



SLOVENSKI STANDARD
SIST ISO 16269-4:2014

01-januar-2014

Statistično tolmačenje podatkov - 4. del: Zaznavanje in obravnava osamelcev

Statistical interpretation of data - Part 4: Detection and treatment of outliers

Interprétation statistique des données - Partie 4: Détection et traitement des valeurs aberrantes

(standards.iteh.ai)

Ta slovenski standard je istoveten z: ISO 16269-4:2010

<https://standards.iteh.ai/catalog/standards/sist/5d77af0d-2eea-4faa-a649-8143b7c76500/sist-iso-16269-4-2014>

ICS:

03.120.30	Uporaba statističnih metod	Application of statistical methods
-----------	----------------------------	------------------------------------

SIST ISO 16269-4:2014

en

iTeh STANDARD PREVIEW
(standards.iteh.ai)

[SIST ISO 16269-4:2014](#)

<https://standards.iteh.ai/catalog/standards/sist/5d77af0d-2eea-4faa-a649-8143b7c76500/sist-iso-16269-4-2014>

**Statistical interpretation of data —
Part 4:
Detection and treatment of outliers**

Interprétation statistique des données —

Partie 4: Détection et traitement des valeurs aberrantes

iTeh STANDARD PREVIEW
(standards.iteh.ai)

[SIST ISO 16269-4:2014](https://standards.iteh.ai/catalog/standards/sist/5d77af0d-2eea-4faa-a649-8143b7c76500/sist-iso-16269-4-2014)

<https://standards.iteh.ai/catalog/standards/sist/5d77af0d-2eea-4faa-a649-8143b7c76500/sist-iso-16269-4-2014>



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.

iTeh STANDARD PREVIEW (standards.iteh.ai)

[SIST ISO 16269-4:2014](https://standards.iteh.ai/catalog/standards/sist/5d77af0d-2eea-4faa-a649-8143b7c76500/sist-iso-16269-4-2014)

<https://standards.iteh.ai/catalog/standards/sist/5d77af0d-2eea-4faa-a649-8143b7c76500/sist-iso-16269-4-2014>

**COPYRIGHT PROTECTED DOCUMENT**

© ISO 2010

All rights reserved. Unless otherwise specified, no part of this publication may be reproduced or utilized in any form or by any means, electronic or mechanical, including photocopying and microfilm, without permission in writing from either ISO at the address below or ISO's member body in the country of the requester.

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	iv
Introduction.....	v
1 Scope	1
2 Terms and definitions	1
3 Symbols.....	10
4 Outliers in univariate data	11
4.1 General	11
4.1.1 What is an outlier?	11
4.1.2 What are the causes of outliers?	11
4.1.3 Why should outliers be detected?	11
4.2 Data screening.....	12
4.3 Tests for outliers	14
4.3.1 General	14
4.3.2 Sample from a normal distribution.....	14
4.3.3 Sample from an exponential distribution.....	16
4.3.4 Samples taken from some known non-normal distributions.....	18
4.3.5 Sample taken from unknown distributions.....	19
4.3.6 Cochran's test for outlying variance	21
4.4 Graphical test of outliers.....	22
5 Accommodating outliers in univariate data.....	23
5.1 Robust data analysis.....	23
5.2 Robust estimation of location.....	24
5.2.1 General	24
5.2.2 Trimmed mean	24
5.2.3 Biweight location estimate	25
5.3 Robust estimation of dispersion	25
5.3.1 General	25
5.3.2 Median-median absolute pair-wise deviation	25
5.3.3 Biweight scale estimate	26
6 Outliers in multivariate and regression data	26
6.1 General	26
6.2 Outliers in multivariate data	26
6.3 Outliers in linear regression.....	28
6.3.1 General	28
6.3.2 Linear regression models.....	29
6.3.3 Detecting outlying Y observations.....	31
6.3.4 Identifying outlying X observations.....	31
6.3.5 Detecting influential observations.....	32
6.3.6 A robust regression procedure.....	35
Annex A (informative) Algorithm for the GESD outliers detection procedure	36
Annex B (normative) Critical values of outliers test statistics for exponential samples	37
Annex C (normative) Factor values of the modified box plot	44
Annex D (normative) Values of the correction factors for the robust estimators of the scale parameter	47
Annex E (normative) Critical values of Cochran's test statistic	48
Annex F (informative) A structured guide to detection of outliers in univariate data	51
Bibliography.....	54

ISO 16269-4:2010(E)

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.

[SIST ISO 16269-4:2014](https://standards.iteh.ai/catalog/standards/sist/5d77af0d-2eea-4faa-a649-8143b7c76500/sist-iso-16269-4-2014)

<https://standards.iteh.ai/catalog/standards/sist/5d77af0d-2eea-4faa-a649-8143b7c76500/sist-iso-16269-4-2014>

iTeh STANDARD PREVIEW
(standards.iteh.ai)

[SIST ISO 16269-4:2014](#)

<https://standards.iteh.ai/catalog/standards/sist/5d77af0d-2eea-4faa-a649-8143b7c76500/sist-iso-16269-4-2014>

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

SIST ISO 16269-4:2014
<https://standards.iteh.ai/catalog/standards/sist/5d77af0d-2eea-4faa-a649-8143b7c76500/sist-iso-16269-4-2014>

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

ISO 16269-4:2010(E)

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 $\{x_1, x_2, \dots, x_n\}$. Reorder the observed values in non-decreasing order designated as $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(k)} \leq \dots \leq x_{(n)}$; then $x_{(k)}$ is the observed value of the k th 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.

iTeh STANDARD PREVIEW
(standards.iteh.ai)

ISO 16269-4:2010(E)

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.

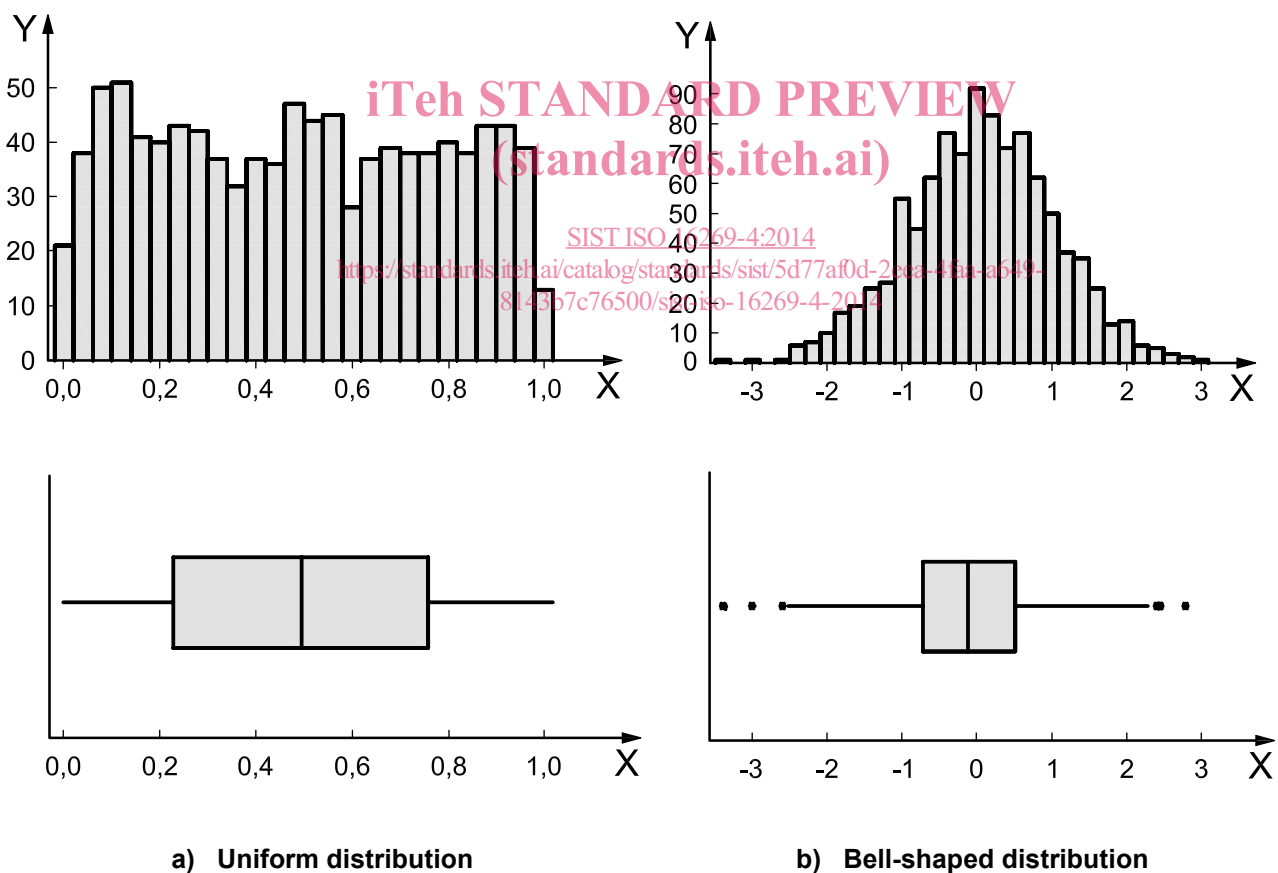
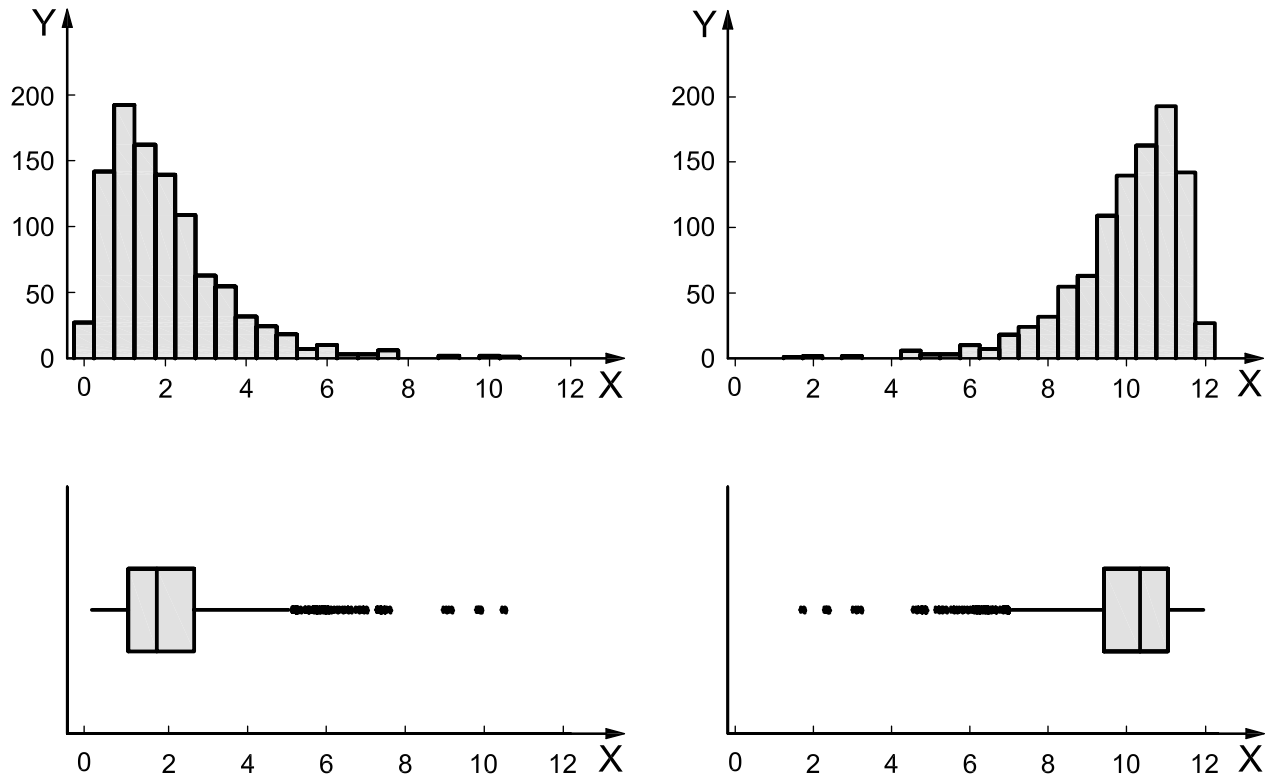


Figure 1 (continued)



c) Right-skewed distribution (standards.itech.ai) d) Left-skewed distribution

Key

X data values

Y frequency

SIST ISO 16269-4:2014
<https://standards.itech.ai/catalog/standards/sist/5d77af0d-2eea-4faa-a649-8143b7c76500/sist-iso-16269-4-2014>

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

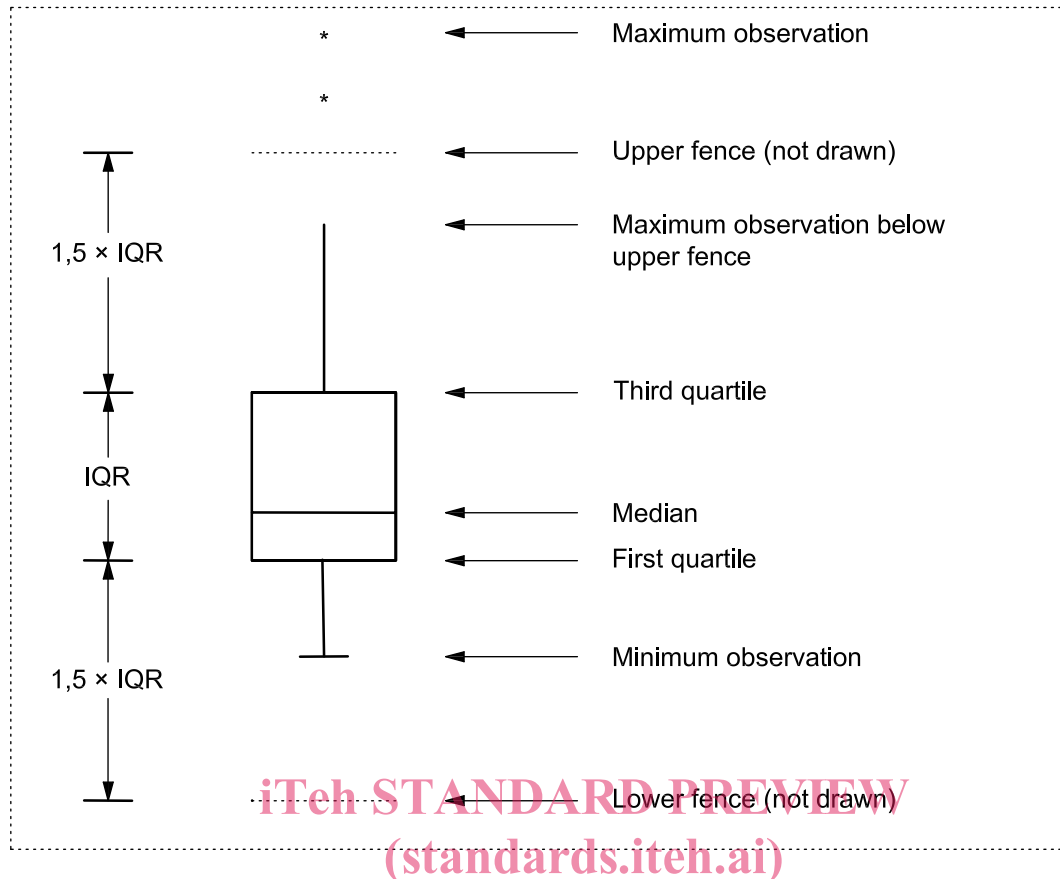


Figure 2 — Modified box plot with lower and upper fences

<https://standards.iteh.ai/catalog/standards/sist/5d77af0d-2eea-4faa-a649-8143b7c76500/sist-iso-16269-4-2014>

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** $x_{L:n}$

for a set $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$ 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	$[x_{(3)} + x_{(4)}]/2$	$[x_{(3)} + x_{(4)}]/2$

iTeh STANDARD PREVIEW
(standards.iteh.ai)

2.20**upper fourth** $x_{U:n}$

for a set $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$ 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	$[x_{(9)} + x_{(10)}]/2$	$[x_{(9)} + x_{(10)}]/2$