ФЕДЕРАЛЬНОЕ АГЕНТСТВО

ПО ТЕХНИЧЕСКОМУ РЕГУЛИРОВАНИЮ И МЕТРОЛОГИИ

НАЦИОНАЛЬНЫЙ СТАНДАРТ РОССИЙСКОЙ ФЕДЕРАЦИИ

ГОСТ Р ИСО 16269-4— 2017

Статистические методы

СТАТИСТИЧЕСКОЕ ПРЕДСТАВЛЕНИЕ ДАННЫХ

Часть 4

Выявление и обработка выбросов

(ISO 16269-4:2010, Statistical interpretation of data — Part 4: Detection and treatment of outliers, IDT)

Издание официальное

Москва
Стандартинформ
2017

Предисловие

1 ПОДГОТОВЛЕН Открытым акционерным обществом «Научно-исследовательский центр контроля и диагностики технических систем» (АО «НИЦ КД») на основе собственного перевода на русский язык англоязычной версии международного стандарта, указанного в пункте 4

2 ВНЕСЕН Техническим комитетом по стандартизации ТК 125 «Применение статистических методов»

3 УТВЕРЖДЕН И ВВЕДЕН В ДЕЙСТВИЕ Приказом Федерального агентства по техническому регулированию и метрологии от 10 августа 2017 г. № 865-ст

4 Настоящий стандарт идентичен международному стандарту ИСО 16269-4:2010 «Статистическое представление данных. Часть 4. Выявление и обработка выбросов» (ISO 16269-4:2010 «Statistical interpretation of data — Part 4: Detection and treatment of outliers», IDT).

Международный стандарт разработан Техническим комитетом ISO/TC 69.

Наименование настоящего стандарта изменено относительно наименования указанного международного стандарта для приведения в соответствие с ГОСТ Р 1.5—2012 (пункт 3.5).

При применении настоящего стандарта рекомендуется использовать вместо ссылочных международных стандартов соответствующие им национальные стандарты Российской Федерации, сведения о которых приведены в дополнительном приложении ДА

5 ВВЕДЕН ВПЕРВЫЕ

Правила применения настоящего стандарта установлены в статье 26 Федерального закона от 29 июня 2015 г. № 162-ФЗ «О стандартизации в Российской Федерации». Информация об изменениях к настоящему стандарту публикуется в ежегодном (по состоянию на 1 января текущего года) информационном указателе «Национальные стандарты», а официальный текст изменений и поправок — в ежемесячном информационном указателе «Национальные стандарты». В случае пересмотра (замены) или отмены настоящего стандарта соответствующее уведомление будет опубликовано в ближайшем выпуске ежемесячного информационного указателя «Национальные стандарты». Соответствующая информация, уведомление и тексты размещаются также в информационной системе общего пользования — на официальном сайте Федерального агентства по техническому регулированию и метрологии в сети Интернет (www.gost.ru)

Содержание

Введение

Выявление выбросов — одна из старейших проблем анализа данных. Причинами появления выбросов могут быть ошибки измерений, ошибки отбора выборки, преднамеренное искажение или некорректная фиксация результатов анализа выборки, ошибочные предположения о распределении данных или модели, малое количество наблюдений и т. д.

Выбросы могут искажать и сокращать информацию, содержащуюся в источнике данных или процедуре их генерации. В производстве наличие выбросов снижает результативность производственных процессов, качество продукции, а также процедур контроля продукции. Выбросы не всегда следует трактовать как «плохие» или «ошибочные» данные. В некоторых случаях выбросы дают важную информацию, которую необходимо учитывать в процессе исследований.

Выявление и анализ выбросов в процессе измерения ведут к более полному пониманию изучаемых процессов и более глубокому анализу данных, и как следствие, к более достоверным выводам.

Так как проблеме обнаружения и обработки выбросов посвящено большое количество литературных публикаций, важной задачей является определение и стандартизация (на международном уровне) этих методов.

Настоящий стандарт содержит шесть приложений. В приложении А приведен алгоритм вычисления статистик и критических значений для выявления выбросов в выборке из нормально распределения. В приложениях В, D и Е приведены таблицы, необходимые для применения рекомендованных в стандарте процедур. В приложении С приведено статистическое обоснование построения диаграмм, помогающих в решении задачи отслеживания выбросов. В приложении F приведено поэтапное руководство по применению процедур, установленных в настоящем стандарте, и представлена блок-схема соответствующих действий.

НАЦИОНАЛЬНЫЙ СТАНДАРТ РОССИЙСКОЙ ФЕДЕРАЦИИ

Статистические методы

СТАТИСТИЧЕСКОЕ ПРЕДСТАВЛЕНИЕ ДАННЫХ

Часть 4

Выявление и обработка выбросов

Statistical methods. Statistical data presentation. Part 4. Detection and treatment of outliers

Дата введения — 2018—12—01

1 Область применения

В настоящем стандарте установлены статистические критерии и методы графического анализа данных, полученные в результате измерений. В настоящем стандарте приведены рекомендации по методам определения робастных оценок и процедурам проверки наличия выбросов в данных.

Методы, представленные в настоящем стандарте, предназначены главным образом для выявления и обработки выбросов одномерных данных. Однако в настоящем стандарте представлены также некоторые рекомендации по работе с многомерными данными и данными регрессионного анализа.

2 Термины и определения

В настоящем стандарте применены следующие термины с соответствующими определениями:

2.1 **выборка, набор данных** (sample, data set): Подмножество генеральной совокупности, состоящее из одной или нескольких выборочных единиц.

Примечание 1 — В зависимости от исследуемой генеральной совокупности выборочными единицами могут быть объекты, числовые значения, а также абстрактные элементы.

Примечание 2 — Выборку из генеральной совокупности, подчиняющуюся нормальному распределению (2.22), гамма-распределению (2.23), экспоненциальному распределению (2.24), распределению Вейбулла (2.25), логнормальному распределению (2.26) или распределению экстремальных значений типа I (2.27) часто называют выборкой из нормального распределения, гамма-распределения, экспоненциального распределения, распределения Вейбулла, логнормального распределения или распределения экстремальных значений типа I соответственно.

2.2 **выброс** (outlier): Элемент маломощного подмножества выборки, существенно отличающийся от остальных элементов выборки (2.1).

Примечание 1 — Классификация наблюдения или подмножество выборки как выброс (или выбросы) зависит от выбранной модели генеральной совокупности, из которой отобрана выборка. Выброс не рассматривают как истинный элемент генеральной совокупности.

Примечание 2 — Выброс может появиться из другой генеральной совокупности, быть результатом некорректной регистрации данных или общей ошибкой измерений.

Примечание 3 — Подмножество может содержать одно или несколько наблюдений.

2.3 **маскировка** (masking): Наличие более одного выброса (2.2), затрудняющее обнаружение каждого выброса.

2.4 **вероятность ложного обнаружения выбросов** (some-outside rate): Вероятность того, что одно или несколько наблюдений незагрязненной выборки, ошибочно классифицированы как выбросы (2.2).

2.5 **метод коррекции выбросов** (outlier accommodation method): Метод нечувствительный к наличию выбросов (2.2) при принятии решения о генеральной совокупности.

Издание официальное

2.6 **устойчивая оценка** (resistant estimation): Оценка, подверженная лишь малым изменениям при замене небольшой доли набора данных (2.1), элементами, возможно, имеющими значительное отличие от замененных элементов.

2.7 **робастная оценка** (robust estimation): Оценка, нечувствительная к небольшим отклонениям от предполагаемой вероятностной модели данных.

Примечание — Примером может быть оценка, полученная методом, предназначенным для нормального распределения (2.2), при применении к близким распределениям, но имеющим некоторую асимметрию или тяжелые хвосты функции распределения. Группа таких оценок включает в себя L-оценки (взвешенное среднее арифметическое порядковых статистик (2.10)) и М-оценки (см. [9]).

2.8 **ранг** (rank): Положение наблюдаемого значения в упорядоченном наборе наблюдаемых значений.

Примечание 1 — Наблюдаемые значения упорядочивают в неубывающем (ведя отсчет от наименьшего элемента) или в невозрастающем (ведя отсчет от наибольшего элемента) порядке.

Примечание 2 — В соответствии с целями настоящего стандарта одинаковым наблюдаемым значениям присваивают разные, но последовательные ранги.

2.9 **глубина** (depth): Наименьший из двух рангов (2.8), присвоенных элементу при упорядочивании выборки (2.1) в неубывающем и невозрастающем порядках.

Примечание 1 — Значение глубины может быть не целым числом (см. приложение А).

Примечание 2 — Для всех полученных значений, отличных от медианы (2.11), глубина определяет два значения — одно ниже медианы, другое выше медианы. Например, два значения с глубиной 1 представляют собой минимальное и максимальное значение в выборке (2.1).

2.10 **порядковая статистика** (order statistic): Статистика, определяемая рангом при упорядочивании набора данных в неубывающем порядке.

[ИСО 3534-1:2006, п. 1.9]

Примечание 1 — Пусть $\{x_1, x_2, ..., x_n\}$ — неупорядоченная выборка. После ее упорядочивания, обозначенные заново элементы составляют упорядоченную выборку, где $x_{(1)} \le x_{(2)} \le ... \le x_{(k)} \le ... \le x_{(n)}$, тогда $x_{(k)}$ — наблюдаемое значение k-й порядковой статистики в выборке объема n.

Примечание 2 — На практике для определения порядковых статистик данных в выборке (2.1) производят их упорядочивание в соответствии с примечанием 1.

2.11 **медиана, выборочная медиана, медиана набора чисел Q_2** (median, sample median, median of a set of numbers, Q_2): k-я порядковая статистика, где $k = \left[\frac{n+1}{2}\right]$, если объем выборки — нечетное число или полусумма $\left[\frac{n}{2}\right]$-й и $\left[\frac{n+1}{2}\right]$-й порядковых статистик, если n — четное число.

[ИСО 3534-1:2006, п. 1.13]

Примечание — Медиана является вторым квартилем (Q_2).

2.12 **первый квартиль, нижний выборочный квартиль Q_1** (first quartile sample lower quartile, Q_1): Медиана (2.11) первых наименьших $(n — 1)/2$ значений для нечетного числа наблюдений; медиана первых наименьших $n/2$ значений для четного числа наблюдений.

Примечание 1 — В литературе встречается много различных определений выборочного квартиля, что приводит в некоторой степени к различным выводам. В настоящем стандарте приведено определение, которое широко распространено и удобно в применении.

Примечание 2 — Популярными вариантами квартиля являются «сгибы» и «четверти» (2.19 и 2.20). В некоторых случаях (см. примечание 3 в 2.19) первый квартиль и нижняя четверть (2.19) идентичны.

2.13 **третий квартиль, верхний выборочный квартиль Q_3** (third quartile, sample upper quartile, Q_3): Медиана (2.11) последних наибольших $(n — 1)/2$ значений для нечетного числа наблюдений или медиана последних наибольших $n/2$ значений для четного числа наблюдений.

Примечание 1 — В литературе встречается много различных определений выборочного квартиля, что приводит в некоторой степени к различным выводам. В настоящем стандарте приведено определение, которое широко распространено и удобно в применении.

Примечание 2 — Популярными вариантами квартиля являются «сгибы» и «четверти» (2.19 и 2.20). В некоторых случаях (см. примечание 3 в 2.20) третий квартиль и верхняя четверть (2.20) идентичны.

2.14 **межквартильный размах IQR** (interquartile range, IQR): Разность третьего квартиля (2.13) и первого квартиля (2.12).

Примечание 1 — Межквартильный размах — широко применяемая статистика для описания рассеяния данных.

Примечание 2 — Иногда вместо межквартильного размаха используют разность верхней четверти (2.20) и нижней четверти (2.19), называемую «четвертным разбросом».

2.15 **сводка пяти чисел** (five-number summary): Набор значений выборочного минимума, первого квартиля (2.12), медианы (2.11), третьего квартиля (2.13) и выборочного максимума.

Примечание — Сводка пяти чисел дает краткую количественную информацию о положении, рассеянии и размахе данных.

2.16 **диаграмма ящик с усами** (box plot): Графическое представление (горизонтальное или вертикальное) сводки пяти чисел (2.15).

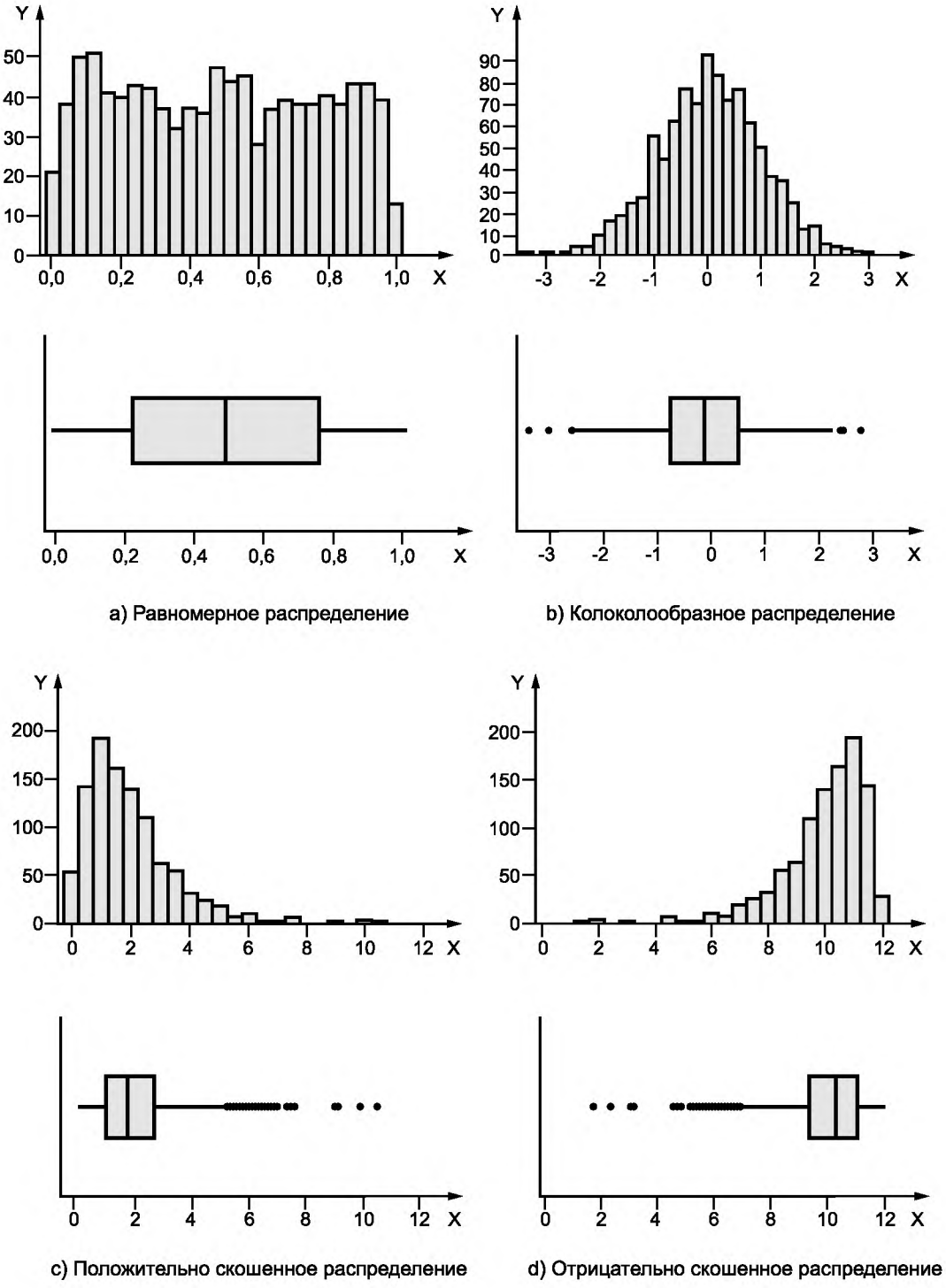
Примечание 1 — В случае горизонтального представления диаграммы ящик с усами, первый квартиль (2.12) и третий квартиль (2.13) наносят на диаграмму как левую и правую боковые стороны ящика, медиану (2.11) наносят как вертикальную линию, перерезающую ящик; левый ус идет от первого квартиля к наименьшему значению в выборке, не выходящему за нижнюю границу (2.17), правый ус идет от третьего квартиля к наибольшему значению, не выходящему за верхнюю границу (2.18); значения за пределами контрольных границ рассматривают как выбросы. В случае вертикального представления диаграммы, первый и третий квартили наносят на диаграмму, как нижнюю и верхнюю стороны ящика, медиану наносят как горизонтальную линию, перерезающую ящик; нижний ус идет от первого квартиля к наименьшему значению в выборке, не выходящему за нижнюю границу, верхний ус идет от третьего квартиля к наибольшему значению, не выходящему за верхнюю границу; значения за пределами контрольных границ рассматривают как выбросы.

Примечание 2 — Ширина ящика и длина уса — графические параметры диаграммы, характеризующие данные, например, параметр положения, разброс, асимметрию, длину хвостов и выбросы. На рисунке 1 для сравнения представлена диаграмма ящик с усами и функция плотности для a) равномерного, b) колоколообразного, c) положительно скошенного и d) отрицательно скошенного распределений. Для каждого распределения над диаграммой ящик с усами приведена соответствующая гистограмма.

Примечание 3 — Диаграмму ящик с усами с нижней (2.17) и верхней (2.18) границами, вычисленными с использованием коэффициента k, рассчитанного на основе объема выборки n и предположении о виде распределения данных, называют модифицированной диаграммой ящик с усами (см. рисунок 2). Построение модифицированной диаграммы ящик с усами представлено в 4.4.

2.17 **нижняя граница, нижняя граница отделяющая выбросы, нижнее предельное значение** (lower fence, lower outlier cut-off, lower adjacent value): Значение, указанное на диаграмме ящик с усами (2.16), находящееся ниже первого квартиля (2.12) на заданное число k межквартильных размахов (2.14).

Примечание — В специализированных пакетах программ статистической обработки данных нижнюю границу обычно вычисляют как $Q_1 - k(Q_3 - Q_1)$, где k берут равным 1,5 или 3,0. В классическом подходе, при $k = 1,5$ нижнюю границу называют «внутренней нижней границей», а при $k = 3,0$ нижнюю границу называют «внешней нижней границей».



X — значение случайной величины; Y — частота появления X

Рисунок 1 — Диаграммы ящик с усами и соответствующие гистограммы для: a) равномерного, b) колоколообразного, c) положительно скошенного и d) отрицательно скошенного распределения

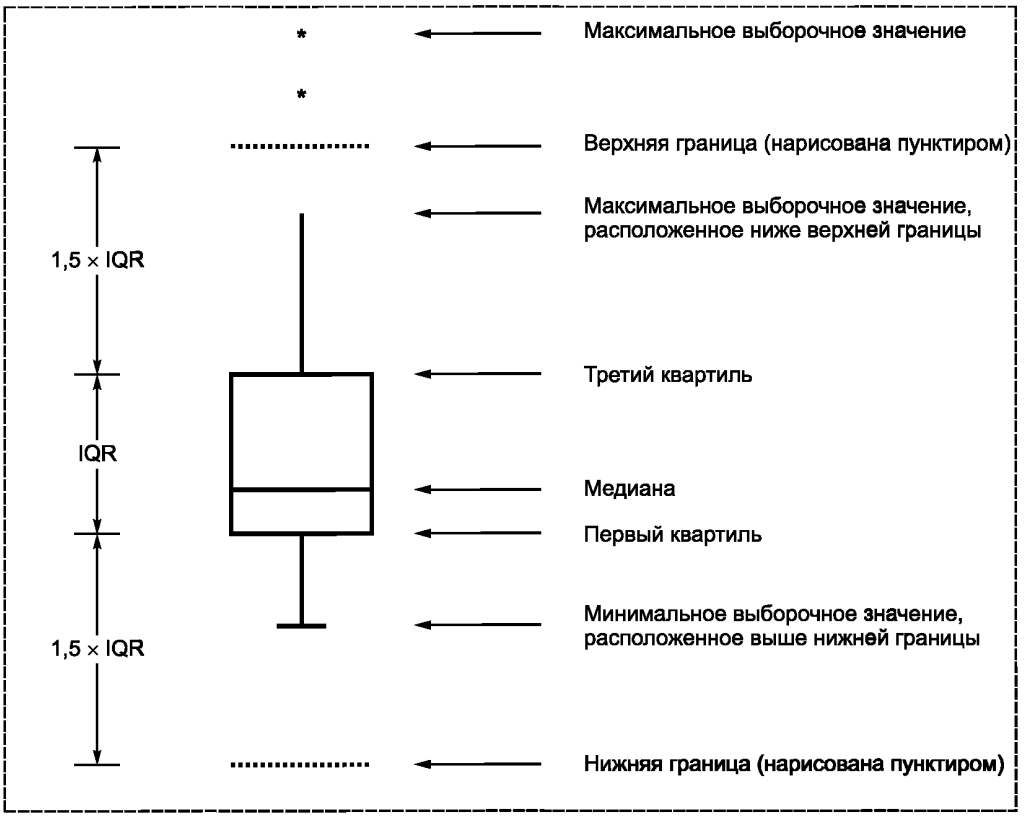


Рисунок 2 — Модифицированная диаграмма ящик с усами с указанными нижней и верхней границами

2.18 **верхняя граница, верхняя граница отделяющая выбросы, верхнее предельное значение** (upper fence, upper outlier cut-off, upper adjacent value): Значение, указанное на диаграмме ящик с усами, расположенное выше третьего квартиля (2.13) на заданное число k межквартильных размахов (2.14).

Примечание — В специализированных пакетах программ статистической обработки данных верхнюю границу обычно вычисляют как $Q_1 + k(Q_3 - Q_1)$, где k берут равным 1,5 или 3,0. В классическом подходе, при k = 1,5 верхнюю границу называют «внутренней верхней границей», а при k = 3,0 верхнюю границу называют «внешней верхней границей».

2.19 **нижняя четверть $x_{L:n}$** (lower fourth, $x_{L:n}$): Для набора наблюдаемых значений $x_{(1)} \le x_{(2)} \le ... \le x_{(n)}$ величина, равная $0{,}5[x_{(i)} + x_{(i+1)}]$ при $f = 0$ или $x_{(i+1)}$ при $f > 0$, где i — целая часть $n/4$, а f — дробная часть $n/4$.

Примечание 1 — Данное определение нижней четверти используют для вычисления рекомендуемых значений k_L и k_U (см. приложение С); во многих программных продуктах статистической обработки данных вычисление нижней четверти по умолчанию или в качестве выбираемой опции производится так, как указано в определении.

Примечание 2 — Нижнюю четверть и верхнюю четверть (2.20) вместе иногда называют сгибами.

Примечание 3 — Нижнюю четверть иногда рассматривают как первый квартиль (2.12).

Примечание 4 — При $f = 0$, $f = 0{,}5$ или $f = 0{,}75$ нижняя четверть тождественно равна первому квартилю, например:

Объем выборки n	i = целая часть $n/4$	f = дробная часть $n/4$	Первый квартиль	Нижняя четверть
9	2	0,25	$[x_{(2)} + x_{(3)}]/2$	$x_{(3)}$
10	2	0,50	$x_{(3)}$	$x_{(3)}$
11	2	0,75	$x_{(3)}$	$x_{(3)}$
12	3	0	$[x_{(3)} + x_{(4)}]/2$	$[x_{(3)} + x_{(4)}]/2$

2.20 **верхняя четверть $x_{U:n}$** (lower fourth, $x_{U:n}$): Для набора наблюдаемых значений $x_{(1)} \le x_{(2)} \le ... \le x_{(n)}$ величина, равная $0{,}5[x_{(n-i)} + x_{(n-i+1)}]$ при $f = 0$ или $x_{(n-i)}$ при $f > 0$, где i — целая часть $n/4$, а f — дробная часть $n/4$.

Примечание 1 — Данное определение верхней четверти используют для вычисления рекомендуемых значений k_U и k_L (см. приложение С); во многих программных продуктах статистической обработки данных вычисление верхней четверти по умолчанию или в качестве выбираемой опции производится так, как указано в определении.

Примечание 2 — Нижнюю четверть (2.19) и верхнюю четверть вместе иногда называют сгибами.

Примечание 3 — Верхнюю четверть иногда рассматривают как третий квартиль (2.13).

Примечание 4 — При $f = 0$, $f = 0{,}5$ или $f = 0{,}75$ верхняя четверть тождественно равна третьему квартилю, например:

Объем выборки n	i = целая часть $n/4$	f = дробная часть $n/4$	Третий квартиль	Верхняя четверть
9	2	0,25	$[x_{(7)} + x_{(8)}]/2$	$x_{(7)}$
10	2	0,50	$x_{(8)}$	$x_{(8)}$
11	2	0,75	$x_{(9)}$	$x_{(9)}$
12	3	0	$[x_{(9)} + x_{(10)}]/2$	$[x_{(9)} + x_{(10)}]/2$

2.21 **ошибка первого рода** (Type I error): Отклонение нулевой гипотезы, когда она истинна.
[ISO 3534-1:2006, п. 1.46]

Примечание 1 — Ошибка первого рода — это принятие неверного решения. Поэтому, желательно поддерживать вероятность принятия такого ошибочного решения была столь малой, насколько это возможно.

Примечание 2 — Возможно в некоторых ситуациях (например, при определении параметра биномиального распределения p), заданный уровень значимости, например, 0,05, не достижим для дискретных данных.

2.22 **нормальное распределение, распределение Гаусса** (normal distribution, Gaussian distribution): Распределение непрерывной случайной величины с функцией плотности вероятностей

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

где x — переменная $-\infty < x < \infty$; μ, σ — параметры $-\infty < \mu < \infty$, $\sigma > 0$.
[ISO 3534-1:2006, п. 2.50]

Примечание 1 — Математическое ожидание μ — параметр положения, стандартное отклонение σ — параметр рассеяния данных.

Примечание 2 — Нормальная выборка является случайной выборкой (2.1), отобранной из генеральной совокупности, подчиняющейся нормальному распределению.

2.23 **гамма-распределение** (gamma distribution): Распределение непрерывной случайной величины с функцией плотности вероятностей

$$f(x) = \frac{x^{\alpha-1}\exp(-x/\beta)}{\beta^{\alpha}\Gamma(\alpha)},$$

где x — переменная, $x > 0$; α, β — параметры, $\alpha > 0$, $\beta > 0$.
[ISO 3534-1:2006, п. 2.56]

Примечание 1 — Гамма-распределение используют при исследовании безотказности для моделирования наработки до отказа. Оно включает экспоненциальное распределение (2.24), а также другие распределения, у которых интенсивность отказов увеличивается во времени.

Примечание 2 — Математическое ожидание гамма-распределения равно $\alpha\beta$, дисперсия равна $\alpha\beta^2$.

Примечание 3 — Выборка гамма-распределения является случайной выборкой (2.1), отобранной из генеральной совокупности, подчиняющейся гамма-распределению.

2.24 **экспоненциальное распределение** (exponential distribution): Распределение непрерывной случайной величины с функцией плотности вероятностей

$$f(x) = \beta^{-1}\exp(-x/\beta),$$

где x — переменная, $x > 0$; β — параметр, $\beta > 0$.
[ISO 3534-1:2006, п. 2.58]

Примечание 1 — Экспоненциальное распределение является основополагающим при исследовании безотказности в ситуациях отсутствия старения или «памяти».

Примечание 2 — Математическое ожидание экспоненциального распределения равно β. Дисперсия экспоненциального распределения равна β^2.

Примечание 3 — Выборка экспоненциального распределения является случайной выборкой (2.1), отобранной из генеральной совокупности, подчиняющейся экспоненциальному распределению.

2.25 **распределение Вейбулла, распределение экстремальных значений типа III** (Weibull distribution, type III extreme-value distribution): Распределение непрерывной случайной величины с функцией распределения

$$F(x) = 1 - \exp\left\{-\left(\frac{x-\theta}{\beta}\right)^k\right\},$$

где x — переменная, $x > 0$; θ, β, k — параметры $-\infty < \theta < \infty$, $\beta > 0$, $k > 0$.

[ISO 3534-1:2006, п. 2.63]

Примечание 1 — Помимо того, что распределение Вейбулла является одним из трех возможных предельных распределений экстремальных значений порядковых статистик, оно также имеет ряд других важных применений, особенно в теории надежности и инженерии. Существует много ситуаций, когда полученные данные могут быть описаны распределением Вейбулла.

Примечание 2 — Параметр θ является параметром положения или пороговым параметром, это минимальное значение, которое может принимать случайная величина. Параметр β — параметр масштаба (связан со стандартным отклонением случайной величины). Параметр k — параметр формы.

Примечание 3 — Выборка из распределения Вейбулла является случайной выборкой (2.1), отобранной из генеральной совокупности, подчиняющейся распределению Вейбулла.

2.26 **логнормальное распределение** (lognormal distribution): Распределение случайной величины с функцией плотности вероятностей

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

где x — переменная, $x > 0$; μ, σ — параметры $-\infty < \mu < \infty$ и $\sigma > 0$.

[ИСО 3534-1:2006, п. 2.52]

2.27 **распределение экстремальных значений типа I, распределение Гумбеля** (type I extreme-value distribution, Gumbel distribution): Распределение случайной величины с функцией распределения

$$F(x) = \exp\{-e^{-(x-\mu/\sigma)}\},$$

где x — переменная, $x > 0$; μ, σ — параметры $-\infty < \mu < \infty$ и $\sigma > 0$.

Примечание — Распределения экстремальных значений позволяют получить соответствующие распределения для экстремальных порядковых статистик (2.10) $x_{(1)}$ и $x_{(n)}$.

[ISO 3534-1:2006, п. 2.61]

3 Обозначения

В настоящем стандарте использованы следующие обозначения и сокращения:

GESD — обобщенное экстремальное стьюдентизированное отклонение;

G_E — статистика Гринвуда;

$g_{E;n}$ — критическое значение статистики критерия Гринвуда для объема выборки n;

I_l — редуцированная выборка объема $n - l$, полученная после удаления из исходной выборки I_0 объема n, самого экстремального элемента $x^{(0)}$, затем удаления самого экстремального элемента $x^{(1)}$ из редуцированной выборки I_1 объема $n - 1$, удаления самого экстремального элемента $x^{(l-1)}$ из редуцированной выборки I_{l-1} объема $n - l + 1$;

$F_{p;\nu_1,\nu_2}$ — процентиль F-распределения уровня p с ν_1 и ν_2 степенями свободы;

λ_l — критическое значение статистики GESD—критерия при проверке того, что $x^{(l)}$ является выбросом;

L_F — нижняя граница модифицированной диаграммы ящик с усами;

U_F — верхняя граница модифицированной диаграммы ящик с усами;

M или Q_2 — выборочная медиана;

M_{ad} — медиана абсолютного отклонения от медианы;

Q_1 — первый квартиль;

Q_3 — третий квартиль;

R_i — контрольная статистика критерия GESD при проверке того, что $x^{(i)}$ является выбросом;

$s(I_i)$ — стандартное отклонение, вычисленное по редуцированной выборке I_i;

T_M — медиана;

T_n — дважды взвешенная оценка параметра положения для выборки объема n;

$T_n^{(i)}$ — оценка T_n в i-й итерации, при объеме выборки, равном n;

$t_{p;\nu}$ — перцентиль уровня p t — распределения с v степенями свободы;

$\chi^2_{p;\nu}$ — перцентиль уровня p распределения хи-квадрат с ν степенями свободы;

x_i — i-й элемент в упорядоченном наборе данных;

$x^{(i)}$ — наиболее экстремальное значение редуцированной выборки I_i;

$\bar{x}(I_i)$ — выборочное среднее редуцированной выборки I_i;

$\bar{x}_T(\alpha)$ — α — усеченное среднее;

$x_{L:n}$ — нижняя четверть диаграммы ящик с усами, построенной по выборке объема n;

$x_{U:n}$ — верхняя четверть диаграммы ящик с усами, построенной по выборке объема n.

4 Выбросы в одномерных данных

4.1 Общие положения

4.1.1 Понятие выброса

В простейшем случае выброс представляет собой наблюдение, несовместимое с остальными наблюдениями набора данных. В общем случае набор данных может содержать более одного выброса, расположенных, как с одной, так и с двух сторон упорядоченного набора данных. Основная проблема выявления выбросов состоит в определении того, действительно ли наблюдения, не совместимые с остальными данными являются выбросами. Эту задачу решают посредством заданного критерия значимости с учетом предполагаемого распределения данных. Наблюдения, для которых получены значимые результаты, рассматривают как выбросы из предполагаемого распределения.

Важность правильного выбора соответствующего распределения данных нельзя переоценить. На практике часто в качестве распределения данных часто рассматривают нормальное распределение, даже если данные получены из другого источника. Однако ошибочное предположение о распределении данных может приводить к некорректному отнесению элементов выборки к выбросам.

4.1.2 Причины выбросов

Появление выбросов обычно связано с одной или несколькими причинами (детальное рассмотрение приведено в [9]).

a) Ошибки измерений и регистрации данных. Сюда относят ошибки в точности измерений, некорректно проведенные наблюдения, некорректную регистрацию данных или их введения в базу данных.

b) Загрязнение данных. Загрязнения данных происходит в том случае, когда данные принадлежат двум или более распределениям, т. е. имеется одно основное распределение и одно или несколько дополнительных распределений (примесей), загрязняющих данные. Если загрязняющие распределения имеют значительно отличающиеся от основного истинные средние, большие значения стандартных отклонений и/или более тяжелые хвосты распределений, чем у основного распределения, то существует возможность того, что экстремальные наблюдения, принадлежащие распределениям-примесям, могут появиться как выбросы основного распределения.

Примечание 1 — Причиной загрязнения может быть ошибка при отборе выборки, когда небольшую часть данных считают полученной из другой совокупности или если было осуществлено преднамеренное искажение (завышение или занижение) результатов эксперимента или опроса.

c) Ошибочное предположение о распределении данных. Набор данных считают полученным из конкретного распределения, но он получен из другого распределения.

Пример — Набор данных считают отобранным из нормального распределения, но он может иметь сильно ассиметричное распределение (например, экспоненциальное или логнормальное) или быть симметричным, но иметь тяжелые хвосты (например, t-распределение). Поэтому наблюдения, далеко отстоящие от медианы распределения, могут быть ошибочно приняты за выбросы, даже если это достоверные данные, принадлежащие ассиметричному распределению или распределению с тяжелыми хвостами.

d) Редкие наблюдения. В выборках, отобранных (как предполагается) из заданных распределений маловероятные наблюдения могут появиться в очень редких случаях. Экстремальные наблюдения в этом случае обычно принимают за выбросы, но они не являются выбросами.

Примечание 2 — Если генеральная совокупность имеет симметричное распределение с тяжелыми хвостами, то редко поступающие наблюдения могут приводить к ошибочным предположениям о распределении.

4.1.3 Необходимость обнаружения выбросов

Выбросы не всегда являются «плохими» или «ошибочными» данными. Они могут быть рассмотрены как индикаторы проявления редких явлений, требующих дальнейшего изучения. Например, если выброс вызван исключительно особенностями промышленной обработки, то важное значение имеет изучение причин выброса.

Многие методы статистической обработки данных и многие получаемые статистики чувствительны к наличию выбросов. Например, выборочные среднее и стандартное отклонения могут изменить свои значения при наличии даже одного выброса, что впоследствии может привести к неверным выводам.

4.2 Проверка данных

Проверку данных начинают с простого визуального контроля полученного набора данных. Для этого строят простые графики, такие как: точечная диаграмма, диаграмма рассеяния, гистограмма, диаграмма стебель—листья, график вероятности, диаграмма ящик с усами; график данных о времени или в порядке не убывания значений. Это может привести к обнаружению новых источников изменчивости и появлению экстремальных значений в наборе данных. Например, бимодальное распределение данных, обнаруженное с помощью гистограммы или диаграммы стебель—листья, может свидетельствовать о загрязнении выборки или смеси данных из двух разных совокупностей. График вероятности и диаграмму ящик с усами рекомендуется использовать для идентификации выбросов. Эти выбросы в дальнейшем необходимо исследовать с помощью методов, представленных в 4.3 или 4.4.

График вероятности позволяет не только осуществлять графическую проверку соответствия наблюдений или большей части наблюдений предполагаемому распределению, но может быть использован для выявления выбросов в наборе данных. Точки на графике вероятности, заметно отклоняющиеся от прямой, вокруг которой лежат все остальные наблюдения, следует рассматривать как возможные выбросы. Графики вероятности используют во многих пакетах программ статистического анализа данных.

Диаграмма ящик с усами — один из наиболее популярных инструментов графического представления данных. Ее используют для определения параметров положения, рассеяния и формы распределения данных. Нижние и верхние границы диаграммы ящик с усами определяют следующим образом

нижняя граница $Q_1 - k(Q_3 - Q_1)$,

верхняя граница $Q_1 + k(Q_3 - Q_1)$, (1)

где Q_1 и Q_3 — первый и третий квартиль выборки; k — константа.

В работе Тьюки [2] наблюдения, лежащие за пределами верхней и нижней границ, при $k = 1{,}5$ рассматривают как возможные выбросы, при $k = 3$ их рассматривают как явные выбросы.

Примечание 1 — Вероятностная бумага для нормального, логнормального, экспоненциального распределения и распределения Вейбулла может быть загружена с интернет-ресурса http://www.weibull.com/GPaper/index.htm.

Примечание 2 — График вероятности зависит от предположений о виде распределения генеральной совокупности. Например, график вероятности для экспоненциального распределения следует использовать при наличии предположений или априорных знаний о том, что выборка отобрана из генеральной совокупности, подчиняется экспоненциальному закону.

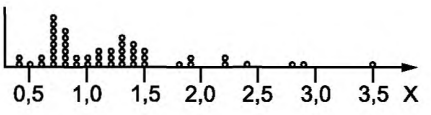
Примечание 3 — При анализе диаграммы ящик с усами, для которой верхняя и нижняя границы определены с помощью (1), большое количество наблюдений может быть ошибочно отнесено к возможным выбросам,

если выборка получена из асимметричного распределения. Данная проблема может быть устранена посредством применения модифицированной диаграммы ящик с усами (см. 4.4).

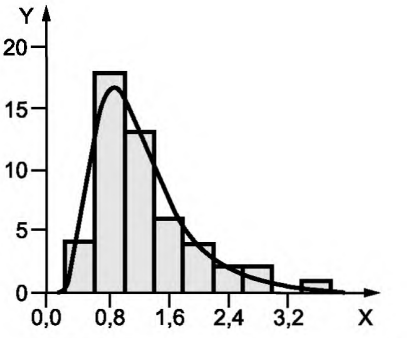
Пример — Точечная диаграмма, гистограмма, диаграмма ящик с усами и диаграмма стебель — листья для приведенной ниже выборки, представлены на рисунках 3 a), 3 b), 3 c) и 3 d) соответственно.

0,745	0,883	0,351	0,806	2,908	1,096	1,310	1,261	0,637	1,226
1,418	0,430	1,870	0,543	0,718	1,229	1,312	1,544	0,965	1,034
1,818	1,409	2,773	1,293	0,842	1,469	0,804	2,219	0,892	1,864
1,214	1,093	0,727	1,527	3,463	2,158	1,448	0,725	0,699	2,435
0,724	0,551	0,733	0,793	0,701	1,323	1,067	0,763	1,375	0,763

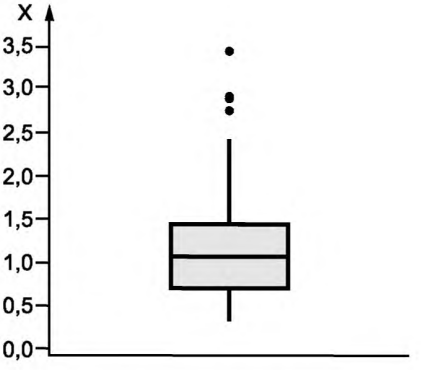
Данные диаграммы показывают, что распределение выборки имеет более длинный правый хвост, чем левый. По рисункам 3 a), 3 b) и 3 d) очевидно, что наибольшее значение 3,463 выглядит как возможный выброс, тогда как диаграмма ящик с усами на рисунке 3 c) определяет три наибольших значения, расположенных над верхней границей, как выбросы. Первая колонка, представленная на рисунке 3 d) диаграммы стебель — листья, показывает глубину, вторая колонка содержит стебли и третья колонка — листья. Значения в колонке глубины содержат суммарное количество листьев снизу или сверху, за исключением значения в скобках, представляющего медиану. Единичный лист указывает на позицию десятичной точки. Единичный лист 0,1 означает, что единичная точка идет перед листом, так первое представленное число равно 0,3, второе и третье 0,4 и 0,5, соответственно. Данный пример рассмотрен также в 4.3.5.



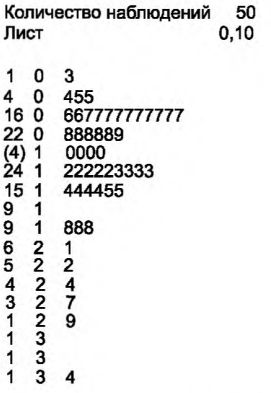
a) Точечная диаграмма



b) Гистограмма
Логонормальное распределение



c) Диаграмма ящик с усами



d) Диаграмма стебель-листья

X — значение случайной величины; Y — частота появления X

Рисунок 3 — Диаграммы, построенные по набору данных

4.3 Выявление выбросов

4.3.1 Общие положения

Существует большое количество методов выявления выбросов (см. [1]). В ИСО 5725-2 (см. [3]) приведены критерии Граббса и Кохрена для идентификации выбросов данных лабораторий. Критерий Грабса применим к отдельным наблюдениям или к выборочным средним наборов данных из нормальных распределений; критерий может быть использован только для выявлений двух наибольших и/или наименьших наблюдений в качестве выбросов в наборе данных. Более общая процедура анализа, представленная в 4.3.2, способна обнаруживать множественные выбросы при анализе отдельных наблюдений или средних арифметических наборов данных, отобранных из нормального распределения. Процедуры, приведенные в 4.3.3 и в 4.3.4, способны обнаруживать множественные выбросы для данных, отобранных из экспоненциального распределения, распределения экстремальных значений типа I, распределения Вейбулла или гамма-распределения. Процедуру, приведенную в 4.3.5, следует применять для обнаружения выбросов в выборках, отобранных из совокупностей с неизвестным законом распределения. Процедура обнаружения выбросов по набору дисперсий, полученных из набора выборок, приведена в 4.3.6.

4.3.2 Выборка из нормального распределения

Один или более выбросов с обеих сторон набора данных из нормального распределения могут быть выявлены при помощи процедуры, известной как обобщенное экстремальное стьюдентизированное отклонение (GESD) (см. [4]). Процедура GESD пригодна для контроля ошибки первого рода при обнаружении более чем l выбросов с уровнем значимости α и $1 \le l \le m$, где m — установленное максимальное количество выбросов.

Перед применением данной процедуры следует удостовериться, что большую часть выборочных данных согласуется с нормальным распределением. График вероятности для нормального распределения, приведенный в ИСО 5479 (см. [18]), может быть использован для проверки справедливости предположения о нормальности распределения.

Этапы процедуры GESD

Этап 1. Точки, соответствующие данным выборки $x_1, x_2, ..., x_n$, наносят на график на нормальной вероятностной бумаге. Подсчитывают количество точек, значимо отклоняющихся от прямой линии, которой соответствуют остальные точки графика. Таким образом, получают количество возможных (предполагаемых) выбросов.

Этап 2. Выбирают уровень значимости α и устанавливают количество выбросов m как число большее или равное числу возможных выбросов, полученному на шаге 1. Следующие этапы начинают, считая $l = 0$.

Этап 3. Вычисляют контрольную статистику

$$R_l = \frac{\max\limits_{x_i \in I_l} |x - \bar{x}(I_l)|}{s(I_l)}, \qquad (2)$$

где

I_0 — исходный набор данных;

I_l — редуцированная выборка объема $n - l$, полученная исключением элемента $x^{(l-1)}$ выборки I_{l-1}, что дает значение R_{l-1};

$\bar{x}(I_l)$ — выборочное среднее выборки I_l;

$s(I_l)$ — выборочное стандартное отклонение выборки I_l.

Примечание 1 — В случае $l = 0$ $\bar{x}(I_0)$ и $s(I_0)$ — выборочное среднее и выборочное стандартное отклонение исходной выборки $I_0 = \{x_1, x_2, ..., x_n\}$ объема n, где наибольшим значением среди значений $x_1 - \bar{x}(I_0)$, $x_2 - \bar{x}(I_0)$, ..., $x_n - \bar{x}(I_0)$ является, например, значение $x_2 - \bar{x}(I_0)$ далее $R_0 = [x_2 - \bar{x}(I_0)] / s(I_0)$ и $x^{(0)} = x_2$. Соответственно, $I_1 = I_0 / \{x^{(0)}\} = \{x_1, x_3, ..., x_n\}$ — редуцированная выборка размера $n - 1$, полученная исключением элемента $x^{(0)}$, т. е. x_2 из I_0.

Этап 4. Вычисляют критическое значение

$$\lambda_1 = \frac{(n-l-1)t_{p;n-l-2}}{\sqrt{(n-l-2+t^2_{p;n-l-2})(n-1)}}, \qquad (3)$$

где $p = (1 - \alpha/2)^{1/(n-l)}$ и $t_{p;\nu}$ — процентиль уровня $100p$ t-распределения с ν степенями свободы. Поскольку выбросы могут быть только среди верхних или нижних экстремальных значений, α заменяют на $\alpha/2$.

Этап 5. Пусть $l = l + 1$

Этап 6. Повторяют этапы 2—4 до тех пор, пока l не станет равно m.

Этап 7. Если $R_l \le \lambda_l$ для всех $l = 0,1,2,\ldots,m$, то считают, что выбросы не обнаружены. В противном случае n_{out} наиболее экстремальных наблюдений $x^{(0)}, x^{(1)}, \ldots x^{(n_{out}-1)}$ редуцированных выборок считают выбросами, при этом $n_{out} = 1 + \max\limits_{0 \le l \le m} \{l : R_l > \lambda_l\}$.

В приложении А приведен алгоритм программной реализации процедуры выявления выбросов GESD.

Примечание 2 — Применение процедуры GESD эквивалентно применению критерия Граббса для проверки того, является ли наибольшее или наименьшее наблюдение выбросом. Критические значения критерия Граббса приведены в таблице 5 ИСО 5725-2:1994 [3], они также могут быть аппроксимированы значением λ_l при $l = 0$ (см. этап 4).

Примечание 3 — На практике, выбирают небольшое значение количества возможных выбросов m. Если в выборке ожидается наличие большого количества выбросов, то в этом случае прекращают рассматривать проблему обнаружения выбросов, и для изучения ситуации применяют другие методы. Однако m не должно быть слишком маленьким, в противном случае может присутствовать эффект маскировки.

Пример — Рассмотрим набор данных из 20 наблюдений:

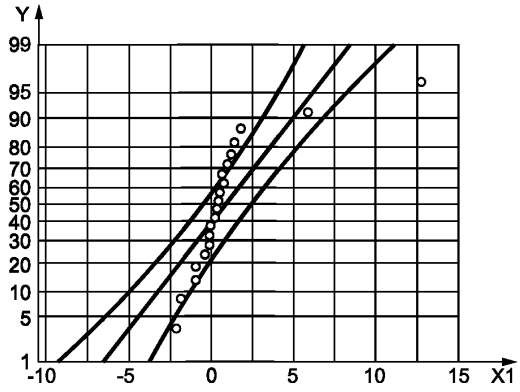
-2,21 -1,84 -0,95 -0,91 -0,36 -0,19 -0,11 -0,10 0,18 0,30
0,43 0,51 0,64 0,67 0,93 1,22 1,35 1,73 5,80 12,6,

где последние два наблюдения первоначально составляли 0,58 и 1,26, но при регистрации данных запятые, отделяющие десятичные разряды, были ошибочно сдвинуты. Перед применением процедуры GESD для обнаружения выбросов необходимо проверить, что наблюдения соответствуют нормальному распределению. Точки на графике вероятности на нормальной вероятностной бумаге (см. рисунок 4 a)) расположены вблизи прямой линии, за исключением двух точек с наибольшими значениями, заметно отклоняющихся от прямой. Данный график показывает, что набор данных, за исключением двух экстремальных значений, можно считать принадлежащим нормальной совокупности. Данное предположение подтверждает рисунок 4 b), где на графике вероятности все данные, за исключением двух крайних значений, расположены внутри границы с уровнем доверия 95 % доверительного интервала. Таким образом, на этапе 2 можно выбрать m = 2. Статистика критерия GESD (R_l) и ее критическое значение λ_l для $l = 0,1,2$ и уровня значимости $\alpha = 0,05$ представлены ниже.

l	0	1	2
R_l	3,6559	3,2634	2,1761
λ_l	2,7058	2,6785	2,6992
x_l	12,60	5,80	-2,21

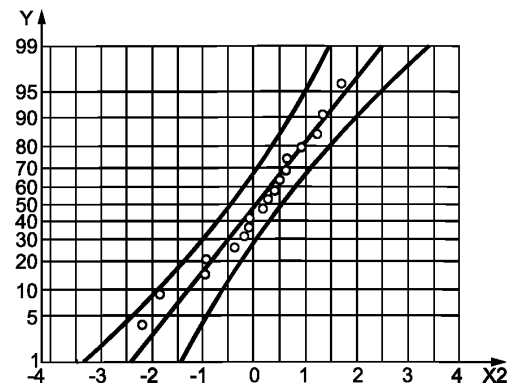
Так как $R_0 = 3,6559 > \lambda_0 = 2,7058$, $R_1 = 3,2634 > \lambda_1 = 2,6785$ и $R_2 = 2,1761 \le \lambda_2 = 2,6992$, следовательно, $\max\limits_{0 \le l \le 2} \{l : R_l > \lambda_l\} = 1$ и $n_{out} = 1 + \max\limits_{0 \le l \le 2} \{l : R_l > \lambda_l\} = 2$. Таким образом обнаружено два выброса, это два наиболее экстремальных значения: $x^{(0)} = 12,60$ и $x^{(0)} = 5,80$.

Примечание 4 — В этом и в следующем примерах не указаны единицы, в которых выполнены измерения данных, так как они не требуются для графической интерпретации и анализа данных, проводимого в рамках настоящего стандарта.



Выборочное среднее	0,9845
Выборочное стандартное отклонение	3,177
Объем выборки	20
Статистика Андерсона-Дарлинга	2,474
p-значение	< 0,005

a) График вероятности для исходного набора данных, нормальное распределение, доверительный интервал уровня 95 %



Выборочное среднее	0,07167
Выборочное стандартное отклонение	1,049
Объем выборки	18
Статистика Андерсона-Дарлинга	0,299
p-значение	< 0,547

b) График вероятности редуцированной выборки, нормальное распределение, доверительный интервал уровня 95 %

X1 — значения исходной выборки; X2 — значения редуцированной выборки; Y — проценты

Рисунок 4 — Графики вероятности

4.3.3 Экспоненциальная выборка

4.3.3.1 Общие положения

Для выявления выбросов в выборках из генеральной совокупности, подчиняющейся экспоненциальному закону распределения, рекомендуется использовать критерий Гринвуда (см. 4.3.3.2). Однако данный критерий позволяет лишь обнаружить наличие выбросов в выборке, но не позволяет идентифицировать конкретные выбросы и определить количество выбросов в выборке. В 4.3.3.3 и 4.3.3.4 представлены два альтернативных последовательных критерия, позволяющих идентифицировать до m возможных верхних или m возможных нижних выбросов в выборке из экспоненциального распределения.

4.3.3.2 Критерий наличия выбросов Гринвуда

Критерий Гринвуда — мощный критерий, позволяющий обнаружить наличие выбросов в выборке, отобранной из экспоненциального распределения с функцией плотности вероятности, $f(x) = \lambda^{-1}\exp[-(x - a)/\lambda]$, $x \geq a$, где λ — параметр масштаба и a — параметр положения или пороговый параметр. Для выборки $x_1, x_2, ..., x_n$ объема n, из генеральной совокупности, подчиняющейся экспоненциальному закону распределения с известным параметром a, статистика критерия имеет вид (см. [1]).

$$G_E = \frac{\sum_{i=1}^{n}(x_i - a)^2}{\left(\sum_{i=1}^{n} x_i - na\right)^2}. \quad (4)$$

Высокое значение G_E свидетельствует о наличие некоторого (неизвестного) количества возможных выбросов среди экстремально высоких значений элементов выборки, однако, низкое значение G_E свидетельствует о наличии некоторого (неизвестного) количества возможных выбросов как среди экстремально низких значений, так и представляющих собой комбинацию экстремально низких и экстремально высоких элементов выборки. Нижние и верхние критические значения $g_{E;n}$ статистики G_E уровней 2,5 % и 1 % соответственно для заданных значений n представлены в таблице В.1. При неизвестном изначальном параметре a, в качестве его оценки используют наименьшее значение в выборке $x_{(1)}$, а в качестве оценки критического значения G_E используют $g_{E;n}$.

4.3.3.3 Последовательные критерии выявления *m* возможных выбросов среди наибольших значений выборки

Статистики критерия для выявления *m* возможных выбросов среди наибольших значений выборки объема *n* из экспоненциального распределения при известном параметре положения *a* (см. [5]).

$$S_j^U = (x_{(n-j+1)} - a) / \sum_{i=1}^{n-j+1} (x_{(i)} - a),\ j = 1, 2, \ldots, m, \quad (5)$$

где $x_{(1)} \le x_{(2)} \le \ldots \le x_{(n)}$ — порядковые статистики выборки. Значимо большие значения S_j^U указывают на то, что высокие экстремальные значения являются выбросами. Верхние, соответствующие уровням 5 % и 1 %, критические значения $s_{j;n}^U$ статистики S_j^U, соответствующие уровням 5 % и 1 % для заданных наблюдений *n* и *m* = 2,3 и 4 представлены в таблице В.2. Если $S_m^U > s_{m;n}^U$, то *m* наибольших наблюдений считают выбросами; если $S_j^U \le s_{j;n}^U$ для $j = m, m-1, \ldots, l+1$, но $S_l^U > s_{l;n}^U$, то *l* наибольших значений считают выбросами; если $S_j^U \le s_{j;n}^U$ для всех $j = 1,2,\ldots,m$, считают, что выбросы в выборке отсутствуют.

В случае, когда параметр *a* неизвестен, в качестве его оценки используют наименьшее значение в выборке $x_{(1)}$, а в качестве оценки критического значения S_j^U используют $s_{j;n-1}^U$.

4.3.3.4 Последовательные критерии выявления *m* возможных выбросов среди наименьших значений выборки

Статистики критерия выявления *m* возможных выбросов среди наименьших значений выборки из экспоненциального распределения объема *n* при известном параметре положения *a* (см. [5])

$$S_j^L = (x_{(j+1)} - a) / \sum_{i=1}^{j+1} (x_{(i)} - a),\ j = 1, 2, \ldots, m, \quad (6)$$

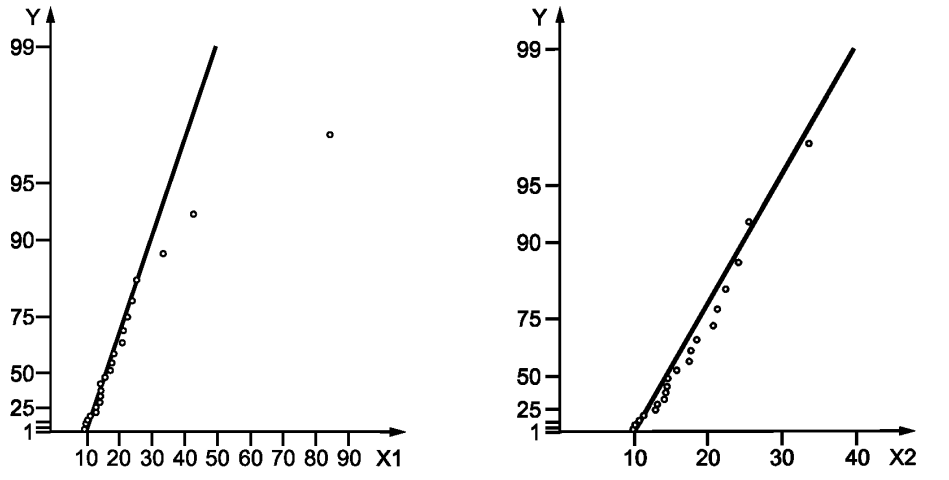
где $x_{(1)} \le x_{(2)} \le \ldots \le x_{(n)}$ — порядковые статистики выборки. Значимо большие значения S_j^L указывают на то, что низкие экстремальные значения являются выбросами. Нижние и верхние критические значения $s_{j;n}^L$ статистики S_j^L уровня 5 % и 1 % соответственно для заданных значений *n* и *m* = 2,3 и 4 представлены в таблице В.3. Если $S_m^L > s_{m;n}^L$, то *m* наименьших наблюдений считают выбросами; если $S_j^L \le s_{j;n}^L$ для $j = m, m-1, \ldots, l+1$, но $S_l^L > s_{l;n}^L$, то *l* наименьших значений считают выбросами; если $S_j^L \le s_{j;n}^L$ для всех $j = 1,2,\ldots,m$ считают, что выбросы отсутствуют.

Данный критерий может быть использован только для выявления выбросов в выборке из экспоненциального распределения с известным параметром *a*. Для выборок с неизвестным параметром *a*, для обнаружения выбросов может быть использована процедура, установленная в 4.4.

Пример — Даны упорядоченные в порядке возрастания наблюдения объема n = 22.

10,10	***10,27***	***10,85***	***11,38***	***12,85***	***13,13***	***14,07***	***14,26***	***14,51***	***14,55***	***15,73***
17,43	***17,72***	***18,49***	***20,75***	***21,37***	***22,50***	***24,22***	***25,61***	***33,84***	***43,00***	***84,94***

На первом этапе использования критерия Гринвуда для определения выбросов следует убедиться, что выборка отобрана из экспоненциального распределения. По графику вероятности с данными выборки, приведенному на рисунке 5 a), видно, что точки данных расположены вблизи прямой линии, за исключением одной или двух точек с наибольшими значениями. Данный график показывает, что выборка, за исключением одного или двух экстремальных значений согласуется с экспоненциальным распределением. Эти выводы подтверждает рисунок 5 b), где на графике вероятности все элементы выборки, за исключением двух крайних значений, расположены вблизи прямой линии. Значения оценки параметра положения a = 10,10 статистик критерия Гринвуда $G_E = 8386,326/(249,37)^2 = 0,13486$. В соответствии с таблицей В.1 нижние и верхние критические значения $g_{E;21}$ статистики G_E, соответствующие уровню 2,5 %, составляют соответственно 0,0673 и 0,1338. Таким образом, вычисленное значение $G_E = 0,13486$ выше верхнего критического значения, равного 0,1338, что позволяет сделать заключение о том, что одно или несколько экстремально высоких значений выборки являются выбросами.



a) График вероятности экспоненциального распределения, построенный по исходной выборке

b) График вероятности экспоненциального распределения, построенный по редуцированной выборке

X1 — значения исходной выборки; X2 — значения редуцированной выборки;
Y — вероятность экспоненциального распределения

Рисунок 5 — Графики вероятности экспоненциального распределения

Так как возможными выбросами являются два верхних экстремальных значения, критерии, представленные в 4.3.3.3, могут быть использованы для проверки того, что выборка содержит два выброса. При $m = 2$, $S_2^U = (43{,}0 - 10{,}01)/174{,}53 = 0{,}188{,}5$ и. После сравнения этих значений с соответствующими критическим $S_1^U = (84{,}94 - 10{,}1)/249{,}37 = 0{,}3001$ и значениями $s_{2;21}^U = 0{,}2313$ и $s_{1;21}^U = 0{,}2834$, определенными по таблице В.2 для $\alpha = 0{,}05$, только наибольшее значение (84,94) можно считать выбросом при уровне значимости 5 %.

4.3.4 Выборки не из нормальных распределений

4.3.4.1 Общие положения

Большое практическое значение имеет выявление выбросов в выборках, взятых не из нормального распределения. Задача выявления выбросов в выборках из экспоненциальных и гамма-распределений стоит, например, при проведении ресурсных испытаний транспортных и речных потоков и т. п.; выборки из распределений экстремальных значений возникают при изучении экстремумов, например, максимальной скорости ветра или максимальных спортивных достижений. Логнормальное распределение и распределение Вейбулла часто используют в задачах надежности. В случае, когда семейство распределений известно и является семейством логнормальных распределений, распределений экстремальных значений, гамма-распределений или распределений Вейбулла, рекомендуется выполнять представленные ниже преобразования данных для приведения их к необходимому распределению.

4.3.4.2 Для выборки $x_1, x_2, ..., x_n$ из логнормального распределения с функцией плотности вероятности

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

преобразованные значения $\ln x_1, \ln x_2, ..., \ln x_n$ представляют собой выборку из нормального распределения с математическим ожиданием μ и дисперсией σ^2. Для обнаружения выбросов в преобразованной выборке может быть применена процедура, представленная в 4.3.2 и/или 4.4.

4.3.4.3 Для выборки x_1, x_2, ..., x_n, взятой из распределения экстремальных значений типа I с функцией распределения

$$P(X \le x) = \exp\{-\exp[-(x - a)/b]\}, \; -\infty < x < \infty,$$

преобразованные значения $\exp(-x_1/b)$, ..., $\exp(-x_n/b)$ представляют собой выборку из экспоненциального распределения с математическим ожиданием $\exp(-a/b)$. Для обнаружения выбросов в преобразованной выборке может быть применена процедура, приведенная в 4.3.3 и/или 4.4.

4.3.4.4 Для выборки x_1, x_2, ..., x_n, из распределения Вейбулла с функцией распределения

$$P(X \le x) = 1 - \exp\{-[(x - a)/b]^r\}, \; x > a, \; b > 0, \; r > 0$$

преобразованные значения $(x_1 - a)^r$, $(x_2 - a)^r$, ..., $(x_n - a)^r$ представляют собой выборку из экспоненциального распределения с математическим ожиданием b^r. Для обнаружения выбросов к преобразованной выборке может быть применена процедура, приведенная в 4.3.3 и/или 4.4.

Примечание — Если x подчиняется экспоненциальному распределению, то $\sqrt[4]{x}$ подчиняется распределению, близкому к нормальному (см. [6]).

4.3.4.5 Для выборки x_1, x_2, ..., x_n из гамма-распределения с функцией плотности вероятностей

$$f(x) = [b^r \Gamma(r)]^{-1} x^{r-1} \exp(-x/b), \; x > 0, \; b > 0$$

преобразованные значения $\sqrt[3]{x_1}$, $\sqrt[3]{x_2}$, ..., $\sqrt[3]{x_n}$ представляют собой выборку из распределения, близкого к нормальному. Для обнаружения выбросов в преобразованной выборке может быть применена процедура, приведенная в 4.3.2 и/или 4.4.

4.3.5 Выборка из неизвестного распределения

При решении задачи выявления выбросов в выборках из генеральной совокупности с неизвестным асимметричным распределением, общий подход состоит в преобразовании данных из ненормального распределения к такому виду, в котором они будут подчиняться распределению, близкому к нормальному. Затем для обнаружения выбросов к преобразованной выборке может быть применена процедура, приведенная в 4.3.3. Для преобразования исходных данных часто применяют преобразование Бокса-Кокса и преобразование Джонсона.

Семейство преобразований Бокса-Кокса имеет форму (см. [7]):

$$y = \begin{cases} (x + m)^{\lambda}, & \text{если } \lambda \ne 0; \\ \log(x + m), & \text{если } \lambda = 0, \end{cases}$$

где если $\lambda \ne 0$, значение m выбирают так, чтобы значение $x + m$ было положительным;

если $\lambda = 1$, в качестве значения m выбирают ноль, в результате чего исходные данные не изменяются.

В некоторых пакетах программ статистической обработки данных выбор оптимального параметра λ происходит автоматически.

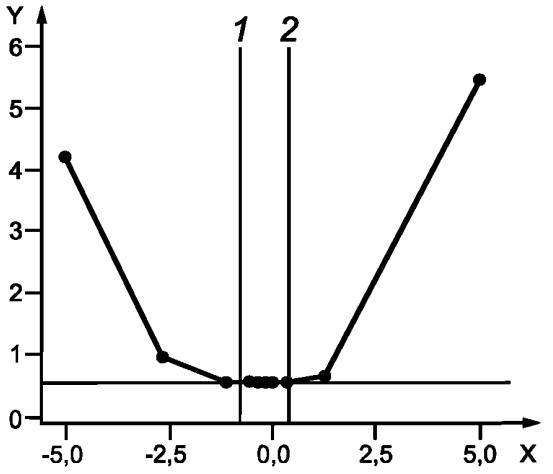
Преобразование Джонсона с помощью семейства распределений Джонсона [8] приводит данные к виду, в котором они подчиняются распределению, близкому к нормальному.

Примечание 1 — Преобразование Бокса-Кокса и преобразование Джонсона могут быть выполнены с помощью соответствующих программных средств обработки данных.

Примечание 2 — Преобразование Бокса-Кокса достаточно просто и понятно. Однако преобразование Джонсона применимо к исходным данным, содержащим отрицательные значения.

Пример — Рассматриваемая выборка отобрана из генеральной совокупности с неизвестным распределением (выборка приведена в 4.2). Построенные по ней диаграмма рассеяния, гистограмма, диаграмма ящик с усами и диаграмма стебель—листья (см. рисунок 3) показывают, что данные взяты из асимметричного распределения. Требуется, чтобы распределение преобразованных данных было близко к нормальному. График Бокса-Кокса и график вероятности, представленные на рисунках 6 и 7,

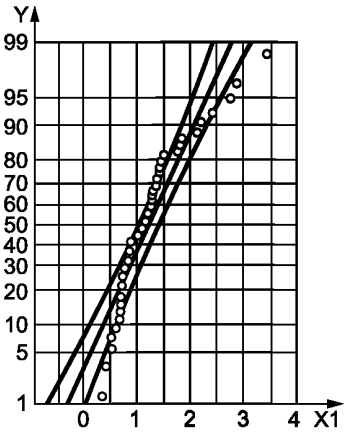
получены с помощью типового пакета программ статистической обработки данных. График, представленный на рисунке 6, соответствует оценке λ, равной минус 0,19, и округленное значение оценки λ, равное 0,00, использованное при проведении преобразования. На графике также приведена нижняя граница доверительного интервала с уровнем доверия 95 %, равную минус 0,77 и соответствующая верхняя граница доверительного интервала, равная 0,36 (границы на графике показаны вертикальными линиями). На практике следует использовать значения λ, полученные с применением общепринятых преобразований, таких как извлечение квадратного корня (λ = 0,5) или вычисление натурального алгоритма (λ = 0). В настоящем примере оценка значения λ, равная нулю, представляет собой разумный выбор, так как попадает в доверительный интервал с уровнем доверия 95 %. Таким образом, преобразование с помощью натурального логарифма может быть более предпочтительным, чем преобразование, обеспечивающее определение наилучшей оценки λ. Графики вероятности исходных и преобразованных данных представлены на рисунке 7. На рисунке 7 b) указано p-значение, вычисленное с помощью статистики критерия Андерсона-Дарлинга, равное 0,318, что говорит о том, что преобразованные данные подчиняются распределению, близкому к нормальному.



Характеристики λ доверительный интервал уровня 95 %	
Оценка	-0,19
Нижняя доверительная граница	0,77
Верхняя доверительная граница	0,36
Округленное значение	0,00

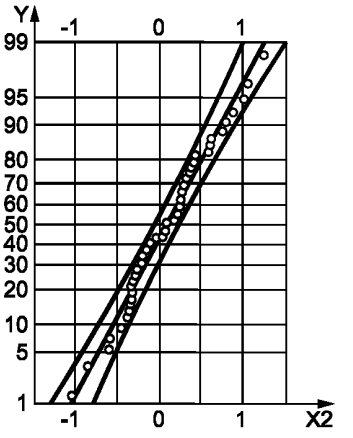
X — λ, Y — стандартное отклонение; *1* — нижняя доверительная граница;
2 — верхняя доверительная граница

Рисунок 6 — График Бокса-Кокса



Выборочное среднее	1,239
Выборочное стандартное отклонение	0,6601
Объем выборки	50
Статистика Андерсона-Дарлинга	54
p-значение	< 0,005

a) График вероятности для исходной выборки



Выборочное среднее	0,09259
Выборочное стандартное отклонение	0,4924
Объем выборки	50
Статистика Андерсона-Дарлинга	0,417
p-значение	0,318

b) График вероятности для преобразованной выборки

X1 — значения исходной выборки; X2 — значения преобразованной выборки; Y — проценты

Рисунок 7 — Графики вероятности для исходных и преобразованных данных

4.3.6 Критерий Кохрена для выявления выбросов дисперсий

Важной задачей является обнаружение выбросов в наборе дисперсий, вычисленных по наборам выборочных данных, в частности, при определении точности методов измерений [3] посредством межлабораторных исследований. Критерий Кохрена широко используют для определения того, является ли действительно значимым отличие наибольших дисперсий от остальных дисперсий в исследуемом наборе дисперсий.

Для набора дисперсий s_1^2, ..., s_p^2, вычисленных по p выборкам, каждая из которых имеет объем n, статистика критерия Кохрена имеет вид

$$C = \frac{s_{max}^2}{\sum_{i=1}^{p} s_i^2}, \quad (7)$$

где s_{max}^2 — наибольшее значение дисперсии в наборе из p дисперсий. В таблицах приложения Е приведены критические значения статистики критерия Кохрена (C) с уровнем 5 %, 1 % и 0,01 % для всех значений p от 2 до 40, при этом предполагается, что дисперсии вычислены по выборкам объема n с n от 2 до 10. Если вычисленное значение C превышает критическое значение, то наибольшую дисперсию в исследуемом наборе дисперсий считают выбросом.

Примечание — Критические значения статистики критерия Кохрена, приведенные в приложении Е, в идеале применяют тогда, когда все стандартные отклонения получены по выборкам одинакового объема n.

Пример — Пять лабораторий принимали участие в проведении исследований по определению показателей поглощения влаги. Каждая лаборатория провела восемь экспериментов в условиях повторяемости и в соответствии со стандартным методом измерений. Был получен следующий набор дисперсий.

Номер лаборатории i	1	2	3	4	5
Дисперсия s_i^2	12,134	2,303	3,594	3,319	3,455

В соответствии с таблицей Е.1 критическое значение критерия Кохрена с уровнем доверия 5 % для p = 5 и n = 8 составляет 0,4564. Так как значение статистики критерия Кохрена

C = 12,134/(12,134+2,3033,594+3,319+3,455)=0,4892 превышает это критическое значение, то можно считать, что дисперсия, вычисленная по результатам лаборатории 1, значительно превышает дисперсии, полученные по результатам остальных лабораторий.

4.4 Графический критерий выявления выбросов

Для обнаружения выбросов рекомендуется применять модифицированную диаграмму ящик с усами, если распределение совокупности является нормальным или экспоненциальным. В отличие от процедур проверки гипотез, приведенных в 4.3, графический критерий выявления выбросов, основанный на диаграмме ящик с усами, не требует предварительного знания о количестве выбросов или расположении выбросов.

При использовании модифицированной диаграммы ящик с усами для определения нижней границы L_F и верхней границы U_F вместо первого квартиля Q_1 и третьего квартиля Q_3 используют соответственно нижнюю четверть $x_{L:n}$ и верхнюю четверть $x_{U:n}$

$$\begin{aligned} L_F &= x_{L:n} - k_L(x_{U:n} - x_{L:n}) \\ U_F &= x_{U:n} - k_U(x_{U:n} - x_{L:n}) \end{aligned}, \qquad (8)$$

где n — объем выборки;

k_L и k_U — показатели, зависящие от предполагаемого распределения данных и объема выборки n;

$x_{L:n}$ — нижняя четверть на диаграмме ящик с усами

$$x_{L:n} = \begin{cases} [x_{(i)} + x_{(i+1)}]/2, & \text{если } f = 0; \\ x_{(i+1)}, & \text{если } f > 0, \end{cases}$$

$x_{U:n}$ — верхняя четверть на диаграмме ящик с усами

$$x_{U:n} = \begin{cases} [x_{(n-i)} + x_{(n-i+1)}]/2, & \text{если } f = 0; \\ x_{(n-i)}, & \text{если } f > 0, \end{cases}$$

при этом $n/4 = i + f$, где i — целая часть $n/4$, а f — дробная часть $n/4$ и $x_{(1)} \le x_{(1)} \le ... \le x_{(n)}$ — порядковые статистики выборки.

Примечание 1 — Данное определение нижней и верхней четвертей используют для определения значений k_L и k_U (см. приложение С), его определяют по умолчанию в большинстве широко используемых пакетах программ статистической обработки данных.

Элементы выборки, расположенные выше верхней границы или ниже нижней границы, рассматривают как возможные выбросы. Характерной особенностью модифицированной диаграммы ящик с усами является определение констант k_L и k_U с учетом того, что для выборки, не содержащей выбросов, определена вероятность того, что один или более элементов выборки могут быть ошибочно классифицированы как выбросы, равная установленному малому значению α. При $k_L = k_U = 1,5$ модифицированная диаграмма ящик с усами является обычной диаграммой ящик с усами, рассмотренной в 4.2. Для выборок из нормального и экспоненциального распределения при выбранном значении α и $9 \le n \le 500$ значения k_L и k_U могут быть определены в соответствии с С.2 (см. приложение С).

Примечание 2 — Нижняя граница модифицированной диаграммы ящик с усами, построенной в предположении об экспоненциальном распределении данных, может принимать отрицательные значения, если данные не подчиняются экспоненциальному распределению.

Пример 1 — Для выборки объема n = 20 из примера, рассмотренного в 4.3.2, n/4 = 20/4 = 5 , т.е. i = 5 и f = 0, таким образом, оценки нижней и верхней четвертей имеют вид

$$x_{L:n} = [x_{(5)} + x_{(6)}] / 2 = 0,5(-0,36 - 0,19) = -0,275,$$
$$x_{U:n} = [x_{(15)} + x_{(16)}] / 2 = 0,5(-0,93 - 1,22) = 1,075.$$

Для выборок из нормального распределения с α = 0,05 нижняя и верхняя границы построены для $k_L = k_U = 2,238$ (см. пример 1 в приложении С)

$$L_F = x_{L:n} - k_L(x_{U:n} - x_{L:n}) = -0,275 - 2,2382(1,075 + 0,275) = -3,297,$$
$$U_F = x_{U:n} - k_U(x_{U:n} - x_{L:n}) = 1,075 - 2,2382(1,075 + 0,275) = 4,097.$$

Таким образом, два наиболее экстремальных значения 5,80 и 12,60, лежащие выше верхней границы, следует считать выбросами.

Пример 2 — Для выборки объема n = 22 из примера, рассмотренного в 4.3.3.4, n/4 = 22/4 = 5 + 1/2 таким образом, оценки нижней и верхней четвертей $x_{L:n} = x_{(6)} = 13,13$ и $x_{U:n} = x_{(17)} = 22,50$.
Для данной выборки с $\alpha = 0,05$ верхняя и нижняя границы имеют вид:

$$L_F = x_{L:n} - k_L(x_{U:n} - x_{L:n}) = 13,3 - 0,6650(22,50 + 13,13) = 6,899,$$
$$U_F = x_{U:n} - k_U(x_{U:n} - x_{L:n}) = 22,50 + 6,2313(22,50 - 13,13) = 80,887.$$

Таким образом, экстремальное значение 84,94, лежащее выше верхней границы, следует рассматривать как выброс. Значения $k_L = 0,6650$ и $k_U = 6,2313$ получены в примере 2 приложения С.

Пример 3 — Предположим, что второе по величине значение в выборке из примера, приведенного в 4.3.3.4 (43,0), было ошибочно записано как 4,30. Так как значение 4,30 лежит на диаграмме ящик с усами ниже нижней границы $L_F = 6,899$, то его следует признать выбросом. Однако из-за эффекта маскировки, формальная процедура проверки в соответствии с 4.3 не расценивает нижнее экстремальное значение 4,30 и верхнее экстремальное значение 84,94, как выбросы.

5 Коррекция влияния выбросов в одномерной выборке

5.1 Робастный анализ данных

Каждый обнаруженный выброс должен быть исследован и объяснен. Если выброс вызван ошибкой, причина которой может быть обнаружена (например, канцелярская ошибка, ошибка получения раствора, ошибка измерений и т. д.), то его значение должно быть скорректировано, если истинное значение известно или, в противном случае, удалено. Если наличие выбросов не может быть разумно объяснено, то данные значения не следует удалять; они должны быть обработаны как достоверные наблюдения и использованы в последующем анализе данных с использованием робастных процедур, устойчивых к наличию выбросов. Методы коррекции влияния выбросов, представленные в 5.2 и 5.3, могут снижать влияние выбросов на результат анализа данных без удаления значений, которые распознаны как выбросы. Альтернативный способ состоит в проведении анализа дважды при наличии выбросов и без выбросов.

5.2 Робастная оценка параметра положения

5.2.1 Общие положения

Выборочное среднее является оптимальной оценкой параметра положения нормального распределения. Однако эта оценка не является устойчивой и робастной оценкой. В литературе предложено большое количество разнообразных процедур получения робастной оценки параметра положения. Усеченное среднее, рассмотренное в 5.2.2, широко используют для снижения искажения оценки параметра положения при наличии выбросов в выборке из симметричного распределения. Для выборок совокупности с асимметричным распределением рекомендуется определять оценку параметра положения в соответствии с 5.2.3.

5.2.2 Усеченное среднее

Если в выборке из симметричного распределения возможно наличие выбросов, в качестве оценки центра распределения рекомендуется использовать усеченное среднее.

Пусть $x_{(1)} \le x_{(1)} \le ... \le x_{(n)}$ — порядковые статистики выборки объема n.

Пусть $r = [\alpha n]$ — наибольшее целое, меньшее или равное αn и $g = \alpha n - r$ — дробная часть αn, где $0 \le \alpha \le 0,5$ — доля выбросов в выборке.

Значение α — усеченного среднего [9], обозначаемого $\bar{x}_T(\alpha)$, вычисляют как среднее без учета r наименьших и r наибольших значений выборки, но включая в вычисления два ближайших сохраненных значения $x_{(r+1)}$ и $x_{(n-r)}$ с уменьшенным весом $(1 - g)$, например,

$$\bar{x}_T(\alpha) = \frac{1}{n(1-2\alpha)}\left[(1-g)(x_{(r+1)} + x_{(n-r)}) + \sum_{i=r+2}^{n-r-1} x_{(i)}\right]. \quad (9)$$

Примечание 1 — Если αn — целое, то $g = 0$, таким образом усеченное среднее является выборочным средним усеченной выборки.

Примечание 2 — Обычно предварительно значение α задают меньше 0,25. Классическое выборочное среднее — это 0-усеченное среднее, тогда как выборочная медиана представляет собой приближенно 0,5-усеченное среднее.

Примечание 3 — Другой распространенной оценкой параметра положения является α-винсоризованное среднее, в котором $r = \alpha n$ наименьших наблюдений, отброшено до значения $x_{(r+1)}$ и r наибольших наблюдений отброшено до $x_{(n-r)}$, т. е. произведена замена r наибольших и r наименьших значений на значение $(1 - g)\ \bar{x}_T(\alpha)$.

Пример — Для выборки объема n = 20, представленной в 4.3.2, вычислены выборочные среднее и медиана, а также усеченные средние с долей усечения 5 %, 10 %, 15 %, 18 % и 20 %. Получены следующие значения:

Выборочное среднее $= \frac{1}{20}\sum_{i=1}^{20} x_i = \frac{1}{20}(19{,}69) = 0{,}9845.$

Выборочная медиана $= \frac{1}{2}[x_{(10)} + x_{(11)}] = \frac{1}{2}(0{,}30 + 0{,}43) = 0{,}365$

$$\bar{x}_T(0{,}05) = \frac{1}{20(1-2\cdot 0{,}05)}\sum_{i=2}^{19} x_{(i)} = \frac{1}{18}(9{,}3) = 0{,}5167,$$

$$\bar{x}_T(0{,}10) = \frac{1}{20(1-2\cdot 0{,}10)}\sum_{i=3}^{18} x_{(i)} = \frac{1}{16}(5{,}34) = 0{,}33375,$$

$$\bar{x}_T(0{,}15) = \frac{1}{20(1-2\cdot 0{,}15)}\sum_{i=4}^{17} x_{(i)} = \frac{1}{14}(4{,}56) = 0{,}3257,$$

$$\bar{x}_T(0{,}18) = \frac{1}{20(1-2\cdot 0{,}18)} = \left[(1-0{,}06)(x_{(4)} + x_{(17)}) + \sum_{i=5}^{16} x_{(i)}\right] = \frac{1}{12{,}8}(0{,}176 + 4{,}12) = 0{,}3356,$$

$$\bar{x}_T(0{,}20) = \frac{1}{20(1-2\cdot 0{,}20)}\sum_{i=5}^{16} x_{(i)} = \frac{1}{12}(4{,}12) = 0{,}3433.$$

Данные результаты предполагают, что относительно большое выборочное среднее соответствует наличию двух выбросов, тогда как усеченные средние стабилизируются от 10 % до 20 % усечения набора данных.

5.2.3 Дважды взвешенная оценка параметра положения

Дважды взвешенная оценка параметра положения [9] является устойчивой к наличию выбросов в выборках из асимметричных распределений и робастной по отношению к небольшим отклонениям от нормального распределения. Для данной выборки $x_1, x_2, ..., x_n$ объема n, дважды взвешенная оценка параметра положения имеет вид

$$T_n = \frac{\sum_{|u_i|<1} x_i (1-u_i^2)^2}{\sum_{|u_i|<1} (1-u_i^2)^2}, \qquad (10)$$

где $u_i = (x_i - T_n) / cM_{ad}$, $c = 6{,}0$, $M_{ad} = \text{Median}(|x_i - M|, i = 1, 2, ..., n)$ M-выборочная медиана. Оценку T_n вычисляют итеративно. Значения $T_n^{(k)}$ и $u_{i,k} = (x_i - T_n^{(k)}) / cM_{ad}$ являются оценкой T_n и u_i на k-й итерации, оценка T_n на $(k+1)$-й итерации

$$T_n^{(k+1)} = \frac{\sum_{|u_i|<1} x_i (1-u_{i,k}^2)^2}{\sum_{|u_i|<1} (1-u_{i,k}^2)^2}.$$

Итеративный процесс следует продолжать до тех пор, пока последовательность оценок не станет сходиться с требуемой точностью. Например, итерации могут быть прекращены, если $|T_n^{(k+1)} - T_n^{(k)}| < 10^{-5}$. Подходящим устойчивым начальным значением $T_n^{(0)}$ является выборочная медиана M.

Примечание — В предположении нормальности распределения данных, дважды взвешенная оценка при $c = 6{,}0$ означает взвешенное среднее, в котором значениям, отклоняющимся от медианы более чем на четыре стандартных отклонения, присвоен нулевой весовой коэффициент.

Пример — Дважды взвешенная оценка параметра положения для выборки, представленной в 4.3.2, T_n = 0,176. Она близка к выборочному среднему (0,1565) при замещении двух экстремальных значений (5,80 и 12,8) корректными значениями (0,58 и 1,28).

5.3 Робастная оценка дисперсии

5.3.1 Общие положения

Ниже представлены две широко используемые оценки параметра масштаба, устойчивые к выбросам и используемые вместо оценки стандартного отклонения выборки.

5.3.2 Попарное абсолютное отклонение медиан

$$S_n = s_n \text{Median}_i\,(\text{Median}_j\,|x_i - x_j|,\ i \neq j,\ i, j = 1, 2, ..., n). \quad (11)$$

Постоянная s_n — корректирующий множитель, выбираемый так, чтобы гарантировать, что S_n является несмещенной оценкой параметра масштаба предполагаемого распределения (нормального, экспоненциального и т. д.). Для больших выборок из нормального распределения значение s_n = 1,1926 (см. [10]), тогда как для больших выборок экспоненциального распределения s_n = 1,6982. Значения s_n для ряда объемов n выборок из нормального распределения приведены в таблице D.1.

5.3.3 Дважды взвешенная оценка параметра масштаба

Дважды взвешенная оценка параметра масштаба для выборки $x_1, x_2, ..., x_n$ соответствует представленной в [9] и имеет вид:

$$S_{bi} = s_{bi} \frac{n}{\sqrt{n-1}} \frac{\sqrt{\sum_{|u_i|<1} (x_i - M)^2 (1-u_i^2)^4}}{\left|\sum_{|u_i|<1} (1-u_i^2)(1-5u_i^2)\right|}, \quad (12)$$

где M — выборочная медиана, $u_i = (x_i - M) / (cM_{ad})$ и M_{ad} = Median($|x_i - M|$, i = 1, 2, ..., n) для выборки объема n из нормального распределения. Рекомендуемое значение для c составляет 9,0. Значения S_{bi}, основанные на c = 9,0 для ряда объемов n выборок из нормального распределения, приведены в таблице D.1.

Примечание — В предположении нормальности распределения данных, дважды взвешенная оценка для c = 9,0 означает взвешенное среднее, в котором значениям, отклоняющимся от медианы более чем на шесть стандартных отклонений, присваивают нулевой весовой коэффициент.

Пример — Для выборки, представленной в 4.3.2, классическое выборочное стандартное отклонение s, робастные оценки масштаба S_n (см. 5.3.2) и S_{bi} (см. 5.3.3 выше) заданы следующим образом

s = 3,1772, S_n = 1,015, S_{bi} = 1,1565.

Эти результаты показывают, что стандартное отклонение s существенно увеличено за счет двух наибольших наблюдений. Соответствующие робастные оценки S_n и S_{bi} имеют относительно небольшие, близкие друг к другу значения.

6 Выбросы многомерных и регрессионных наборов данных

6.1 Общие положения

Задача обнаружения выбросов в наборе многомерных и регрессионных данных является более сложной, чем задача обнаружения выбросов в наборе одномерных данных. Многомерный выброс — это выброс по любой из компонент наблюдения или многомерных координат. Многомерные выбросы также могут быть в некоторой степени скрыты механизмом их появления, и их присутствие обнаруживается только после анализа структуры данных. Выброс регрессионных данных может не быть просто экстремальным значением, а быть наблюдением, которое значимо отклоняется от основной регрессионной модели.

6.2 Выбросы многомерных данных

Общая идея методов выявления выбросов в многомерных наборах данных заключается в преобразовании многомерных данных к одномерным статистикам. Одной из широко используемых статистик является расстояние Махаланобиса, являющееся мерой расстояния от многомерного наблюдения до выборочного среднего набора данных, нормированного при помощи выборочной ковариационной матрицы. Из p переменных (случайных величин) $X_1, X_2, ..., X_p$ составлен упорядоченный набор p-мерный вектор $\boldsymbol{X} = (X_1, X_2, ..., X_p)^T$.

Пусть $\mu = (\mu_1, \mu_2, ..., \mu_p)^T$ — вектор средних p случайных переменных $\boldsymbol{X}$, а матрица $p \times p$ — матрица ковариаций Σ, где диагональные элементы являются дисперсиями, а остальные — ковариациями элементов вектора $\boldsymbol{X}$.

Расстояние Махаланобиса от случайного вектора $\boldsymbol{X}$ до вектора средних значений μ определено следующим образом

$$M_D = \sqrt{(\boldsymbol{X}-\mu)^T \Sigma^{-1} (\boldsymbol{X}-\mu)} \ . \quad (13)$$

Выбросы в выборке многомерных наблюдений объема n: $X_1, X_2, ..., X_p$ могут быть выявлены посредством определения n соответствующих расстояний Махаланобиса $M_{Di} = \sqrt{(x_i-\mu)^T \Sigma^{-1} (x_i-\mu)}$ $i = 1, 2, ..., n$. Если вектор $\boldsymbol{X}$ может быть многомерным нормальным распределением со средним μ и матрицей ковариации Σ квадрат расстояния Махаланобиса M_D^2, подчиняется распределению хи-квадрат с p степенями свободы.

Приведенная выше формула для вычисления расстояния Махаланобиса зависит от знаний μ и Σ. На практике требуется найти оценки μ и Σ по выборочным данным. При наличии выбросов робастные оценки для μ и Σ должны быть получены с помощью метода минимального определителя ковариации (MCD). Метод MCD находит среди n данных наблюдений, те h наблюдений, которые придают определителю матрицы ковариации наименьшее значение. В предположении, что выборка содержит не более 100α % выбросов, значение h следует выбирать близким к $(1 - \alpha)n$, однако оно должно быть больше целой части числа $[(n + p + 1)/2]$. Тогда среднее и матрица ковариаций, определенным по этим найденным h наблюдениям являются MCD-оценками μ_{MCD} и Σ_{MCD} для μ и Σ соответственно. Робастное расстояние для наблюдения x_i имеет вид:

$$D_{Ri} = \sqrt{(x_i - \hat{\mu}_{MCD})^T \hat{\Sigma}_{MCD}^{-1} (x_i - \hat{\mu}_{MCD})} \ . \quad (14)$$

В предположении нормальности распределения данных, консервативный критерий [11] объявляет выбросами те наблюдения, которые имеют робастное расстояние, превышающее критическое значение $\sqrt{\chi^2_{0,975;p}}$, где $\chi^2_{0,975;p}$ — процентиль уровня 97,5 % распределения хи-квадрат с p степенями свободы.

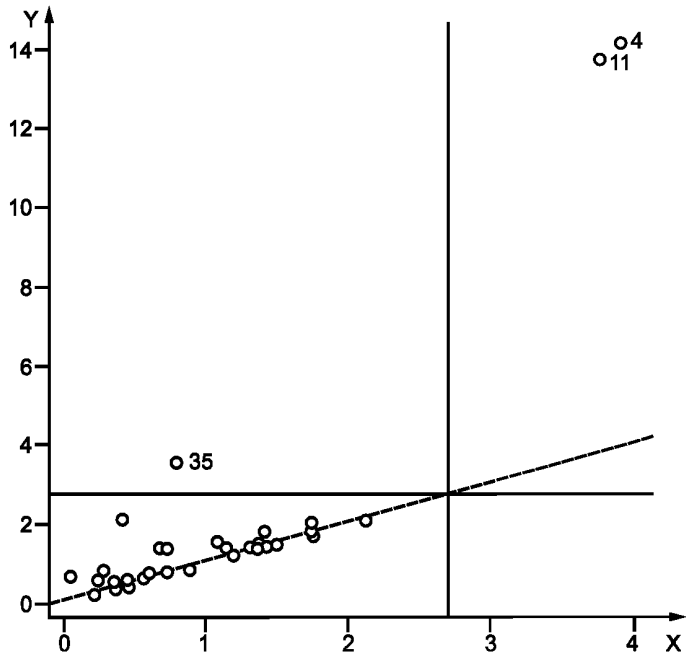
Визуальное сопоставление расстояния Махаланобиса с робастным расстоянием, а также результативность использования робастного расстояния в выявлении выбросов показано на примере.

Пример — Дана выборка объема n = 35, содержащая двумерные наблюдения (x_1, x_2).

Номер наблюдения i	x_{1i}	x_{2i}	Номер наблюдения i	x_{1i}	x_{2i}	Номер наблюдения i	x_{1i}	x_{2i}
1	12,00	12,60	13	12,90	12,95	25	15,60	15,64
2	9,30	10,20	14	12,90	13,50	26	13,25	12,85
3	15,00	14,50	15	13,10	13,80	27	16,83	16,85
4	10,15	19,30	16	16,00	16,25	28	12,00	11,70
5	10,45	10,80	17	13,45	13,00	29	17,30	17,25
6	17,45	16,90	18	13,55	15,20	30	10,65	10,80
7	10,80	11,95	19	14,30	15,10	31	17,55	17,70
8	10,80	10,85	20	14,40	14,55	32	18,20	18,35
9	10,75	11,65	21	13,60	14,35	33	19,10	19,30
10	17,00	17,50	22	14,80	14,99	34	13,55	14,00
11	8,25	17,20	23	10,15	9,90	35	12,55	15,10
12	12,66	13,30	24	15,10	15,15			

Для каждого наблюдения вычислены расстояние Махаланобиса и робастное расстояние, и нанесены на график, представленный на рисунке 8; при этом был использован метод MCD при h, равном 32 наблюдениям. Данный график построен при помощи свободно распространяемого пакета программ статистической обработки данных LIBRA [11]. При помощи штриховой линии представлено множество значений, для которых расстояние Махаланобиса равно робастному расстоянию. Горизонтальная и вертикальная линии пересекаются в точке, соответствующей критической точке

$\sqrt{\chi^2_{0,975;p}} = \sqrt{7,378} = 2,716$. Точки, расположенные за этими линиями, могут быть рассмотрены как выбросы. Робастное расстояние на данном графике выявляет то, что точки 4, 11 и 35 являются выбросами. Однако расстояние Махаланобиса выявляет в качестве выбросов только точки 4 и 11. То, что расстояние Махаланобиса выявляет в качестве выбросов только точки 4 и 11, может выглядеть как эффект маскировки, рассмотренный в 2.3. При вычислении расстояния Махаланобиса без учета наблюдений 4 и 11, наблюдение 35 также выявлено как выброс.



X — расстояние Махаланобиса; Y — робастное расстояние

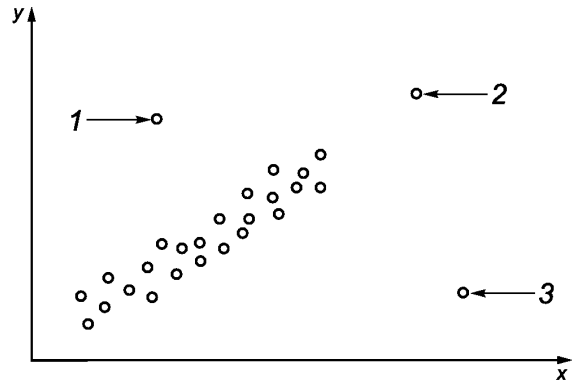
Для данных, представленных на рисунке 11, наблюдения 4, 11 и 35 обозначены своими номерами.

Рисунок 8 — График расстояния Махаланобиса и робастного расстояния

6.3 Выбросы в линейной регрессии

6.3.1 Общие положения

При анализе простой линейной регрессии, случайная точка (Y,X) может быть выбросом как по компоненте Y, так и по компоненте X или по обеим. На рисунке 9 представлен график рассеяния точек с координатами (y_i, x_i), точка 1 удалена от линии по координате y и, таким образом является выбросом по координате y, но не является выбросом по координате x; точка 3 удалена от остальных точек по координате x, но по координате y не является выбросом; точка 2 является выбросом как по координате x, так и по координате y.



1, *2*, *3* — выбросы

Рисунок 9 — График рассеяния точек (*Y*,*X*)

По графику, представленному на рисунке 9, видно, что не все выбросы оказывают различное влияние на положение линии регрессии. Точка 1 имеет координату *x*, близкую к координатам *x* других элементов выборки, поэтому оказывает влияние только по координате *y*. Аналогично координата *y* точки 3 соответствует координатам *y* других точек выборки, эта точка оказывает влияние на линию регрессии по координате *x*. Точка 2 оказывает влияние на линию регрессии, как по координате *x*, так и по координате *y*.

6.3.2 Модели линейной регрессии

В моделях линейной регрессии случайную величину *Y* рассматривают как зависимую от единственной переменной *X*, линию регрессии строят по точкам (y_i, x_i), $i = 1,2,\ldots n$, принадлежащим выборке объема *n*, в соответствии с моделью:

$$\hat{y}_i = b_0 + b_1 x_i \tag{15}$$

при этом, определяют *i*-й остаток, как разность между наблюдаемым значением y_i и соответствующим приближенным значением $\hat{y}_i$, т. е.

$$e_i = \hat{y}_i,\ i = 1, 2, \ldots, n.$$

С помощью обычного метода наименьших квадратов можно определить значения b_0 и b_1, так чтобы минимизировать сумму квадратов остатков $\sum_{i=1}^{n} e_i^2$

$$b_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})\, y_i}{\sum_{i=1}^{n} (x_i - \bar{x})^2}, \tag{16}$$

$$b_0 = \bar{y} - b_1 \bar{x},$$

где $\bar{x}$ и $\bar{y}$ — выборочные средние соответственно для компонент x_i и y_i.

Влияние выбросов по *X* и/или *Y* на построение линии регрессии методом наименьших квадратов может быть проанализировано с помощью оценки значений

$$\hat{y}_i = \bar{y} + b_1 (x_i - \bar{x}) = \bar{y} + (x_i - \bar{x}) \frac{\sum_{j=1}^{n} (x_j - \bar{x})\, y_i}{\sum_{k=1}^{n} (x_k - \bar{x})^2}$$

или эквивалентно

$$\hat{y}_i = \sum_{j=1}^{n} h_{ij} y_j, \tag{17}$$

где значения

$$h_{ij} = \frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{\sum_{k=1}^{n} (x_k - \bar{x})^2}$$

вычисляют только на основании независимой переменной X. Значения h_{ij} являются элементами симметричной матрицы $\boldsymbol{H} = (h_{ij})$ размера $n \times n$, называемой проекционной матрицей. Из равенства $\hat{y}_i = \sum_{j=1}^{n} h_{ij} y_j$ следует, что значения h_{ij} являются показателем того, как значения X влияют на то, насколько существенна роль y_j в получении приближенного значения $\hat{y}_i$.

Подобным образом, рассматривают случайную величину Y, зависящую от p случайных величин X_1, X_2, ..., X_p, для которой значение регрессионной функции для выборки из n элементов $(y_i, x_{i1}, x_{i2}, ..., x_{ip})$, $i = 1, 2, ..., n$ представляет собой

$$\hat{y}_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + ... + b_1 x_{ip},$$

где b_j является j-м коэффициентом регрессионной функции, а x_{ij} — i-е частное значение j-й случайной независимой переменной x_j. Как и в случае с одной независимой переменной, i-й остаток приближения имеет вид $e_i = y_i - \hat{y}_i$. В матричном виде модель многомерной регрессии записывают следующим образом:

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\boldsymbol{b}, \tag{18}$$

где $\hat{\boldsymbol{y}} = (\hat{y}_1, ..., \hat{y}_n)^T$ — n-мерный вектор, $\boldsymbol{b} = (b_0, b_1, ..., b_p)^T$ вектор, размерности $(p + 1)$, $\boldsymbol{X}$ — матрица $n \times (p + 1)$

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix}.$$

Вектор коэффициентов находят методом наименьших квадратов

$$\boldsymbol{b} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{y} \tag{19}$$

и вектор значений $\hat{\boldsymbol{y}}$ может быть получен непосредственно в терминах проекционной матрицы $\boldsymbol{H}$:

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\boldsymbol{b} = \boldsymbol{X}(\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{y} = \boldsymbol{H}\boldsymbol{y},$$

где $\boldsymbol{y} = (y_1, ..., y_n)^T$ — вектор размерности n, состоящий из n значений y.

$$\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T$$

$\boldsymbol{H}$ — матрица $n \cdot n$.

6.3.3 Обнаружение выбросов по компоненте Y

Робастная процедура обнаружения выбросов по компоненте Y в выборке объема n анализирует стьюдентизированные ошибки r_i, которые являются ошибками построения регрессионной функции, вычисленными без использования i-го наблюдения. Стьюдентизированные ошибки рассчитывают по формуле (см. [12]).

$$r_i = e_i \sqrt{\frac{n - p - 2}{(1 - h_{ii}) R_{SSE} - e_i^2}}, \; i = 1, 2, ..., n, \tag{20}$$

где $e_i = y_i - \hat{y}_i$ — i-й остаток;

h_{ij} — диагональный элемент матрицы $\boldsymbol{H}$;

$R_{SSE} = \sum_{i=1}^{n} e_i^2$ — сумма квадратов остатков, полученных при построении регрессионной функции на основе n наблюдений, при этом количество оцениваемых параметров функции регрессии равно p+1.

Примечание — Выражение для стьюдентизированной ошибки r_i (см. [12]) основано на том, что i-е наблюдение (y_i, x_{i1}, x_{i2}, ..., x_{ip}) не включено в построение функции регрессии по оставшимся n-1 точкам. Такая ошибка может быть подсчитана для каждой i-й точки без изменения регрессионной функции в соответствии с уравнением (20).

Стьюдентизированные ошибки r_i имеют t-распределение с n–p–2 степенями свободы, наблюдения для которых стьюдентизированные ошибки, больше чем $t_{1-\alpha/2n;n-p-2}$, следует рассматривать как выбросы по компоненте Y.

6.3.4 Обнаружение выбросов по компоненте $\boldsymbol{X}$

Диагональные элементы матрицы $\boldsymbol{H}$ также могут быть использованы для определения выбросов по компоненте X. Некоторые полезные свойства элементов h_{ij} проекционной матрицы

$$\frac{1}{n} \le h_{ii} \le 1,$$

$$\sum_{i=1}^{n} h_{ii} = p + 1,$$

если $h_{ij} = 0$ или $h_{ij} = 1$, то $h_{ij} = 0$ для всех $j \ne i$,

где p+1— количество параметров регрессионной модели, включающей постоянный член.

В частном случае, линейной регрессии с единственной независимой переменной (p=1) и постоянным членом, диагональные элементы h_{ij} проекционной матрицы $\boldsymbol{H}$ имеют вид

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{k=1}^{n} (x_k - \bar{x})^2}. \qquad (21)$$

Это выражение показывает, что h_{ij} характеризует расстояние между значением, принимаемым случайной величиной $\boldsymbol{X}$ в i-й точке и средним арифметическим всех n значений, принимаемых $\boldsymbol{X}$. Большие значения h_{ij} говорят о том, что значение x_i значительно отклоняется от соответствующих значений большинства наблюдений, о чем свидетельствует то, что при $j \ne i$ значения $|x_i - \bar{x}|$ меньше, чем при $j = i$. Диагональные элементы h_{ij} проекционной матрицы в данном контексте называют влиянием i-го наблюдения. В общем случае h_{ij} считают существенным, если h_{ij} более чем в два раза превосходит среднее $\bar{h} = \frac{1}{n}\sum_{i=1}^{n} h_{ij} = (p + 1) / n$. Данное правило означает, что если $h_{ij} \ge \frac{2(p+1)}{n}$, то i-е наблюдение по координате $\boldsymbol{X}$ следует считать выбросом. В соответствии с другим простым критерием (см. [13]):

- данные с h_{ij} менее 0,2 можно безопасно использовать в регрессионном анализе;
- данные с h_{ij} от 0,2 до 0,5 могут быть включены в регрессионный анализ;
- данные с h_{ij} более 0,5 должны быть исключены из регрессионного анализа.

6.3.5 Обнаружение влияющих наблюдений

Следующим шагом после выявления выбросов по компоненте Y и/или X является установление того, ведет ли удаление точек, соответствующих выявленным выбросам к значительным изменениям построенной регрессионной модели. Широко используют два показателя влияния выявленных выбросов: значение DFFITS и расстояние Кука (см. [12], [14]).

Значение DFFITS

Обозначение DFFITS представляет собой аббревиатуру английского выражения, означающего «различие приближений». Для i-го наблюдаемого значения DFFITS определяют как

$$(\text{DFFITS})_i = e_i \left[\frac{n-p-2}{R_{SSE}(1-h_{ii}) - e_i^2}\right]^{1/2} \left(\frac{h_{ii}}{1-h_{ii}}\right)^{1/2} = r_i \left(\frac{h_{ii}}{1-h_{ii}}\right)^{1/2}, \qquad (22)$$

где r_i — стьюдентизированная ошибка, определяемая по формуле (20). Наблюдаемое значение с номером i считают влияющим элементом выборки, если абсолютное значение $(DFFITS)_i$, превышает 1 для малых и средних выборок и превышает $2\sqrt{(p+1)/n}$ для больших выборок.

Расстояние Кука

Расстояние Кука, обозначаемое D_i, определяют следующим образом

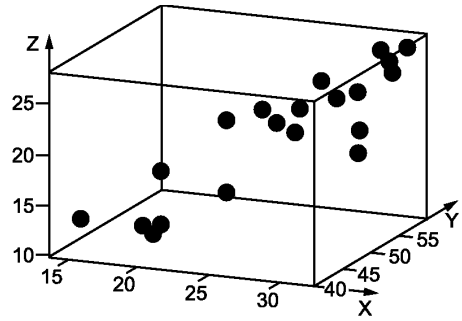
$$D_i = \frac{(n-p-1)e_i^2}{(p+1)R_{SSE}}\left[\frac{h_{ii}}{(1-h_{ii})^2}\right], \quad (23)$$

где большие значения e_i или h_{ii} дают большие значения D_i. Таким образом, большие значения D_i свидетельствуют о влияющих наблюдениях. В [14] сделано предположение, что наблюдения, для которых расстояние Кука превышает значение $F_{0,50;p+1,n-p-1}$, соответствующее процентилю уровня 50 % F-распределения, можно считать влияющими выбросами; здесь n — объем выборки, p+1— количество параметров регрессионной модели (включая свободный член), показывающий количество степеней свободы, связанных с числителем n–p–1 — число степеней свободы, связанных со знаменателем. Наблюдения со значениями расстояния Кука, превышающими $F_{0,50;p+1,n-p-1}$, следует изучить на предмет наличия ошибок при записи полученных данных или других возможных причин появления экстремальных значений.

Примечание — Приведенные методы не эффективны, если два или более влияющих наблюдения расположены близко друг к другу. Дополнительные процедуры, направленные на выявления двух или более влияющих наблюдений, расположенных близко друг к другу, требуют выполнения значительного количества вычислений.

Пример — Проведено исследование по определению связи общего содержания жира в организме человека (Y) с толщиной кожной складки над трицепсом (X_1) и обхватом бедра (X_2) (см. столбцы 2, 3 и 4 в таблице ниже). Данные исследования представлены в [12]. Трехмерный график для точек (Y, X_1, X_2) приведен на рисунке 10.

Номер наблюдения	Толщина кожной складки над трицепсом	Обхват бедра	Общее содержание жира	Ошибка	Значение влияния	Стьюдентизированный остаток
i	X_{1i}	X_{2i}	Y_i,	e_i	h_{ii}	r_i,
1	19,5	43,1	11,9	-1,683	0,201	-0,730
2	24,7	49,8	22,8	3,643	0,059	1,534
3	30,7	51,9	18,7	-3,176	0,372	-1,656
4	29,8	54,3	20,1	-3,158	0,111	-1,348
5	19,1	42,2	12,9	0,000	0,248	0,000
6	25,6	53,9	21,7	-0,361	0,129	-0,148
7	31,4	58,5	27,1	0,716	0,156	0,298
8	27,9	52,1	25,4	4,015	0,096	1,760
9	22,1	49,9	21,3	2,655	0,115	1,117
10	25,5	53,5	19,3	-2,475	0,110	-1,034
11	31,1	56,6	25,4	0,336	0,120	0,137
12	30,4	56,7	27,2	2,226	0,109	0,923
13	18,7	46,5	11,7	-3,947	0,178	-1,825
14	19,7	44,2	17,8	3,447	0,148	1,524
15	14,6	42,7	12,8	0,571	0,333	0,267
16	29,5	54,4	23,9	0,642	0,095	0,258
17	27,7	55,3	22,6	-0,851	0,106	0,344
18	30,2	58,6	25,4	-0,783	0,197	0,335
19	22,7	48,2	14,8	-2,857	0,067	-1,176
20	25,2	51,0	21,1	1,040	0,050	0,409



X — толщина кожной складки над трицепсом; Y — обхват бедра; Z — общее содержание жира

Рисунок 10 — График рассеяния для связи общего содержания жира с обхватом бедра и толщиной кожной складки над трицепсом

Методом наименьших квадратов получена функция регрессии

$$\hat{y}_i = -19{,}174 + 0{,}2224x_{1i} + 0{,}6594x_{2i}$$

при этом сумма квадратов остатков $R_{SSE} = \sum_{i=1}^{20} e_i^2 = 109{,}95$; h_{ii} и r_i для полученной функции регрессии представлены соответственно в столбцах 5, 6 и 7 таблицы выше.

Так как $n = 20$ и $p = 2$, то при установленном уровне значимости $\alpha = 0{,}05$

$$t_{1-\alpha/2n;n-p-2} = t_{0{,}99875;16} = 3{,}5802.$$

Так как $|r_i| \le 3{,}5802$ для всех i, то по компоненте Y не выявлено выбросов.

При выявлении выбросов по компоненте X, получено, что $h_{33} = 0{,}372$ и $h_{15,15} = 0{,}33$ превышают значение

$$2\bar{h} = 2(p + 1) / n = 2(2 + 1) / 20 = 0{,}3,$$

т. е. наблюдения 3 и 15 являются выбросами по компоненте X.

Для определения влияния наблюдений 3 и 15 на построенную линию регрессии подсчитаны соответствующие значения расстояния Кука

$$D_3 = \frac{17(-3{,}176)^2}{3(109{,}95)}\left[\frac{0{,}372}{(1-0{,}372)^2}\right] = 0{,}490$$

и $D_{15} = 0{,}212$. Так как оба значения меньше значения $F_{0{,}50;3,17} = 0{,}8212$, то наблюдения 3 и 15 не объявлены влияющими выбросами.

Функция регрессии при исключении наблюдения 3 представляет собой

$$\hat{y} = -12{,}248 + 0{,}5641x_{1i} + 0{,}3635x_{2i}$$

здесь значения оценок параметров существенно отличаются от соответствующих оценок, полученных с учетом 3-го наблюдения.

6.3.6 Робастная регрессионная процедура

Альтернативный подход к выявлению выбросов в регрессионном анализе состоит в построении робастной регрессионной модели для большей части данных и дальнейшем определении выбросов как точек, имеющих наибольшие остатки. Широко используют робастную регрессионную модель, получаемую методом усеченных наименьших квадратов (LTS) [15]. Регрессионные коэффициенты LTS-регрессии получают путем минимизации суммы m наименьших квадратов регрессионных остатков. Также рассматривают выборку объема n (y_i, x_{i1}, x_{i2}, ..., x_{ip}), i = 1, 2, ..., n, где приближенные значения и остатки находят по формулам

$$\hat{y}_i = b_0 + b_1 x_{i1} + \ldots + b_p x_{ip},$$

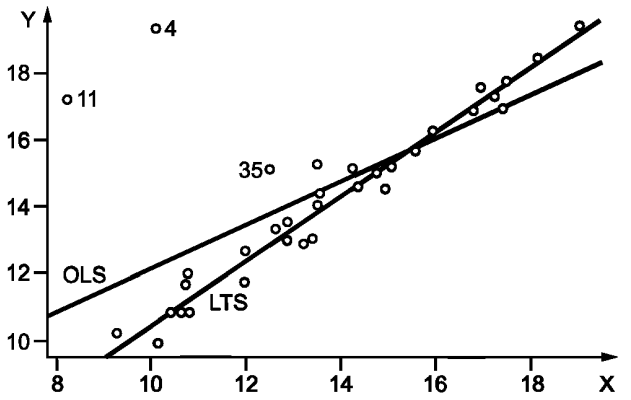
$$e_i = y_i - \hat{y}_i$$

соответственно.

В данном случае коэффициенты $b_0, b_1, \ldots, b_p$, LTS-регрессии представляют собой значения, минимизирующие сумму квадратов остатков $\sum_{i=1}^{m} e_{(i)}^2$, где $e_{(i)}^2$ является i-й порядковой статистикой квадратов остатков (т. е. остатки сначала возводят в квадрат, а затем упорядочивают), m — количество наблюдений (из n наблюдений), в отношении которых делается предположение о том, что они хорошо соответствуют регрессионной модели, полученной методом LTS. Если предполагается, что выборка содержит не более 100α % выбросов, значение m следует брать близким к $(1-\alpha)n$, но не менее целой части значения $[(n+p+1)/2]$. Наблюдения, содержащие большие остатки, считают выбросами.

Примечание — Оценку коэффициентов LTS-регрессии можно получить с помощью патентованных программных средств статистической обработки данных.

Пример — Для двумерных данных, составляющих выборку, рассмотренную в 6.2, на рисунке 11 представлены две линии регрессии, одна, соответствующая построению с помощью обычного метода наименьших квадратов (OLS), друга с помощью метода усеченных наименьших квадратов (LTS) при m = [0,9n] .



В соответствии с 6.2, точки 4, 11 и 35 представляют собой выбросы.

Рисунок 11 — Сравнение линий регрессии, построенных с помощью LTS и OLS

Две наиболее влиятельные точки, расположенные в левом верхнем углу, вызывают отклонение линии регрессии, полученной OLS-методом от основной массы элементов выборки, при этом метод LTS позволяет определить линию регрессии очень хорошо соответствующую данным. Робастная процедура LTS-регрессии по существу игнорирует две влияющих точки, в построение регрессионной модели входит только около 90 % выборочных данных.

Приложение А
(обязательное)

Алгоритм GESD-процедуры обнаружения выбросов

Пусть выборка $x_1, x_2, ..., x_n$ объема n отобрана из нормального распределения. Следующий алгоритм описывает необходимые этапы обнаружения m возможных выбросов с помощью процедуры обобщенных стьюдентизированных экстремальных отклонений (GESD) с уровнем значимости α.

Считывают значения α, m.

Устанавливают $l = 0$.

Устанавливают $I_0 = \{x_1, x_2, ..., x_n\}$.

ПОВТОРЯЮТ

Вычисляют выборочное среднее $\bar{x}(I_l)$ и выборочное стандартное отклонение для выборки I_l

Вычисляют статистику $R_l \dfrac{\max\limits_{x_i \in I_l} |x - \bar{x}(I_l)|}{s(I_l)}$.

Вычисляют $t_{p,n-l-2}$ процентиль t-распределения с $(n-l-2)$ степенями свободы, где $p = (1 - \alpha/2)^{1/(n-l)}$ уровня $100p$.

Вычисляют критическое значение $\lambda_l = \dfrac{(n-l-1)t_{p;n-l-2}}{\sqrt{(n-l-2+t^2_{p;n-l-2})(n-l)}}$.

Устанавливают $I_{l+1} = I_l \setminus \{x^{(l)}\}$ (см. примечание 1).

Устанавливают $l = l + 1$.

ДО ТЕХ ПОР ПОКА НЕ ВЫПОЛНЕНО $l = l + 1$.

Устанавливают $l = 0$.

ПОВТОРЯЮТ

Если $(R_l > \lambda_i)$, то $x(l)$ (значение x в I_l, участвующее в вычислении значения R_l) считают выбросом.

Устанавливают $l = l + 1$.

ДО ТЕХ ПОР ПОКА НЕ ВЫПОЛНЕНО $l = l + 1$.

Примечание 1 — I_{l+1} представляет собой редуцированную выборку объема $n-l$, полученную удалением точки $x^{(l)}$, участвующей в вычислении значения R_l, из выборки I_l.

Примечание 2 — Если $R_l > \lambda_i$ для всех $l = 0, 1, 2, ..., m$, то делают заключение о том, что в выборке нет выбросов.

Приложение В
(обязательное)

Критические значения статистик для критерия наличия выбросов в выборке из экспоненциального распределения

Таблица В.1 — Нижние и верхние критические уровни 2,5 % и 1 % значения $g_{E;n}$ для статистики G_E критерия Гринвуда для выборки из экспоненциального распределения

n	Нижнее 1 %	Нижнее 2,5 %	Верхнее 2,5 %	Верхнее 1 %	n	Нижнее 1 %	Нижнее 2,5 %	Верхнее 2,5 %	Верхнее 1 %	n	Нижнее 1 %	Нижнее 2,5 %	Верхнее 2,5 %	Верхнее 1 %
2	0,5000	0,5003	0,9754	0,9901	34	0,042 8	0,0443	0,0790	0,0863	82	0,0195	0,0201	0,0301	0,031 9
3	0,3360	0,3402	0,8314	0,8901	35	0,0417	0,0431	0,0765	0,0835	84	0,0191	0,0196	0,0293	0,0311
4	0,2585	0,2658	0,6828	0,7563	36	0,0407	0,0421	0,0742	0,0809	86	0,0187	0,0192	0,0286	0,0302
5	0,2137	0,2217	0,5680	0,6400	37	0,0397	0,0411	0,0720	0,0784	88	0,0183	0,0188	0,0279	0,0295
6	0,1838	0,1914	0,4821	0,5474	38	0,0388	0,0401	0,0699	0,0761	90	0,0179	0,0184	0,0272	0,0288
7	0,1620	0,1689	0,4173	0,4749	39	0,0379	0,0392	0,0680	0,0738	92	0,0176	0,0180	0,0266	0,0281
8	0,1452	0,1514	0,3667	0,4173	40	0,0371	0,0383	0,0661	0,0717	94	0,0173	0,0177	0,0260	0,0274
9	0,1318	0,1374	0,3263	0,3710	41	0,0363	0,0375	0,0643	0,0698	96	0,0169	0,0174	0,0254	0,0268
10	0,1208	0,1260	0,2934	0,3331	42	0,0355	0,0367	0,0626	0,0679	98	0,0166	0,0170	0,0248	0,0262
11	0,1116	0,1164	0,2661	0,3016	43	0,0348	0,0359	0,0610	0,0661	100	0,0163	0,0167	0,0243	0,0256
12	0,1039	0,1082	0,2431	0,2751	44	0,0341	0,0352	0,0595	0,0644	105	0,0156	0,0160	0,0230	0,0242
13	0,0972	0,1012	0,2236	0,2525	45	0,0334	0,0345	0,0581	0,0628	110	0,0149	0,0153	0,0219	0,0230
14	0,0913	0,0951	0,2068	0,2330	46	0,0328	0,0338	0,0567	0,0612	115	0,0143	0,0147	0,0209	0,0219
15	0,0862	0,0897	0,1922	0,2161	47	0,0322	0,0332	0,0554	0,0597	120	0,0138	0,0141	0,0199	0,0209
16	0,0816	0,0849	0,1794	0,2013	48	0,0316	0,0326	0,0541	0,0583	125	0,0133	0,0136	0,0191	0,0200
17	0,0776	0,0807	0,1681	0,1883	49	0,0310	0,0320	0,0529	0,0570	130	0,0128	0,0131	0,0183	0,0191
18	0,0739	0,0768	0,1581	0,1768	50	0,0305	0,0314	0,0517	0,0557	135	0,0124	0,0127	0,0176	0,0184
19	0,0706	0,0734	0,1491	0,1664	52	0,0294	0,0303	0,0496	0,0533	140	0,0120	0,0122	0,0169	0,0176
20	0,0676	0,0702	0,1411	0,1572	54	0,0284	0,0293	0,0475	0,0511	145	0,0116	0,0118	0,0163	0,0170
21	0,0648	0,0673	0,1338	0,1488	56	0,0275	0,0284	0,0457	0,0490	150	0,0112	0,0115	0,0157	0,0163
22	0,0623	0,0647	0,1272	0,1412	58	0,0267	0,0275	0,0440	0,0471	155	0,0109	0,0111	0,0152	0,0158
23	0,0600	0,0623	0,1212	0,1343	60	0,0259	0,0267	0,0424	0,0453	160	0,0106	0,0108	0,0146	0,0152
24	0,0578	0,0600	0,1157	0,1280	62	0,0251	0,0259	0,0409	0,0437	165	0,0103	0,0105	0,0142	0,0147
25	0,0558	0,0579	0,1107	0,1223	64	0,0244	0,0251	0,0395	0,0421	170	0,0100	0,0102	0,0137	0,0143
26	0,0540	0,0560	0,1060	0,1170	66	0,0238	0,0244	0,0382	0,0407	175	0,0097	0,0099	0,0133	0,0138
27	0,0522	0,0542	0,1017	0,1121	68	0,0231	0,0238	0,0369	0,0394	180	0,0095	0,0097	0,0129	0,0134
28	0,0506	0,0525	0,0978	0,1076	70	0,0225	0,0232	0,0358	0,0381	185	0,0092	0,0094	0,0125	0,0130
29	0,0491	0,0509	0,0941	0,1034	72	0,0220	0,0226	0,0347	0,0369	190	0,0090	0,0092	0,0122	0,0126
30	0,0477	0,0494	0,0906	0,0995	74	0,0214	0,0220	0,0337	0,0358	195	0,0088	0,0090	0,0119	0,0123
31	0,0464	0,0480	0,0874	0,0958	76	0,0209	0,0215	0,0327	0,0347	200	0,0086	0,0087	0,0115	0,0120
32	0,0451	0,0467	0,0844	0,0924	78	0,0204	0,0210	0,031 8	0,0337	225	0,0077	0,0078	0,0102	0,0105
33	0,0439	0,0454	0,081 6	0,0893	80	0,0200	0,0205	0,030 9	0,0328	250	0,0070	0,0071	0,0091	0,0094

Примечание 1 — Каждое критическое значение основано на обработке данных, полученных при исследовании ста миллионов модельных выборок объема n.

Примечание 2 — Каждое значение в таблице округлено вверх до четвертой цифры после запятой, что гарантирует требуемый уровень значимости.

Таблица В.2 — Верхние критические значения уровня 5 % и 1 % для последовательных критериев обнаружения верхних выбросов в выборке из экспоненциального распределения при $m = 2$

$m = 2$									
n	5 %		1 %		n	5 %		1 %	
	$S^{U}_{2:n}$	$S^{U}_{t:n}$	$S^{U}_{2:n}$	$S^{U}_{t:n}$		$S^{U}_{2:n}$	$S^{U}_{t:n}$	$S^{U}_{2:n}$	$S^{U}_{t:n}$
10	0,4348	0,4834	0,5143	0,5696	46	0,1187	0,1522	0,1376	0,1830
11	0,4010	0,4533	0,4748	0,5363	48	0,1145	0,1470	0,1327	0,1769
12	0,3724	0,4269	0,4412	0,5066	50	0,1106	0,1421	0,1282	0,1708
13	0,3480	0,4033	0,4125	0,4793	55	0,1020	0,1314	0,1179	0,1578
14	0,3268	0,3827	0,3868	0,4555	60	0,0946	0,1222	0,1092	0,1467
15	0,3082	0,3639	0,3647	0,4345	65	0,0884	0,1143	0,1020	0,1371
16	0,2916	0,3473	0,3447	0,4149	70	0,0830	0,1074	0,0955	0,1287
17	0,2770	0,3320	0,3273	0,3972	75	0,0783	0,1013	0,0899	0,1214
18	0,2637	0,3183	0,3114	0,3813	80	0,0741	0,0960	0,0849	0,1150
19	0,2519	0,3058	0,2971	0,3667	85	0,0703	0,0912	0,0807	0,1092
20	0,2413	0,2941	0,2845	0,3529	90	0,0670	0,0869	0,0767	0,1039
21	0,2313	0,2834	0,2723	0,3403	95	0,0639	0,0830	0,0732	0,0992
22	0,2224	0,2735	0,2618	0,3286	100	0,0612	0,0794	0,0700	0,0949
23	0,2142	0,2644	0,2519	0,3175	110	0,0564	0,0732	0,0644	0,0873
24	0,2065	0,2558	0,2426	0,3074	120	0,0524	0,0679	0,0596	0,0810
25	0,1995	0,2478	0,2340	0,2980	130	0,0489	0,0634	0,0556	0,0755
26	0,1929	0,2403	0,2263	0,2888	140	0,0458	0,0595	0,0521	0,0708
27	0,1868	0,2333	0,2190	0,2805	150	0,0432	0,0560	0,0491	0,0666
28	0,1812	0,2268	0,2123	0,2729	160	0,0409	0,0530	0,0464	0,0629
29	0,1757	0,2207	0,2058	0,2654	170	0,0388	0,0503	0,0440	0,0596
30	0,1708	0,2148	0,1998	0,2584	180	0,0369	0,0478	0,0418	0,0567
32	0,1617	0,2041	0,1890	0,2457	190	0,0353	0,0456	0,0399	0,0540
34	0,1535	0,1944	0,1792	0,2339	200	0,0337	0,0436	0,0381	0,0516
36	0,1462	0,1857	0,1705	0,2235	220	0,0312	0,0404	0,0351	0,0474
38	0,1397	0,1777	0,1627	0,2139	240	0,0289	0,0373	0,0325	0,0439
40	0,1337	0,1706	0,1555	0,2051	260	0,0269	0,0347	0,0303	0,0409
42	0,1283	0,1639	0,1491	0,1972	280	0,0252	0,0325	0,0284	0,0382
44	0,1233	0,1578	0,1432	0,1898	300	0,0238	0,0306	0,0267	0,0359

Таблица В.3 — Верхние критические значения уровня 5 % и 1 % для последовательных критериев обнаружения верхних выбросов в выборке из экспоненциального распределения при $m = 3$

$m = 3$													
n	5 %			1 %			n	5 %			1 %		
	$S^U_{3:n}$	$S^U_{2:n}$	$S^U_{t:n}$	$S^U_{3:n}$	$S^U_{2:n}$	$S^U_{t:n}$		$S^U_{3:n}$	$S^U_{2:n}$	$S^U_{t:n}$	$S^U_{3:n}$	$S^U_{2:n}$	$S^U_{t:n}$
15	0,3058	0,3210	0,3803	0,3577	0,3775	0,4497	55	0,0931	0,1052	0,1367	0,1056	0,1214	0,1635
16	0,2875	0,3035	0,3630	0,3360	0,3569	0,4296	60	0,0863	0,0976	0,1271	0,0975	0,1124	0,1520
17	0,2712	0,2881	0,3470	0,3165	0,3387	0,4112	65	0,0804	0,0912	0,1189	0,0908	0,1048	0,1421
18	0,2570	0,2743	0,3326	0,2994	0,3222	0,3949	70	0,0754	0,0855	0,1117	0,0849	0,0981	0,1333
19	0,2441	0,2619	0,3195	0,2837	0,3074	0,3798	75	0,0710	0,0806	0,1054	0,0799	0,0924	0,1257
20	0,2325	0,2507	0,3072	0,2698	0,2945	0,3658	80	0,0671	0,0762	0,0997	0,0754	0,0872	0,1190
21	0,2221	0,2403	0,2962	0,2579	0,2817	0,3525	85	0,0637	0,0724	0,0947	0,0715	0,0829	0,1130
22	0,2125	0,2309	0,2857	0,2462	0,2707	0,3404	90	0,0606	0,0689	0,0902	0,0679	0,0787	0,1076
23	0,2040	0,2224	0,2761	0,2362	0,2605	0,3290	95	0,0578	0,0658	0,0862	0,0648	0,0752	0,1026
24	0,1961	0,2142	0,2672	0,2268	0,2507	0,3186	100	0,0553	0,0629	0,0824	0,0619	0,0718	0,0981
25	0,1890	0,2068	0,2587	0,2181	0,2419	0,3087	110	0,0509	0,0580	0,0760	0,0569	0,0660	0,0903
26	0,1823	0,2000	0,2509	0,2104	0,2338	0,2993	120	0,0472	0,0538	0,0705	0,0527	0,0612	0,0837
27	0,1761	0,1937	0,2436	0,2029	0,2263	0,2907	130	0,0441	0,0502	0,0658	0,0491	0,0570	0,0780
28	0,1703	0,1878	0,2368	0,1962	0,2191	0,2829	140	0,0413	0,0471	0,0616	0,0460	0,0535	0,0731
29	0,1649	0,1821	0,2303	0,1897	0,2125	0,2749	150	0,0390	0,0444	0,0581	0,0433	0,0503	0,0688
30	0,1600	0,1770	0,2241	0,1840	0,2063	0,2680	160	0,0368	0,0420	0,0549	0,0409	0,0475	0,0650
32	0,1509	0,1674	0,2129	0,1730	0,1951	0,2546	170	0,0350	0,0398	0,0521	0,0388	0,0451	0,0616
34	0,1428	0,1589	0,2028	0,1637	0,1849	0,2426	180	0,0333	0,0379	0,0495	0,0369	0,0428	0,0585
36	0,1356	0,1513	0,1936	0,1552	0,1758	0,2318	190	0,0318	0,0362	0,0472	0,0352	0,0409	0,0557
38	0,1292	0,1444	0,1853	0,1476	0,1679	0,2218	200	0,0304	0,0346	0,0452	0,0336	0,0390	0,0533
40	0,1234	0,1382	0,1778	0,1409	0,1603	0,2125	220	0,0280	0,0318	0,0415	0,0309	0,0359	0,0489
42	0,1182	0,1326	0,1708	0,1348	0,1537	0,2044	240	0,0260	0,0295	0,0385	0,0287	0,0332	0,0453
44	0,1134	0,1274	0,1644	0,1291	0,1474	0,1969	260	0,0242	0,0276	0,0359	0,0267	0,0310	0,0421
46	0,1091	0,1226	0,1585	0,1240	0,1418	0,1898	280	0,0227	0,0258	0,0336	0,0250	0,0290	0,0394
48	0,1050	0,1182	0,1531	0,1193	0,1367	0,1834	300	0,0214	0,0243	0,0316	0,0236	0,0273	0,0370
50	0,1013	0,1142	0,1480	0,1150	0,1320	0,1769							

Таблица В.4 — Верхние критические значения уровня 5 % и 1 % для последовательных критериев обнаружения верхних выбросов в выборке из экспоненциального распределения при $m = 4$

$m = 4$								
n	5 %				1 %			
	$S^U_{4:n}$	$S^U_{3:n}$	$S^U_{2:n}$	$S^U_{t:n}$	$S^U_{4:n}$	$S^U_{3:n}$	$S^U_{2:n}$	$S^U_{t:n}$
20	0,231 9	0,238 1	0,257 3	0,316 4	0,267 5	0,275 8	0,301 3	0,374 7
21	0,220 8	0,227 4	0,246 5	0,304 9	0,254 4	0,263 5	0,288 3	0,360 7
22	0,210 4	0,217 5	0,236 9	0,294 1	0,242 0	0,251 5	0,277 0	0,348 5
23	0,201 3	0,208 8	0,228 0	0,284 2	0,231 0	0,241 2	0,266 2	0,336 8
24	0,192 8	0,200 7	0,219 6	0,275 0	0,221 1	0,231 6	0,256 3	0,326 3
25	0,185 2	0,193 2	0,212 0	0,266 2	0,212 1	0,222 7	0,247 3	0,316 3
26	0,178 1	0,186 3	0,204 9	0,258 1	0,203 7	0,214 8	0,239 0	0,306 5
27	0,171 6	0,180 0	0,198 4	0,250 7	0,196 1	0,207 2	0,231 3	0,297 6
28	0,165 6	0,174 0	0,192 4	0,243 6	0,189 0	0,200 2	0,223 8	0,289 7
29	0,160 2	0,168 5	0,186 6	0,236 9	0,182 5	0,193 4	0,217 1	0,281 7
30	0,154 9	0,163 4	0,181 1	0,230 5	0,176 4	0,187 6	0,210 9	0,274 5
32	0,145 6	0,154 1	0,171 3	0,219 0	0,165 4	0,176 3	0,199 3	0,260 7
34	0,137 5	0,145 8	0,162 6	0,208 5	0,155 9	0,166 8	0,188 9	0,248 3
36	0,130 2	0,138 4	0,154 7	0,199 0	0,147 3	0,158 1	0,179 5	0,237 3
38	0,123 8	0,131 8	0,147 7	0,190 5	0,140 0	0,150 4	0,171 4	0,227 0
40	0,118 0	0,125 9	0,141 3	0,182 7	0,133 0	0,143 5	0,163 6	0,217 7
42	0,112 8	0,120 5	0,135 5	0,175 5	0,127 1	0,137 2	0,156 7	0,209 2
44	0,108 0	0,115 6	0,130 2	0,168 9	0,121 5	0,131 4	0,150 4	0,201 5
46	0,103 7	0,111 1	0,125 2	0,162 8	0,116 6	0,126 2	0,144 6	0,194 3
48	0,099 7	0,107 0	0,120 8	0,157 2	0,112 0	0,121 4	0,139 3	0,187 8
50	0,096 0	0,103 2	0,116 6	0,151 9	0,107 7	0,117 0	0,134 5	0,181 1
55	0,088 1	0,094 8	0,107 4	0,140 4	0,098 6	0,107 3	0,123 7	0,167 2
60	0,081 4	0,087 8	0,099 6	0,130 5	0,090 9	0,099 2	0,114 5	0,155 5
65	0,075 8	0,081 8	0,093 0	0,122 0	0,084 5	0,092 3	0,106 8	0,145 4
70	0,070 9	0,076 7	0,087 2	0,114 6	0,078 9	0,086 3	0,099 9	0,136 3
75	0,066 7	0,072 2	0,082 2	0,108 0	0,074 1	0,081 1	0,094 1	0,128 6
80	0,063 0	0,068 2	0,077 7	0,102 3	0,069 9	0,076 5	0,088 8	0,121 7
85	0,059 7	0,064 7	0,073 8	0,097 2	0,066 2	0,072 6	0,084 3	0,115 5
90	0,056 8	0,061 6	0,070 2	0,092 5	0,062 9	0,068 9	0,080 1	0,109 9
95	0,054 1	0,058 7	0,067 0	0,088 3	0,059 8	0,065 7	0,076 5	0,105 0
100	0,051 7	0,056 2	0,064 1	0,084 5	0,057 2	0,062 8	0,073 0	0,100 3
110	0,047 6	0,051 7	0,059 0	0,077 8	0,052 5	0,057 7	0,067 2	0,092 3
120	0,044 1	0,047 9	0,054 7	0,072 2	0,048 6	0,053 4	0,062 2	0,085 5
130	0,041 1	0,044 7	0,051 1	0,067 3	0,045 2	0,049 8	0,057 9	0,079 7
140	0,038 6	0,042 0	0,047 9	0,063 1	0,042 4	0,046 6	0,054 3	0,074 6
150	0,036 3	0,039 5	0,045 1	0,059 5	0,039 8	0,043 9	0,051 1	0,070 2
160	0,034 3	0,037 4	0,042 7	0,056 2	0,037 6	0,041 4	0,048 3	0,066 4
170	0,032 6	0,035 5	0,040 5	0,053 3	0,035 7	0,039 3	0,045 8	0,062 9
180	0,031 0	0,033 7	0,038 5	0,050 7	0,033 9	0,037 4	0,043 5	0,059 7
190	0,029 6	0,032 2	0,036 8	0,048 3	0,032 3	0,035 6	0,041 5	0,056 9
200	0,028 3	0,030 8	0,035 2	0,046 2	0,030 9	0,034 0	0,039 6	0,054 3
220	0,026 1	0,028 4	0,032 4	0,042 5	0,028 4	0,031 3	0,036 4	0,049 9
240	0,024 2	0,026 3	0,030 0	0,039 3	0,026 4	0,029 0	0,033 7	0,046 2
260	0,022 6	0,024 6	0,028 0	0,036 6	0,024 6	0,027 0	0,031 4	0,043 0
280	0,021 2	0,023 0	0,026 2	0,034 3	0,023 0	0,025 3	0,029 4	0,040 2
300	0,020 0	0,021 7	0,024 7	0,032 3	0,021 7	0,023 9	0,027 7	0,037 8

Таблица В.5 — Верхние критические значения уровня 5 % и 1 % для последовательных критериев обнаружения нижних выбросов в выборке из экспоненциального распределения при $m = 2$

$m = 2$									
n	5 %		1 %		n	5 %		1 %	
	$S^{L}_{2:n}$	$S^{L}_{t:n}$	$S^{L}_{2:n}$	$S^{L}_{t:n}$		$S^{L}_{2:n}$	$S^{L}_{t:n}$	$S^{L}_{2:n}$	$S^{L}_{t:n}$
10	0,836 7	0,977 5	0,921 6	0,995 5	29	0,822 4	0,975 9	0,913 0	0,995 2
11	0,834 4	0,977 3	0,920 0	0,995 5	30	0,822 4	0,975 8	0,912 8	0,995 2
12	0,832 6	0,977 0	0,919 1	0,995 5	35	0,821 2	0,975 7	0,912 2	0,995 2
13	0,831 4	0,976 9	0,917 7	0,995 4	40	0,820 4	0,975 6	0,911 7	0,995 2
14	0,830 3	0,976 7	0,917 4	0,995 4	45	0,819 8	0,975 5	0,911 4	0,995 1
15	0,829 2	0,976 6	0,917 3	0,995 3	50	0,819 1	0,975 5	0,911 1	0,995 1
16	0,828 3	0,976 5	0,916 3	0,995 3	60	0,818 9	0,975 5	0,910 8	0,995 1
17	0,827 0	0,976 4	0,915 7	0,995 3	70	0,817 9	0,975 4	0,910 2	0,995 1
18	0,826 6	0,976 4	0,915 7	0,995 3	80	0,817 9	0,975 3	0,909 9	0,995 1
19	0,826 1	0,976 3	0,915 1	0,995 3	90	0,817 2	0,975 3	0,909 9	0,995 1
20	0,825 4	0,976 3	0,914 6	0,995 3	100	0,817 2	0,975 2	0,910 0	0,995 1
21	0,824 8	0,976 2	0,914 5	0,995 2	120	0,816 6	0,975 2	0,909 5	0,995 0
22	0,824 5	0,976 2	0,914 1	0,995 2	140	0,816 6	0,975 2	0,909 1	0,995 0
23	0,824 1	0,976 1	0,914 0	0,995 2	160	0,816 6	0,975 1	0,909 1	0,995 0
24	0,823 6	0,976 1	0,914 0	0,995 2	180	0,816 2	0,975 1	0,908 9	0,995 0
25	0,823 6	0,976 0	0,913 7	0,995 2	200	0,815 9	0,975 1	0,908 9	0,995 0
26	0,823 1	0,976 0	0,913 5	0,995 2	300	0,815 7	0,975 1	0,909 2	0,995 0
27	0,822 8	0,975 9	0,913 2	0,995 2					
28	0,822 5	0,976 0	0,913 0	0,995 2					

Таблица В.6 — Верхние критические значения уровня 5 % и 1 % для последовательных критериев обнаружения нижних выбросов в выборке из экспоненциального распределения при $m = 3$

$m = 3$														
n	5 %			1 %			n	5 %			1 %			
	$S^{U}_{3:n}$	$S^{U}_{2:n}$	$S^{U}_{t:n}$	$S^{U}_{3:n}$	$S^{U}_{2:n}$	$S^{U}_{t:n}$		$S^{U}_{3:n}$	$S^{U}_{2:n}$	$S^{U}_{t:n}$	$S^{U}_{3:n}$	$S^{U}_{2:n}$	$S^{U}_{t:n}$
15	0,7051	0,8555	0,9840	0,8073	0,9314	0,9969	40	0,6888	0,8472	0,9833	0,7937	0,9266	0,9968
16	0,7035	0,8544	0,9840	0,8062	0,9306	0,9969	50	0,6871	0,8462	0,9832	0,7922	0,9260	0,9967
17	0,7019	0,8536	0,9839	0,8050	0,9300	0,9968	60	0,6852	0,8459	0,9832	0,7911	0,9257	0,9967
18	0,7007	0,8532	0,9839	0,8034	0,9300	0,9968	70	0,6843	0,8449	0,9832	0,7904	0,9253	0,9967
19	0,6990	0,8527	0,9838	0,8027	0,9296	0,9968	80	0,6838	0,8449	0,9831	0,7895	0,9251	0,9967
20	0,6980	0,8520	0,9838	0,8015	0,9290	0,9968	90	0,6830	0,8443	0,9831	0,7895	0,9250	0,9967
21	0,6970	0,8517	0,9837	0,8011	0,9288	0,9968	100	0,6832	0,8444	0,9830	0,7887	0,9253	0,9967
22	0,6964	0,8511	0,9837	0,7995	0,9286	0,9968	120	0,6827	0,8438	0,9830	0,7885	0,9247	0,9967
23	0,6956	0,8507	0,9837	0,7995	0,9285	0,9968	140	0,6821	0,8434	0,9830	0,7882	0,9244	0,9967
24	0,6948	0,8502	0,9836	0,7988	0,9285	0,9968	160	0,6821	0,8437	0,9830	0,7877	0,9245	0,9967
25	0,6939	0,8503	0,9836	0,7978	0,9281	0,9968	180	0,6817	0,8436	0,9829	0,7874	0,9242	0,9967
26	0,6935	0,8499	0,9836	0,7980	0,9283	0,9968	200	0,6813	0,8437	0,9830	0,7866	0,9242	0,9967
27	0,6929	0,8495	0,9835	0,7970	0,9280	0,9968	250	0,6812	0,8432	0,9829	0,7869	0,9239	0,9967
28	0,6924	0,8493	0,9835	0,7972	0,9279	0,9968	300	0,6804	0,8431	0,9829	0,7863	0,9243	0,9966
29	0,6919	0,8491	0,9835	0,7969	0,9278	0,9968							
30	0,6915	0,8491	0,9834	0,7965	0,9276	0,9968							

Таблица В.7 — Верхние критические значения уровня 5 % и 1 % для последовательных критериев обнаружения нижних выбросов в выборке из экспоненциального распределения при $m = 4$

$m = 4$								
n	5 %				1 %			
	$S^{L}_{4:n}$	$S^{L}_{3:n}$	$S^{L}_{2:n}$	$S^{L}_{t:n}$	$S^{L}_{4:n}$	$S^{L}_{3:n}$	$S^{L}_{2:n}$	$S^{L}_{t:n}$
20	0,596 1	0,717 0	0,868 3	0,987 6	0,693 5	0,816 4	0,937 7	0,997 6
21	0,594 6	0,716 3	0,868 2	0,987 5	0,691 6	0,815 7	0,937 7	0,997 6
22	0,593 1	0,715 2	0,867 3	0,987 5	0,691 1	0,814 4	0,937 4	0,997 6
23	0,592 0	0,714 5	0,867 0	0,987 5	0,689 6	0,814 2	0,937 3	0,997 6
24	0,591 6	0,713 8	0,866 6	0,987 5	0,688 9	0,813 8	0,937 2	0,997 6
25	0,590 3	0,713 0	0,866 6	0,987 5	0,687 3	0,812 6	0,937 0	0,997 6
26	0,589 1	0,712 5	0,866 4	0,987 4	0,685 9	0,812 8	0,937 1	0,997 6
28	0,587 8	0,711 6	0,865 8	0,987 4	0,684 9	0,812 4	0,936 6	0,997 6
30	0,586 7	0,710 6	0,865 5	0,987 3	0,683 7	0,811 3	0,936 6	0,997 6
35	0,584 2	0,709 3	0,864 6	0,987 3	0,682 2	0,809 6	0,936 0	0,997 6
40	0,582 3	0,707 8	0,863 6	0,987 1	0,680 1	0,808 9	0,935 7	0,997 5
45	0,580 8	0,706 3	0,863 1	0,987 1	0,678 4	0,807 9	0,935 4	0,997 5
50	0,579 7	0,706 1	0,862 6	0,987 1	0,677 8	0,807 5	0,935 3	0,997 5
70	0,577 4	0,703 3	0,861 7	0,987 1	0,674 6	0,805 3	0,934 6	0,997 5
100	0,574 9	0,702 1	0,861 1	0,986 9	0,672 8	0,804 4	0,934 4	0,997 5
150	0,573 3	0,701 2	0,860 0	0,987 0	0,671 6	0,803 2	0,933 5	0,997 5
200	0,572 8	0,700 3	0,860 5	0,986 9	0,670 6	0,801 7	0,933 4	0,997 5

Приложение С
(обязательное)

Значения коэффициентов модифицированной диаграммы ящик с усами

Когда параметр положения θ и параметр масштаба σ предполагаемого распределения $F_{\theta,\sigma}(x)$ неизвестны, первый и третий квартили функции распределения оценивают посредством нижней четверти $X_{L;n}$ и верхней четверти $X_{U;n}$ выборки объема n, из распределения $F_{\theta,\sigma}(x)$. Существует много определений глубины выборочных четвертей. Рекомендуемое определение глубины следующее

$$\text{глубина четверти} = \begin{cases} i + 0{,}5, & \text{если } f = 0 \\ i + 1, & \text{если } f > 0 \end{cases},$$

где i — целая часть, а f — дробная часть значения $n/4$. Два значения, имеющие данную глубину, а именно нижнюю выборочную четверть $x_{L;n}$ и верхнюю выборочную четверть $x_{U;n}$ данной выборки объема n анализируют в соответствии с 4.4.

Точное выражение, которое может быть применено для оценки коэффициентов k_L и k_U, используемых при построении диаграммы ящик с усами для выборки из предполагаемого распределения $F_{\theta,\sigma}(x)$ приведено в [16].

$$\int_{-\infty}^{\infty}\int_{z_{l:n}}^{\infty}\{1 - I_{G_u(y_u)}(n-u,1)[1 - I_{G_l(y_l)}(1,l-1)]\}f_{z_{l:n},z_{u:n}}(z_{l:n},z_{u:n})dz_{u:n}dz_{l:n} = \alpha, \qquad (C.1)$$

где

a) α — установленная вероятность того, что в выборке, не содержащей выбросов, одно или более наблюдений будут ошибочно идентифицированы как выбросы;

b) $y_l = z_{l:n} - k_L(z_{u:n} - z_{l:n})$ и $y_l = z_{u:n} - k_U(z_{u:n} - z_{l:n})$;

c) $f_{z_{l:n},z_{u:n}}(z_{l:n},z_{u:n})$ — совместная функция плотности вероятностей для $z_{l:n}$ и $z_{u:n}$

$$f_{z_{l:n},z_{u:n}}(x,y) = \frac{n!}{(l-1)!(u-l-1)!(n-u)!}f(x)f(y)F^{l-1}(x)[F(y) - F(x)]^{u-l-1}[1 - F(y)]^{n-u}.$$

d) $Z_{r:n} = (X_{r:n} - \theta) / \sigma$ — r-я порядковая статистика для нормализованной случайной величины $Z - (X - \theta)/\sigma$ с функцией распределения $F(x)$;

e) $G_l(y) = F(y) / F(z_{l:n})$ и $G_u(y) = [F(y) - F(z_{u:n})] / [1 - F(z_{u:n})]$;

f) $I_p(a.b) = \frac{1}{B(a,b)}\int_0^p t^{a-1}(1-t)^{b-1}dt$ — неполная бета-функция.

Для определения значений k_L и k_U, удовлетворяющих двойному интегральному уравнению (С.1), может быть использован прямой алгоритм поиска.

В случае симметричного распределения в уравнении (С.1) используют $k_L = k_U = k$. Для асимметричного распределения значения k_L и k_U определяют отдельно при $P(X < L_F) = 1 - \Pr(X > U_F)$, т.е. $I_{G_l(y_l)}(1,l-1) = I_{G_u(y_u)}(n-u,1)$ в уравнении (С.1).

Значения $k_L = k_U = k$ для выборок объема $9 \le n \le 500$, отобранных из стандартного нормального распределения, могут быть аппроксимированы следующей функцией

$$k = \exp\{b_0 + b_1\ln(n) + b_2\ln^2(n) + b_3\ln^3(n) + b_4\ln^4(n) + b_5\ln^5(n)\}, \qquad (C.2)$$

где $b_5 = 0$, а коэффициенты b_i, $i = 0, 1, 2, 3, 4$ приведены в таблице С.1.

Значения k_L и k_U для выборки из асимметричного распределения или распределения экстремальных значений, также могут быть определены с помощью уравнения (С.2) с коэффициентами b_i, $i = 0, 1, 2, 3, 4$ из таблицы С.2.

В случае большого объема выборки значения k_L и k_U могут быть аппроксимированы следующим образом

$$k_L \approx \frac{F^{-1}(1/4) - F^{-1}(\alpha_n/2)}{F^{-1}(3/4) - F^{-1}(1/4)} \text{ и } k_U \approx \frac{F^{-1}(1-(\alpha_n/2) - F^{-1}(3/4)}{F^{-1}(3/4) - F^{-1}(1/4)},$$

где $\alpha_n = 1 - (1 - \alpha)^{1/n}$ может быть интегрирована, как вероятность того, что некоторое наблюдение из выборки объема n может быть ошибочно признано выбросом.

Пример 1 — Для выявления выбросов в выборке объема n = 20 из нормального распределения, значение $k_L = k_U = k$ для α = 0,05 определяют следующим образом

$$k = exp\{0,83707 + 0,07596 \times ln(20) - 0,06119 \times ln^2(20) + 0,01328 \times ln^3(20) - 0,00083 \times ln^4(20)\} = exp(0,80567) \approx 2,2382.$$

Пример 2 — Для выявления выбросов в выборке объема n = 22 из экспоненциального распределения, значения k_L и k_U для α = 0,05 определяют следующим образом

$$k_L = exp\{2,20604 - 1,41752 \cdot ln(20) - 0,24170 \cdot ln^2(20) - 0,02057 \cdot ln^3(20) + 0,00072 \cdot ln^4(20)\} = exp(-0,40802) \approx 0,6650,$$

$$k_U = exp\{2,74179 - 0,77067 \cdot ln(22) + 0,22688 \cdot ln^2(22) - 0,02853 \cdot ln^3(22) + 0,00170 \cdot ln^4(22) - 0,00004 \cdot ln^5(22)\} = exp(1,82958) \approx 6,2313.$$

Таблица С.1 — Коэффициенты функции аппроксимации коэффициентов *k*, используемых при построении диаграммы ящик с усами для выборок объема $9 \le n \le 500$ из нормального распределения с неизвестными параметрами

α	Нормальное распределение							
	mod(n,4)	b_0	b_1	b_2	b_3	b_4	b_5	δ
0,05	1	4,01761	-2,35363	0,64618	-0,07893	0,00368	—	0,01457
	2	2,06429	-0,88523	0,22237	-0,02391	0,00099	—	0,00064
	3	0,48006	0,25854	-0,09622	0,01620	-0,00092	—	0,00407
	0	0,83707	0,07596	-0,06119	0,01328	-0,00083	—	0,00462
0,01	1	6,37902	-3,84770	1,04438	-0,12813	0,00601	—	0,04183
	2	3,98772	-2,00630	0,50277	-0,05677	0,00248	—	0,00634
	3	2,14895	-0,65278	0,11985	-0,00796	0,00013	—	0,00417
	0	2,28507	-0,66052	0,10264	-0,00393	-0,00013	—	0,00686

Таблица С.2 — Коэффициенты для функции аппроксимации коэффициентов *k*, используемых при построении диаграммы ящик с усами для выборок объема $9 \le n \le 500$ из экспоненциального распределения с неизвестным параметром

α	Экспоненциальное распределение								
	коэффициент	mod(n,4)	b_0	b_1	b_2	b_3	b_4	b_5	δ
0,10	k_L	1	3,99024	-3,24052	0,95534	-0,15995	0,01440	-0,00054	0,00022
		2	1,13059	-0,72169	0,02306	0,01804	-0,00290	0,00014	0,00019
		3	-1,54986	1,60282	-0,82526	0,17801	-0,01829	0,00074	0,00047
		0	-1,95058	2,26133	-1,14744	0,24930	-0,02581	0,00105	0,00067
	k_U	1	3,58501	-1,56711	0,46464	-0,05769	0,00271	—	0,02172
		2	1,79740	-0,22367	0,07684	-0,00733	0,00024	—	0,00345
		3	0,33262	0,83429	-0,21797	0,02979	-0,00153	—	0,01154
		0	1,08640	0,33192	-0,08635	0,01396	-0,00080	—	0,00807
0,05	k_L	1	5,18220	-4,05528	1,22229	-0,20833	0,01901	-0,00072	0,00033
		2	2,20604	-1,41752	0,24170	-0,02057	0,00072	—	0,00011
		3	-0,57542	1,02024	-0,65689	0,15043	-0,01586	0,00065	0,00048
		0	-1,19027	1,86402	-1,04428	0,23327	-0,02440	0,00099	0,00088
	k_U	1	5,18029	-2,96781	1,04743	-0,18511	0,01683	-0,00063	0,00385
		2	2,74179	-0,77067	0,22688	-0,02853	0,00170	-0,00004	0,00131
		3	0,53026	1,19859	-0,50210	0,10967	-0,01158	0,00048	0,00544
		0	1,31043	0,60192	-0,30396	0,07456	-0,00832	0,00035	0,00437
0,02	k_L	1	6,72983	-5,17448	1,60518	-0,27980	0,02596	-0,00099	0,00052
		2	3,53662	-2,31042	0,53046	-0,07255	0,00566	-0,00019	0,00006
		3	0,56897	0,32976	-0,45563	0,11723	-0,01292	0,00054	0,00049
		0	-0,38125	1,48550	-0,96254	0,22351	-0,02380	0,00098	0,00126
	k_U	1	5,90497	-2,95227	0,83153	-0,10310	0,00486	—	0,06900
		2	3,79484	-1,32856	0,35393	-0,04015	0,00174	—	0,00715
		3	2,17127	-0,13525	0,01652	0,00286	-0,00033	—	0,01278
		0	2,67762	-0,43984	0,08873	-0,00507	0,00001	—	0,01325

Примечание — δ—максимум абсолютного отклонения значения *k* от его приближения для каждого класса значений *n* по модулю 4 (mod(*n*,4)) и объема выборки $9 \le n \le 500$.

Приложение D
(обязательное)

Значения коэффициентов коррекции для определения робастной оценки параметра масштаба

Т а б л и ц а D.1 — Коэффициенты коррекции s_n и s_{bi} для определения робастных оценок параметра масштаба S_n и S_{bi} соответственно

Объем выборки n	Коэффициент		Объем выборки n	Коэффициент	
	s_n	s_{bi}		s_n	s_{bi}
2	0,8866	1,1912	18	1,1961	1,0025
3	2,2051	1,3821	19	1,2438	1,0252
4	1,1385	1,1272	20	1,1951	1,0006
5	1,6081	1,1855	30	1,1927	0,9962
6	1,1858	1,0650	40	1,1921	0,9944
7	1,4297	1,1111	50	1,1920	0,9935
8	1,1989	1,0369	60	1,1920	0,9929
9	1,3500	1,0762	70	1,1921	0,9925
10	1,2015	1,0219	80	1,1921	0,9923
11	1,3074	1,0567	90	1,1922	0,9921
12	1,2006	1,0136	100	1,1923	0,9920
13	1,2814	1,0444	120	1,1924	0,9918
14	1,1994	1,0086	150	1,1925	0,9915
15	1,2647	1,0360	200	1,1926	0,9914
16	1,1978	1,0050	300	1,1927	0,9912
17	1,2526	1,0299	500	1,1927	0,9910

Приложение Е
(обязательное)

Критические значения статистики критерия Кохрена

Т а б л и ц а Е.1 — Критические значения статистики критерия Кохрена уровня 5 %

p	*n* = 2	*n* = 3	*n* = 4	*n* = 5	*n* = 6	*n* = 7	*n* = 8	*n* = 9	*n* = 10
2	0,998 5	0,975 1	0,939 2	0,905 8	0,877 3	0,853 4	0,833 2	0,816 0	0,801 1
3	0,967 0	0,871 0	0,797 8	0,745 7	0,707 0	0,677 1	0,653 1	0,633 4	0,616 8
4	0,906 5	0,768 0	0,683 9	0,628 8	0,589 5	0,559 9	0,536 5	0,517 6	0,501 8
5	0,841 3	0,683 8	0,598 1	0,544 1	0,506 4	0,478 3	0,456 4	0,438 8	0,424 2
6	0,780 8	0,616 2	0,532 2	0,480 4	0,444 8	0,418 5	0,398 1	0,381 7	0,368 2
7	0,727 0	0,561 2	0,480 0	0,430 8	0,397 2	0,372 6	0,353 6	0,338 4	0,325 9
8	0,679 9	0,515 7	0,437 8	0,391 0	0,359 4	0,336 3	0,318 5	0,304 3	0,292 7
9	0,638 5	0,477 5	0,402 8	0,358 4	0,328 5	0,306 8	0,290 1	0,276 8	0,266 0
10	0,602 1	0,445 0	0,3734	0,331 1	0,302 8	0,282 3	0,266 6	0,254 1	0,243 9
11	0,569 8	0,416 9	0,348 2	0,308 0	0,281 1	0,261 6	0,246 8	0,235 0	0,225 4
12	0,541 0	0,392 4	0,326 5	0,288 0	0,262 4	0,244 0	0,229 9	0,218 7	0,209 6
13	0,515 2	0,370 9	0,307 5	0,270 7	0,246 2	0,228 6	0,215 2	0,204 6	0,196 0
14	0,492 0	0,351 8	0,290 7	0,255 4	0,232 0	0,215 2	0,202 4	0,192 3	0,184 1
15	0,470 9	0,334 7	0,275 8	0,241 9	0,219 5	0,203 4	0,191 2	0,181 5	0,173 7
16	0,451 7	0,319 3	0,262 4	0,229 8	0,208 3	0,192 9	0,181 1	0,171 9	0,164 4
17	0,434 2	0,305 3	0,250 4	0,219 0	0,198 3	0,183 4	0,172 2	0,163 3	0,156 1
18	0,418 1	0,292 7	0,239 5	0,209 2	0,189 2	0,174 9	0,164 1	0,155 6	0,148 6
19	0,403 2	0,281 1	0,229 6	0,200 2	0,181 0	0,167 2	0,156 8	0,148 6	0,141 9
20	0,389 5	0,270 5	0,220 5	0,192 1	0,173 5	0,160 2	0,150 1	0,142 2	0,135 8
21	0,376 7	0,260 7	0,212 1	0,184 6	0,166 6	0,153 8	0,144 0	0,136 4	0,130 2
22	0,364 9	0,251 6	0,204 4	0,177 8	0,160 3	0,147 9	0,138 4	0,131 0	0,125 0
23	0,353 8	0,243 2	0,197 3	0,171 4	0,154 5	0,142 4	0,133 3	0,126 1	0,120 3
24	0,343 4	0,235 4	0,190 7	0,165 5	0,149 1	0,137 4	0,128 5	0,121 6	0,116 0
25	0,333 7	0,228 1	0,184 6	0,160 1	0,144 1	0,132 7	0,124 1	0,117 4	0,111 9
26	0,324 6	0,221 3	0,178 8	0,155 0	0,139 4	0,128 4	0,120 0	0,113 5	0,108 2
27	0,316 0	0,214 9	0,173 5	0,150 2	0,135 1	0,124 3	0,116 2	0,109 8	0,104 7
28	0,307 9	0,208 9	0,168 4	0,145 8	0,131 0	0,120 5	0,112 6	0,106 4	0,101 4
29	0,300 2	0,203 2	0,163 7	0,141 6	0,127 2	0,116 9	0,109 2	0,103 2	0,098 3
30	0,292 9	0,197 9	0,159 2	0,137 6	0,123 6	0,113 6	0,106 1	0,100 2	0,095 4
31	0,286 0	0,192 9	0,155 0	0,133 9	0,120 2	0,110 5	0,103 1	0,097 4	0,092 7
32	0,279 5	0,188 1	0,151 1	0,130 4	0,117 0	0,107 5	0,100 3	0,094 7	0,090 2
33	0,273 3	0,183 6	0,147 3	0,127 1	0,114 0	0,104 7	0,097 7	0,092 2	0,087 8
34	0,267 3	0,179 3	0,143 7	0,124 0	0,111 1	0,102 0	0,095 2	0,089 8	0,085 5
35	0,261 7	0,175 2	0,140 4	0,121 0	0,108 4	0,099 5	0,092 8	0,087 6	0,083 3

Окончание таблицы Е.1

p	$n = 2$	$n = 3$	$n = 4$	$n = 5$	$n = 6$	$n = 7$	$n = 8$	$n = 9$	$n = 10$
36	0,256 3	0,171 3	0,137 1	0,118 1	0,105 8	0,097 1	0,090 6	0,085 4	0,081 3
37	0,251 1	0,167 6	0,134 1	0,115 5	0,103 4	0,094 9	0,088 4	0,083 4	0,079 4
38	0,246 2	0,164 0	0,131 2	0,112 9	0,101 1	0,092 7	0,086 4	0,081 5	0,077 5
39	0,241 4	0,160 7	0,128 4	0,110 4	0,098 8	0,090 6	0,084 5	0,079 6	0,075 8
40	0,236 9	0,157 4	0,125 7	0,108 1	0,096 7	0,088 7	0,082 6	0,077 9	0,074 1

Примечание 1 — p — размерность вектора случайных величин (количество дисперсий); n — объем выборки (количество репликаций при определении дисперсии).

Примечание 2 — У каждого значения в таблице последний десятичный знак округлен вверх, что обеспечивает требуемый уровень значимости.

Примечание 3 — Каждое значение в таблице получено по результатам моделирования 50 миллионов выборок.

Таблица Е.2 — Критические значения статистики критерия Кохрена уровня 1 %

p	$n = 2$	$n = 3$	$n = 4$	$n = 5$	$n = 6$	$n = 7$	$n = 8$	$n = 9$	$n = 10$
2	0,999 94	0,995 1	0,979 4	0,958 6	0,937 3	0,917 2	0,898 9	0,882 3	0,867 4
3	0,993 4	0,942 3	0,883 2	0,833 5	0,793 4	0,760 7	0,733 6	0,710 8	0,691 2
4	0,967 6	0,864 3	0,781 5	0,721 3	0,676 2	0,641 1	0,612 9	0,589 8	0,570 3
5	0,927 9	0,788 6	0,695 8	0,632 9	0,587 6	0,553 1	0,525 9	0,503 8	0,485 4
6	0,882 9	0,721 8	0,625 9	0,563 5	0,519 6	0,486 6	0,460 9	0,440 1	0,423 0
7	0,837 7	0,664 5	0,568 5	0,508 0	0,466 0	0,434 8	0,410 6	0,391 2	0,375 2
8	0,794 5	0,615 2	0,521 0	0,462 7	0,422 7	0,393 2	0,370 5	0,352 3	0,337 4
9	0,754 4	0,572 8	0,481 0	0,425 1	0,387 1	0,359 2	0,337 8	0,320 8	0,306 8
10	0,717 5	0,535 9	0,446 9	0,393 4	0,357 2	0,330 9	0,310 6	0,294 6	0,281 4
11	0,683 7	0,503 6	0,417 6	0,366 3	0,331 8	0,306 8	0,287 7	0,272 5	0,260 1
12	0,652 8	0,475 2	0,392 0	0,342 9	0,310 0	0,286 2	0,268 0	0,253 6	0,241 9
13	0,624 5	0,449 9	0,369 5	0,322 4	0,290 9	0,268 2	0,251 0	0,237 3	0,226 2
14	0,598 6	0,427 3	0,349 6	0,304 3	0,274 2	0,252 5	0,236 0	0,223 0	0,212 5
15	0,574 7	0,406 9	0,331 8	0,288 2	0,259 4	0,238 6	0,222 9	0,210 4	0,200 4
16	0,552 8	0,388 6	0,315 8	0,273 9	0,246 1	0,226 2	0,211 1	0,199 3	0,189 6
17	0,532 5	0,371 9	0,301 4	0,260 9	0,234 2	0,215 1	0,200 6	0,189 3	0,180 0
18	0,513 7	0,356 6	0,288 3	0,249 2	0,223 5	0,205 1	0,191 2	0,180 2	0,171 4
19	0,496 2	0,342 6	0,276 4	0,238 6	0,213 7	0,196 0	0,182 6	0,172 1	0,163 5
20	0,479 9	0,329 8	0,265 5	0,228 8	0,204 8	0,187 7	0,174 8	0,164 7	0,156 4
21	0,464 8	0,317 9	0,255 4	0,219 9	0,196 7	0,180 1	0,167 7	0,157 9	0,149 9
22	0,450 6	0,306 9	0,246 1	0,211 7	0,189 2	0,173 2	0,161 1	0,151 7	0,144 0
23	0,437 3	0,296 7	0,237 5	0,204 1	0,182 3	0,166 8	0,155 1	0,145 9	0,138 5
24	0,424 8	0,287 1	0,229 5	0,197 0	0,175 9	0,160 8	0,149 5	0,140 6	0,133 4
25	0,413 0	0,278 2	0,222 1	0,190 5	0,169 9	0,155 3	0,144 3	0,135 7	0,128 8
26	0,401 9	0,269 9	0,215 1	0,184 4	0,164 4	0,150 2	0,139 5	0,131 1	0,124 4
27	0,391 5	0,262 1	0,208 6	0,178 7	0,159 2	0,145 4	0,135 0	0,126 9	0,120 3
28	0,381 6	0,254 8	0,202 5	0,173 3	0,154 3	0,140 9	0,130 8	0,122 9	0,116 5
29	0,372 2	0,247 8	0,196 8	0,168 3	0,149 8	0,136 7	0,126 9	0,119 2	0,113 0

Окончание таблицы Е.2

p	$n = 2$	$n = 3$	$n = 4$	$n = 5$	$n = 6$	$n = 7$	$n = 8$	$n = 9$	$n = 10$
30	0,363 3	0,241 3	0,191 4	0,163 6	0,145 5	0,132 8	0,123 2	0,115 7	0,109 6
31	0,354 8	0,235 1	0,186 3	0,159 1	0,141 5	0,129 0	0,119 7	0,112 4	0,106 5
32	0,346 8	0,229 3	0,181 5	0,154 9	0,137 7	0,125 5	0,116 4	0,109 3	0,103 5
33	0,339 1	0,223 7	0,176 9	0,150 9	0,134 1	0,122 2	0,113 3	0,106 4	0,100 8
34	0,331 8	0,218 4	0,172 6	0,147 2	0,130 7	0,119 1	0,110 4	0,103 6	0,098 1
35	0,324 8	0,213 4	0,168 5	0,143 6	0,127 5	0,116 1	0,107 6	0,101 0	0,095 6
36	0,318 1	0,208 6	0,164 6	0,140 2	0,124 4	0,113 3	0,105 0	0,098 5	0,093 3
37	0,311 7	0,204 1	0,160 9	0,136 9	0,121 5	0,110 6	0,102 5	0,096 1	0,091 0
38	0,305 6	0,199 7	0,157 3	0,133 9	0,118 7	0,108 1	0,100 1	0,093 9	0,088 9
39	0,299 7	0,195 6	0,153 9	0,130 9	0,116 1	0,105 7	0,097 8	0,091 7	0,086 8
40	0,294 1	0,191 6	0,150 7	0,128 1	0,113 6	0,103 3	0,095 7	0,089 7	0,084 9

Примечание 1 — p — размерность вектора случайных величин (количество дисперсий); n — объем выборки (количество репликаций при определении дисперсии).

Примечание 2 — У каждого значения в таблице последний десятичный знак округлен вверх, что обеспечивает требуемый уровень значимости.

Примечание 3 — Каждое значение в таблице получено по результатам моделирования 50 миллионов выборок.

Таблица Е.3 — Критические значения статистики критерия Кохрена уровня 0,1 %

p	$n = 2$	$n = 3$	$n = 4$	$n = 5$	$n = 6$	$n = 7$	$n = 8$	$n = 9$	$n = 10$
2	0,999 999 4	0,999 6	0,995 6	0,987 1	0,975 5	0,962 5	0,949 2	0,936 1	0,923 6
3	0,999 4	0,981 8	0,946 3	0,907 9	0,872 6	0,841 4	0,814 2	0,790 3	0,769 3
4	0,993 0	0,937 1	0,870 3	0,813 2	0,766 8	0,728 8	0,697 3	0,670 8	0,648 1
5	0,977 0	0,881 1	0,794 6	0,728 8	0,678 4	0,638 8	0,606 8	0,580 3	0,558 0
6	0,952 9	0,824 5	0,727 1	0,657 9	0,606 8	0,567 6	0,536 4	0,510 9	0,489 7
7	0,923 8	0,771 4	0,668 5	0,598 7	0,548 5	0,510 5	0,480 6	0,456 4	0,436 3
8	0,892 3	0,723 1	0,618 0	0,549 1	0,500 3	0,463 9	0,435 4	0,412 5	0,393 6
9	0,860 2	0,679 6	0,574 4	0,507 0	0,460 0	0,425 2	0,398 1	0,376 5	0,358 7
10	0,828 5	0,640 7	0,536 4	0,471 0	0,425 8	0,392 5	0,366 9	0,346 4	0,329 6
11	0,798 0	0,605 7	0,503 2	0,439 8	0,396 4	0,364 7	0,340 3	0,320 9	0,305 0
12	0,768 8	0,574 3	0,473 9	0,412 6	0,371 0	0,340 6	0,317 4	0,298 9	0,283 9
13	0,741 2	0,545 9	0,447 8	0,388 6	0,348 7	0,319 6	0,297 4	0,279 9	0,265 6
14	0,715 2	0,520 2	0,424 6	0,367 4	0,329 0	0,301 1	0,279 9	0,263 2	0,249 5
15	0,690 6	0,496 9	0,403 7	0,348 4	0,311 4	0,284 7	0,264 5	0,248 4	0,235 4
16	0,667 6	0,475 6	0,384 8	0,331 4	0,295 7	0,270 1	0,250 6	0,235 3	0,222 8
17	0,645 9	0,456 1	0,367 7	0,315 9	0,281 6	0,256 9	0,238 2	0,223 5	0,211 6
18	0,625 5	0,438 1	0,352 1	0,302 0	0,268 8	0,245 0	0,227 0	0,212 9	0,201 4
19	0,606 3	0,421 6	0,337 8	0,289 2	0,257 2	0,234 2	0,216 9	0,203 3	0,192 2
20	0,588 2	0,406 3	0,324 6	0,277 5	0,246 5	0,224 4	0,207 6	0,194 5	0,183 9
21	0,571 1	0,392 1	0,312 5	0,266 8	0,236 7	0,215 3	0,199 2	0,186 5	0,176 2

Окончание таблицы Е.3

p	*n* = 2	*n* = 3	*n* = 4	*n* = 5	*n* = 6	*n* = 7	*n* = 8	*n* = 9	*n* = 10
22	0,555 0	0,378 9	0,301 3	0,256 9	0,227 7	0,207 0	0,191 4	0,179 1	0,169 2
23	0,539 8	0,366 6	0,290 9	0,247 7	0,219 4	0,199 3	0,184 2	0,172 3	0,162 8
24	0,525 4	0,355 1	0,281 2	0,239 2	0,211 7	0,192 2	0,177 6	0,166 1	0,156 8
25	0,511 8	0,344 3	0,272 1	0,231 2	0,204 6	0,185 6	0,171 4	0,160 3	0,151 3
26	0,498 8	0,334 2	0,263 7	0,223 8	0,197 9	0,179 5	0,165 7	0,154 8	0,146 1
27	0,486 5	0,324 6	0,255 8	0,216 9	0,191 6	0,173 7	0,160 3	0,149 8	0,141 3
28	0,474 9	0,315 7	0,248 3	0,210 4	0,185 8	0,168 4	0,155 3	0,145 1	0,136 9
29	0,463 8	0,307 2	0,241 3	0,204 3	0,180 3	0,163 3	0,150 6	0,140 7	0,132 7
30	0,453 2	0,299 2	0,234 7	0,198 6	0,175 2	0,158 6	0,146 2	0,136 5	0,128 7
31	0,443 1	0,291 6	0,228 5	0,193 2	0,170 3	0,154 1	0,142 1	0,132 6	0,125 0
32	0,433 4	0,284 4	0,222 6	0,188 0	0,165 7	0,149 9	0,138 1	0,128 9	0,121 5
33	0,424 2	0,277 6	0,217 0	0,183 2	0,161 4	0,146 0	0,134 4	0,125 5	0,118 2
34	0,415 4	0,271 1	0,211 7	0,178 6	0,157 3	0,142 2	0,131 0	0,122 2	0,115 1
35	0,406 9	0,264 9	0,206 7	0,174 3	0,153 4	0,138 6	0,127 6	0,119 1	0,112 2
36	0,398 8	0,259 0	0,201 9	0,170 1	0,149 7	0,135 3	0,124 5	0,116 1	0,109 4
37	0,391 0	0,253 4	0,197 3	0,166 2	0,146 1	0,132 0	0,121 5	0,113 3	0,106 7
38	0,383 6	0,248 0	0,192 9	0,162 4	0,142 8	0,129 0	0,118 7	0,110 6	0,104 2
39	0,376 4	0,242 9	0,188 8	0,158 8	0,139 6	0,126 1	0,116 0	0,108 1	0,101 8
40	0,369 5	0,238 0	0,184 8	0,155 4	0,136 5	0,123 3	0,113 4	0,105 7	0,099 5

Примечание 1 — *p* — размерность вектора случайных величин (количество дисперсий); *n* — объем выборки (количество репликаций при определении дисперсии).

Примечание 2 — У каждого значения в таблице последний десятичный знак округлен вверх, что обеспечивает требуемый уровень значимости.

Примечание 3 — Каждое значение в таблице получено по результатам моделирования 50 миллионов выборок.

Приложение F
(справочное)

Руководство по выявлению выбросов в одномерной выборке

Пусть имеется партия, выборка наблюдений или набор выборочных средних или дисперсий. Целью является выявление и идентификация выбросов в наборе данных. В данном приложении приведено руководство для пользователей настоящего стандарта. Данное руководство представляет собой набор этапов, выполнение которых соответствует содержанию определенных разделов и подразделов настоящего стандарта. Используемые в данном приложении обозначения соответствуют обозначениям, примененным в настоящем стандарте.

Этап 1. Представление точек, соответствующих набору данных на графике рассеяния, диаграмме стебель — листья, диаграмме ящик с усами или упорядочивание данных в порядке неубывания

$$x_{(1)} \le x_{(2)} \le \ldots \le x_{(k)} \le \ldots \le x_{(n)},$$

где $x_{(i)}$ i-е от наименьшего наблюдения.

Этап 2. Проверка графического представления данных или упорядоченных данных визуального вычисления возможных (предполагаемых) выбросов. При отсутствии сомнений о том, что предполагаемые выбросы действительно являются выбросами, переходят к выполнению этапа 5. Если одно или более наблюдений расположены достаточно далеко от других данных, переходят к выполнению этапа 3; в противном случае принимают решение о том, что выборка не содержит выбросов и может быть использована для дальнейшего анализа данных.

Этап 3. Подтверждают соответствие данных выборочному закону распределения или изменяют вид распределения:

a) предположение о нормальном распределении данных подтверждают с помощью графика нормальной вероятности на вероятностной бумаге;

b) предположение об экспоненциальном распределении данных подтверждают с помощью графика экспоненциальной вероятности на вероятностной бумаге;

c) при предположении о логнормальном распределении данных выполняют преобразование исходных данных к данным, распределение которых близко к нормальному распределению, используя процедуру, проведенную в 4.3.4.2 и с последующей проверкой соответствия преобразованных данных графику нормального распределения на вероятностной бумаге;

d) при предположении о том, что распределение является распределением экстремальных значений, выполняют преобразование исходных данных к данным, распределение которых близко к экспоненциальному распределению, используя процедуру, приведенную в 4.3.4.3 с последующей проверкой соответствия преобразованных данных графику экспоненциального распределения на вероятностной бумаге;

e) при предположении о том, что распределение является распределением Вейбулла, выполняют преобразование исходных данных к данным, распределение которых близко к экспоненциальному распределению, используя процедуру, приведенную в 4.3.4.4 с последующей проверкой соответствия преобразованных данных графику экспоненциального распределения на вероятностной бумаге;

f) при предположении о гамма-распределении данных, выполняют преобразование исходных данных к данным, распределение которых близко к нормальному распределению, используя процедуру, приведенную в 4.3.4.5 с последующей проверкой соответствия преобразованных данных к графику нормального распределения на вероятностной бумаге;

g) если распределение совокупности, из которой отобрана выборка, неизвестно или предполагаемое распределение не соответствует данным, или распределение не является одним из указанных выше распределений, выполняют преобразование исходных данных к данным, распределение которых близко к нормальному распределению, используя преобразование Бокса-Кокса или преобразование Джонсона с последующей проверкой соответствия преобразованных данных к графику нормального распределения на вероятностной бумаге. Если нормальное распределение не соответствует преобразованным данным, следует перейти к выполнению этапа 6 и провести анализ данных, используя робастные процедуры, приведенные в 5.

Этап 4. Выполняют проверку того, что предполагаемые выбросы, выявленные на этапе 2, действительно являются выбросами:

a) если исходные или преобразованные данные согласуются с нормальным распределением, следует использовать процедуру, описанную в 4.3.2 и/или в 4.4;

b) если исходные данные или преобразованные данные согласуются с экспоненциальным распределением, следует использовать процедуру, описанную в 4.3.3 и/или в 4.4;

c) если один или несколько предполагаемых выбросов идентифицированы в качестве выбросов, следует перейти к выполнению этапа 5, в противном случае, принимают решение об отсутствии выбросов и используют исходные или преобразованные данные для дальнейшего анализа.

Этап 5. Устанавливают причины появления выявленных выбросов.

Этап 6. Если причины появления выбросов могут быть установлены, удаляют выявленные выбросы из набора данных, а оставшиеся данные используют для последующего анализа, в противном случае используют робастные процедуры для анализа данных.

В блок-схеме, представленной на рисунке F.1, приведены рекомендуемые этапы выявления и обработки выбросов.

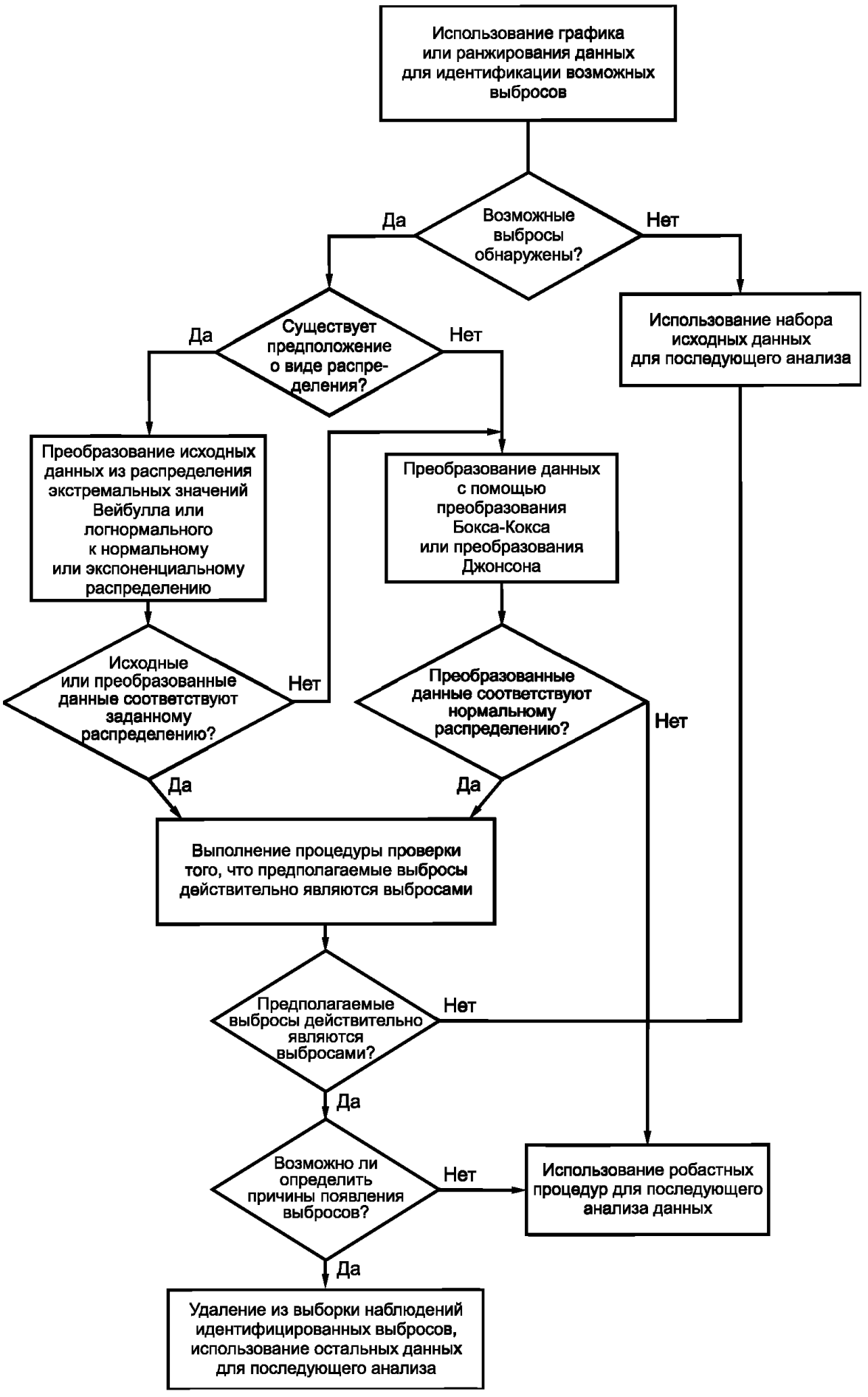


Рисунок F.1 — Блок схема для процедуры выявления и обработки выбросов

Библиография

[1] BARNETT, V. and LEWIS, T. Outliers in Statistical data. 3rd edition. New York: Wiley, 1994

[2] TUKEY, J.W. Exploratory data analysis. Reading, Massachusetts: Addison-Wesley, 1977

[3] ISO 5725-2:1994, Accuracy (trueness and precision) of measurement methods and results — Part 2: Basic method for the determination of repeatability and reproducibility of a standard measurement method

[4] ROSNER, B. Percentage Points for a Generalized ESD Many-Outlier Procedure. Technometrics, 25, 1983, pp. 165-172

[5] KIMBER, A.C., Tests for many outliers in an exponential sample. Applied Statistics, 31, 1982, pp. 263-271

[6] KITTLITZ, R.G. Transforming the exponential for SPC applications. Journal of Quality Technology, 31, 1999, pp. 301-308

[7] BOX, G.E.P. and COX, D.R. An analysis of transformations. Journal of the Royal Statistical Society, Series B 26, 1964, pp. 211-246

[8] CHOU, Y., POLANSKY, A.M. and MASON, R.L. Transforming Nonnormal Data to Normality in Statistical Process Control. Journal of Quality Technology, 30, 1998, pp. 133-141

[9] HOAGLIN, D.C., MOSTELLER, F. and TUKEY, J.W. Understanding robust and exploratory data analysis. New York: Wiley, 1983

[10] ROUSSEEUW, P.J. and CROUX, C. Alternatives to the median absolute deviation. Journal of the American Statistical Association, 88, 1993, pp. 1273-1283

[11] VERBOVEN, S. and HUBERT, M. LIBRA: a MATLAB Library for Robust Analysis, Chemometrics and Intelligent Laboratory Systems, 75, 2005, pp. 127-136

[12] KUTNER, M.H., NACHTSHEIM, C.J., NETER, J. and LI, W. Applied linear statistical models. Singapore: McGraw-Hill, 2005

[13] HUBER, P.J. Robust Statistics. New York: Wiley, 1981

[14] COOK, R.D. and WEISBERG, S. Residuals and influence in regression. London: Chapman & Hall, 1982

[15] ROUSSEEUW, P.J. and LEROY, A.M. Robust Regression and Outlier Detection. New York: John Wiley, 1987

[16] SIM, C.H., GAN, F.F. and CHANG, T.C. Outlier Labeling with Boxplot Procedures. Journal of the American Statistical Association, 100, 2005, pp. 642-652

[17] ISO 3534-1:2006, Statistics — Vocabulary and symbols — Part 1: General statistical terms and terms used in probability

[18] ISO 5479, Statistical interpretation of data — Tests for departure from the normal distribution

УДК 658.562.012.7:65.012.122:006.354 ОКС 03.120.30 Т59

Ключевые слова: выборка, распределение, выброс, устойчивая процедура, робастная процедура, робастная оценка, порядковая статистика, глубина

БЗ 9—2017/115

Редактор *И.М. Сазонкина*
Технический редактор *И.Е. Черепкова*
Корректор *С.И. Фирсова*
Компьютерная верстка *А.А. Ворониной*

Сдано в набор 14.08.2017. Подписано в печать 24.08.2017. Формат 60×84 1/8. Гарнитура Ариал.
Усл. печ. л. 6,05. Уч.-изд. л. 5,45. Тираж 23 экз. Зак. 1515.
Подготовлено на основе электронной версии, предоставленной разработчиком стандарта

Издано и отпечатано во ФГУП «СТАНДАРТИНФОРМ», 123001 Москва, Гранатный пер., 4.
www.gostinfo.ru info@gostinfo.ru