Outliers

An outlier is an observation clearly differing from other observations of the same group, and thus able to alter one or more parameters like mean, variance and symmetry.

The search for outliers may have different goals:

- ✓ estimating the real mean or variance of a phenomenon, once the outlier(s) has(have) been eliminated
- ✓ studying the causes that led to their generation.

The difficulty in the identification of an outlier is due to at least three reasons:

1) presence of masking effects: sometimes an outlier can be “masked” by the presence of another one;
2) dependence of the outlier recognition on the sample dimension
3) incorrect hypothesis on the data distribution.

As an example, let us consider the following data series:

97	98	98	95	86	99
98	98	97	99	98	95

If only the first row is considered, statistical parameters are significantly different if the datum 86 is included or not.

If all the 12 data are considered, mean and standard deviation become closer to those obtained for the first row when the datum 86 is excluded:

Data	Row 1		Rows 1 + 2
	Without 86	With 86	
Dimension n	5	6	12
Mean	97.4	95.5	96.5
Standard dev.	1.5	4.8	3.5
Range (max. diff.)	4	11	11
Median	98	97.5	98

It would thus be less likely to consider datum 86 as an outlier when the sample size is increased.

As a further example, if the following series is obtained for the counts of insects collected in a trap:

3	3	4	5	7	11
12	15	18	24	51	54
84	120	560			

Datum 560 would likely be considered as an outlier by almost all statistical tests for outlier recognition, since most of the latter assume a normal distribution.
However, the distribution of insect counts might be highly asymmetric (*e.g.*, because of the possibility of collecting swarms occasionally).

Great care must then be posed before discarding an observation as an outlier.

Among the several tests available for outlier recognition, the following will be considered in next slides:

1) Dixon's Q-test
2) Grubbs' test
3) Tukey's Box-and-Whiskers plot (also called box-plot)
4) Median Absolute Deviation (MAD)

Dixon's Q test for a single outlier

Dixon's Q test for a single outlier, that is one of a set of tests proposed by the American statistician Wilfrid Joseph Dixon in the 1950s, can be applied to small ($n \leq 30$) and normally-distributed data sets.

The test requires that data are ordered in increasing or decreasing order, based on the tail where the suspected datum is located. Afterwards, the following ratio is calculated and then compared to appropriate critical values:

$$r_{10} = \frac{X_2 - X_1}{X_n - X_1}$$

Here the subscript used for each X value is referred to its order in the series, thus the suspect datum is X_1.

In the notation adopted by Dixon for r_{xy} values, x indicates the number of suspected outliers on the same end of the dataset as the value being tested, whereas y indicates the number of possible outliers on the opposite end of the data set.

r_{10} thus represents the statistic to be adopted when the presence of only one outlier is suspected.

Different compilations of critical values to be compared to Dixon's r_{10} parameter are present in the statistical literature.

Sometimes, discrepancies can be found between different sources, due to:

1) an erroneous use of values calculated for a one-tailed test when a two-tailed test is actually employed

2) small differences in interpolations performed to find critical values for specific n values.

One of the most recently checked set of critical values for different confidence levels and sample sizes is that reported by David D. Rorabacher in a paper published on *Analytical Chemistry* (63, 1991, 139-146), shown as a table in the next slide.

The set includes data for sample sizes up to n = 30, since the Q-test can be used reliably even for such a sample size, provided that the presence of only one outlier is suspected.

Critical values reported in the following table should not be used if the presence of more than one outlier is suspected.

Table I. Critical Values of Dixon's r_{10} (Q) Parameter As Applied to a Two-Tailed Test at Various Confidence Levels, Including the 95% Confidence Level[a]

	confidence level					
N[b]	80% (α = 0.20)	90% (α = 0.10)	**95%** (α = 0.05)	96% (α = 0.04)	98% (α = 0.02)	99% (α = 0.01)
3	0.886	0.941	**0.970**	0.976	0.988	0.994
4	0.679	0.765	**0.829**	0.846	0.889	0.926
5	0.557	0.642	**0.710**	0.729	0.780	0.821
6	0.482	0.560	**0.625**	0.644	0.698	0.740
7	0.434	0.507	**0.568**	0.586	0.637	0.680
8	0.399	0.468	**0.526**	0.543	0.590	0.634
9	0.370	0.437	**0.493**	0.510	0.555	0.598
10	0.349	0.412	**0.466**	0.483	0.527	0.568
11	0.332	0.392	**0.444**	0.460	0.502	0.542
12	0.318	0.376	**0.426**	0.441	**0.482**	0.522
13	0.305	0.361	**0.410**	0.425	0.465	0.503
14	0.294	0.349	**0.396**	0.411	0.450	0.488
15	0.285	0.338	**0.384**	0.399	0.438	0.475
16	0.277	0.329	**0.374**	0.388	0.426	0.463
17	0.269	0.320	**0.365**	0.379	0.416	0.452
18	0.263	0.313	**0.356**	0.370	0.407	0.442
19	0.258	0.306	**0.349**	0.363	0.398	0.433
20	0.252	0.300	**0.342**	0.356	0.391	0.425
21	0.247	0.295	**0.337**	0.350	0.384	0.418
22	0.242	0.290	**0.331**	0.344	0.378	0.411
23	0.238	0.285	**0.326**	0.338	0.372	0.404
24	0.234	0.281	**0.321**	0.333	0.367	0.399
25	0.230	0.277	**0.317**	0.329	0.362	0.393
29	0.227	0.273	**0.312**	0.324	0.357	0.388
27	0.224	0.269	**0.308**	0.320	0.353	0.384
28	0.220	0.266	**0.305**	0.316	0.349	0.380
29	0.218	0.263	**0.301**	0.312	0.345	0.376
30	0.215	0.260	**0.298**	0.309	0.341	0.372

[a] In this and the other accompanying tables, the newly generated or corrected values are indicated in boldface. [b] Sample size.

Numerical example of Dixon's Q-test for a single outlier

The following six data have been drawn from a normally-distributed variable:

0.505	0.511	0.519	0.478	0.357	0.506

The test is applied to check if value 0.357 is an outlier.

In this case the suspect datum is the lowest, thus data are ordered in increasing order:

X_1	X_2	X_3	X_4	X_5	X_6
0.357	0.478	0.505	0.506	0.511	0.519

Since n = 6, the following calculation is done:

$$r_{10} = \frac{X_2 - X_1}{X_6 - X_1} = \frac{0{,}478 - 0{,}357}{0{,}519 - 0{,}357} = \frac{0{,}121}{0{,}162} = 0{,}747$$

In the Dixon's Q-test r_{10} critical values table, shown in the previous slide, the critical value for n = 6 is equal to 0.740 for α = 0.01 (the lowest value for which critical values are usually reported). Since 0.747 > 0.740, datum 0.357 is considered an outlier at 99% confidence.

Dixon's Q-test when more than one outlier is present

In its research Dixon considered also cases in which more than one outlier could be present, on one or on both tails of the dataset.

Appropriate parameters, to be compared with further set of critical values, were then defined:

one outlier on both tails $$r_{11} = \frac{x_2 - x_1}{x_{n-1} - x_1}$$

one outlier on the tail under evaluation and two on the other tail $$r_{12} = \frac{x_2 - x_1}{x_{n-2} - x_1}$$

two outliers on the tail under evaluation $$r_{20} = \frac{x_3 - x_1}{x_n - x_1}$$

two outliers on the tail under evaluation and one on the other tail $$r_{21} = \frac{x_3 - x_1}{x_{n-1} - x_1}$$

two outliers on both tails $$r_{22} = \frac{x_3 - x_1}{x_{n-2} - x_1}$$

As apparent, if x_1 is the datum under evaluation as a possibile outlier the nearest neighbouring value in the dataset to be considered is x_2 if this is not a potential outlier, otherwise x_3 is considered.

By analogy, the last value in the dataset, x_n, can be considered only if it is not a potential outlier itself, like for r_{10} or r_{20} parameters.
It this is not the case, the last but one, x_{n-1}, or even the last but two, x_{n-2}, value has to be considered in the calculation.

In its treatment Dixon generated critical values for all the parameters cited in the previous slide and for all sample sizes from 4-5 to 30, according to the case:

A summary of those values for a two-tailed Dixon's Q test at a 95% confidence level, taken from the Rorabacher's paper, is reported in the figure:

N	r_{11}	r_{12}	r_{20}	r_{21}
4	0.977		0.983	
5	0.863	0.980	0.890	0.987
6	0.748	0.878	0.786	0.913
7	0.673	0.773	0.716	0.828
8	0.615	0.692	0.657	0.763
9	0.570	0.639	0.614	0.710
10	0.534	0.594	0.579	0.664
11	0.505	0.559	0.551	0.625
12	0.481	0.529	0.527	0.592
13	0.461	0.505	0.506	0.565
14	0.445	0.485	0.489	0.544
15	0.430	0.467	0.473	0.525
16	0.417	0.452	0.460	0.509
17	0.406	0.438	0.447	0.495
18	0.396	0.426	0.437	0.482
19	0.386	0.415	0.427	0.469
20	0.379	0.405	0.418	0.460
21	0.371	0.396	0.410	0.450
22	0.364	0.388	0.402	0.441
23	0.357	0.381	0.395	0.434
24	0.352	0.374	0.390	0.427
25	0.346	0.368	0.383	0.420
26	0.341	0.362	0.379	0.414
27	0.337	0.357	0.374	0.407
28	0.332	0.352	0.370	0.402
29	0.328	0.347	0.365	0.396
30	0.324	0.343	0.361	0.391

Grubbs's test

Grubbs's test, proposed by the American statistician Frank Ephraim Grubb in 1950, is another approach for the identification of outliers in relatively small, normally-distributed samples.

As for the Dixon's test, data have to be ordered and one of the following statistics has to be used, implying the calculation of the sampling mean and standard deviation:

$$T = \frac{\overline{X} - X_1}{S}$$ when the potential outlier is the first datum in the sample

$$T = \frac{X_n - \overline{X}}{S}$$ when the potential outlier is the last datum in the sample

The realization of the appropriate statistic is then compared to a critical value, depending, as usual, on the significance level α and on the sample size.

A table of critical values referred to a unilateral test, *i.e.*, a test performed when the tail in which the potential outlier is located is known, is shown in the next slide.

n	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.025$	$\alpha = 0.01$	$\alpha = 0.005$	n
3	1.148	1.153	1.155	1.155	1.155	3
4	1.425	1.463	1.481	1.492	1.496	4
5	1.602	1.672	1.715	1.749	1.764	5
6	1.729	1.822	1.887	1.944	1.973	6
7	1.828	1.938	2.020	2.097	2.139	7
8	1.909	2.032	2.126	2.221	2.274	8
9	1.977	2.110	2.215	2.323	2.387	9
10	2.036	2.176	2.290	2.410	2.482	10
11	2.088	2.234	2.355	2.485	2.564	11
12	2.134	2.285	2.412	2.550	2.636	12
13	2.175	2.331	2.462	2.607	2.699	13
14	2.213	2.371	2.507	2.659	2.755	14
15	2.247	2.409	2.549	2.705	2.806	15
16	2.279	2.443	2.585	2.747	2.852	16
17	2.309	2.475	2.620	2.785	2.894	17
18	2.335	2.504	2.651	2.821	2.932	18
19	2.361	2.532	2.681	2.854	2.968	19
20	2.385	2.557	2.709	2.884	3.001	20
21	2.408	2.580	2.733	2.912	3.051	21
22	2.429	2.603	2.758	2.939	3.060	22
23	2.448	2.624	2.781	2.963	3.087	23
24	2.467	2.644	2.802	2.987	3.112	24
25	2.486	2.663	2.822	3.009	3.135	25
26	2.502	2.681	2.841	3.029	3.157	26
27	2.519	2.698	2.859	3.049	3.178	27
28	2.534	2.714	2.876	3.068	3.199	28
29	2.549	2.730	2.893	3.085	3.218	29
30	2.563	2.745	2.908	3.103	3.236	30

Numerical example of Grubbs's test

The following data set has been obtained after a series of 15 measurements:

99,3	99,7	98,6	99,0	99,1	99,3	99,5	98,0	98,9	99,4	99,0	99,4	99,2	98,8	99,2

Grubbs's test can be used to evaluate if value 98.0, the first in the set, in increasing order, is an outlier.

In this case: $\overline{X} = 99{,}09$ and $S = 0{,}41$

then: $$T = \frac{\overline{X} - X_1}{S} = \frac{99{,}09 - 98{,}00}{0{,}41} = \frac{1{,}09}{0{,}41} = 2{,}66$$

The critical value for n = 15 is 2.549, for α = 0.025, and 2.705, for α = 0.01.

Grubbs's test indicates that the value 98.0 is an outlier at a significance level of 2.5% but not at a significance level of 1%.

Notably, the Grubbs's test can be used also when the presence of more than one outlier is suspected.

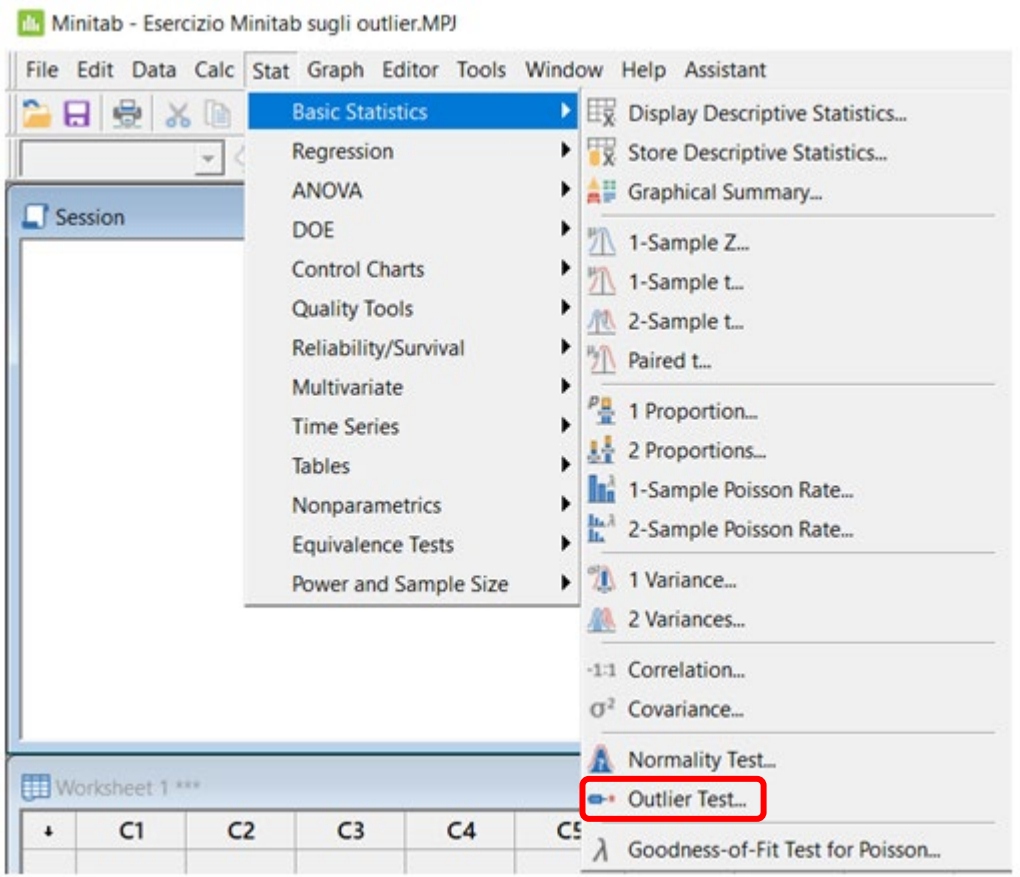
Use of Minitab 18 for outlier detection

The Minitab 18 software enables the detection of outliers based on the Dixon's and on the Grubbs's tests, once data are transferred into the program Worksheet.
The two tests can be accessed from the Basic Statistics sub-menu of the Stat menu:

Minitab® 18

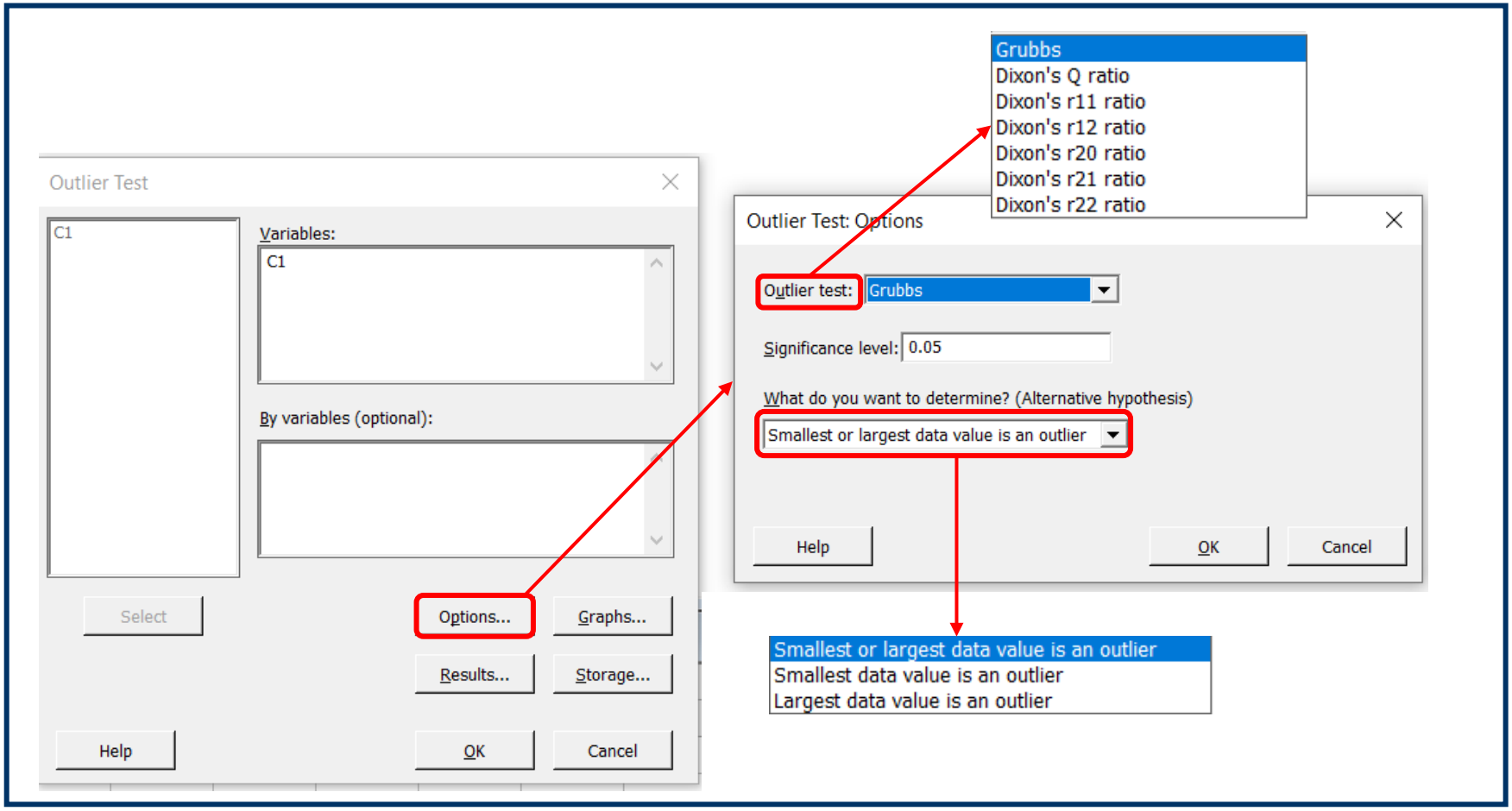
Worksheet 1 ***

	C1
1	99.7
2	98.6
3	99.0
4	99.1
5	99.3
6	99.5
7	98.0
8	98.9
9	99.4
10	99.0
11	99.2
12	98.8
13	99.2



Once the Outlier Test window is opened the worksheet column including data to be checked is selected, then the Options… button has to be pressed.

A new window is opened, where the type of test, significance level and alternative hypothesis (*i.e.*, the hypothesis opposite to the null one, corresponding to the absence of outliers) can be chosen:



Note that all the different types of Dixon's test described before can be performed.

The output can be read in the program window named Session and can be also reported as a plot:

Outlier Test: C1

Method

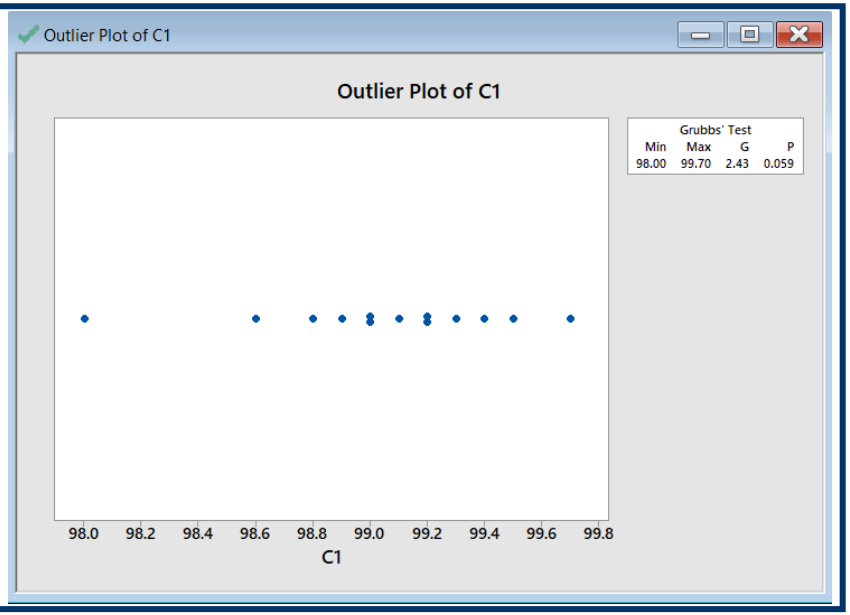
Null hypothesis	All data values come from the same normal population
Alternative hypothesis	Smallest or largest data value is an outlier
Significance level	α = 0.05

Grubbs' Test

Variable	N	Mean	StDev	Min	Max	G	P
C1	13	99.054	0.433	98.000	99.700	2.43	0.059

* NOTE * No outlier at the 5% level of significance

Outlier Plot of C1



Outlier Test: C1

Method

Null hypothesis	All data values come from the same normal population
Alternative hypothesis	Smallest or largest data value is an outlier
Significance level	α = 0.05

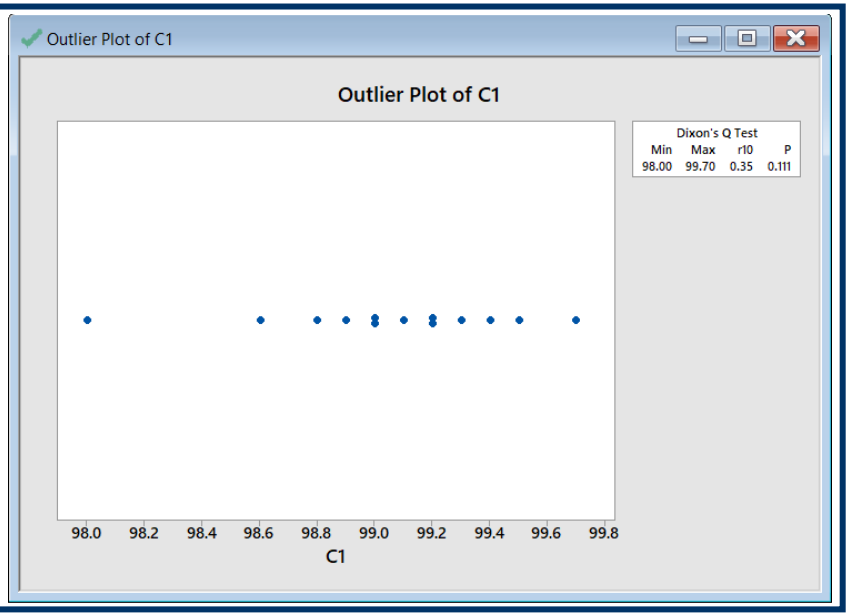
Dixon's Q Test

Variable	N	Min	x[2]	x[N-1]	Max	r10	P
C1	13	98.000	98.600	99.500	99.700	0.35	0.111

x[i] denotes the ith smallest observation.

* NOTE * No outlier at the 5% level of significance

Outlier Plot of C1



It is worth noting that the outcome of the two tests can be inferred indirectly also from the P value reported at the end of the respective tables:

Grubbs' Test

Variable	N	Mean	StDev	Min	Max	G	P
C1	13	99.054	0.433	98.000	99.700	2.43	0.059

Dixon's Q Test

Variable	N	Min	x[2]	x[N-1]	Max	r10	P
C1	13	98.000	98.600	99.500	99.700	0.35	0.111

x[i] denotes the ith smallest observation.

Indeed, if the P value is higher than the significance level (in this case 0.05, *i.e.*, 5%) the null hypothesys (no outlier is present) is accepted.
If the P value is lower than the significance level the alternative hypothesys (an outlier is present) is accepted.

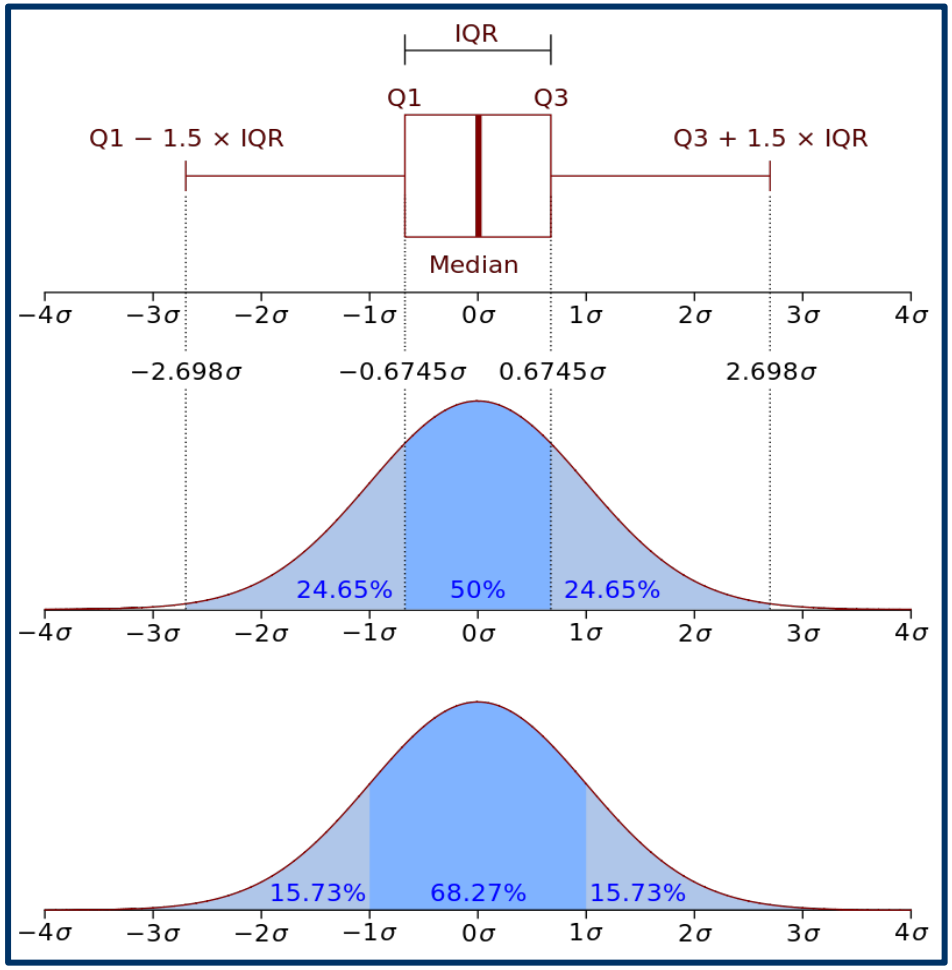
This way of reporting a statistical test's outcome is very common and will be encountered many times in future lessons.

Tukey's Box-and-Whiskers plot

Tukey's Box-and-Whiskers plot, introduced by the American mathematician John Tukey in 1970, is one of the most typical examples of box-plots, *i.e.*, graphical representations of the dispersion of a dataset with respect to the median and to quartiles, enabling a relatively easy evaluation of symmetry and of the presence of eventual outliers.

A rectangle including the median value and extending from the first (Q1) to the third (Q3) quartile is drawn as the «box».
«Whiskers» are segments extending externally from Q1 and Q3, with a length usually corresponding to 1.5 times the inter-quartile range (IQR = Q3 - Q1).

In the figure, the comparison between typical representations of a normal probability density function and the Tukey's box-and-whiskers plot is evidenced.
In this case the box is obviously symmetric with respect to the median.

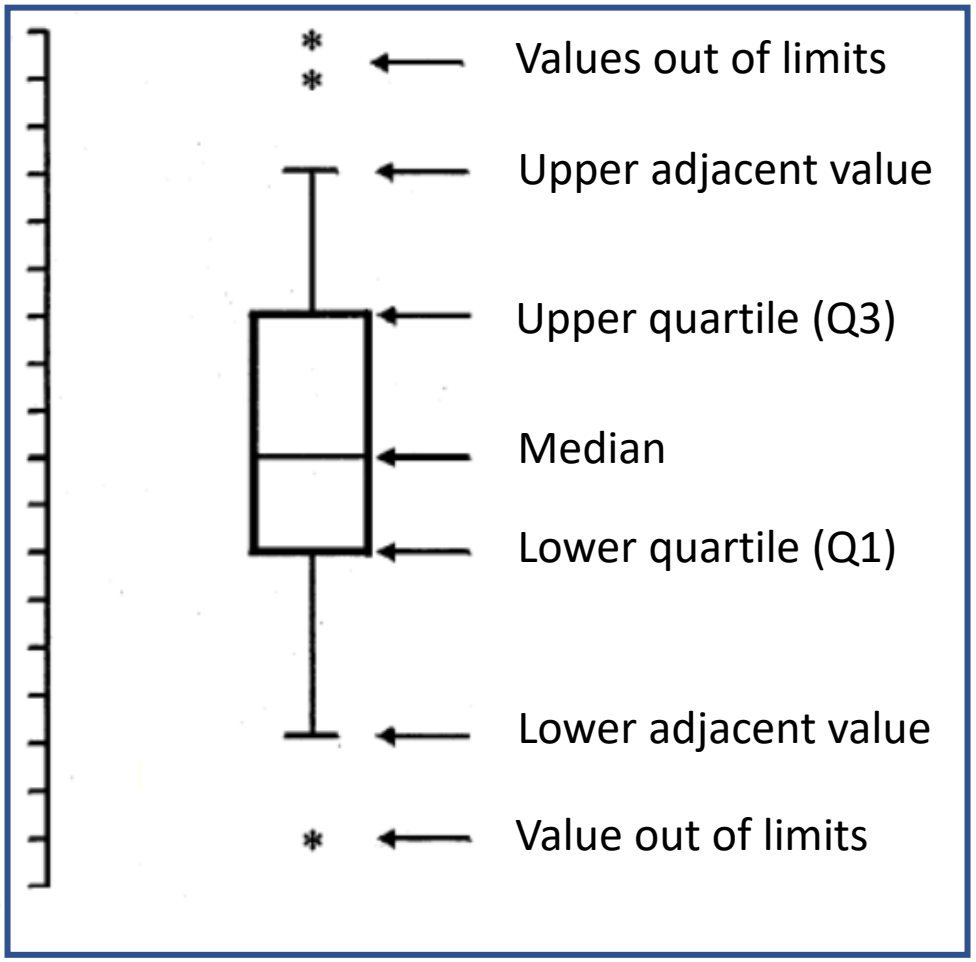


It is worth noting that lower and upper adjacent values represent values located at Q1 – 1.5 IQR and Q3 + 1.5 IQR, respectively.

Moreover, data suspected to be outliers can be drawn as individual points or asterisks, located outside the whiskers.

Alternative Box-and-Whiskers plot can be drawn with whiskers extending to the minimum and maximum observed values.

The eventual asymmetry of data distribution is easily inferred from the Box-and-Whiskers plot, since in this case the box is not symmetric with respect to the median.



It is important to point out that no assumption on the underlying statistical distribution is made when a Tukey's Box-and-Whiskers plot is drawn.

Consequently, by definition, this plot is an example of non-parametric approach.

An example of Box-and-whiskers plot

Let us suppose that the following dataset (n = 20) was obtained:

60, 69, 28, 51, 112, 80, 73, 103, 40, 47, 58, 58, 74, 56, 64, 68, 56, 54, 63, 60

Once data are re-ordered in ascending order, the following set is obtained, with intervals representing quartile intervals drawn with different colors:

28, 40, 47, 51, 54, 56, 56, 58, 58, 60, 60, 63, 64, 68, 69, 73, 74, 80, 103, 112

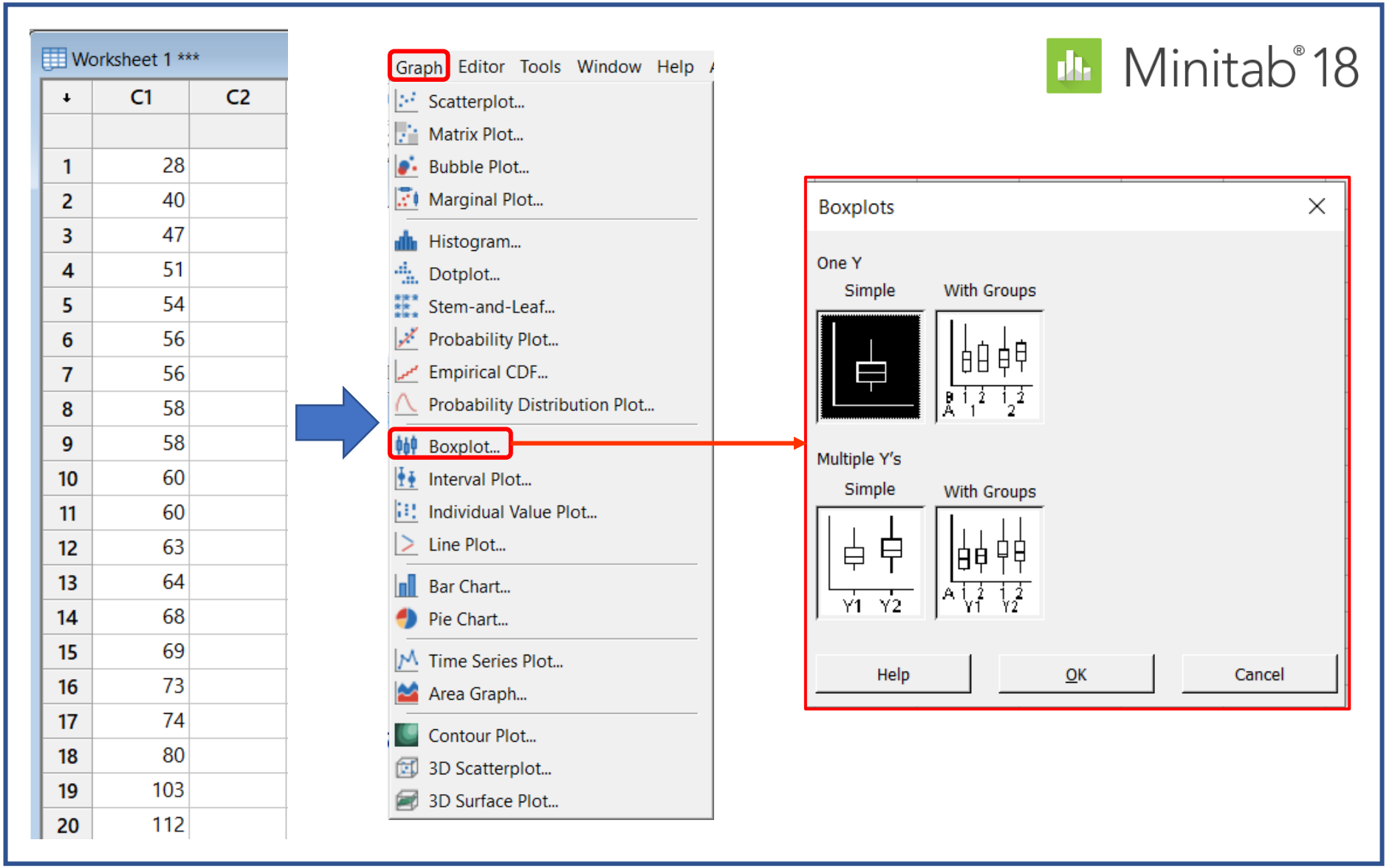
These are the parameters required to draw the corresponding box-and-whiskers plot:

Median = 60
First quartile (Q1) is comprised between 54 and 56 $\Rightarrow$ (54 + 56) / 2 = 55
Third quartile (Q3) is comprised between 69 and 73 $\Rightarrow$ (69 + 73) / 2 = 71
Inter-quartile range (IQR) = 71 – 55 = 16
Lower adjacent value = Q1 – 1.5 IQR = 55 – 24 = 31
Upper adjacent value = Q3 + 1.5 IQR = 71 + 24 = 95

Based on the lower and upper adjacent values, datum 28, on the lower side of the dataset, and data 103 and 112, on the upper side of the dataset, can be considered as outliers.

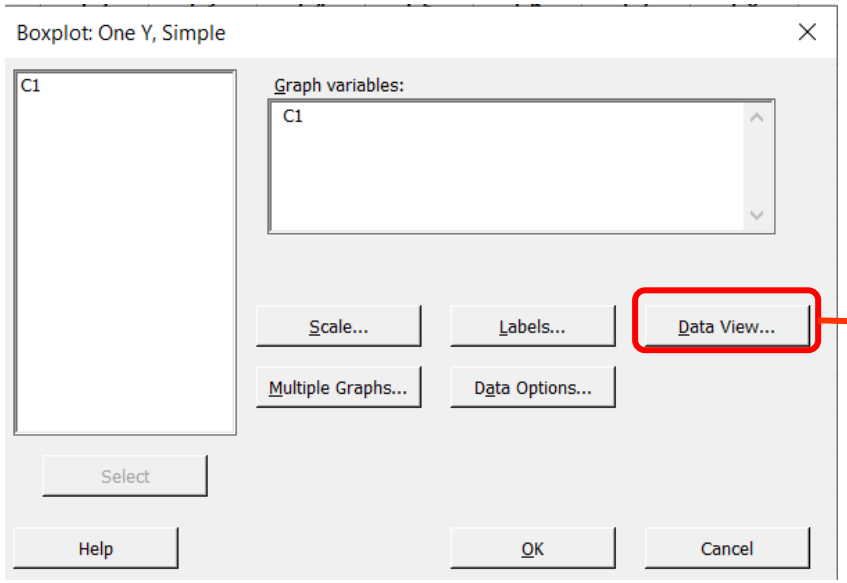
The Minitab 18 software can be used to rapidly draw a Box-and-Whiskers plot.

The procedure starts by inserting data in the program Worksheet, then the Boxplot… option in the Graph menu is selected. The Simple option for Boxplots is selected in this case.

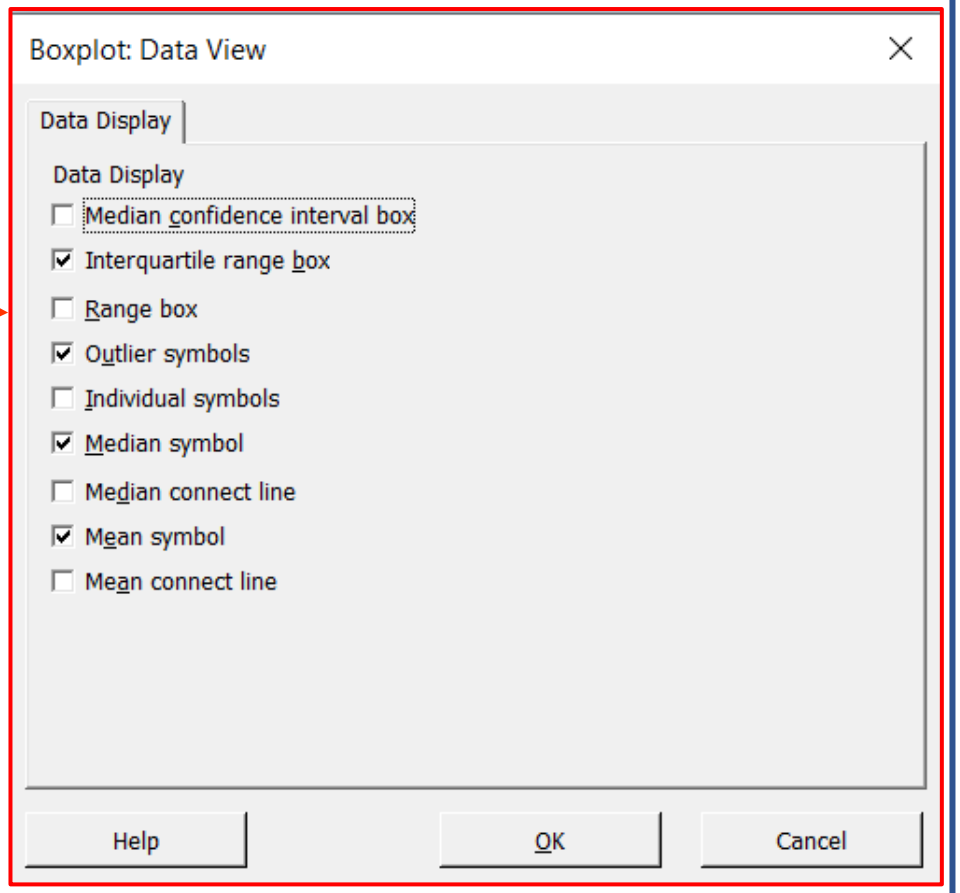


Several parameters can be set inside the Minitab 18's Simple boxplot window.

The most important ones are accessible through the Data View sub-menu:

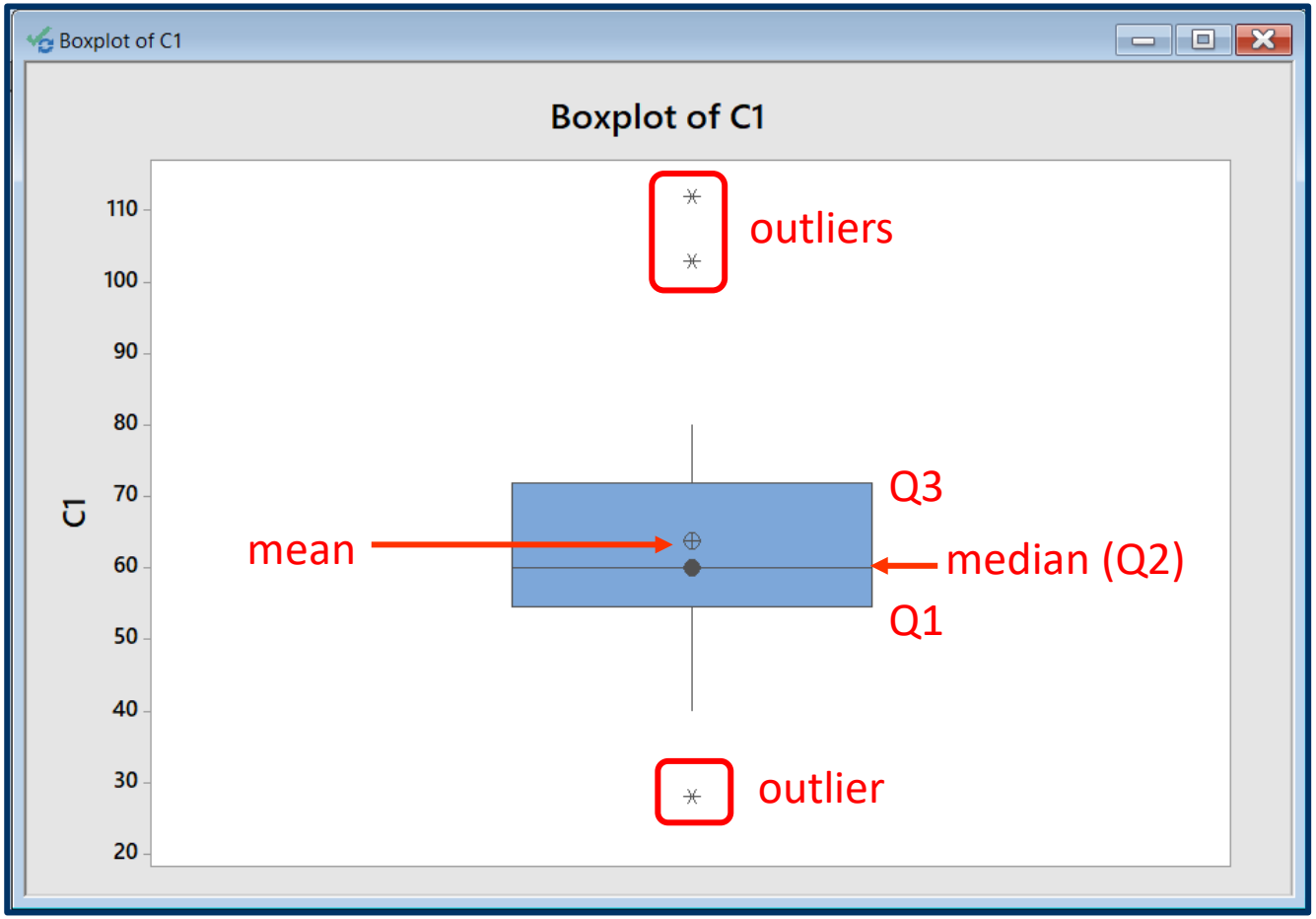


C1 is the only column including values in the Worksheet, thus it has to be selected in the Graph variables window.



In the figure, most typical settings for Box-and-Whiskers plots are selected, *i.e.*, the box represents the Interquartile (Q3-Q1) range and outliers are represented individually (the *Outlier symbols* option is selected). The appearance in the plot of Median and Mean symbols is also selected in the specific case.

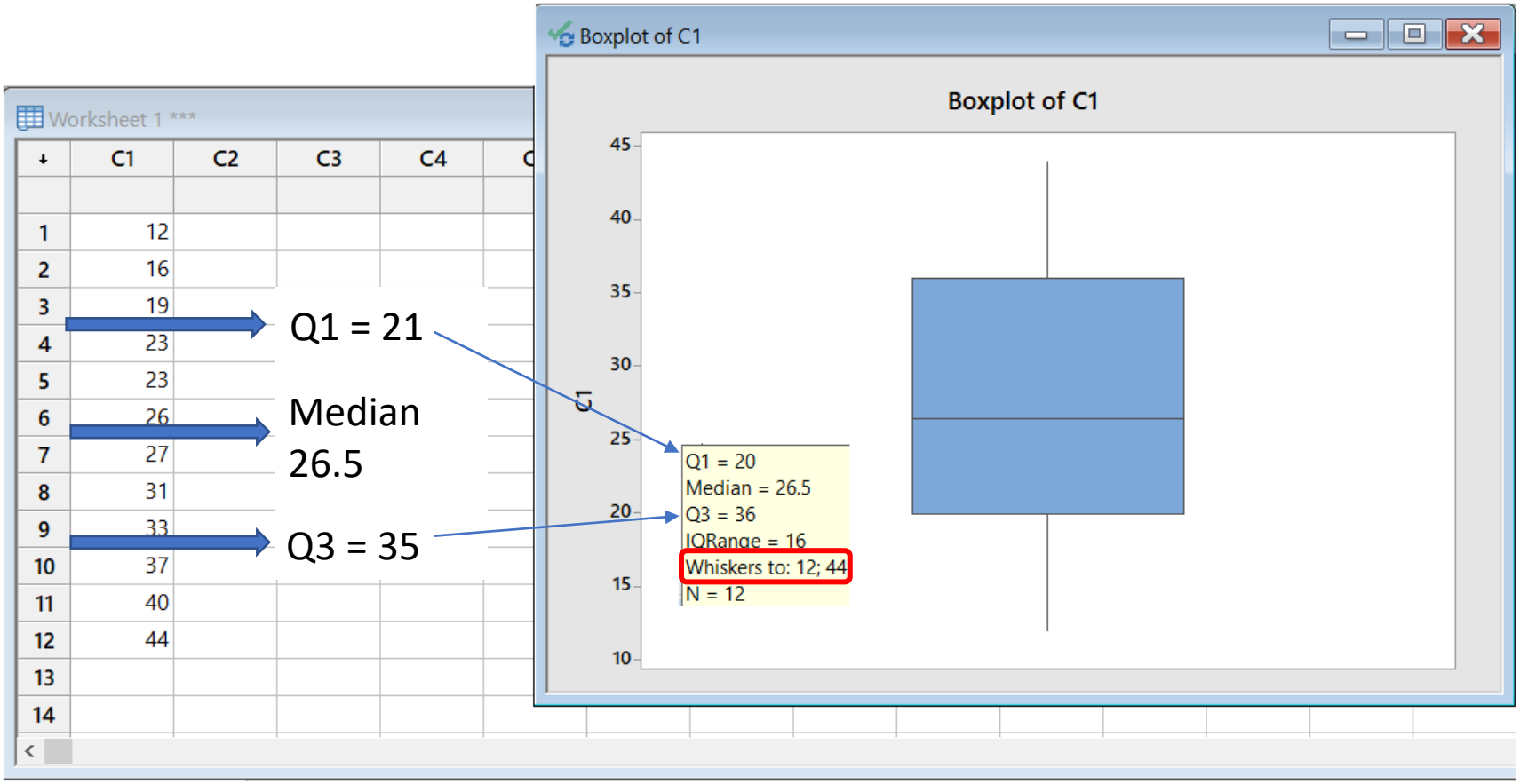
The following plot is obtained using data shown before:



The presence of three outliers, corresponding to values 28, at the lower end of dataset, and 103 and 112, at the upper end, is easily inferred from the plot.

Note that median (corresponding to the Q2 quartile) is always reported as a transversal segment inside the box. Its location closer to the Q1 limit (or, equivalently, further from the Q3 limit) implies an asymmetry of data distribution.

Moreover, if outliers are not present, Minitab 18 considers as whiskers ends the highest and the lowest datum. Its calculation of Q1 and Q3 quartiles is also based on a specific algorithm, as shown in the following example:



In particular, the position of Q1 is considered as (N+1)/4 = 3.25, where N is the number of data. In terms of value, this means that Q1 is far from 19, the 3rd number in the series, 0.25 times the distance between 19 and 23, *i.e.*, 0.25 *4 = 1, thus Q1 = 19 + 1 = 20. Similarly, the position of Q3 is (N+1)*3/4 = 9.75, thus Q3 = 33 + 0.75 * (37-33) = 33 + 3 = 36.

Median Absolute Deviation (MAD)

The procedure based on the Median Absolute Deviation (MAD) is another example of a non-parametric approach (*i.e.*, not implying any prior knowledge about data distribution) to the detection of outliers.

By definition, MAD is the median of absolute values of deviations of single data from their median:

$$\mathrm{MAD} = \mathrm{median}(|X_i - \tilde{X}|) \quad \text{where:} \quad \tilde{X} = \mathrm{median}(X)$$

As an example, given the following 11 observations:

8,9	6,2	7,2	5,4	3,7	2,8	22,2	12,7	6,9	3,1	29,8

data are first ordered in increasing order:

2,8	3,1	3,7	5,4	6,2	**6,9**	7,2	8,9	12,7	22,2	29,8

The median, 6.9, is thus easily recognized and deviations of data from it are easily calculated:

4,1	3,8	3,2	1,5	0,7	**0,0**	0,3	2,0	5,8	15,3	22,9

Data expressing the deviations from the median are then ordered themselves:

0,0	0,3	0,7	1,5	2,0	**3,2**	3,8	4,1	5,8	15,3	22,9

It is thus easy to recognize their median, 3.2.

In order to identify an outlier, the absolute value of the difference existing between the suspected datum and the median of original data is ratioed to the MAD value:

$$|X_{\text{susp outl.}} - \text{Median}(X)| \,/\, \text{MAD}$$

The resulting ratio is finally compared with a critical value, which is usually put equal to 5.

In the cited example, the ratio for the suspected outlier, 29.8, is 7.156.

Since this number is greater than 5 the value 29.8 is discarded as an outlier.

The procedure can be repeated for the second most distant value (from the median), *i.e.*, 22.2. In this case the resulting ratio is 4.78; since this value is lower than 5, the datum 22.2 cannot be discarded as an outlier.
It is easy to see that the first value in the original dataset, 2.8, cannot be considered an outlier, since the above ratio (4.1/3.2 = 1.281) is much lower than 5.

Treatment of outliers once their presence has been assessed

Once the presence of an outlier in a dataset has been assessed, different approaches can be adopted:

1) transforming data, to reconstruct the normality of distribution
2) discarding the outlier
3) keeping the outlier
4) reaching a compromise (*i.e.*, keeping the outlier but reducing its incidence on the information arising from all data)

The most common methods are the following:

1) Using the median instead of the mean, as a measurement of central tendency
2) Trimming
3) Winsorization

Trimming and Winsorization are examples of transformation performed on outliers.

Trimming

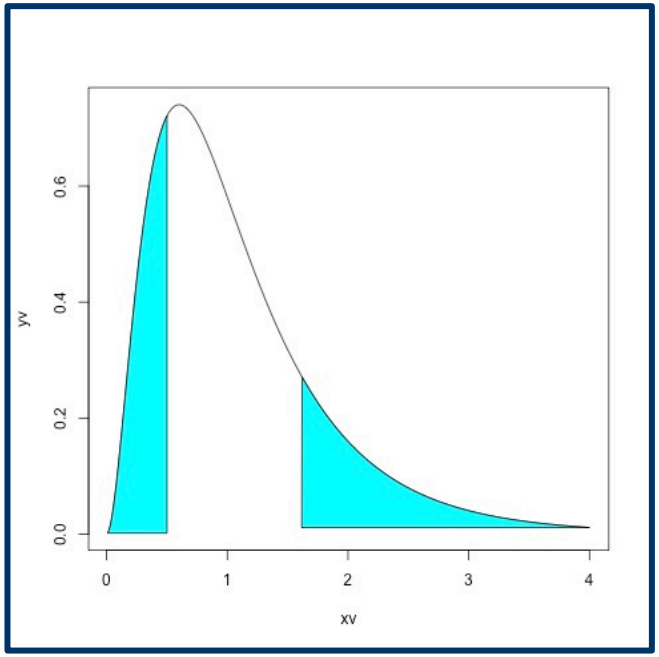
Trimming (also called truncation) consists in eliminating a fixed percentage of extreme values in a dataset, considering one or both tails.

Different trimming approaches can be used:

1) discarding the highest and the lowest values
2) discarding values included in the first and last 5% of probability density
3) discarding values included in the first and in the last quartiles (25% of probability)

The mean calculated when approach 3), which is one of the most frequently adopted, is performed, is called "interquartile mean".

As shown in the figure on the right, when a skewed distribution (an *F* distribution in the specific case) is considered, there is more variability on one side. Since the same amount is trimmed on each side, trimming removes a longer portion of the distribution on one side than on the other.
As a consequence, the mean of the remaining points is more representative of the location of the bulk of the observations.



Winsorization

Winsorization, that is named after the American engineer, physiologist and biostatistician Charles Paine Winsor, who proposed the procedure at the end of the 1940s, consists in the replacement of extreme values in a dataset with less extreme values, with the aim of attenuating the effect of possible outliers.

As an example, let us consider the following series of 13 data:

0	1	12	13	15	16	18	20	22	25	26	154	322

Potential outliers (0, 1, 154 and 322) can be seen on both tails of this set; winsorization replaces those data with closest ones, thus leading to the following dataset:

12	12	12	13	15	16	18	20	22	25	26	26	26

This operation has a remarkable effect on the mean, that is decreased from 49.5 to 18.7, whereas the median remains the same (18).