительная техника. — 2022. — № 12. — С. 12-19.

Предметный указатель