INTERNATIONAL STANDARD

ISO 16269-4

First edition
2010-10-15

Statistical interpretation of data —

Part 4: Detection and treatment of outliers

Interprétation statistique des données —

Partie 4: Détection et traitement des valeurs aberrantes

Reference number
ISO 16269-4:2010(E)

PDF disclaimer

This PDF file may contain embedded typefaces. In accordance with Adobe's licensing policy, this file may be printed or viewed but shall not be edited unless the typefaces which are embedded are licensed to and installed on the computer performing the editing. In downloading this file, parties accept therein the responsibility of not infringing Adobe's licensing policy. The ISO Central Secretariat accepts no liability in this area.

Adobe is a trademark of Adobe Systems Incorporated.

Details of the software products used to create this PDF file can be found in the General Info relative to the file; the PDF-creation parameters were optimized for printing. Every care has been taken to ensure that the file is suitable for use by ISO member bodies. In the unlikely event that a problem relating to it is found, please inform the Central Secretariat at the address given below.

ISO copyright office
Case postale 56 • CH-1211 Geneva 20
Tel. + 41 22 749 01 11
Fax + 41 22 749 09 47
E-mail copyright@iso.org
Web www.iso.org

Published in Switzerland

Contents

Page

Foreword

ISO (the International Organization for Standardization) is a worldwide federation of national standards bodies (ISO member bodies). The work of preparing International Standards is normally carried out through ISO technical committees. Each member body interested in a subject for which a technical committee has been established has the right to be represented on that committee. International organizations, governmental and non-governmental, in liaison with ISO, also take part in the work. ISO collaborates closely with the International Electrotechnical Commission (IEC) on all matters of electrotechnical standardization.

International Standards are drafted in accordance with the rules given in the ISO/IEC Directives, Part 2.

The main task of technical committees is to prepare International Standards. Draft International Standards adopted by the technical committees are circulated to the member bodies for voting. Publication as an International Standard requires approval by at least 75 % of the member bodies casting a vote.

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights. ISO shall not be held responsible for identifying any or all such patent rights.

ISO 16269-4 was prepared by Technical Committee ISO/TC 69, *Applications of statistical methods*.

ISO 16269 consists of the following parts, under the general title *Statistical interpretation of data*:

— *Part 4: Detection and treatment of outliers*

— *Part 6: Determination of statistical tolerance intervals*

— *Part 7: Median — Estimation and confidence intervals*

— *Part 8: Determination of prediction intervals*

Introduction

Identification of outliers is one of the oldest problems in interpreting data. Causes of outliers include measurement error, sampling error, intentional under- or over-reporting of sampling results, incorrect recording, incorrect distributional or model assumptions of the data set, and rare observations, etc.

Outliers can distort and reduce the information contained in the data source or generating mechanism. In the manufacturing industry, the existence of outliers will undermine the effectiveness of any process/product design and quality control procedures. Possible outliers are not necessarily *bad* or *erroneous*. In some situations, an outlier may carry essential information and thus it should be identified for further study.

The study and detection of outliers from measurement processes leads to better understanding of the processes and proper data analysis that subsequently results in improved inferences.

In view of the enormous volume of literature on the topic of outliers, it is of great importance for the international community to identify and standardize a sound subset of methods used in the identification and treatment of outliers. The implementation of this part of ISO 16269 enables business and industry to recognize the data analyses conducted across member countries or organizations.

Six annexes are provided. Annex A provides an algorithm for computing the test statistic and critical values of a procedure in detecting outliers in a data set taken from a normal distribution. Annexes B, D and E provide the tables needed to implement the recommended procedures. Annex C provides the tables and statistical theory that underlie the construction of modified box plots in outlier detection. Annex F provides a structured guide and flow chart to the procedures recommended in this part of ISO 16269.

Statistical interpretation of data —

Part 4: Detection and treatment of outliers

1 Scope

This part of ISO 16269 provides detailed descriptions of sound statistical testing procedures and graphical data analysis methods for detecting outliers in data obtained from measurement processes. It recommends sound robust estimation and testing procedures to accommodate the presence of outliers.

This part of ISO 16269 is primarily designed for the detection and accommodation of outlier(s) from univariate data. Some guidance is provided for multivariate and regression data.

2 Terms and definitions

For the purposes of this document, the following terms and definitions apply.

2.1
sample
data set
subset of a population made up of one or more sampling units

NOTE 1 The sampling units could be items, numerical values or even abstract entities depending on the population of interest.

NOTE 2 A sample from a **normal** (2.22), a **gamma** (2.23), an **exponential** (2.24), a **Weibull** (2.25), a **lognormal** (2.26) or a **type I extreme value** (2.27) population will often be referred to as a normal, a gamma, an exponential, a Weibull, a lognormal or a type I extreme value sample, respectively.

2.2
outlier
member of a small subset of observations that appears to be inconsistent with the remainder of a given **sample** (2.1)

NOTE 1 The classification of an observation or a subset of observations as outlier(s) is relative to the chosen model for the population from which the data set originates. This or these observations are not to be considered as genuine members of the main population.

NOTE 2 An outlier may originate from a different underlying population, or be the result of incorrect recording or gross measurement error.

NOTE 3 The subset may contain one or more observations.

2.3
masking
presence of more than one **outlier** (2.2), making each outlier difficult to detect

2.4
some-outside rate
probability that one or more observations in an uncontaminated sample will be wrongly classified as **outliers** (2.2)

2.5
outlier accommodation method
method that is insensitive to the presence of **outliers** (2.2) when providing inferences about the population

2.6
resistant estimation
estimation method that provides results that change only slightly when a small portion of the data values in a **data set** (2.1) is replaced, possibly with very different data values from the original ones

2.7
robust estimation
estimation method that is insensitive to small departures from assumptions about the underlying probability model of the data

NOTE An example is an estimation method that works well for, say, a **normal distribution** (2.22), and remains reasonably good if the actual distribution is skew or heavy-tailed. Classes of such methods include the L-estimation [weighted average of **order statistics** (2.10)] and M-estimation methods (see Reference [9]).

2.8
rank
position of an observed value in an ordered set of observed values

NOTE 1 The observed values are arranged in ascending order (counting from below) or descending order (counting from above).

NOTE 2 For the purposes of this part of ISO 16269, identical observed values are ranked as if they were slightly different from one another.

2.9
depth
⟨box plot⟩ smaller of the two **ranks** (2.8) determined by counting up from the smallest value of the **sample** (2.1), or counting down from the largest value

NOTE 1 The depth may not be an integer value (see Annex C).

NOTE 2 For all summary values other than the **median** (2.11), a given depth identifies two (data) values, one below the median and the other above the median. For example, the two data values with depth 1 are the smallest value (minimum) and largest value (maximum) in the given **sample** (2.1).

2.10
order statistic
statistic determined by its ranking in a non-decreasing arrangement of random variables

[ISO 3534-1:2006, definition 1.9]

NOTE 1 Let the observed values of a random sample be $\{x_1, x_2, \ldots, x_n\}$. Reorder the observed values in non-decreasing order designated as $x_{(1)} \leqslant x_{(2)} \leqslant \ldots \leqslant x_{(k)} \leqslant \ldots \leqslant x_{(n)}$; then $x_{(k)}$ is the observed value of the kth order statistic in a sample of size n.

NOTE 2 In practical terms, obtaining the order statistics for a **sample** (2.1) amounts to sorting the data as formally described in Note 1.

2.11
median
sample median
median of a set of numbers
Q_2
$[(n + 1)/2]$th **order statistic** (2.10), if the sample size n is odd; sum of the $[n/2]$th and the $[(n/2) + 1]$th order statistics divided by 2, if the sample size n is even

[ISO 3534-1:2006, definition 1.13]

NOTE The sample median is the second quartile (Q_2).

2.12
first quartile
sample lower quartile
Q_1
for an odd number of observations, **median** (2.11) of the smallest $(n - 1)/2$ observed values; for an even number of observations, median of the smallest $n/2$ observed values

NOTE 1 There are many definitions in the literature of a sample quartile, which produce slightly different results. This definition has been chosen both for its ease of application and because it is widely used.

NOTE 2 Concepts such as hinges or **fourths** (2.19 and 2.20) are popular variants of quartiles. In some cases (see Note 3 to 2.19), the first quartile and the **lower fourth** (2.19) are identical.

2.13
third quartile
sample upper quartile
Q_3
for an odd number of observations, median of the largest $(n - 1)/2$ observed values; for an even number of observations, median of the largest $n/2$ observed values

NOTE 1 There are many definitions in the literature of a sample quartile, which produce slightly different results. This definition has been chosen both for its ease of application and because it is widely used.

NOTE 2 Concepts such as hinges or **fourths** (2.19 and 2.20) are popular variants of quartiles. In some cases (see Note 3 to 2.20), the third quartile and the upper fourth (2.20) are identical.

2.14
interquartile range
IQR
difference between the **third quartile** (2.13) and the **first quartile** (2.12)

NOTE 1 This is one of the widely used statistics to describe the spread of a data set.

NOTE 2 The difference between the **upper fourth** (2.20) and the **lower fourth** (2.19) is called the fourth-spread and is sometimes used instead of the interquartile range.

2.15
five-number summary
the minimum, **first quartile** (2.12), **median** (2.11), **third quartile** (2.13), and maximum

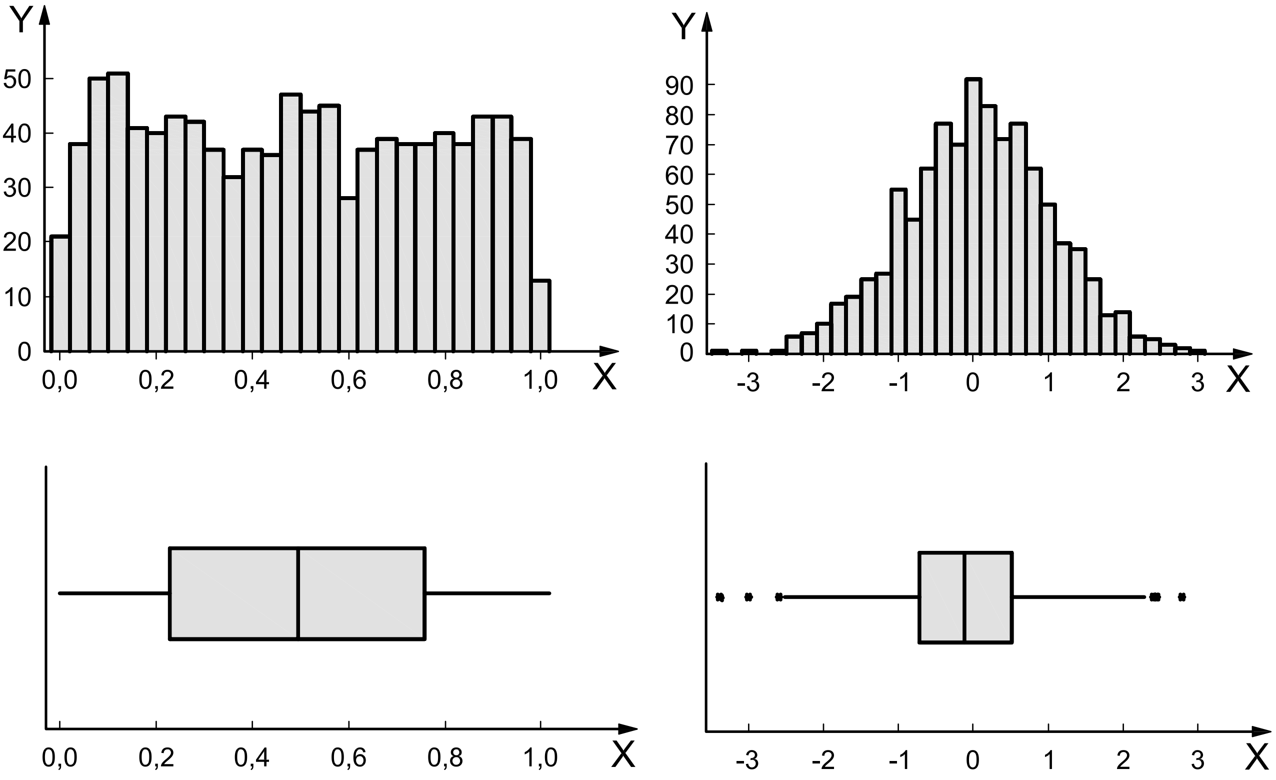
NOTE The five-number summary provides numerical information about the location, spread and range.

2.16
box plot
horizontal or vertical graphical representation of the **five-number summary** (2.15).

NOTE 1 For the horizontal version, the **first quartile** (2.12) and the **third quartile** (2.13) are plotted as the left and right sides, respectively, of a box, the **median** (2.11) is plotted as a vertical line across the box, the whiskers stretching downwards from the first quartile to the smallest value at or above the **lower fence** (2.17) and upwards from the third quartile to the largest value at or below the **upper fence** (2.18), and value(s) beyond the lower and upper fences are marked separately as **outlier(s)** (2.2). For the vertical version, the first and third quartiles are plotted as the bottom and the top, respectively, of a box, the median is plotted as a horizontal line across the box, the whiskers stretching downwards from the first quartile to the smallest value at or above the lower fence and upwards from the third quartile to the largest value at or below the upper fence and value(s) beyond the lower and upper fences are marked separately as outlier(s).

NOTE 2 The box width and whisker length of a box plot provide graphical information about the location, spread, skewness, tail lengths, and outlier(s) of a sample. Comparisons between box plots and the density function of a) uniform, b) bell-shaped, c) right-skewed, and d) left-skewed distributions are given in the diagrams in Figure 1. In each distribution, a histogram is shown above the boxplot.

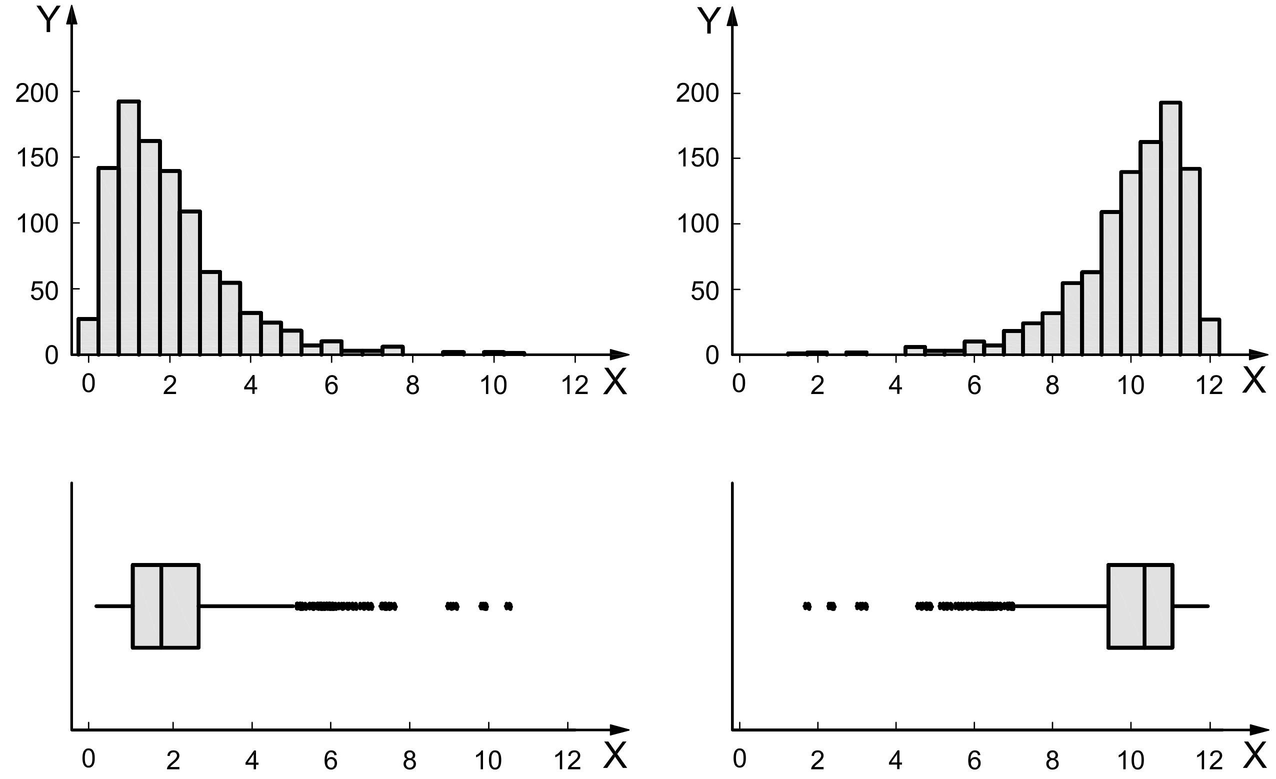
NOTE 3 A box plot constructed with its **lower fence** (2.17) and **upper fence** (2.18) evaluated by taking k to be a value based on the sample size n and the knowledge of the underlying distribution of the sample data is called a modified box plot (see example, Figure 2). The construction of a modified box plot is given in 4.4.



a) Uniform distribution

b) Bell-shaped distribution

Figure 1 (*continued*)



c) Right-skewed distribution

d) Left-skewed distribution

Key

X data values

Y frequency

In each distribution, a histogram is shown above the box plot.

Figure 1 — Box plots and histograms for a) uniform, b) bell-shaped, c) right-skewed, and d) left-skewed distributions

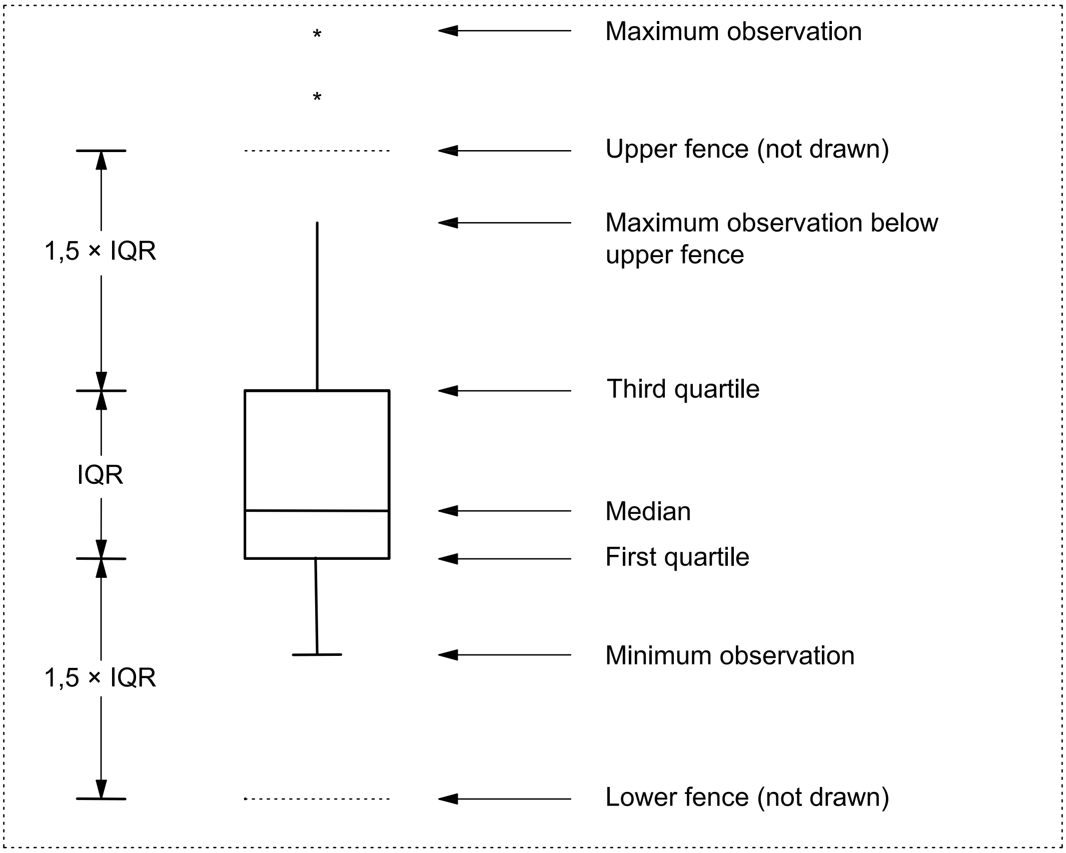


Figure 2 — Modified box plot with lower and upper fences

2.17
lower fence
lower outlier cut-off
lower adjacent value
value in a **box plot** (2.16) situated k times the **interquartile range** (2.14) below the **first quartile** (2.12), with a predetermined value of k

NOTE In proprietary statistical packages, the lower fence is usually taken to be $Q_1 - k\,(Q_3 - Q_1)$ with k taken to be either 1,5 or 3,0. Classically, this fence is called the “inner lower fence” when k is 1,5, and “outer lower fence” when k is 3,0.

2.18
upper fence
upper outlier cut-off
upper adjacent value
value in a box plot situated k times the **interquartile range** (2.14) above the **third quartile** (2.13), with a predetermined value of k

NOTE In proprietary statistical packages, the upper fence is usually taken to be $Q_3 + k\,(Q_3 - Q_1)$, with k taken to be either 1,5 or 3,0. Classically, this fence is called the “inner upper fence” when k is 1,5, and the “outer upper fence” when k is 3,0.

2.19
lower fourth

$x_{\mathrm{L}:n}$

for a set $x_{(1)} \leqslant x_{(2)} \leqslant \ldots \leqslant x_{(n)}$ of observed values, the quantity 0,5 $[x_{(i)} + x_{(i+1)}]$ when $f = 0$ or $x_{(i+1)}$ when $f > 0$, where i is the integral part of $n/4$ and f is the fractional part of $n/4$

NOTE 1 This definition of a lower fourth is used to determine the recommended values of k_{L} and k_{U} given in Annex C and is the default or optional setting in some widely used statistical packages.

NOTE 2 The lower fourth and the **upper fourth** (2.20) as a pair are sometimes called hinges.

NOTE 3 The lower fourth is sometimes referred to as the **first quartile** (2.12).

NOTE 4 When $f = 0$, 0,5 or 0,75, the lower fourth is identical to the first quartile. For example:

Sample size n	i = **integral part of** n**/4**	f = **fractional part of** n**/4**	**First quartile**	**Lower fourth**
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
upper fourth

$x_{\mathrm{U}:n}$

for a set $x_{(1)} \leqslant x_{(2)} \leqslant \ldots \leqslant x_{(n)}$ of observed values, the quantity 0,5 $[x_{(n-i)} + x_{(n-i+1)}]$ when $f = 0$ or $x_{(n-i)}$ when $f > 0$, where i is the integral part of $n/4$ and f is the fractional part of $n/4$

NOTE 1 This definition of an upper fourth is used to determine the recommended values of k_{L} and k_{U} given in Annex C and is the default or optional setting in some widely used statistical packages.

NOTE 2 The **lower fourth** (2.19) and the upper fourth as a pair are sometimes called hinges.

NOTE 3 The upper fourth is sometimes referred to as the **third quartile** (2.13).

NOTE 4 When $f = 0$, 0,5 or 0,75, the upper fourth is identical to the third quartile. For example:

Sample size n	i = **integral part of** n**/4**	f = **fractional part of** n**/4**	**Third quartile**	**Upper fourth**
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
Type I error

rejection of the null hypothesis when in fact it is true

[ISO 3534-1:2006, definition 1.46]

NOTE 1 A Type I error is an incorrect decision. Hence, it is desired to keep the probability of making such an incorrect decision as small as possible.

NOTE 2 It is possible in some situations (for example, testing the binomial parameter p) that a pre-specified significance level such as 0,05 is not attainable due to discreteness in outcomes.

2.22
normal distribution

Gaussian distribution

continuous distribution having the probability density function

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

where $-\infty < x < \infty$ and with parameters $-\infty < \mu < \infty$ and $\sigma > 0$

[ISO 3534-1:2006, definition 2.50]

NOTE 1 The location parameter μ is the mean and the scale parameter σ is the standard deviation of the normal distribution.

NOTE 2 A normal sample is a random **sample** (2.1) taken from a population that follows a normal distribution.

2.23
gamma distribution

continuous distribution having the probability density function

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

where $x > 0$ and parameters $\alpha > 0$, $\beta > 0$

[ISO 3534-1:2006, definition 2.56]

NOTE 1 The gamma distribution is used in reliability applications for modelling time to failure. It includes the **exponential distribution** (2.24) as a special case as well as other cases with failure rates that increase with age.

NOTE 2 The mean of the gamma distribution is $\alpha\beta$. The variance of the gamma distribution is $\alpha\beta^2$.

NOTE 3 A gamma sample is a random **sample** (2.1) taken from a population that follows a gamma distribution.

2.24
exponential distribution

continuous distribution having the probability density function

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

where $x > 0$ and with parameter $\beta > 0$

[ISO 3534-1:2006, definition 2.58]

NOTE 1 The exponential distribution provides a baseline in reliability applications, corresponding to the case of “lack of ageing” or memory-less property.

NOTE 2 The mean of the exponential distribution is β. The variance of the exponential distribution is β^2.

NOTE 3 An exponential sample is a random **sample** (2.1) taken from a population that follows an exponential distribution.

2.25
Weibull distribution
type III extreme-value distribution
continuous distribution having the distribution function

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

where $x > \theta$ with parameters $-\infty < \theta < \infty$, $\beta > 0$, $\kappa > 0$

[ISO 3534-1:2006, definition 2.63]

NOTE 1 In addition to serving as one of the three possible limiting distributions of extreme order statistics, the Weibull distribution occupies a prominent place in diverse applications, particularly reliability and engineering. The Weibull distribution has been demonstrated to provide usable fits to a variety of data sets.

NOTE 2 The parameter θ is a location or threshold parameter in the sense that it is the minimum value that a Weibull variate can achieve. The parameter β is a scale parameter (related to the standard deviation of a Weibull variate). The parameter κ is a shape parameter.

NOTE 3 A Weibull sample is a random **sample** (2.1) taken from a population that follows a Weibull distribution.

2.26
lognormal distribution
continuous distribution having the probability density function

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

where $x > 0$ and with parameters $-\infty < \mu < \infty$ and $\sigma > 0$

[ISO 3534-1:2006, definition 2.52]

2.27
type I extreme-value distribution
Gumbel distribution
continuous distribution having the distribution function

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

where $-\infty < x < \infty$ and with parameters $-\infty < \mu < \infty$ and $\sigma > 0$

NOTE Extreme-value distributions provide appropriate reference distributions for the extreme **order statistics** (2.10) $x_{(1)}$ and $x_{(n)}$.

[ISO 3534-1:2006, definition 2.61]

3 Symbols

The symbols and abbreviated terms used in this part of ISO 16269 are as follows:

GESD generalized extreme studentized deviate

G_{E} Greenwood's statistic

$g_{\mathrm{E};n}$ critical value of the Greenwood's test statistic for sample size n

I_l reduced sample of size $n - l$ after removing the most extreme observation $x^{(0)}$ in the original sample I_0 of size $n,$ removing the most extreme observation $x^{(1)}$ in the reduced sample I_1 of size $n - 1,\ldots.,$ and removing the most extreme observation $x^{(l-1)}$ in the reduced sample I_{l-1} of size $n - l + 1$

$F_{p;\nu_1,\nu_2}$ pth percentile of a F-distribution with ν_1 and ν_2 degrees of freedom

λ_l critical value of the GESD test in testing whether the value $x^{(l)}$ is an outlier

L_{F} lower fence of a modified box plot

U_{F} upper fence of a modified box plot

M or Q_2 sample median

M_{ad} median absolute deviation about the median

Q_1 first quartile

Q_3 third quartile

R_l GESD test statistic for testing whether the value $x^{(l)}$ is an outlier

$s(I_l)$ standard deviation of the reduced sample I_l

T_{M} total median

T_n biweight location estimate from a sample of size n

$T_n^{(i)}$ estimate of T_n at the ith iteration based on a sample of size n

$t_{p;\,\nu}$ pth percentile of a t-distribution with ν degrees of freedom

$\chi^2_{p;\nu}$ pth percentile of a chi-square distribution with ν degrees of freedom

$x_{(i)}$ ith observation in the ordered data set

$x^{(l)}$ most extreme value in the reduced sample I_l

$\bar{x}(I_l)$ mean of the reduced sample I_l

$\bar{x}_{\mathrm{T}}(\alpha)$ α-trimmed mean

$x_{\mathrm{L}:n}$ lower fourth of a box plot for a sample of size n

$x_{\mathrm{U}:n}$ upper fourth of a box plot for a sample of size n

4 Outliers in univariate data

4.1 General

4.1.1 What is an outlier?

In the simplest case, an outlier is an observation that appears to be inconsistent with the rest of a given data set. In general, there may be more than one outlier at one or both ends of the data set. The problem is to determine whether or not apparently inconsistent observations are in fact outliers. This determination is performed by means of a pre-specified significance test with respect to a presumed underlying distribution. Observations that lead to a significant result are deemed to be outliers with respect to that distribution.

The importance of using the correct underlying distribution in an outlier test cannot be over-stressed. Often in practice, an underlying normal distribution is assumed when the data arise from a different distribution. Such an erroneous assumption can lead to observations being incorrectly classified as outliers.

4.1.2 What are the causes of outliers?

Outlying observations or outliers typically are attributable to one or more of the following causes (see Reference [1] for more detail and perspective):

a) *Measurement or recording error*. The measurements are imprecisely generated, incorrectly observed, incorrectly recorded, or incorrectly entered into the database.

b) *Contamination*. The data arise from two or more distributions, i.e. the basic one and one or more contaminating distributions. If the contaminating distributions have significantly different means, larger standard deviations and/or heavier tails than the basic distribution, then there is a possibility that extreme observations coming from the contaminating distributions may appear as outliers in the basic distribution.

NOTE 1 The cause of contamination can be due to sampling error where a small portion of sample data is inadvertently regarded as having been drawn from a different population than the rest of sample data; or intentional under- or over-reporting of experiments or sampling surveys.

c) *Incorrect distributional assumption*. The data set is regarded as drawn from a particular distribution, but it should have been regarded as drawn from another distribution.

EXAMPLE The data set is regarded as drawn from a normal distribution, but it should have been regarded as drawn from a highly skewed distribution (e.g. exponential or lognormal) or a symmetric but heavier-tailed distribution (e.g. a t-distribution). Therefore, observations that deviate far from the central location can be incorrectly labelled as outliers even though they are valid observations with respect to a highly skewed or heavy-tailed distribution.

d) *Rare observations*. Highly improbable observations might occur on rare occasions, in samples regarded as drawn from an assumed probability distribution. These extreme observations are usually incorrectly labelled as outliers due to their rare occurrence, but they are not truly outliers.

NOTE 2 The occurrence of rare observations when the underlying distribution is symmetric but heavy-tailed may lead to incorrect distributional assumptions.

4.1.3 Why should outliers be detected?

Outliers are not necessarily *bad* or *erroneous*. They can be taken as an indication of the existence of rare phenomena that could be a reason for further investigation. For example, if an outlier is caused exclusively by a particular industrial treatment, important discoveries may be made by investigating the cause.

Many statistical techniques and summary statistics are sensitive to the presence of outliers. For example, the sample mean and sample standard deviation are easily influenced by the presence of even a single outlier that could subsequently lead to invalid inferences.

The study of the nature and frequency of outliers in a particular problem can lead to appropriate modifications of the distributional or model assumptions regarding the data set, and also lead to appropriate choices of robust methods that can accommodate the presence of possible outliers in subsequent data analyses and thus result in improved inferences (see Clause 6).

4.2 Data screening

Data screening can begin with a simple visual inspection of the given data set. Simple data plots, such as dot plot, scatter diagram, histogram, stem-and-leaf plot, probability plot, box plot, time series plot or arranging data in non-decreasing order of magnitude, can reveal unanticipated sources of variability and extreme/outlying data points. For example, a bimodal distribution of a data set revealed by the histogram or stem-and-leaf plot might be evidence of a contaminated sample or mixture of data regarded as drawn from two different populations. Probability plots and box plots are recommended for identifying extreme/outlying data points. These possible outliers can then be further investigated using the methods given in 4.3 or 4.4.

A probability plot not only provides a graphical test of whether the observations, or the majority of the observations, can be regarded as following an assumed distribution; it also reveals outlying observations in the data set. Data points that deviate markedly from a straight line fitted by eye to the points on a probability plot can be considered as possible outliers. Probability plot facilities for a wide range of distributions are available in proprietary software.

The box plot is one of the most popular graphical tools for exploring data. It is useful for displaying the central location, spread and shape of the distribution of a data set. The lower and upper fences of the box plot are defined as

$$\begin{aligned}\text{lower fence} &= Q_1 - k\,(Q_3 - Q_1)\\ \text{upper fence} &= Q_3 + k\,(Q_3 - Q_1)\end{aligned} \qquad (1)$$

where Q_1 and Q_3 are the first and third quartiles of the data set and k is a constant value.

Tukey[2] labelled data values that lie outside the lower and upper fences with $k = 1{,}5$ as suspected (possible) outliers, and those that lie outside the fences with $k = 3{,}0$ as extreme outliers.

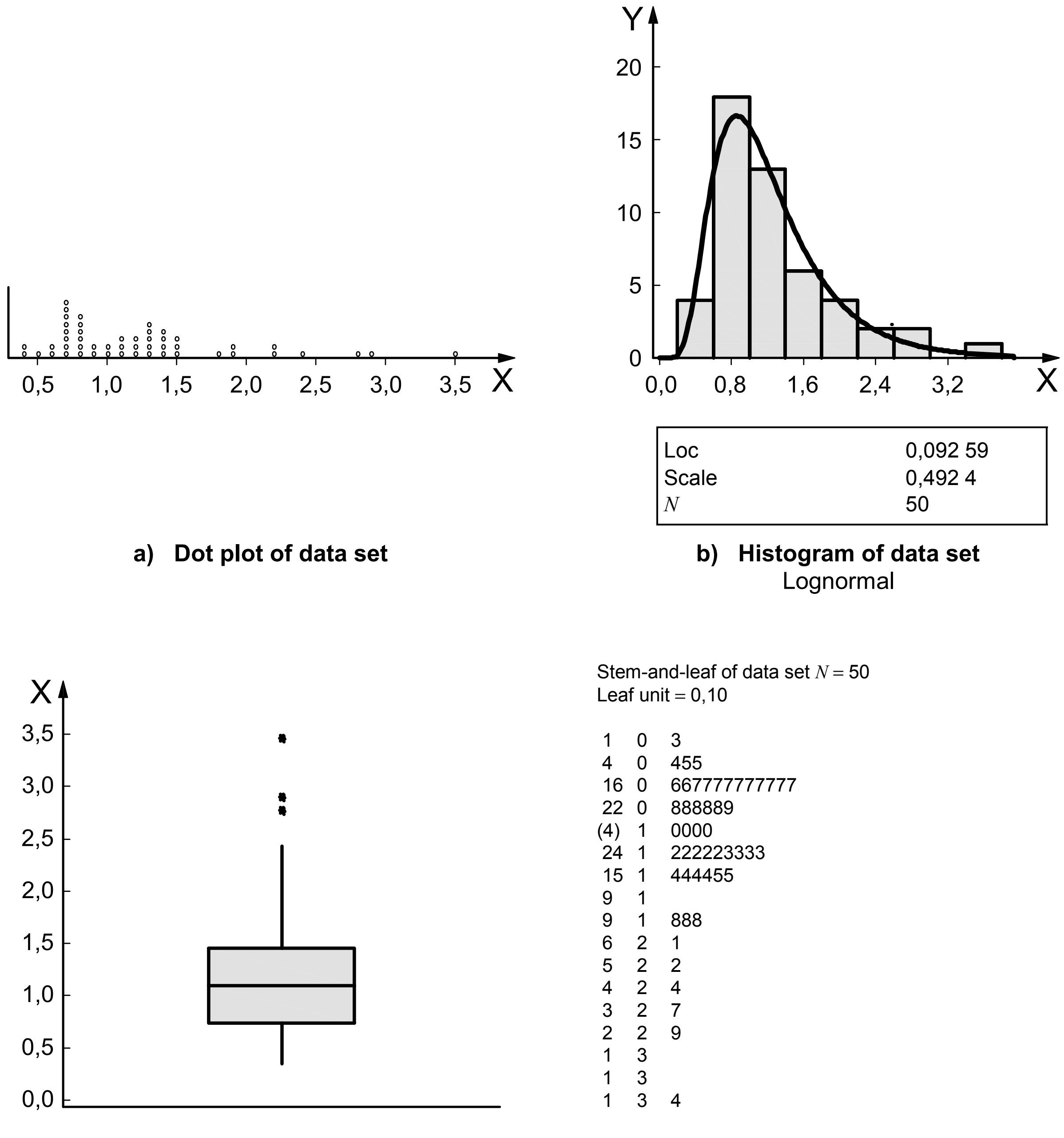
NOTE 1 Probability plotting paper for the normal, exponential, lognormal and Weibull distributions may be obtained at the time of publication from http://www.weibull.com/GPaper/index.htm.

NOTE 2 The type of probability plot should depend on the distributional assumption of the population. For example, the exponential probability plot should be used if it is assumed, or there is *a priori* knowledge, that the data set can be regarded as drawn from an exponential population.

NOTE 3 A large number of observations may incorrectly be identified as potential outliers by the box plot with its lower and upper fences defined in Equation (1) when the data set can be regarded as sampled from skewed distributions. The recommended modified box plot that is able to handle this problem is given in 4.4.

EXAMPLE The dot plot, histogram, box plot and stem-and-leaf plot of the following data values are plotted in Figures 3 a), 3 b), 3 c) and 3 d), respectively.

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



Loc	0,092 59
Scale	0,492 4
N	50

a) Dot plot of data set

b) Histogram of data set
Lognormal

Stem-and-leaf of data set $N = 50$
Leaf unit = 0,10

```
  1  0  3
  4  0  455
 16  0  667777777777
 22  0  888889
(4)  1  0000
 24  1  222223333
 15  1  444455
  9  1
  9  1  888
  6  2  1
  5  2  2
  4  2  4
  3  2  7
  2  2  9
  1  3
  1  3
  1  3  4
```

c) Box plot of data set

d) Stem-and-leaf display of data set

Key

X data set

Y frequency

Figure 3 — Plots of the data set

These plots reveal that the given data set has a longer right tail than left tail. Figures 3 a), 3 b) and 3 d) indicate that its largest value (3,463) appears to be a potential outlier, whereas the box plot in Figure 3 c) classifies the three largest values that fall above the upper fence as outliers. The first column of the stem-and-leaf display in Figure 3 d) is called the depth, the second column contains the stems, and the third column contains the leaves. The rows of the depth column give the cumulative count of leaves from the top and from the bottom except for the row that contains the median in parentheses. The leaf unit indicates the position of decimal points. Leaf unit = 0,1 means that the decimal point goes before the leaf, thus the first number in the display is 0,3, the second and third numbers are 0,4 and 0,5, respectively. (This example is considered further in 4.3.5.)

4.3 Tests for outliers

4.3.1 General

There are a large number of outlier tests (see Reference [1]). ISO 5725-2[3] provides the Grubbs and Cochran tests to identify outlying laboratories that yield unexplained abnormal test results. The Grubbs test is applicable to individual observations or to the means of sets of data taken from normal distributions, and it can only be used to detect up to the two largest and/or smallest observations as outliers in the data set. The testing procedure given in 4.3.2 is more general, being capable of detecting multiple outliers from individual observations or from the means of sets of data taken from a normal distribution. The procedures given in 4.3.3 and 4.3.4 are capable of detecting multiple outliers for data taken from an exponential, type I extreme-value, Weibull or gamma distribution. The procedure given in 4.3.5 should be used to detect outliers in samples regarded as taken from populations with unknown distribution. A test procedure that detects outliers from a given set of variances evaluated from sets of samples is given in 4.3.6.

4.3.2 Sample from a normal distribution

One or more outliers on either side of a normal data set can be detected by using a procedure known as the generalized extreme studentized deviate (GESD) many-outlier procedure (see Reference [4]). The GESD procedure is able to control the Type I error of detecting more than l outliers at a significance level α when there are l outliers present in the data set ($1 \leqslant l < m$), where m is a prescribed maximum number of outliers.

Before adopting this outlier detection method, it should be verified that the majority of the sample data approximately follow the normal distribution. The graphical normal probability plot of ISO 5479[18] can be used to test the validity of the normality assumption.

Steps to follow when using the GESD many-outlier procedure

Step 1. Plot the given sample data $x_1, x_2, \ldots, x_n$ on normal probability paper. Count the number of points that appear to deviate significantly from a straight line that fits the remaining data points. This is the suspected number of outliers.

Step 2. Select a significance level α and prescribe the number of outliers m to be larger than or equal to the suspected number of outliers from step 1. Start the following steps with $l = 0$.

Step 3. Compute the test statistic

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

where

I_0 denotes the original sample data set;

I_l denotes the reduced sample of size $n - l$ obtained by deleting the point $x^{(l-1)}$ in I_{l-1} that yields the value R_{l-1};

$\bar{x}(I_l)$ is the sample mean of the sample I_l;

$s(I_l)$ is the standard deviation of the sample I_l.

NOTE 1 For the case when $l = 0$: $\bar{x}(I_0)$ and $s(I_0)$ are the sample mean and sample standard deviation obtained from the original sample $I_0 = \{x_1, x_2, \ldots, x_n\}$ of size n, when the largest value among the values $x_1 - \bar{x}(I_0)$, $x_2 - \bar{x}(I_0)$, $\cdots$, $x_n - \bar{x}(I_0)$ is $x_2 - \bar{x}(I_0)$ (say), we then have $R_0 = [x_2 - \bar{x}(I_0)]/s(I_0)$ and $x^{(0)} = x_2$. Subsequently, $I_1 = I_0 \setminus \{x^{(0)}\} = \{x_1, x_3, \ldots, x_n\}$ is the reduced sample of size n–1 obtained by deleting the data value $x^{(0)}$, i.e. x_2, in I_0.

Step 4. Compute the critical value

$$\lambda_l = \frac{(n-l-1)t_{p;n-l-2}}{\sqrt{(n-l-2+t_{p;n-l-2}^2)(n-l)}} \tag{3}$$

where $p = (1 - \alpha/2)^{1/(n-l)}$ and $t_{p;\nu}$ represents the 100pth percentile of a t-distribution with ν degrees of freedom. Note that if one has the additional information that the outliers occur only on either the upper or the lower extreme, substitute α for $\alpha/2$ in the equation.

Step 5. Set $l = l + 1$.

Step 6. Repeat step 2 to step 4 until $l = m$.

Step 7. If $R_l \leqslant \lambda_l$ for all $l = 0, 1, 2, \ldots, m$, then no outliers are declared. Otherwise, the n_{out} most extreme observations $x^{(0)}, x^{(1)}, \ldots, x^{(n_{\text{out}}-1)}$ in the successively reduced samples are declared as outliers when $n_{\text{out}} = 1 + \max\limits_{0 \leqslant l \leqslant m} \{l : R_l > \lambda_l\}$.

A computer algorithm that describes the necessary steps in implementing the GESD many-outlier procedure is given in Annex A.

NOTE 2 The GESD test is equivalent to the Grubbs test when it is used to test whether the largest or the smallest outlying observation is an outlier. The critical values of the Grubbs test are given in Table 5 of ISO 5725-2:1994[3], and can also be approximated from λ_l of step 4 by taking $l = 0$.

NOTE 3 In practice, the number of outliers m envisaged in the sample should be small. If many outlying observations are expected in the sample, then it ceases to be an outlier detection problem and different approaches are needed. However, m should not be too small, otherwise there is a possibility of a masking effect.

EXAMPLE Consider a data set of 20 observations:

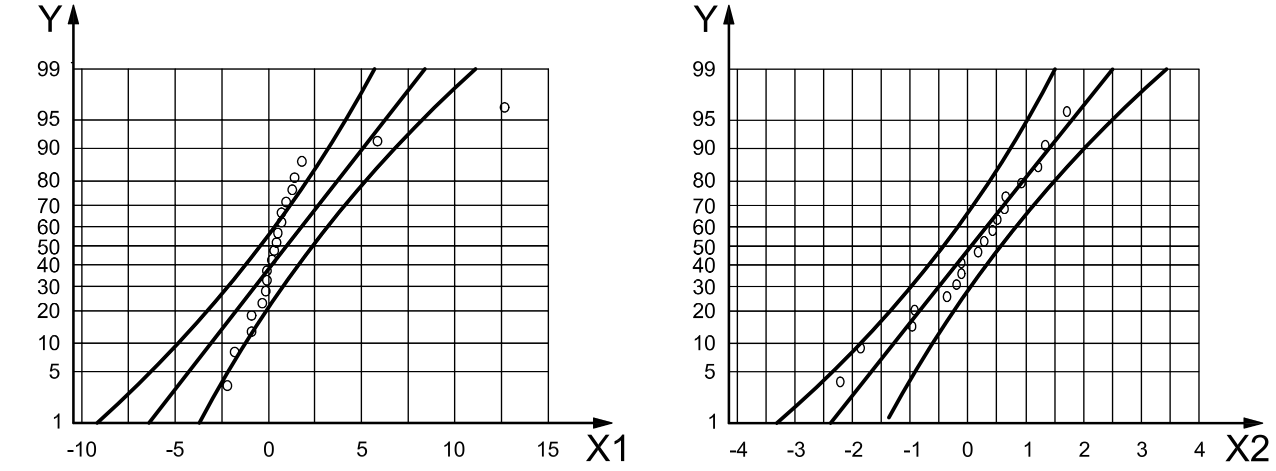
−2,21	−1,84	−0,95	−0,91	−0,36	−0,19	−0,11	−0,10	0,18	0,30
0,43	0,51	0,64	0,67	0,93	1,22	1,35	1,73	5,80	12,6

where the latter two observations were originally 0,58 and 1,26, but the decimal commas were entered at the wrong place. In detecting outliers using the GESD procedure, we shall first verify that the given observations are taken from a normal distribution. The data points of the normal probability plot given in Figure 4 a) appear to be scattered around a straight line, with the exception of the two largest values which distinctly depart from the straight line. This plot reveals that the data set, with the exception of the two extreme data values, can be assumed to come from a normal distribution. This assumption is confirmed in Figure 4 b) in which the data values, without the two extreme values, all plot inside the 95 % confidence band of the normal probability plot. Accordingly, we can then select the number of outliers to be $m = 2$ in step 2. The GESD test statistics R_l and its respective critical value λ_l for $l = 0, 1, 2$ with significance level $\alpha = 0{,}05$ are given in the table below.

l	0	1	2
R_l	3,655 9	3,263 4	2,176 1
λ_l	2,705 8	2,678 5	2,699 2
$x^{(l)}$	12,60	5,80	−2,21

As $R_0 = 3{,}6559 > \lambda_0 = 2{,}705\,8$, $R_1 = 3{,}2634 > \lambda_1 = 2{,}678\,5$ and $R_2 = 2{,}1761 \leqslant \lambda_2 = 2{,}699\,2$, we have $\max\limits_{0 \leqslant l \leqslant 2} \{l : R_l > \lambda_l\} = 1$ and $n_{\text{out}} = 1 + \max\limits_{0 \leqslant l \leqslant 2} \{l : R_l > \lambda_l\} = 2$. Thus, we declare the two most extreme values $x^{(0)} = 12{,}60$ and $x^{(1)} = 5{,}80$ as outliers.

NOTE 4 In this and in the following examples, the units for the observations are omitted because they are not relevant for the graphical plots and tests in this part of ISO 16269.



Mean	0,984 5
Standard deviation	3,177
N	20
AD (Anderson-Darling)	2,474
p-value	< 0,005

a) Probability plot of original data set
Normal – 95 % CI

Mean	0,071 67
Standard deviation	1,049
N	18
AD (Anderson-Darling)	0,299
p-value	0,547

b) Probability plot of reduced data set
Normal – 95 % CI

Key

X1 original data set

X2 reduced data set

Y percent

Figure 4 — Probability plots

4.3.3 Sample from an exponential distribution

4.3.3.1 General

Greenwood's test (see 4.3.3.2) is the recommended test for outliers in samples regarded as having been drawn from an exponential distribution. However, this test only indicates the presence of outliers but cannot identify the individual outliers and the number of outliers in the sample. Two alternative consecutive tests that can identify up to m possible upper or m possible lower outliers in exponential samples are given in 4.3.3.3 and 4.3.3.4, respectively.

4.3.3.2 Greenwood's test for the existence of outliers

This is a powerful test for outliers in samples regarded as having been drawn from an exponential distribution with probability density function $f(x) = \lambda^{-1} \exp [-(x - a)/\lambda]$, $x \geqslant a$, where λ is the scale parameter and a is the location or threshold parameter. For a given exponential sample $x_1, x_2, \ldots, x_n$ of size n regarded as drawn from an exponential distribution with known parameter value a, the test statistic is given as (Reference [1]):

$$G_{\mathrm{E}} = \frac{\sum_{i=1}^{n} (x_i - a)^2}{\left(\sum_{i=1}^{n} x_i - na\right)^2} \tag{4}$$

A significantly high value of G_{E} indicates the likely presence of an unknown number of outliers that are the high extreme values in the sample; however, a significantly low value of G_{E} indicates the presence of outliers that are the low extreme values or the combination of low and high extreme values. The lower and upper 2,5 % and 1 % critical values $g_{\mathrm{E};n}$ of G_{E} are given in Table B.1 for selected sample sizes n. For the case when the origin a is unknown, it is estimated by the value of the smallest observation $x_{(1)}$, and the critical value of G_{E} is then $g_{\mathrm{E};n-1}$.

4.3.3.3 Consecutive tests for m possible upper outliers

The test statistics that can be used to declare up to the m largest observations as outliers in an exponential sample of size n with known location parameter a are given as (Reference [5]):

$$S_j^{\mathrm{U}} = \left(x_{(n-j+1)} - a\right) \Big/ \sum\nolimits_{i=1}^{n-j+1}\left(x_{(i)} - a\right), \quad j = 1, 2, \ldots, m \tag{5}$$

where $x_{(1)} \leqslant x_{(2)} \leqslant \ldots \leqslant x_{(n)}$ are the order statistics of the given sample. Significantly large values of S_j^{U} indicate that the high extreme values are outliers. The upper 5 % and 1 % critical values $s_{j:n}^{\mathrm{U}}$ of S_j^{U} are given in Table B.2 for selected values of n with $m = 2$, 3 and 4. If $S_m^{\mathrm{U}} > s_{m;n}^{\mathrm{U}}$, declare the m largest observations as outliers; if $S_j^{\mathrm{U}} \leqslant s_{j;n}^{\mathrm{U}}$ for $j = m, m - 1, \ldots, l + 1$, but $S_l^{\mathrm{U}} > s_{l;n}^{\mathrm{U}}$, declare the l largest observations as outliers; if $S_j^{\mathrm{U}} \leqslant s_{j;n}^{\mathrm{U}}$ for all $j = 1, 2, \ldots, m$, declare there to be no outliers in the sample.

For the case when the parameter a is unknown, it can be estimated by the value of the smallest observation $x_{(1)}$ and the critical value of S_j^{U} is then $s_{j;n-1}^{\mathrm{U}}$.

4.3.3.4 Consecutive tests of m possible lower outliers

The test statistics that can be used to declare up to the m smallest observations as outliers in an exponential sample of size n with known location parameter a are given as (Reference [5]):

$$S_j^{\mathrm{L}} = \left(x_{(j+1)} - a\right) \Big/ \sum\nolimits_{i=1}^{j+1}\left(x_{(i)} - a\right), \quad j = 1, 2, \ldots, m \tag{6}$$

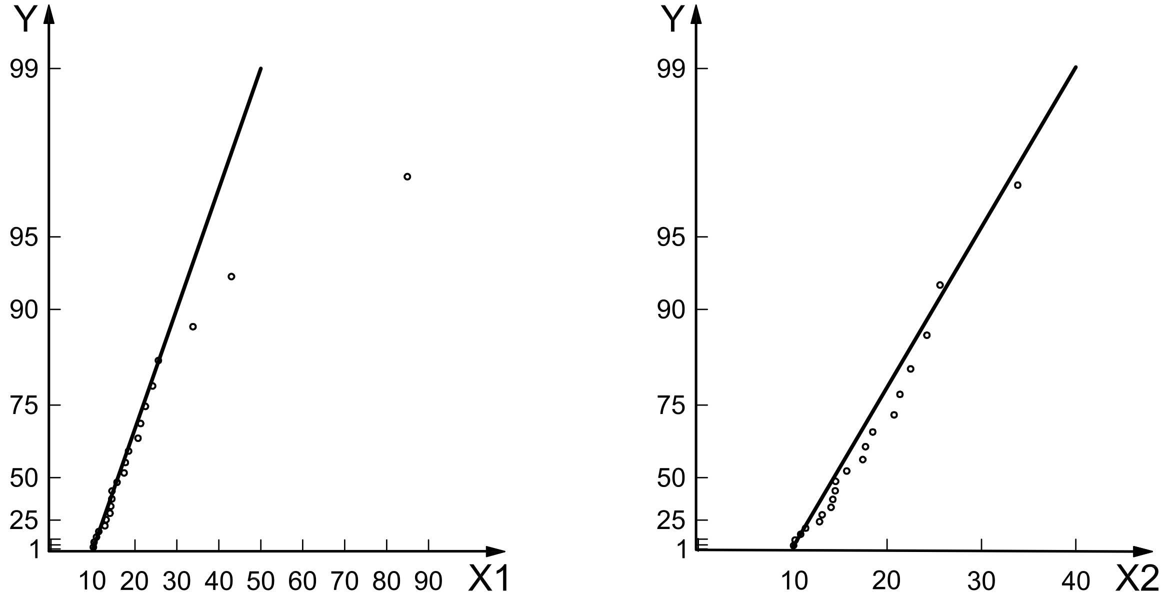
where $x_{(1)} \leqslant x_{(2)} \leqslant \ldots \leqslant x_{(n)}$ are the order statistics of the given sample. Significantly high values of S_j^{L} indicate that the low extreme values are outliers. The lower and upper 5 % and 1 % critical values $s_{j;n}^{\mathrm{L}}$ of S_j^{L} are given in Table B.3 for selected values of n with $m = 2$, 3 and 4. If $S_m^{\mathrm{L}} > s_{m;n}^{\mathrm{L}}$, declare the m smallest observations as outliers; if $S_j^{\mathrm{L}} \leqslant s_{j;n}^{\mathrm{L}}$ for $j = m, m - 1, \ldots, l + 1$, but $S_l^{\mathrm{L}} > s_{l;n}^{\mathrm{L}}$, declare the l smallest observations as outliers; if $S_j^{\mathrm{L}} \leqslant s_{j;n}^{\mathrm{L}}$ for all $j = 1, 2, \ldots, m$, declare there to be no outliers in the sample.

This test can only be used to detect outliers from exponential samples with known parameter a. For exponential samples with unknown a, the procedure discussed in 4.4 can be used to detect outliers from the sample data.

EXAMPLE Consider the following 22 observations that are arranged in ascending order:

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

In detecting outliers using the Greenwood statistic, the first step is to verify that the given observations are regarded as drawn from an exponential distribution. The data points of the exponential probability plot given in Figure 5 a) appear to be scattered around a straight line, with the exception of the largest or the two largest values. This plot reveals that the data set, with the exception of one or two extreme data values, can be assumed to come from an exponential distribution. This assumption is confirmed in Figure 5 b) in which the data values, without the two largest values, are scattered around a straight line. With an estimated location parameter $a = 10{,}10$, the Greenwood statistic is $G_{\mathrm{E}} = 8\,386{,}326/(249{,}37)^2 = 0{,}134\,86$. From Table B.1, the lower and upper 2,5 % critical values $g_{\mathrm{E};21}$ of G_{E} are 0,067 3 and 0,133 8, respectively. Thus, the calculated G_{E} value of 0,134 86 falls above the upper critical value of 0,133 8 and we conclude that one or more of the high extreme value(s) in the given data set are outliers.



a) Exponential probability plot of original data set

b) Exponential probability plot of reduced data set

Key

X1 original data set

X2 reduced data set

Y exponential probabilities

Figure 5 — Exponential probability plots

As the questionable data points are the two high extreme values, the tests of 4.3.3.3 can be used to check for up to two possible outliers in the sample. Taking $m = 2$, we have S_2^U = (43,0 – 10,1)/174,53 = 0,188 5 and S_1^U = (84,94 – 10,1)/249,37 = 0,300 1. After comparing these values with the respective critical values of $s_{2;21}^U$ = 0,231 3 and $s_{1;21}^U$ = 0,283 4 taken from Table B.2 at $\alpha = 0,05$, only the largest value (84,94) is declared as an outlier at the 5 % significance level.

4.3.4 Samples taken from some known non-normal distributions

4.3.4.1 General

Detection of outliers in samples taken from non-normal distributions is of considerable practical importance. Outliers in exponential and gamma samples arise in the study of life testing, traffic and river flows, etc., whereas the extreme-value sample arises in the study of extremes, such as maximum wind speeds, or sporting achievements. The lognormal and Weibull distributions often arise in reliability applications. In cases when the non-normal family of distributions is known and is either the lognormal, extreme-value, Weibull or gamma distribution, the following transformations are recommended to transform the data to resemble the required distribution.

4.3.4.2 For a sample of data $x_1, x_2, \ldots, x_n$ regarded as drawn from a lognormal distribution with probability density function

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

the transformed values $\ln x_1, \ln x_2, \ldots, \ln x_n$ are a sample from a normal distribution with mean μ and variance σ^2. The test procedure of 4.3.2 and/or 4.4 can then be used to detect outliers among the transformed values.

4.3.4.3 For a sample of data $x_1, x_2, \ldots, x_n$ taken from a type 1 extreme-value distribution with distribution function

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

the transformed sample values exp($-x_1/b$), exp($-x_2/b$), …, exp($-x_n/b$) follow the exponential distribution with mean exp($-a/b$). The test procedures of 4.3.3 and/or 4.4 can then be used to detect outliers from the transformed values.

4.3.4.4 For a sample of data $x_1, x_2, \ldots, x_n$ taken from a Weibull distribution with distribution function

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

the transformed sample values $(x_1 - a)^r$, $(x_2 - a)^r$, …, $(x_n - a)^r$ follow an exponential distribution with mean b^r. The test procedures of 4.3.3 and/or 4.4 can then be used to detect outliers among the transformed values.

NOTE Exponentially distributed data x can be transformed to $\sqrt[4]{x}$ to give approximately normally distributed data[6].

4.3.4.5 For a sample of data $x_1, x_2, \ldots, x_n$ regarded as drawn from a gamma distribution with probability density function

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

the transformed values $\sqrt[3]{x_1}$, $\sqrt[3]{x_2}$, ..., $\sqrt[3]{x_n}$ approximately follow a normal distribution. The test procedure in 4.3.2 and/or 4.4 can then be used to detect outliers among the transformed values.

4.3.5 Sample taken from unknown distributions

When detecting outliers in samples regarded as drawn from populations with unknown and skewed distribution, a general method is to transform the non-normal data to resemble a normal distribution. The outlier tests of 4.3.3 for the normal samples can then be applied to the transformed normal sample. Two widely used methods are the Box-Cox transformation and the Johnson transformation. The Box-Cox family of power transformations takes the form[7]:

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

where

if $\lambda \neq 0$, the parameter m is chosen so that $x + m$ is positive, and

if $\lambda = 1$, the parameter m is set equal to zero to ensure that the original data x remain unchanged.

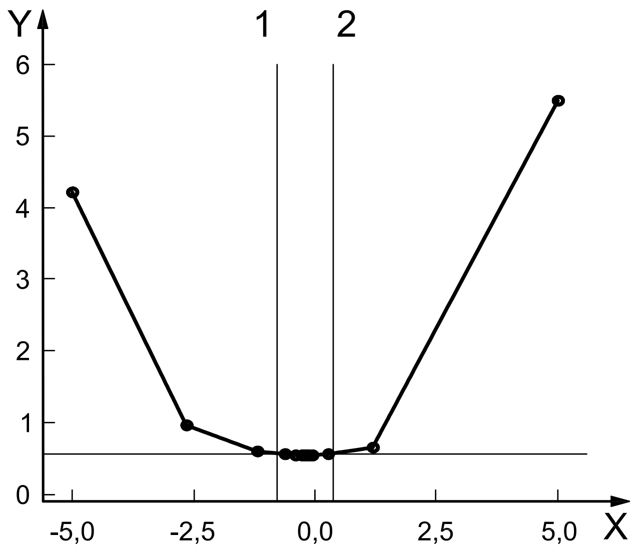
Optimal selection of the transformation parameter λ is provided automatically in some statistical packages.

The Johnson transformation transforms data to resemble a normal distribution using the families of Johnson distributions[8].

NOTE 1 The Box-Cox power transformation and Johnson transformation are available in relevant statistical software packages.

NOTE 2 The Box-Cox transformation is simple and easy to understand. However, the Johnson transformation system is able to accommodate data containing negative values.

EXAMPLE Consider the data set in 4.2 which is taken from a population with unknown distribution. As its dot plot, histogram, box plot and stem-and-leaf plot (shown in Figure 3) indicate that the data are taken from a skewed distribution, a data transformation is required to transform the data values to resemble a normal distribution. The Box-Cox plot and probability plot of the data set given in Figures 6 and 7 were obtained from a readily available statistical package. Figure 6 contains an estimate λ value of –0,19, and the rounded λ value of 0,00 which is the value used in the transformation. The figure also includes the 95 % lower confidence limit of –0,77 and upper confidence limit of 0,36, which are marked on the graph by vertical lines. In practical situations, a value of λ that corresponds to a common transformation, such as the square root ($\lambda = 0{,}5$) or the natural log ($\lambda = 0$), should be used. In this example, taking the value of λ to be zero is a reasonable choice because it falls within the 95 % confidence interval. Therefore, the natural log transformation may be preferred to the transformation defined by the best estimate of λ. The probability plots of the original and transformed data are given in Figure 7. A p-value of 0,318 given in Figure 7(b), evaluated from the Anderson-Darling test statistic, indicates that the transformed data resemble a normal distribution.



Lambda (**using** 95,0 % **confidence**)	
Estimate	–0,19
Lower confidence limit	–0,77
Upper confidence limit	0,36
Rounded value	0,00

Key

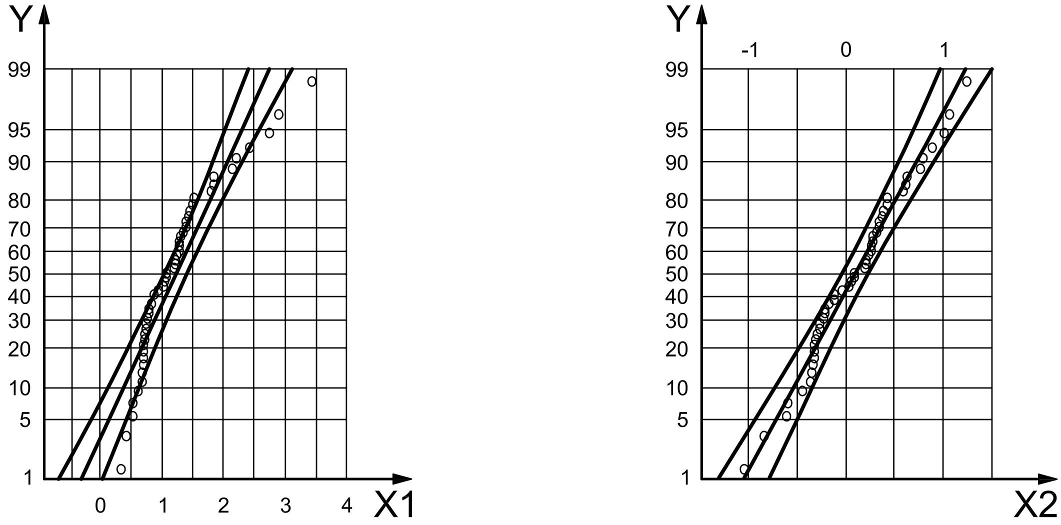
X lambda

Y standard deviation

1 lower confidence limit

2 upper confidence limit

Figure 6 — Box-Cox plot of data set



Mean	1,239
Standard deviation	0,660 1
N	50
AD (Anderson-Darling)	1,954
p-value	< 0,005

a) Probability plot of original data

Mean	0,092 59
Standard deviation	0,492 4
N	50
AD (Anderson-Darling)	0,417
p-value	0,318

b) Probability plot of transformed data

Key

X1 original data set

X2 transformed data set

Y percent

Figure 7 — Probability plots of original and transformed data

4.3.6 Cochran's test for outlying variance

It is of great importance to detect outliers from a given set of variances evaluated from sets of sample data, in particular in estimating the precision of measurement methods[3] by means of a collaborative interlaboratory experiment. Cochran's test is a widely used test for ascertaining whether the largest variance value in a given set of variances is significantly larger than the rest.

Given a set of p variances $s_1^2, ..., s_p^2$ computed from p samples each of size n, Cochran's test statistic is given by

$$C = \frac{s_{\text{max}}^2}{\sum_{i=1}^{p} s_i^2} \tag{7}$$

where s_{max}^2 is the largest variance in the set of p variances.

The 5 %, 1 % and 0,1 % critical values of the test statistic C are given in the tables of Annex E for $p = 2(1)40$ [1] sample variances evaluated from p samples each of size $n = 2(1)10$. The largest variance is then declared as an outlier if the computed value of C exceeds the critical value.

NOTE The critical values of Cochran's test given in Annex E should ideally be applied only when all the standard deviations are obtained from the same number (n) of test results.

[1] The convention 2(1)40 refers to the numbers from 2 to 40 in increments of 1.

EXAMPLE Five laboratories participated in an experiment to determine the absorption of moisture in concrete aggregates. Eight test results are obtained under repeatability conditions and according to a standardized measurement method by each of the laboratories. The set of variances obtained are

Laboratories, i	1	2	3	4	5
Variance, s_i^2	12,134	2,303	3,594	3,319	3,455

From Table E.1, the 5 % critical value of Cochran's test for $p = 5$ laboratories and $n = 8$ replicates is 0,456 4. Since the Cochran's test statistic value $C = 12{,}134/(12{,}134 + 2{,}303 + 3{,}594 + 3{,}319 + 3{,}455) = 0{,}489\,2$ exceeds the critical value, we conclude that the variance of laboratory 1 may be regarded as significantly larger than the rest.

4.4 Graphical test of outliers

The following modified box plot is recommended for detecting outliers when the population distribution of the given data set is assumed to follow a normal or exponential distribution. Unlike the hypothesis testing procedures of 4.3, this graphical test of outliers based on the box plot has no prior requirement on the knowledge of the number of outliers or in which direction the outliers are located.

The lower and upper fourths $x_{\mathrm{L}:n}$ and $x_{\mathrm{U}:n}$ are used instead of the first and third quartiles Q_1 and Q_3 in evaluating the lower fence L_F and upper fence U_F of this distribution-specific modified box plot, i.e.

$$\begin{aligned} L_\mathrm{F} &= x_{\mathrm{L}:n} - k_\mathrm{L}(x_{\mathrm{U}:n} - x_{\mathrm{L}:n}) \\ U_\mathrm{F} &= x_{\mathrm{U}:n} + k_\mathrm{U}(x_{\mathrm{U}:n} - x_{\mathrm{L}:n}) \end{aligned} \tag{8}$$

where

n is the sample size;

k_L and k_U are values that depend upon the underlying distribution of the hypothesized population and the sample size n;

$x_{\mathrm{L}:n}$ is the lower fourth of the box plot evaluated as

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

$x_{\mathrm{U}:n}$ is the upper fourth of the box plot evaluated as

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

in which $n/4 = i + f$ where i is the integral part of $n/4$ and f is the fractional part of $n/4$, and $x_{(1)} \leqslant x_{(2)} \leqslant \ldots \leqslant x_{(n)}$ are the order statistics from the sample.

NOTE 1 This definition of lower and upper fourths is used to determine the recommended values of k_L and k_U given in Annex C and is the default or optional setting in some widely used statistical packages.

Observations that fall above the upper fence or below the lower fence are labelled as possible outliers. A salient feature of this modified box plot is that its constant values k_L and k_U are determined from the requirement that for an outlier-free sample the some-outside rate per sample, i.e. the probability that one or more observations in the sample will be falsely classified as outliers, is equal to a small given value α. This modified box plot reduces to the classical box plot discussed in 4.2 when $k_\mathrm{L} = k_\mathrm{U} = 1{,}5$. The values of k_L and k_U can be determined from Equation (C.2) given in Annex C for samples taken from the normal and exponential distributions for selected values of α when $9 \leqslant n \leqslant 500$.

NOTE 2 The lower fence of a modified box plot constructed under the exponential distribution assumption may take a negative value if the given data set does not closely follow an exponential distribution.

EXAMPLE 1 From the $n = 20$ observations of the example in 4.3.2, we have $n/4 = 20/4 = 5$ which leads to $i = 5$ and $f = 0$. Thus, the lower and upper fourths of the box plot are evaluated as

$$x_{\mathrm{L}:n} = [x_{(5)} + x_{(6)}]/2 = 0{,}5\ (-0{,}36 - 0{,}19) = -0{,}275$$

and

$$x_{\mathrm{U}:n} = [x_{(15)} + x_{(16)}]/2 = 0{,}5\ (0{,}93 + 1{,}22) = 1{,}075$$

For normal samples, the lower and upper fences of the box plot with some-outside rate per sample of $\alpha = 0{,}05$ are constructed using $k_\mathrm{L} = k_\mathrm{U} = 2{,}238\ 2$ (as illustrated in Example 1 in Annex C)

$$L_\mathrm{F} = x_{\mathrm{L}:n} - k_\mathrm{L}\ (x_{\mathrm{U}:n} - x_{\mathrm{L}:n}) = -0{,}275 - 2{,}238\ 2\ (1{,}075 + 0{,}275) = -3{,}297$$

$$U_\mathrm{F} = x_{\mathrm{U}:n} + k_\mathrm{U}\ (x_{\mathrm{U}:n} - x_{\mathrm{L}:n}) = 1{,}075 + 2{,}238\ 2\ (1{,}075 + 0{,}275) = 4{,}097$$

Therefore the two large extreme values 5,80 and 12,60 that fall above the upper fence are declared to be outliers.

EXAMPLE 2 From the $n = 22$ observations of the example in 4.3.3.4, we have $n/4 = 22/4 = 5 + 1/2$, thus the lower and upper fourths of box plot are evaluated as

$$x_{\mathrm{L}:n} = x_{(6)} = 13{,}13 \text{ and } x_{\mathrm{U}:n} = x_{(17)} = 22{,}50$$

For this exponential sample, the lower and upper fences of the box plot with some-outside rate of $\alpha = 0{,}05$ are computed as

$$L_\mathrm{F} = x_{\mathrm{L}:n} - k_\mathrm{L}\ (x_{\mathrm{U}:n} - x_{\mathrm{L}:n}) = 13{,}13 - 0{,}665\ 0\ (22{,}50 - 13{,}13) = 6{,}899$$

$$U_\mathrm{F} = x_{\mathrm{U}:n} + k_\mathrm{U}\ (x_{\mathrm{U}:n} - x_{\mathrm{L}:n}) = 22{,}50 + 6{,}231\ 3\ (22{,}50 - 13{,}13) = 80{,}887$$

Thus, the extreme value 84,94 that falls above the upper fence is declared as an outlier. The values of $k_\mathrm{L} = 0{,}665\ 0$ and $k_\mathrm{U} = 6{,}231\ 3$ are obtained from Annex C, Example 2.

EXAMPLE 3 Suppose that the second largest value 43,0 of the example in 4.3.3.4 has been wrongly recorded as 4,30. As the value 4,30 falls below the lower fence $L_\mathrm{F} = 6{,}899$ of the box plot, it is then declared as an outlier. However, due to the masking effect of the extreme values 4,30 and 84,94, not only are the formal testing procedures of 4.3 incapable of detecting the value 4,30 as an outlier, but they also fail to detect the largest value 84,94 as an outlier.

5 Accommodating outliers in univariate data

5.1 Robust data analysis

Any detected outlier should be investigated for explanations. If it is caused by an error for which the cause can be found (e.g. clerical error, dilution error, measurement error, etc.), its value should be corrected or deleted if the actual value is not known. If the presence of outliers cannot be reasonably explained, then they should not be removed; they should be treated as valid observations and used in subsequent data analysis using robust procedures that are resistant to the influence of outliers. The outlier accommodation methods of 5.2 and 5.3 can reduce the influence of outlying observations on the results of data analysis without deleting them. Another alternative is to conduct analyses both with and without the outliers.

5.2 Robust estimation of location

5.2.1 General

The sample mean is the optimal estimate of centre location for normal data. However, it is not a resistant and robust estimate of centre location. A large assortment of robust estimation methods of location have been proposed in the literature. The trimmed mean given in 5.2.2 has been widely used to alleviate the distortion caused by outlying observations when estimating the centre location from samples taken from symmetrical population distributions. For samples taken from asymmetrical population distributions the location estimator described in 5.2.3 is recommended.

5.2.2 Trimmed mean

When possible outliers are detected in samples taken from symmetric population distributions, the trimmed mean is recommended for estimating the centre of the symmetric distributions.

Let $x_{(1)} \leqslant x_{(2)} \leqslant \dots \leqslant x_{(n)}$ be the order statistics from a sample of size n.

Let $r = [\alpha n]$ denote the greatest integer less than or equal to αn and $g = \alpha n - r$ be the fractional part of αn, where $0 \leqslant \alpha < 0{,}5$ is the proportion of outlying observations in the data set.

The α-trimmed mean[9], denoted by $\bar{x}_\mathrm{T}(\alpha)$, is computed by omitting the r smallest and r largest observations of the given sample, and by including the two nearest retained observations $x_{(r+1)}$ and $x_{(n-r)}$ with reduced weight $(1 - g)$, i.e.

$$\bar{x}_\mathrm{T}(\alpha) = \frac{1}{n(1-2\alpha)}\left[(1-g)(x_{(r+1)} + x_{(n-r)}) + \sum\nolimits_{i=r+2}^{n-r-1} x_{(i)}\right] \qquad (9)$$

NOTE 1 When αn is an integer, we have $g = 0$, thus the α-trimmed mean is the sample mean of the trimmed sample.

NOTE 2 The pre-specified value of α is usually taken to be less than 0,25. The classical sample mean is a 0-trimmed mean, whereas the sample median is approximately a 0,5-trimmed mean.

NOTE 3 The α-Winsorized mean is another popular alternative in which the $r = [\alpha n]$ smallest observations are each truncated to take the value $x_{(r+1)}$ and the r largest observations of a data set are each truncated to $x_{(n-r)}$, i.e. replacing the $(1 - g)$ of $\bar{x}_\mathrm{T}(\alpha)$ by the value r.

EXAMPLE For the data set of $n = 20$ observations given in 4.3.2, we compute the mean, median, 5 %, 10 %, 15 %, 18 % and 20 %-trimmed means. These values are

$$\text{Mean} = \frac{1}{20}\sum\nolimits_{i=1}^{20} x_i = \frac{1}{20}(19{,}69) = 0{,}984\ 5$$

$$\text{Median} = \frac{1}{2}\left[x_{(10)} + x_{(11)}\right] = \frac{1}{2}(0{,}30 + 0{,}43) = 0{,}365$$

$$\bar{x}_\mathrm{T}(0{,}05) = \frac{1}{20(1-2\times 0{,}05)}\sum\nolimits_{i=2}^{19} x_{(i)} = \frac{1}{18}(9{,}3) = 0{,}516\ 7$$

$$\bar{x}_\mathrm{T}(0{,}10) = \frac{1}{20(1-2\times 0{,}10)}\sum\nolimits_{i=3}^{18} x_{(i)} = \frac{1}{16}(5{,}34) = 0{,}333\ 75$$

$$\bar{x}_\mathrm{T}(0{,}15) = \frac{1}{20(1-2\times 0{,}15)}\sum\nolimits_{i=4}^{17} x_{(i)} = \frac{1}{14}(4{,}56) = 0{,}325\ 7$$

$$\bar{x}_{\mathrm{T}}(0{,}18) = \frac{1}{20(1-2\times 0{,}18)}\left[(1-0{,}6)\left(x_{(4)}+x_{(17)}\right)+\sum\nolimits_{i=5}^{16} x_{(i)}\right] = \frac{1}{12{,}8}(0{,}176+4{,}12) = 0{,}335\ 6$$

$$\bar{x}_{\mathrm{T}}(0{,}20) = \frac{1}{20(1-2\times 0{,}20)}\sum\nolimits_{i=5}^{16} x_{(i)} = \frac{1}{12}(4{,}12) = 0{,}343\ 3$$

These results suggest that the relatively large sample mean is due to the presence of the two outliers, whereas the trimmed means stabilize after 10 % to 20 % of the data have been trimmed.

5.2.3 Biweight location estimate

The biweight location estimate[9] is resistant to the presence of outliers for samples taken from asymmetrical distributions and is robust to small departures from the normality assumptions. Given a sample $x_1, x_2, \ldots, x_n$ of size n, the biweight location estimate can be obtained as

$$T_n = \frac{\sum_{|u_i|<1} x_i \left(1-u_i^2\right)^2}{\sum_{|u_i|<1} \left(1-u_i^2\right)^2} \tag{10}$$

where $u_i = \left(x_i - T_n\right)/cM_{\mathrm{ad}}$, with $c = 6{,}0$, $M_{\mathrm{ad}} = \mathrm{Median}\left(\left|x_i - M\right|, i = 1, 2, \ldots, n\right)$ and M is the sample median. The estimate of T_n needs to be computed iteratively. Letting $T_n^{(k)}$ and $u_{i,k} = \left(x_i - T_n^{(k)}\right)/cM_{\mathrm{ad}}$ be the estimate of T_n and u_i at the kth iteration, the estimate of T_n at the $(k+1)$th iteration is

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

This iterative computation should continue until the sequence of estimates converges to within a desired accuracy. For example, the iterations can be terminated if $\left|T_n^{(k+1)} - T_n^{(k)}\right| < 10^{-5}$ (say). An appropriate resistant starting value $T_n^{(0)}$ is the sample median M.

NOTE Under the normality assumption, a biweight estimator with $c = 6{,}0$ implies that observations more than about four standard deviations away from the median will be given zero weight.

EXAMPLE The biweight location estimate of the data set given in 4.3.2 is $T_n = 0{,}176\ 9$. It is close to the mean value (0,156 5) of the data set with the two extreme values (5,80 and 12,8) replaced by their correct values (0,58 and 1,28).

5.3 Robust estimation of dispersion

5.3.1 General

Two of the widely used scale estimators that are resistant to outlying observations and can be used in place of the sample standard deviation are given below.

5.3.2 Median-median absolute pair-wise deviation

$$S_n = s_n \,\mathrm{Median}_i \left(\mathrm{Median}_j \left|x_i - x_j\right|, i \neq j, i, j = 1, 2, \ldots, n\right) \tag{11}$$

The constant s_n is a correction factor chosen to ensure that S_n is an unbiased estimator for the scale parameter of a hypothesized distribution (normal, exponential, etc.). For large normal samples, the value of s_n is taken to be 1,192 6 (see Reference [10]), whereas s_n = 1,698 2 for large exponential samples. The values of s_n are given in Table D.1 for normal samples of size n = 2(1)20(10)100, 120, 150, 200, 300 and 500.

5.3.3 Biweight scale estimate

The biweight estimate of scale in the sample $x_1, x_2, \ldots, x_n$, follows the discussion given in Reference [9], and can be obtained as

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

where M is the sample median, $u_i = (x_i - M)/(cM_{\text{ad}})$ and $M_{\text{ad}} = \text{Median}\left(\left|x_i - M\right|, i = 1, 2, \cdots, n\right)$ for normal samples of size n. A recommended choice for c is the value 9,0. The values of s_{bi} based on $c = 9{,}0$ are given in Table D.1 for normal samples of size $n = 2(1)20(10)100$, 120, 150, 200, 300 and 500.

NOTE Under the normality assumption, a biweight estimator with $c = 9{,}0$ gives zero weight to observations more than about 6 standard deviations away from the median.

EXAMPLE For the data set given in 4.3.2, the classical sample standard deviation s, robust scale estimates S_n of 5.3.2, and S_{bi} of 5.3.3 are given by

$s = 3{,}177\ 2$, $S_n = 1{,}015\ 0$, $S_{bi} = 1{,}156\ 5$

These results clearly reveal that the classical sample standard deviation (s) has been greatly inflated by the two large observations. The two corresponding robust estimates S_n and S_{bi} have relatively smaller values and are close to each other.

6 Outliers in multivariate and regression data

6.1 General

Outliers are much harder to identify in multivariate and regression data than in univariate data. A multivariate outlier need not be an outlier in any of its components or bivariate coordinates. Multivariate outliers can also be cloaked to some extent by the general structure of their generating mechanism and their presence only comes to light after the structure of the data has been modelled. An outlier in regression data may not be a simple extreme value, but an observation that significantly deviates from the general pattern of the regression model.

6.2 Outliers in multivariate data

The general idea behind methods to identify outliers from multivariate data is to transform the multivariate observations into univariate statistics. One widely used statistic is the Mahalanobis distance, which measures the distance of a multivariate observation to the sample mean of the data set, standardized by the sample covariance matrix. Suppose that we have p variables, given by $X_1, X_2, \ldots, X_p$ which are arranged in a p-component vector $\boldsymbol{X} = (X_1, X_2, \ldots, X_p)^T$.

Let $\mu = (\mu_1, \mu_2, \ldots, \mu_p)^T$ be the vector of the means of the p random variables in $\boldsymbol{X}$, and let the variances and covariances of the random variables in $\boldsymbol{X}$ be denoted by a $p \times p$ covariance matrix $\boldsymbol{\Sigma}$ in which the main diagonal elements of $\boldsymbol{\Sigma}$ are the variances and the off-diagonal elements are the covariances of the X's in $\boldsymbol{X}$.

The Mahalanobis distance from $\boldsymbol{X}$ to μ is defined as

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

The outliers for a sample of n multivariate observations $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ can be detected from the corresponding n Mahalanobis distances $M_{Di} = \sqrt{(\boldsymbol{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu})}$, i = 1, 2, …, n. For the case when the vector $\boldsymbol{X}$ follows a multivariate normal distribution with mean μ and covariance matrix $\boldsymbol{\Sigma}$, the squared Mahalanobis distance, M_D^2, is known to follow a chi-square distribution with p degrees of freedom.

The above computation of Mahalanobis distance depends on knowledge of μ and $\boldsymbol{\Sigma}$. In practice, it is usually necessary to estimate the values of μ and $\boldsymbol{\Sigma}$ from the sample data. In the presence of outliers, robust estimates of μ and $\boldsymbol{\Sigma}$ should be obtained by the minimum covariance determinant (MCD) estimator[11]. The MCD method looks for the set of h observations out of the n given observations which yield a covariance matrix that has the smallest possible determinant. If the data set is presumed to contain at most 100α% of outlying observations, the value of h should be taken close to $(1 - \alpha)n$; however, it should be greater than the integer value $[(n + p + 1)/2]$. The mean value and covariance matrix of these h observations is then the MCD estimates $\boldsymbol{\mu}_{\text{MCD}}$ and $\boldsymbol{\Sigma}_{\text{MCD}}$ of μ and $\boldsymbol{\Sigma}$, respectively. The robust distance of the observation $\boldsymbol{x}_i$ is then defined as

$$D_{Ri} = \sqrt{\left(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_{\text{MCD}}\right)^T \hat{\boldsymbol{\Sigma}}_{\text{MCD}}^{-1} \left(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_{\text{MCD}}\right)} \tag{14}$$

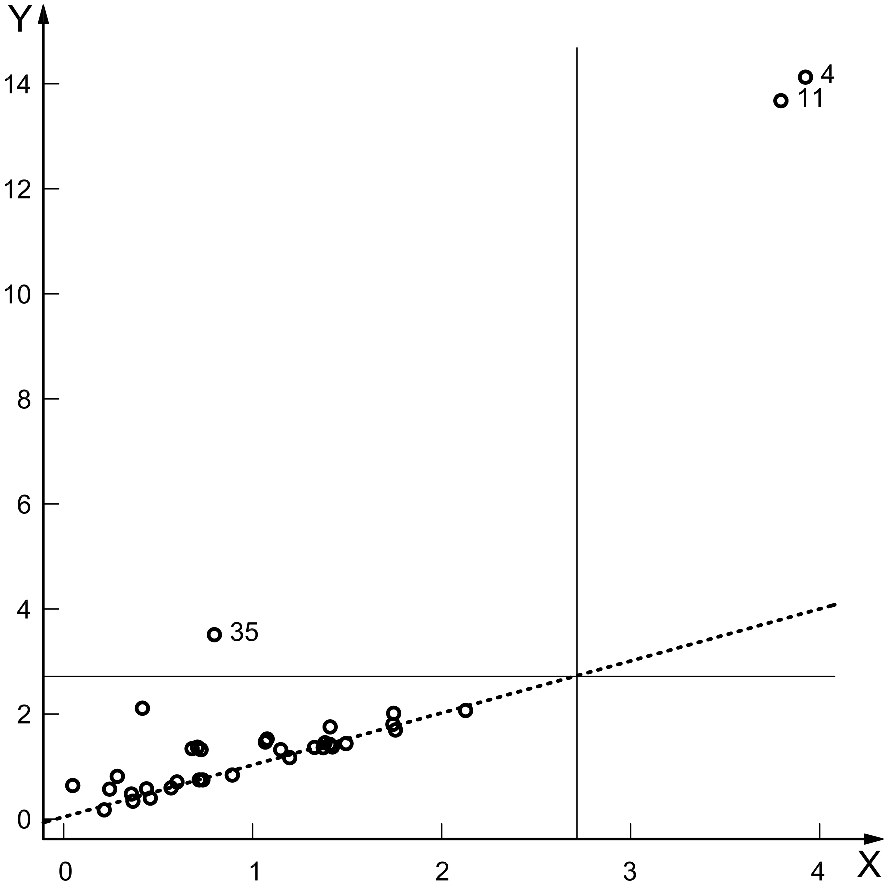
Under the multivariate normality assumption, a conservative criterion[11] is to declare observations that have a robust distance larger than the cutoff-value $\sqrt{\chi^2_{0,975;p}}$ as outliers, where $\chi^2_{0,975;p}$ is the 97,5 % percentile of a chi-squared distribution with p degrees of freedom.

A visual comparison between the Mahalanobis distances and the robust distances, and the effectiveness of using the robust distance in detecting outliers, is given in the following example.

EXAMPLE A set of 35 bivariate observations (x_1, x_2) collected from an experiment were recorded as follows:

Datum number i	x_{1i}	x_{2i}	Datum number i	x_{1i}	x_{2i}	Datum number 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			

The Mahalanobis distance and robust distance of each observation are computed and plotted in Figure 8 using $h = 32$ observations to calculate the MCD estimator. This figure is plotted using the open-source software LIBRA[11]. The dashed line is the set of points where the robust distance is equal to the Mahalanobis distance. The horizontal and vertical lines are drawn at the cutoff-value of $\sqrt{\chi^2_{0,975;\,2}} = \sqrt{7,378} = 2,716$. Points beyond these lines can be declared as outliers. The robust distance in this plot reveals that points 4, 11 and 35 are outliers. However, only points 4 and 11 are declared as outliers when the Mahalanobis distance is used. It can be seen as an example of masking defined in 2.3 that the Mahalanobis distance only declares observations 4 and 11 as outliers. If the Mahalanobis distance is calculated without using observations 4 and 11, then observation 35 is also declared as an outlier.



Key

X Mahalanobis distance

Y robust distance

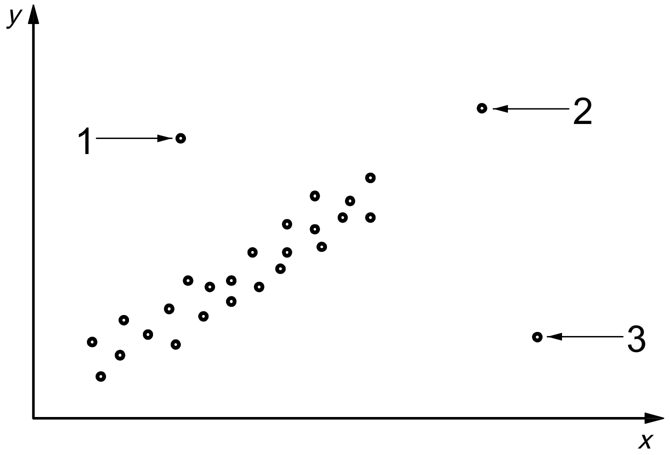
The data are plotted in Figure 11 where the points 4, 11 and 35 are labelled.

Figure 8 — Plot of the Mahalanobis distance against the robust distance of the data set

6.3 Outliers in linear regression

6.3.1 General

In simple linear regression analysis, a data point (Y, X) can be outlying with respect to its Y value, its X value, or both. In the scatter plot of (y_i, x_i) given in Figure 9, point 1 is outlying with respect to its y value as it falls far outside the scatter, although its x value is not an outlying value; point 3 is outlying with respect to its x value as this x value is much larger than the values of other points and its y value is not an outlying value; point 2 is outlying with respect to both its x and y values.



Key

1, 2, 3 outlying points

Figure 9 — Scatter plot of (Y, X)

Figure 9 also reveals that not all outlying points have a strong influence on the fitted regression line. Point 1 may not be too influential because a number of points in the scatter plot have similar x values that will prevent the fitted regression line from being displaced too far by point 1. Point 2 is also not too influential because its value of y is consistent with the linear regression line formed by the majority of the data points. By contrast, point 3 is influential in affecting the fit of the regression line, as not only is its x value an outlier, but its y value is also inconsistent with the linear regression of the other points.

6.3.2 Linear regression models

In relating a response variable Y to a single explanatory variable X, the linear regression line fitted to a sample of n data points (y_i, x_i), $i = 1, 2, \ldots, n$, is given by

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

and the ith residual is defined as the difference between the observed value y_i and the corresponding fitted value $\hat{y}_i$, i.e.

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

The ordinary least squares (OLS) estimates b_0 and b_1 that minimize the error residual sum of square $\sum_{i=1}^{n} e_i^2$ are given by

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

and (16)

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

where $\bar{x}$ and $\bar{y}$ are the means of the x_i and the y_i observations, respectively.

The influence of outlying X and/or Y observations in fitting the linear regression line using the OLS estimates can be diagnosed by examining the fitted OLS regression value

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

or, equivalently,

$$\hat{y}_i = \sum_{j=1}^{n} h_{ij} y_j \qquad (17)$$

where the values

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

involve only the observations on the explanatory variable X. The h_{ij} values form a symmetric $n \times n$ matrix $\boldsymbol{H} = (h_{ij})$, often called a hat matrix. The equation $\hat{y}_i = \sum_{j=1}^{n} h_{ij} y_j$ clearly reveals that the h_{ij} values measure the role of the X values in determining how important the observed value y_j is in influencing the fitted value $\hat{y}_i$.

Similarly, in relating a response variable Y to p explanatory variables $X_1, X_2, \ldots, X_p$, the regression function fitted to a sample of n data points $(y_i, x_{i1}, x_{i2}, \ldots, x_{ip})$, $i = 1, 2, \ldots, n$, can be given by

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

in which b_j refers to the jth fitted regression coefficient, and x_{ij} refers to the ith individual value of the jth explanatory X_j,. As in the case of a single explanatory variable, the ith residual of the fitted regression function is $e_i = y_i - \hat{y}_i$. In matrix notation, the multiple regression model is written as

$$\hat{\boldsymbol{y}} = \boldsymbol{Xb} \qquad (18)$$

where $\hat{\boldsymbol{y}} = (\hat{y}_1, \ldots, \hat{y}_n)^T$ is an $n \times 1$ vector, $\boldsymbol{b} = (b_0, b_1, \ldots, b_p)^T$ is a $(p + 1) \times 1$ vector and $\boldsymbol{X}$ is an $n \times (p + 1)$ matrix of the form

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

The vector of estimated least squares regression coefficients is given as

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

and the vector of the fitted values $\hat{\boldsymbol{y}}$ can be obtained directly in terms of the hat matrix $\boldsymbol{H}$ as

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

where $\boldsymbol{y} = (y_1, ..., y_n)^{\mathrm{T}}$ is an $n \times 1$ vector of n observed y values, and

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

is an $n \times n$ matrix.

6.3.3 Detecting outlying Y observations

A robust procedure in detecting outlying Y observations from a sample of size n is to analyse the studentized deletion residuals r_i which are the studentized residual errors of the regression function fitted without using the ith data point. The studentized deletion residuals r_i can be computed as[12]

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

where

$e_i = y_i - \hat{y}_i$ is the ith residual,

h_{ii} is the ith diagonal element in the hat matrix $\boldsymbol{H}$,

$R_{\mathrm{SSE}} = \sum_{i=1}^{n} e_i^2$ is the residual error sum of squares of the fitted regression function based on the n data points, and the number of estimated parameters in the fitted regression function is $p + 1$.

NOTE The expression for the studentized deletion residual r_i is derived[12] based on the ith point $(y_i, x_{i1}, x_{i2}, \ldots, x_{ip})$ being discarded in fitting the regression function to the remaining $n - 1$ points. It can be calculated without having to fit new regression functions each time a different data point is omitted as can be seen from Equation (20).

By using the result that each of the studentized deletion residuals r_i follows a t distribution with $n - p - 2$ degrees of freedom, data points whose studentized deletion residuals have absolute value greater than $t_{1-\alpha/2n;\, n-p-2}$ would be identified as outlying with respect to the Y value.

6.3.4 Identifying outlying X observations

The diagonal elements of the hat matrix $\boldsymbol{H}$ can also be used to detect outlying X observations. Some useful properties of the elements h_{ii} in the hat matrix of the linear regression model with an intercept parameter are:

— $\frac{1}{n} \leqslant h_{ii} \leqslant 1$

— $\sum_{i=1}^{n} h_{ii} = p + 1$

— if $h_{ii} = 0$ or $h_{ii} = 1$, then $h_{ij} = 0$ for all $j \neq i$

where $p + 1$ is the number of regression parameters in the regression model including the intercept term.

In the special case of linear regression line with one explanatory variable ($p = 1$) and an intercept term the diagonal elements h_{ii} in the hat matrix $\boldsymbol{H}$ can be expressed as

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

The above equation of h_{ii} reveals that it is a measure of the distance between the X value of the ith point and the mean of the X values of all n data points. A large h_{ii} value reveals that the value x_i deviates away from the majority of the X observations and that x_i can be an outlying value compared with the majority of the x_j values that have smaller values of $\left|x_j - \bar{x}\right|$ for $j \neq i$. The diagonal element h_{ii} of the hat matrix in this context is called the leverage of the ith observation. In general, a leverage value h_{ii} is considered to be large if it is more than twice as large as the mean leverage value $\bar{h} = \frac{1}{n}\sum_{i=1}^{n} h_{ii} = (p+1)/n$. This rule implies that if $h_{ii} \geqslant \frac{2(p+1)}{n}$, then the ith observed value is taken to be an outlier with regard to its x value. Another simple guideline suggested by Reference [13] is that

- data with leverage values less than 0,2 can be safely included in the regression analysis,
- data with leverage values between 0,2 and 0,5 may be included in the regression analysis,
- data with leverage values greater than 0,5 should be discarded in the regression analysis.

6.3.5 Detecting influential observations

After identifying data points that are outlying either in their Y values and/or their X values, the next step is to ascertain whether these outlying data are influential by examining if deleting these data points would lead to major changes in the fitted regression model. Two of the widely used measures of influence are the DFFITS value and Cook's distance[12][14].

DFFITS value

The notation DFFITS is an abbreviation for “difference in fits”. The DFFITS value for the ith data point is defined as

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

where r_i is the studentized deletion residual defined in Equation (20). The ith data point is declared as an influential point if the absolute value $|(\text{DFFITS})_i|$ exceeds 1 for small to medium data sets and exceeds $2\sqrt{(p+1)/n}$ for large data sets.

Cook's distance

Cook's distance, denoted by D_i, is defined as

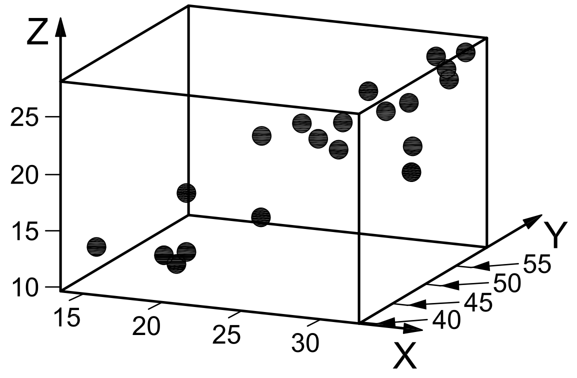
$$D_i = \frac{(n-p-1)e_i^2}{(p+1)R_{\text{SSE}}} \left[\frac{h_{ii}}{(1-h_{ii})^2} \right] \tag{23}$$

where the larger the e_i and/or h_{ii}, the larger the D_i. Therefore, large D_i values signify influential observations. Reference [14] suggests that observations with a Cook's distance greater than the 50th percentile value $F_{0,50;p+1,n-p-1}$ of an F-distribution may be declared as influential outliers, where n is the number of observations, $p + 1$ is the number of parameters in the regression model (including the intercept parameter) and is used to indicate the degrees of freedom associated with the numerator; and $n - p - 1$ is the denominator degrees of freedom. Observations which have a Cook's distance value that is above $F_{0,50;p+1,n-p-1}$ should be examined for typographical errors or other causes for the extremeness of the observation.

NOTE The methods described in this clause will be ineffective if two or more influential outlying data points fall close to each other. Extensions of the above procedures to detect two or more closely grouped influential data points have been developed in which extensive computation is required.

EXAMPLE The data obtained in a study conducted to determine the relationship between the amount of body fat (Y) to two explanatory variables, triceps skinfold thickness (X_1) and thigh circumference (X_2), are given in columns 2, 3 and 4 of the table below. The data are taken from Reference [12]. The three-dimensional plot of (Y, X_1, X_2) is also given in Figure 10.

Data points (subject)	**Triceps skinfold thickness**	**Thigh circumference**	**Body fat**	**Residual**	**Leverage value**	**Studentized deletion residual**
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



Key

X triceps skinfold thickness

Y thigh circumference

Z body fat

Figure 10 — Scatter plot of body fat vs thigh circumference vs triceps skinfold thickness

The regression function fitted by the OLS method is given by

$$\hat{y}_i = -19{,}174 + 0{,}222\,4x_{1i} + 0{,}659\,4x_{2i}$$

with sum of squares error, $R_{\text{SSE}} = \sum_{i=1}^{20} e_i^2 = 109{,}95$, where the residuals e_i, leverage h_{ii} and the studentized deletion residuals r_i of the fitted regression function are given in columns 5, 6 and 7, respectively.

As $n = 20$ and $p = 2$, then by taking the significance level to be $\alpha = 0{,}05$, we have

$$t_{1-\alpha/2n;\, n-p-2} = t_{0{,}998\,75;16} = 3{,}580\,2$$

Since $\left|r_i\right| \leqslant 3{,}580\,2$ for all i, we conclude that none of the data points has an outlying Y value.

In detecting an outlying X value, as both $h_{33} = 0{,}372$ and $h_{15,15} = 0{,}333$ exceed the value

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

we conclude that data points 3 and 15 are outlying in terms of their X value.

Finally, we shall ascertain how influential the data points 3 and 15 are in fitting the regression line by using their respective Cook's distance of

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

and $D_{15} = 0{,}212$. Since these two values are both less than the 50th percentile value $F_{0{,}50;3{,}17} = 0{,}821\,2$ of the F-distribution, both data points 3 and 15 are not influential enough to declare them as influential outliers.

The regression function fitted with data point 3 discarded is given by

$$\hat{y}_i = -12{,}248 + 0{,}564\,1x_{1i} + 0{,}363\,5x_{2i}$$

in which the values of the estimated parameters are substantially different from those fitted with data point 3.

6.3.6 A robust regression procedure

An alternative approach of detecting outliers in regression analysis is to fit a robust regression model to the majority of the data and then discover the outliers as those points having large residuals from the robust equation. A widely used robust regression model is the least trimmed squares (LTS) regression[15]. The regression coefficients of the LTS regression are those that minimize the sum of the m smallest squared regression residuals. Consider again a given sample of n data (y_i, x_{i1}, x_{i2}, …, x_{ip}), i = 1, 2, …, n, with the fitted values and residuals given as

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

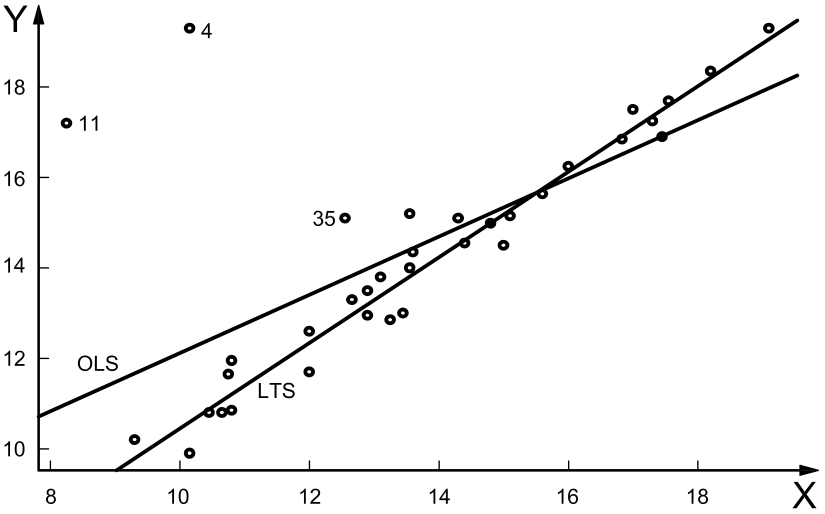
and

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

respectively. In this case, the regression coefficients b_0, b_1, …, b_p of the LTS regression are values that minimize the sum of squares $\sum_{i=1}^{m} \mathrm{e}_{(i)}^2$, where $\mathrm{e}_{(i)}^2$ refers to the ith order statistics of the squared residuals (i.e. the residuals are first squared and then ordered), and m is the number of observations (out of n) that are presumed to be fitted well by the LTS regression model. When the data set is presumed to contain at most 100α % of outlying observations, the value of m should be taken close to (1 − α)n but not less than the integer value [($n + p$ + 1)/2]. Observations with large residuals are then identified as outliers.

NOTE The estimation of LTS regression coefficients is available in proprietary statistical packages.

EXAMPLE The bivariate data of 6.2 are plotted in Figure 11 together with two fitted regression lines of the response variable x_2 (y) to x_1 (x), i.e. the ordinary least squares (OLS) line that minimizes the residual sum of squares, and the least trimmed squared (LTS) line that minimizes the trimmed residual sum of squares with m = [0,9n].



The labelled points 4, 11 and 35 were the points considered to be outliers in 6.2.

Figure 11 — Comparison of the least trimmed squares (LTS) regression line and the ordinary least squares (OLS) regression line

Note that the two most influential points at the top left hand corner pull the OLS line away from the main cluster of the data set that has been fitted extremely well by the LTS line. The robust LTS regression procedure essentially ignores the two influential points as only around 90 % of the data are included in fitting the LTS line.

Annex A
(informative)

Algorithm for the GESD outliers detection procedure

Suppose that a sample $x_1, x_2, \ldots, x_n$ of size n is taken from a normal distribution. The following algorithm describes the necessary steps for detecting m possible outliers using the generalized extreme studentized deviate (GESD) procedure at significance level α.

```
Read $\alpha$, $m$.

Set $l = 0$.

Set $I_0 = \{x_1, x_2, \ldots, x_n\}$.

REPEAT

    Compute the sample mean $\bar{x}(I_l)$ and sample standard deviation $s(I_l)$ from the sample $I_l$.

    Compute the statistic $R_l = \dfrac{\max\limits_{x_i \in I_l} \left| x - \bar{x}(I_l) \right|}{s(I_l)}$.

    Compute the 100$p$th percentile $t_{p;\ n-l-2}$ of the $t$-distribution with $(n - l - 2)$ degrees of freedom, where $p = (1 - \alpha/2)^{1/(n-l)}$

    Compute the critical value $\lambda_l = \dfrac{(n-l-1) t_{p,n-l-2}}{\sqrt{(n-l-2+t_{p,n-l-2}^2)(n-l)}}$.

    Set $I_{l+1} = I_l \backslash \{x^{(l)}\}$. [See Note 1]

    Set $l = l + 1$.

UNTIL $l = m$.

Set $l = 0$.

REPEAT

    If $(R_l > \lambda_i)$, then declare $x^{(l)}$ (the value of $x$ in $I_l$ that yielded the value $R_l$) as an outlier.

    Set $l = l + 1$.

UNTIL $l = m$.
```

NOTE 1 I_{l+1} is the reduced sample of size $n - l$ obtained by deleting the data point $x^{(l)}$ in the sample I_l that yields the value R_l.

NOTE 2 If $(R_l \leqslant \lambda_i)$ for all $l = 0, 1, 2, \ldots, m$, it is concluded that no outliers are present in the sample.

Annex B
(normative)

Critical values of outliers test statistics for exponential samples

Table B.1 — The lower and upper 2,5 % and 1 % critical values $g_{E;n}$ of Greenwood's test statistic G_E for exponential samples

	Lower	Lower	Upper	Upper		Lower	Lower	Upper	Upper		Lower	Lower	Upper	Upper
n	1 %	2,5 %	2,5 %	1 %	n	1 %	2,5 %	2,5 %	1 %	n	1 %	2,5 %	2,5 %	1 %
2	0,500 0	0,500 3	0,975 4	0,990 1	34	0,042 8	0,044 3	0,079 0	0,086 3	82	0,019 5	0,020 1	0,030 1	0,031 9
3	0,336 0	0,340 2	0,831 4	0,890 1	35	0,041 7	0,043 1	0,076 5	0,083 5	84	0,019 1	0,019 6	0,029 3	0,031 1
4	0,258 5	0,265 8	0,682 8	0,756 3	36	0,040 7	0,042 1	0,074 2	0,080 9	86	0,018 7	0,019 2	0,028 6	0,030 2
5	0,213 7	0,221 7	0,568 0	0,640 0	37	0,039 7	0,041 1	0,072 0	0,078 4	88	0,018 3	0,018 8	0,027 9	0,029 5
6	0,183 8	0,191 4	0,482 1	0,547 4	38	0,038 8	0,040 1	0,069 9	0,076 1	90	0,017 9	0,018 4	0,027 2	0,028 8
7	0,162 0	0,168 9	0,417 3	0,474 9	39	0,037 9	0,039 2	0,068 0	0,073 8	92	0,017 6	0,018 0	0,026 6	0,028 1
8	0,145 2	0,151 4	0,366 7	0,417 3	40	0,037 1	0,038 3	0,066 1	0,071 7	94	0,017 3	0,017 7	0,026 0	0,027 4
9	0,131 8	0,137 4	0,326 3	0,371 0	41	0,036 3	0,037 5	0,064 3	0,069 8	96	0,016 9	0,017 4	0,025 4	0,026 8
10	0,120 8	0,126 0	0,293 4	0,333 1	42	0,035 5	0,036 7	0,062 6	0,067 9	98	0,016 6	0,017 0	0,024 8	0,026 2
11	0,111 6	0,116 4	0,266 1	0,301 6	43	0,034 8	0,035 9	0,061 0	0,066 1	100	0,016 3	0,016 7	0,024 3	0,025 6
12	0,103 9	0,108 2	0,243 1	0,275 1	44	0,034 1	0,035 2	0,059 5	0,064 4	105	0,015 6	0,016 0	0,023 0	0,024 2
13	0,097 2	0,101 2	0,223 6	0,252 5	45	0,033 4	0,034 5	0,058 1	0,062 8	110	0,014 9	0,015 3	0,021 9	0,023 0
14	0,091 3	0,095 1	0,206 8	0,233 0	46	0,032 8	0,033 8	0,056 7	0,061 2	115	0,014 3	0,014 7	0,020 9	0,021 9
15	0,086 2	0,089 7	0,192 2	0,216 1	47	0,032 2	0,033 2	0,055 4	0,059 7	120	0,013 8	0,014 1	0,019 9	0,020 9
16	0,081 6	0,084 9	0,179 4	0,201 3	48	0,031 6	0,032 6	0,054 1	0,058 3	125	0,013 3	0,013 6	0,019 1	0,020 0
17	0,077 6	0,080 7	0,168 1	0,188 3	49	0,031 0	0,032 0	0,052 9	0,057 0	130	0,012 8	0,013 1	0,018 3	0,019 1
18	0,073 9	0,076 8	0,158 1	0,176 8	50	0,030 5	0,031 4	0,051 7	0,055 7	135	0,012 4	0,012 7	0,017 6	0,018 4
19	0,070 6	0,073 4	0,149 1	0,166 4	52	0,029 4	0,030 3	0,049 6	0,053 3	140	0,012 0	0,012 2	0,016 9	0,017 6
20	0,067 6	0,070 2	0,141 1	0,157 2	54	0,028 4	0,029 3	0,047 5	0,051 1	145	0,011 6	0,011 8	0,016 3	0,017 0
21	0,064 8	0,067 3	0,133 8	0,148 8	56	0,027 5	0,028 4	0,045 7	0,049 0	150	0,011 2	0,011 5	0,015 7	0,016 3
22	0,062 3	0,064 7	0,127 2	0,141 2	58	0,026 7	0,027 5	0,044 0	0,047 1	155	0,010 9	0,011 1	0,015 2	0,015 8
23	0,060 0	0,062 3	0,121 2	0,134 3	60	0,025 9	0,026 7	0,042 4	0,045 3	160	0,010 6	0,010 8	0,014 6	0,015 2
24	0,057 8	0,060 0	0,115 7	0,128 0	62	0,025 1	0,025 9	0,040 9	0,043 7	165	0,010 3	0,010 5	0,014 2	0,014 7
25	0,055 8	0,057 9	0,110 7	0,122 3	64	0,024 4	0,025 1	0,039 5	0,042 1	170	0,010 0	0,010 2	0,013 7	0,014 3
26	0,054 0	0,056 0	0,106 0	0,117 0	66	0,023 8	0,024 4	0,038 2	0,040 7	175	0,009 7	0,009 9	0,013 3	0,013 8
27	0,052 2	0,054 2	0,101 7	0,112 1	68	0,023 1	0,023 8	0,036 9	0,039 4	180	0,009 5	0,009 7	0,012 9	0,013 4
28	0,050 6	0,052 5	0,097 8	0,107 6	70	0,022 5	0,023 2	0,035 8	0,038 1	185	0,009 2	0,009 4	0,012 5	0,013 0
29	0,049 1	0,050 9	0,094 1	0,103 4	72	0,022 0	0,022 6	0,034 7	0,036 9	190	0,009 0	0,009 2	0,012 2	0,012 6
30	0,047 7	0,049 4	0,090 6	0,099 5	74	0,021 4	0,022 0	0,033 7	0,035 8	195	0,008 8	0,009 0	0,011 9	0,012 3
31	0,046 4	0,048 0	0,087 4	0,095 8	76	0,020 9	0,021 5	0,032 7	0,034 7	200	0,008 6	0,008 7	0,011 5	0,012 0
32	0,045 1	0,046 7	0,084 4	0,092 4	78	0,020 4	0,021 0	0,031 8	0,033 7	225	0,007 7	0,007 8	0,010 2	0,010 5
33	0,043 9	0,045 4	0,081 6	0,089 3	80	0,020 0	0,020 5	0,030 9	0,032 8	250	0,007 0	0,007 1	0,009 1	0,009 4

NOTE 1 Each of these critical values is based on one hundred million simulated exponential samples of size n.

NOTE 2 Each table entry has been rounded upwards in the fourth decimal place in order to guarantee the significance level.

Table B.2 — The upper 5 % and 1 % critical values for consecutive tests of up to $m = 2$ upper outliers for exponential samples

$m = 2$									
	5 %		1 %			5 %		1 %	
n	$s_{2:n}^{\mathrm{U}}$	$s_{1:n}^{\mathrm{U}}$	$s_{2:n}^{\mathrm{U}}$	$s_{1:n}^{\mathrm{U}}$	n	$s_{2:n}^{\mathrm{U}}$	$s_{1:n}^{\mathrm{U}}$	$s_{2:n}^{\mathrm{U}}$	$s_{1:n}^{\mathrm{U}}$
10	0,434 8	0,483 4	0,514 3	0,569 6	46	0,118 7	0,152 2	0,137 6	0,183 0
11	0,401 0	0,453 3	0,474 8	0,536 3	48	0,114 5	0,147 0	0,132 7	0,176 9
12	0,372 4	0,426 9	0,441 2	0,506 6	50	0,110 6	0,142 1	0,128 2	0,170 8
13	0,348 0	0,403 3	0,412 5	0,479 3	55	0,102 0	0,131 4	0,117 9	0,157 8
14	0,326 8	0,382 7	0,386 8	0,455 5	60	0,094 6	0,122 2	0,109 2	0,146 7
15	0,308 2	0,363 9	0,364 7	0,434 5	65	0,088 4	0,114 3	0,102 0	0,137 1
16	0,291 6	0,347 3	0,344 7	0,414 9	70	0,083 0	0,107 4	0,095 5	0,128 7
17	0,277 0	0,332 0	0,327 3	0,397 2	75	0,078 3	0,101 3	0,089 9	0,121 4
18	0,263 7	0,318 3	0,311 4	0,381 3	80	0,074 1	0,096 0	0,084 9	0,115 0
19	0,251 9	0,305 8	0,297 1	0,366 7	85	0,070 3	0,091 2	0,080 7	0,109 2
20	0,241 3	0,294 1	0,284 5	0,352 9	90	0,067 0	0,086 9	0,076 7	0,103 9
21	0,231 3	0,283 4	0,272 3	0,340 3	95	0,063 9	0,083 0	0,073 2	0,099 2
22	0,222 4	0,273 5	0,261 8	0,328 6	100	0,061 2	0,079 4	0,070 0	0,094 9
23	0,214 2	0,264 4	0,251 9	0,317 5	110	0,056 4	0,073 2	0,064 4	0,087 3
24	0,206 5	0,255 8	0,242 6	0,307 4	120	0,052 4	0,067 9	0,059 6	0,081 0
25	0,199 5	0,247 8	0,234 0	0,298 0	130	0,048 9	0,063 4	0,055 6	0,075 5
26	0,192 9	0,240 3	0,226 3	0,288 8	140	0,045 8	0,059 5	0,052 1	0,070 8
27	0,186 8	0,233 3	0,219 0	0,280 5	150	0,043 2	0,056 0	0,049 1	0,066 6
28	0,181 2	0,226 8	0,212 3	0,272 9	160	0,040 9	0,053 0	0,046 4	0,062 9
29	0,175 7	0,220 7	0,205 8	0,265 4	170	0,038 8	0,050 3	0,044 0	0,059 6
30	0,170 8	0,214 8	0,199 8	0,258 4	180	0,036 9	0,047 8	0,041 8	0,056 7
32	0,161 7	0,204 1	0,189 0	0,245 7	190	0,035 3	0,045 6	0,039 9	0,054 0
34	0,153 5	0,194 4	0,179 2	0,233 9	200	0,033 7	0,043 6	0,038 1	0,051 6
36	0,146 2	0,185 7	0,170 5	0,223 5	220	0,031 2	0,040 4	0,035 1	0,047 4
38	0,139 7	0,177 7	0,162 7	0,213 9	240	0,028 9	0,037 3	0,032 5	0,043 9
40	0,133 7	0,170 6	0,155 5	0,205 1	260	0,026 9	0,034 7	0,030 3	0,040 9
42	0,128 3	0,163 9	0,149 1	0,197 2	280	0,025 2	0,032 5	0,028 4	0,038 2
44	0,123 3	0,157 8	0,143 2	0,189 8	300	0,023 8	0,030 6	0,026 7	0,035 9

Table B.3 — The upper 5 % and 1 % critical values for consecutive tests of up to $m = 3$ upper outliers for exponential samples

$m = 3$													
	5 %			1 %				5 %			1 %		
n	$s_{3:n}^{\mathrm{U}}$	$s_{2:n}^{\mathrm{U}}$	$s_{1:n}^{\mathrm{U}}$	$s_{3:n}^{\mathrm{U}}$	$s_{2:n}^{\mathrm{U}}$	$s_{1:n}^{\mathrm{U}}$	n	$s_{3:n}^{\mathrm{U}}$	$s_{2:n}^{\mathrm{U}}$	$s_{1:n}^{\mathrm{U}}$	$s_{3:n}^{\mathrm{U}}$	$s_{2:n}^{\mathrm{U}}$	$s_{1:n}^{\mathrm{U}}$
15	0,305 8	0,321 0	0,380 3	0,357 7	0,377 5	0,449 7	55	0,093 1	0,105 2	0,136 7	0,105 6	0,121 4	0,163 5
16	0,287 5	0,303 5	0,363 0	0,336 0	0,356 9	0,429 6	60	0,086 3	0,097 6	0,127 1	0,097 5	0,112 4	0,152 0
17	0,271 2	0,288 1	0,347 0	0,316 5	0,338 7	0,411 2	65	0,080 4	0,091 2	0,118 9	0,090 8	0,104 8	0,142 1
18	0,257 0	0,274 3	0,332 6	0,299 4	0,322 2	0,394 9	70	0,075 4	0,085 5	0,111 7	0,084 9	0,098 1	0,133 3
19	0,244 1	0,261 9	0,319 5	0,283 7	0,307 4	0,379 8	75	0,071 0	0,080 6	0,105 4	0,079 9	0,092 4	0,125 7
20	0,232 5	0,250 7	0,307 2	0,269 8	0,294 5	0,365 8	80	0,067 1	0,076 2	0,099 7	0,075 4	0,087 2	0,119 0
21	0,222 1	0,240 3	0,296 2	0,257 9	0,281 7	0,352 5	85	0,063 7	0,072 4	0,094 7	0,071 5	0,082 9	0,113 0
22	0,212 5	0,230 9	0,285 7	0,246 2	0,270 7	0,340 4	90	0,060 6	0,068 9	0,090 2	0,067 9	0,078 7	0,107 6
23	0,204 0	0,222 4	0,276 1	0,236 2	0,260 5	0,329 0	95	0,057 8	0,065 8	0,086 2	0,064 8	0,075 2	0,102 6
24	0,196 1	0,214 2	0,267 2	0,226 8	0,250 7	0,318 6	100	0,055 3	0,062 9	0,082 4	0,061 9	0,071 8	0,098 1
25	0,189 0	0,206 8	0,258 7	0,218 1	0,241 9	0,308 7	110	0,050 9	0,058 0	0,076 0	0,056 9	0,066 0	0,090 3
26	0,182 3	0,200 0	0,250 9	0,210 4	0,233 8	0,299 3	120	0,047 2	0,053 8	0,070 5	0,052 7	0,061 2	0,083 7
27	0,176 1	0,193 7	0,243 6	0,202 9	0,226 3	0,290 7	130	0,044 1	0,050 2	0,065 8	0,049 1	0,057 0	0,078 0
28	0,170 3	0,187 8	0,236 8	0,196 2	0,219 1	0,282 9	140	0,041 3	0,047 1	0,061 6	0,046 0	0,053 5	0,073 1
29	0,164 9	0,182 1	0,230 3	0,189 7	0,212 5	0,274 9	150	0,039 0	0,044 4	0,058 1	0,043 3	0,050 3	0,068 8
30	0,160 0	0,177 0	0,224 1	0,184 0	0,206 3	0,268 0	160	0,036 8	0,042 0	0,054 9	0,040 9	0,047 5	0,065 0
32	0,150 9	0,167 4	0,212 9	0,173 0	0,195 1	0,254 6	170	0,035 0	0,039 8	0,052 1	0,038 8	0,045 1	0,061 6
34	0,142 8	0,158 9	0,202 8	0,163 7	0,184 9	0,242 6	180	0,033 3	0,037 9	0,049 5	0,036 9	0,042 8	0,058 5
36	0,135 6	0,151 3	0,193 6	0,155 2	0,175 8	0,231 8	190	0,031 8	0,036 2	0,047 2	0,035 2	0,040 9	0,055 7
38	0,129 2	0,144 4	0,185 3	0,147 6	0,167 9	0,221 8	200	0,030 4	0,034 6	0,045 2	0,033 6	0,039 0	0,053 3
40	0,123 4	0,138 2	0,177 8	0,140 9	0,160 3	0,212 5	220	0,028 0	0,031 8	0,041 5	0,030 9	0,035 9	0,048 9
42	0,118 2	0,132 6	0,170 8	0,134 8	0,153 7	0,204 4	240	0,026 0	0,029 5	0,038 5	0,028 7	0,033 2	0,045 3
44	0,113 4	0,127 4	0,164 4	0,129 1	0,147 4	0,196 9	260	0,024 2	0,027 6	0,035 9	0,026 7	0,031 0	0,042 1
46	0,109 1	0,122 6	0,158 5	0,124 0	0,141 8	0,189 8	280	0,022 7	0,025 8	0,033 6	0,025 0	0,029 0	0,039 4
48	0,105 0	0,118 2	0,153 1	0,119 3	0,136 7	0,183 4	300	0,021 4	0,024 3	0,031 6	0,023 6	0,027 3	0,037 0
50	0,101 3	0,114 2	0,148 0	0,115 0	0,132 0	0,176 9							

Table B.4 — The upper 5 % and 1 % critical values for consecutive tests of up to $m = 4$ upper outliers for exponential samples

$m = 4$								
	5 %				1 %			
n	$s_{4:n}^{\mathrm{U}}$	$s_{3:n}^{\mathrm{U}}$	$s_{2:n}^{\mathrm{U}}$	$s_{1:n}^{\mathrm{U}}$	$s_{4:n}^{\mathrm{U}}$	$s_{3:n}^{\mathrm{U}}$	$s_{2:n}^{\mathrm{U}}$	$s_{1:n}^{\mathrm{U}}$
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

Table B.5 — The upper 5 % and 1 % critical values for consecutive tests of up to $m = 2$ lower outliers for exponential samples

$m = 2$									
	5 %		1 %			5 %		1 %	
n	$s_{2:n}^{\mathrm{L}}$	$s_{1:n}^{\mathrm{L}}$	$s_{2:n}^{\mathrm{L}}$	$s_{1:n}^{\mathrm{L}}$	n	$s_{2:n}^{\mathrm{L}}$	$s_{1:n}^{\mathrm{L}}$	$s_{2:n}^{\mathrm{L}}$	$s_{1:n}^{\mathrm{L}}$
10	0,836 7	0,977 5	0,921 6	0,995 5	29	0,8224	0,975 9	0,9130	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	0,9174	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	0,9140	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					

Table B.6 — The upper 5 % and 1 % critical values for consecutive tests of up to $m = 3$ lower outliers for exponential samples

$m = 3$													
	5 %			1 %				5 %			1 %		
n	$s_{3:n}^{\mathrm{L}}$	$s_{2:n}^{\mathrm{L}}$	$s_{1:n}^{\mathrm{L}}$	$s_{3:n}^{\mathrm{L}}$	$s_{2:n}^{\mathrm{L}}$	$s_{1:n}^{\mathrm{L}}$	n	$s_{3:n}^{\mathrm{L}}$	$s_{2:n}^{\mathrm{L}}$	$s_{1:n}^{\mathrm{L}}$	$s_{3:n}^{\mathrm{L}}$	$s_{2:n}^{\mathrm{L}}$	$s_{1:n}^{\mathrm{L}}$
15	0,705 1	0,855 5	0,984 0	0,807 3	0,931 4	0,996 9	40	0,688 8	0,847 2	0,983 3	0,793 7	0,926 6	0,996 8
16	0,703 5	0,854 4	0,984 0	0,806 2	0,930 6	0,996 9	50	0,687 1	0,846 2	0,983 2	0,792 2	0,926 0	0,996 7
17	0,701 9	0,853 6	0,983 9	0,805 0	0,930 0	0,996 8	60	0,685 2	0,845 9	0,983 2	0,791 1	0,925 7	0,996 7
18	0,700 7	0,853 2	0,983 9	0,803 4	0,930 0	0,996 8	70	0,684 3	0,844 9	0,983 2	0,790 4	0,925 3	0,996 7
19	0,699 0	0,852 7	0,983 8	0,802 7	0,929 6	0,996 8	80	0,683 8	0,844 9	0,983 1	0,789 5	0,925 1	0,996 7
20	0,698 0	0,852 0	0,983 8	0,801 5	0,929 0	0,996 8	90	0,683 0	0,844 3	0,983 1	0,789 5	0,925 0	0,996 7
21	0,697 0	0,851 7	0,983 7	0,801 1	0,928 8	0,996 8	100	0,683 2	0,844 4	0,983 0	0,788 7	0,925 3	0,996 7
22	0,696 4	0,851 1	0,983 7	0,799 5	0,928 6	0,996 8	120	0,682 7	0,843 8	0,983 0	0,788 5	0,924 7	0,996 7
23	0,695 6	0,850 7	0,983 7	0,799 5	0,928 5	0,996 8	140	0,682 1	0,843 4	0,983 0	0,788 2	0,924 4	0,996 7
24	0,694 8	0,850 2	0,983 6	0,798 8	0,928 5	0,996 8	160	0,682 1	0,843 7	0,983 0	0,787 7	0,924 5	0,996 7
25	0,693 9	0,850 3	0,983 6	0,797 8	0,928 1	0,996 8	180	0,681 7	0,843 6	0,982 9	0,787 4	0,924 2	0,996 7
26	0,693 5	0,849 9	0,983 6	0,798 0	0,928 3	0,996 8	200	0,681 3	0,843 7	0,983 0	0,786 6	0,924 2	0,996 7
27	0,692 9	0,849 5	0,983 5	0,797 0	0,928 0	0,996 8	250	0,681 2	0,843 2	0,982 9	0,786 9	0,923 9	0,996 7
28	0,692 4	0,849 3	0,983 5	0,797 2	0,927 9	0,996 8	300	0,680 4	0,843 1	0,982 9	0,786 3	0,924 3	0,996 6
29	0,691 9	0,849 1	0,983 5	0,796 9	0,927 8	0,996 8							
30	0,691 5	0,849 1	0,983 4	0,796 5	0,927 6	0,996 8							

Table B.7 — The upper 5 % and 1 % critical values for consecutive tests of up to $m = 4$ lower outliers for exponential samples

$m = 4$								
	5 %				1 %			
n	$s_{4:n}^{\mathrm{L}}$	$s_{3:n}^{\mathrm{L}}$	$s_{2:n}^{\mathrm{L}}$	$s_{1:n}^{\mathrm{L}}$	$s_{4:n}^{\mathrm{L}}$	$s_{3:n}^{\mathrm{L}}$	$s_{2:n}^{\mathrm{L}}$	$s_{1:n}^{\mathrm{L}}$
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

Annex C
(normative)

Factor values of the modified box plot

When the location parameter θ and scale parameter σ of a hypothesized location-scale distribution $F_{\theta,\sigma}(x)$ are unknown, its first and third quartiles are estimated by the lower fourth $X_{\text{L}:n}$ and upper fourth $X_{\text{U}:n}$ of a sample of n observations drawn from $F_{\theta,\sigma}(x)$. There are many definitions for the depth of sample fourths. The recommended depth is

$$\text{depth of fourth} = \begin{cases} i + 0{,}5 & f = 0; \\ i + 1 & f > 0, \end{cases}$$

where i is the integral part and f is the fractional part of $n/4$. The two data values with this depth, namely the lower sample fourth ($x_{\text{L}:n}$) and upper sample fourth ($x_{\text{U}:n}$), in a given sample of size n are then evaluated as in 4.4.

An exact expression that can be routinely used to evaluate the factors k_{L} and k_{U} of the box plot for samples taken from the hypothesized $F_{\theta,\sigma}(x)$ distribution is given in Reference [16] as

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

where

a) α is the specified some-outside rate per sample, i.e. the probability that one or more outliers in an outlier-free sample will be falsely labelled as an outlier;

b) $y_l = z_{l:n} - k_{\text{L}}\ (z_{u:n} - z_{l:n})$ and $y_u = z_{u:n} - k_{\text{U}}\ (z_{u:n} - z_{l:n})$;

c) $f_{Z_{l:n},Z_{u:n}}\ (z_{l:n}\ , z_{u:n})$ is the joint probability density function of $z_{l:n}$ and $z_{u:n}$ which takes the form

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

d) $Z_{r:n} = (X_{r:n} - \theta)/\sigma$ is the rth-order statistic of the standardized variable $Z = (X - \theta)/\sigma$ with distribution function $F(x)$;

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

f) $I_p(a,b) = \dfrac{1}{\mathrm{B}(a,b)}\displaystyle\int_0^p t^{a-1}(1-t)^{b-1}\mathrm{d}t$ is the incomplete beta function.

A direct search algorithm can be used to search for the values of k_{L} and k_{U} that satisfy the double integral Equation (C.1).

For a symmetric distribution, we take $k_l = k_u = (k)$ in Equation (C.1). For an asymmetric distribution, one can obtain the values of k_l and k_u separately by taking $P(X < L_F) = 1 - \Pr(X > U_F)$, i.e. $I_{G_l(y_l)}(1, l-1) = 1 - I_{G_u(y_u)}(n-u, 1)$ in Equation (C.1).

The values of $k_{\mathrm{L}} = k_{\mathrm{U}}$ $(= k)$ for samples of size $9 \leqslant n \leqslant 500$ taken from the standard normal distribution can be approximated from the following function

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

with coefficients $b_5 = 0$ and b_i, $i = 0, 1, 2, 3, 4$ given in Table C.1.

The values of k_{L} and k_{U} for samples taken from the asymmetric exponential and extreme-value distributions can also be obtained from Equation (C.2) with coefficients b_i, $i = 0, 1, 2, 3, 4, 5$ given in Table C.2.

For cases when the sample size is large, the values of k_{L} and k_{U} can be approximated as

$$k_{\mathrm{L}} \approx \frac{F^{-1}(1/4) - F^{-1}(\alpha_n/2)}{F^{-1}(3/4) - F^{-1}(1/4)} \text{ and } k_{\mathrm{U}} \approx \frac{F^{-1}(1-\alpha_n/2) - F^{-1}(3/4)}{F^{-1}(3/4) - F^{-1}(1/4)}$$

where $\alpha_n = 1 - (1 - \alpha)^{1/n}$ can be interpreted as the error rate that an observation from a random sample of n regular observations is falsely labelled as an outlier.

EXAMPLE 1 To detect outliers from a normal sample of size $n = 20$, the value of $k_{\mathrm{L}} = k_{\mathrm{U}}$ $(= k)$ for a some-outside rate of $\alpha = 0{,}05$ is evaluated as

$$k = \exp\left\{0{,}837\,07 + 0{,}075\,96 \times \ln(20) - 0{,}061\,19 \times \ln^2(20) + 0{,}013\,28 \times \ln^3(20) - 0{,}000\,83 \times \ln^4(20)\right\}$$

$$= \exp(0{,}805\,67) \approx 2{,}238\,2$$

EXAMPLE 2 To detect outliers from an exponential sample of size $n = 22$, the values of k_{L} and k_{U} for a some-outside rate of $\alpha = 0{,}05$ are evaluated as

$$k_{\mathrm{L}} = \exp\left\{2{,}206\,04 - 1{,}417\,52 \times \ln(20) + 0{,}241\,70 \times \ln^2(20) - 0{,}020\,57 \times \ln^3(20) + 0{,}000\,72 \times \ln^4(20)\right\}$$

$$= \exp(-0{,}408\,02) \approx 0{,}665\,0$$

$$k_{\mathrm{U}} = \exp\left\{2{,}741\,79 - 0{,}770\,67 \times \ln(22) + 0{,}226\,88 \times \ln^2(22) - 0{,}028\,53 \times \ln^3(22) + 0{,}001\,70 \times \ln^4(22) - 0{,}000\,04 \times \ln^5(22)\right\}$$

$$= \exp(1{,}829\,58) \approx 6{,}231\,3$$

Table C.1 — Coefficients of the fitted functions for the factors k of the box plot for samples of size $9 \leqslant n \leqslant 500$ taken from the normal distribution with parameters unknown

Normal distribution								
α	mod(n,4)	b_0	b_1	b_2	b_3	b_4	b_5	δ
0,05	1	4,017 61	−2,353 63	0,646 18	−0,078 93	0,003 68	—	0,014 57
	2	2,064 29	−0,885 23	0,222 37	−0,023 91	0,000 99	—	0,000 64
	3	0,480 06	0,258 54	−0,096 22	0,016 20	−0,000 92	—	0,004 07
	0	0,837 07	0,075 96	−0,061 19	0,013 28	−0,000 83	—	0,004 62
0,01	1	6,379 02	−3,847 70	1,044 38	−0,128 13	0,006 01	—	0,041 83
	2	3,987 72	−2,006 30	0,502 77	−0,056 77	0,002 48	—	0,006 34
	3	2,148 95	−0,652 78	0,119 85	−0,007 96	0,000 13	—	0,004 17
	0	2,285 07	−0,660 52	0,102 64	−0,003 93	−0,000 13	—	0,006 86

Table C.2 — Coefficients of the fitted functions for the factors k of the box plot for samples of size $9 \leqslant n \leqslant 500$ taken from the exponential distributions with parameters unknown

Exponential distribution									
α	factor	mod(n, 4)	b_0	b_1	b_2	b_3	b_4	b_5	δ
0,10	k_L	1	3,990 24	–3,240 52	0,955 34	–0,159 95	0,014 40	–0,000 54	0,000 22
		2	1,130 59	–0,721 69	0,023 06	0,018 04	–0,002 90	0,000 14	0,000 19
		3	–1,549 86	1,602 82	–0,825 26	0,178 01	–0,018 29	0,000 74	0,000 47
		0	–1,950 58	2,261 33	–1,147 44	0,249 30	–0,025 81	0,001 05	0,000 67
	k_U	1	3,585 01	–1,567 11	0,464 64	–0,057 69	0,002 71	—	0,021 72
		2	1,797 40	–0,223 67	0,076 84	–0,007 33	0,000 24	—	0,003 45
		3	0,332 62	0,834 29	–0,217 97	0,029 79	–0,001 53	—	0,011 54
		0	1,086 40	0,331 92	–0,086 35	0,013 96	–0,000 80	—	0,008 07
0,05	k_L	1	5,182 20	–4,055 28	1,222 29	–0,208 33	0,019 01	–0,000 72	0,000 33
		2	2,206 04	–1,417 52	0,241 70	–0,020 57	0,000 72	—	0,000 11
		3	–0,575 42	1,020 24	–0,656 89	0,150 43	–0,015 86	0,000 65	0,000 48
		0	–1,190 27	1,864 02	–1,044 28	0,233 27	–0,024 40	0,000 99	0,000 88
	k_U	1	5,180 29	–2,967 81	1,047 43	–0,185 11	0,016 83	–0,000 63	0,003 85
		2	2,741 79	–0,770 67	0,226 88	–0,028 53	0,001 70	–0,000 04	0,001 31
		3	0,530 26	1,198 59	–0,502 10	0,109 67	-0,011 58	0,000 48	0,005 44
		0	1,310 43	0,601 92	–0,303 96	0,074 56	–0,008 32	0,000 35	0,004 37
0,02	k_L	1	6,729 83	–5,174 48	1,605 18	–0,279 80	0,025 96	–0,000 99	0,000 52
		2	3,536 62	–2,310 42	0,530 46	–0,072 55	0,005 66	–0,000 19	0,000 06
		3	0,568 97	0,329 76	–0,455 63	0,117 23	–0,012 92	0,000 54	0,000 49
		0	–0,381 25	1,485 50	–0,962 54	0,223 51	–0,023 80	0,000 98	0,001 26
	k_U	1	5,904 97	–2,952 27	0,831 53	–0,103 10	0,004 86	—	0,069 00
		2	3,794 84	–1,328 56	0,353 93	–0,040 15	0,001 74	—	0,007 15
		3	2,171 27	–0,135 25	0,016 52	0,002 86	–0,000 33	—	0,012 78
		0	2,677 62	–0,439 84	0,088 73	–0,005 07	0,000 01	—	0,013 25

NOTE δ is the maximum absolute deviation between the original and the fitted values of k for each class of mod(n, 4) for $9 \leqslant n \leqslant 500$.

Annex D
(normative)

Values of the correction factors for the robust estimators of the scale parameter

Table D.1 — Correction factors s_n and s_{bi} of the robust scale estimators S_n and S_{bi} respectively

Sample size, n	Factor		Sample size, n	Factor	
	s_n	s_{bi}		s_n	s_{bi}
2	0,886 6	1,191 2	18	1,196 1	1,002 5
3	2,205 1	1,382 1	19	1,243 8	1,025 2
4	1,138 5	1,127 2	20	1,195 1	1,000 6
5	1,608 1	1,185 5	30	1,192 7	0,996 2
6	1,185 8	1,065 0	40	1,192 1	0,994 4
7	1,429 7	1,111 1	50	1,192 0	0,993 5
8	1,198 9	1,036 9	60	1,192 0	0,992 9
9	1,350 0	1,076 2	70	1,192 1	0,992 5
10	1,201 5	1,021 9	80	1,192 1	0,992 3
11	1,307 4	1,056 7	90	1,192 2	0,992 1
12	1,200 6	1,013 6	100	1,192 3	0,992 0
13	1,281 4	1,044 4	120	1,192 4	0,991 8
14	1,199 4	1,008 6	150	1,192 5	0,991 5
15	1,264 7	1,036 0	200	1,192 6	0,991 4
16	1,197 8	1,005 0	300	1,192 7	0,991 2
17	1,252 6	1,029 9	500	1,192 7	0,991 0

Annex E
(normative)

Critical values of Cochran's test statistic

Table E.1 — The 5 % critical values of Cochran's test statistic

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	0,3518	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
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	0,1257	0,108 1	0,096 7	0,088 7	0,082 6	0,077 9	0,074 1

NOTE 1 n is the number of replicate results per variance and p is the number of variances.

NOTE 2 The final decimal place of each table entry has been rounded upwards in order to guarantee the significance level.

NOTE 3 Each table entry is based on 50 million simulations.

Table E.2 — The 1 % critical values of Cochran's test statistic

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
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

NOTE 1 n is the number of replicate results per variance and p is the number of variances.

NOTE 2 The final decimal place of each table entry has been rounded upwards in order to guarantee the significance level.

NOTE 3 Each table entry is based on 50 million simulations.

Table E.3 — The 0,1 % critical values of Cochran's test statistic

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
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

NOTE 1 n is the number of replicate results per variance and p is the number of variances.

NOTE 2 The final decimal place of each table entry has been rounded upwards in order to guarantee the significance level.

NOTE 3 Each table entry is based on 50 million simulations.

Annex F
(informative)

A structured guide to detection of outliers in univariate data

A batch/sample of observations or a set of sample means or variances is available. The object is to detect and identify possible outliers in this data set. This annex is a guide for users of this part of ISO 16269. It leads the users through a number of stages utilizing the different clauses and subclauses of this part of ISO 16269. The notation follows that of this part of ISO 16269.

Step 1. Plot the given data set $x_1, x_2, \ldots, x_n$ graphically using dot plot, stem-and-leaf plot or standard box plot, or rank them numerically in ascending order

$$x_{(1)} \leqslant x_{(2)} \leqslant \cdots \leqslant x_{(k)} \leqslant \cdots \leqslant x_{(n)}$$

where $x_{(i)}$ is the ith smallest observation.

Step 2. Inspect the graphical plot or ranked data of the data set for outlying observations (suspected outliers). If the suspected observations are without doubt outliers, go to step 5. If one or more of the outlying observations is suspiciously far from the main part of the data set, go to step 3, or else declare that there are no outliers and use the given data set in subsequent data analysis.

Step 3. Confirm or transform the distribution of the given data set:

a) if the hypothesized distribution is a normal distribution, confirm it with a normal probability plot;

b) if the hypothesized distribution is an exponential distribution, confirm it with an exponential probability plot;

c) if the hypothesized distribution is a lognormal distribution, transform the given data set to resemble normal data using the procedure in 4.3.4.2, and subsequently confirm it with a normal probability plot;

d) if the hypothesized distribution is an extreme-value distribution, transform the given data set to resemble exponential data using the procedure in 4.3.4.3, and subsequently confirm it with an exponential probability plot;

e) if the hypothesized distribution is a Weibull distribution, transform the given data set to resemble exponential data using the procedure in 4.3.4.4, and subsequently confirm it with an exponential probability plot;

f) if the hypothesized distribution is a gamma distribution, transform the given data set to resemble normal data using the procedure in 4.3.4.5, and subsequently confirm it with a normal probability plot;

g) if the distribution of the given data set is unknown or the assumed distributions cannot be confirmed or it is not one of the above distributions, transform the data set to resemble normal data by using the Box-Cox or Johnson transformations, and subsequently confirm it with a normal probability plot. If the normality of the transformed data cannot be confirmed, then go to step 6 and conduct the data analysis using robust procedures discussed in Clause 5.

Step 4. Conduct a testing procedure(s) to determine whether the outlying observations identified in step 2 are outliers:

a) if the original or transformed data set resembles normal data, use the test procedure in 4.3.2 and/or 4.4;

b) if the original or transformed data set resembles exponential data, use the test procedures in 4.3.3 and/or 4.4.

If one or more outlying observations are declared to be outliers, go to step 5, otherwise declare that there are no outliers and use the original or transformed data set in subsequent data analysis;

Step 5. Identify causes of the declared outliers.

Step 6. If causes of outliers can be identified, remove the declared outliers and use the remaining data in subsequent data analysis, otherwise use robust procedures in subsequent data analysis.

The flow chart in Figure F.1 summarizes the recommended steps in the detection and treatment of outliers.

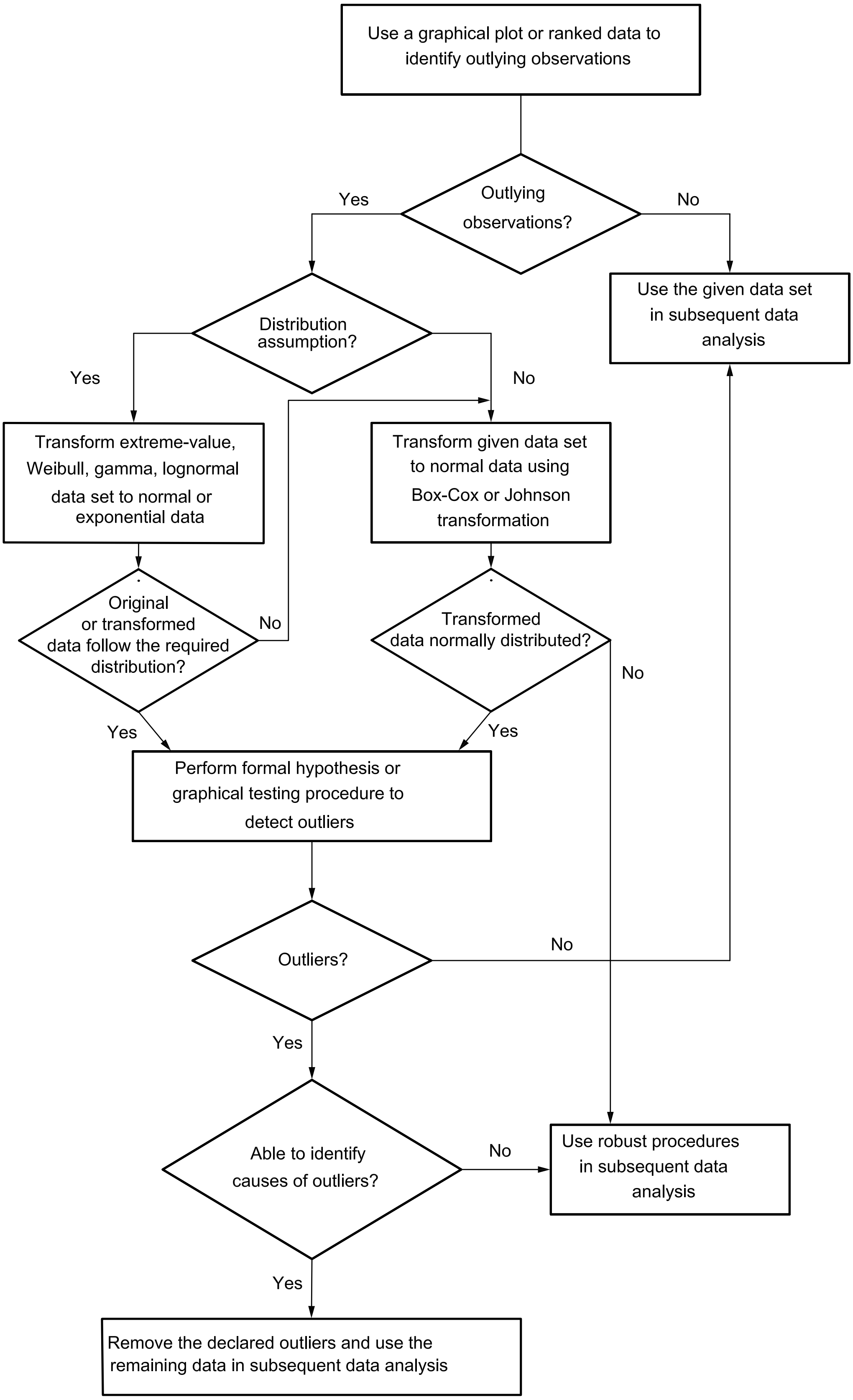

Use a graphical plot or ranked data to identify outlying observations
Outlying observations?
Yes
No
Use the given data set in subsequent data analysis
Distribution assumption?
Yes
No
Transform extreme-value, Weibull, gamma, lognormal data set to normal or exponential data
Transform given data set to normal data using Box-Cox or Johnson transformation
Original or transformed data follow the required distribution?
No
Transformed data normally distributed?
No
Yes
Yes
Perform formal hypothesis or graphical testing procedure to detect outliers
Outliers?
No
Yes
Able to identify causes of outliers?
No
Use robust procedures in subsequent data analysis
Yes
Remove the declared outliers and use the remaining data in subsequent data analysis

Figure F.1 — Flow chart for the detection and treatment of outliers

Bibliography

[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*

ICS 03.120.30

Price based on 54 pages