The quality of analytical measurements

In almost all applications of analysis the results obtained are supplied to a customer or user, thus the latter should be satisfied as much as possible with the quality – the fitness for purpose – of the measurements.

This has many important implications for analytical practice:

1) any assessment of the measurement errors must take into account the whole analytical process – including the sampling steps, which often contribute to the overall error significantly.

2) the performance of the analyses undertaken in each laboratory must be checked internally on a regular basis, usually by applying them to standard or reference materials.

3) in many application areas the results from different laboratories must be compared with each other, so it can be assessed if the performance of the laboratories meets statutory, regulatory and other requirements.

4) the analytical results must be supplied with a realistic estimate of their uncertainty

Quality control methods

If a laboratory has to produce analytical results of a quality acceptable to its customers and perform well in proficiency tests or collaborative trials, it is obviously essential that the results obtained show excellent consistency from day to day.

Quality control methods are statistical techniques developed to show whether time-dependent trends are occurring in the results, together with inevitable random errors.

As an example, let us suppose that a laboratory uses a chromatographic method for determining the level of a pesticide in fruits or vegetables. The results may be used to determine whether a large batch of fruit/vegetable is acceptable or not, and their quality is thus of great importance.

The performance of the method will be checked at regular intervals by applying it, with a small number of replicate analyses, to a standard reference material (SRM), whose pesticide level is certified by a regulatory authority.
Alternatively, an internal quality control (IQC) standard of known composition and high stability can be used.

The SRM or IQC standard should be inserted at random into the sequence of materials analyzed by the laboratory, so that they are analyzed using exactly the same procedures as those used for the routine samples.

The known concentration of the pesticide in the SRM/IQC materials is the target value for the analysis, μ_0.
The laboratory needs to be able to stop and examine the analytical method if it seems to be giving erroneous results.

On the other hand, resources, time and materials would be wasted if the sequence of analyses was halted unnecessarily, so quality control methods should allow its continued use as long as it is working satisfactorily.

Quality control methods are also very widely used to monitor industrial processes.

For example, the weights of pharmaceutical tablets coming off a production line can be monitored by taking small samples of tablets from time to time.

The tablet weights are bound to fluctuate around the target value μ_0 because of random errors, but if these random errors are not too large, and are not accompanied by time-dependent trends, the process is under control.

Control charts are one of the approaches adopted to monitor an industrial process or the performance of an analytical method.

Shewhart charts for mean values

Over a long period, the population standard deviation, σ, related to the concentration of a target analyte which is the object of quality control, will become known from experience.

For this reason, the confidence interval on the mean value of that concentration can be expressed according to Case 1 situation, thus using coefficients related to the N(0,1) distribution:

95% confidence

$$\mu = \overline{x} \pm \frac{2\sigma}{\sqrt{n}}$$

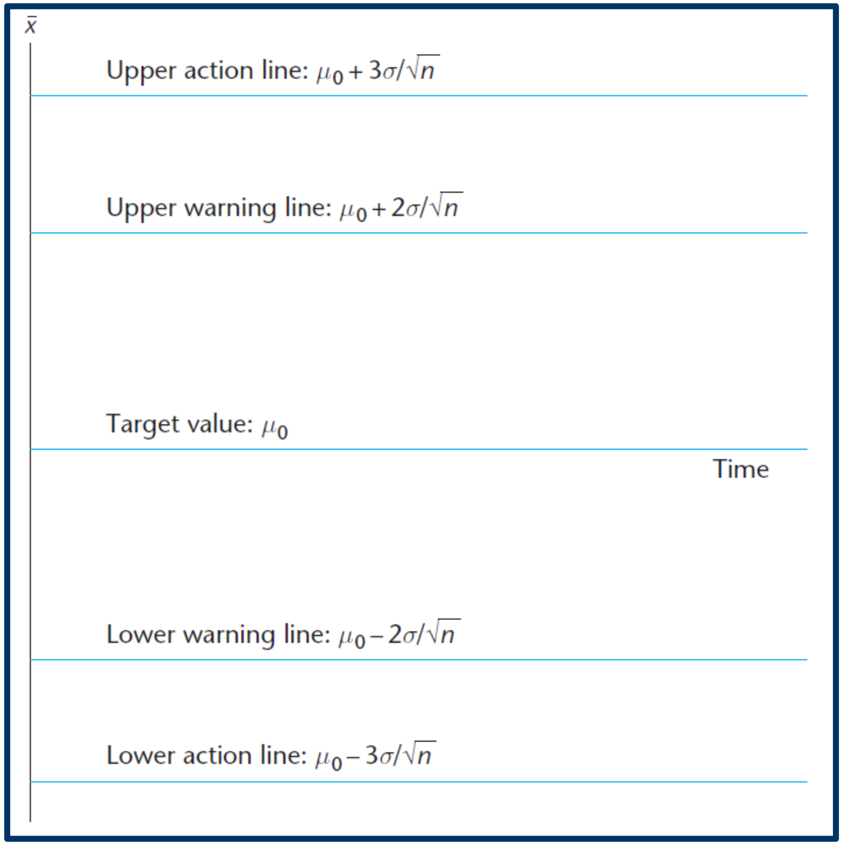
99.7% confidence

$$\mu = \overline{x} \pm \frac{3\sigma}{\sqrt{n}}$$

Note that, for the sake of simplicity, actual values to be multiplied by sampling standard deviation, i.e., 1.96 and 2.97, for 95 and 99.7% confidence, respectively, are usually rounded to 2 and 3.

The equations reported above are exploited in the construction of one of the most common types of control chart, the Shewhart chart, introduced by the American statistician, mathematician and physicist Walter Andrew Shewhart in 1924.

The vertical axis of a Shewhart chart displays the process mean, whereas the horizontal axis represents time.



The target value, μ_0, is marked by a horizontal line.

The chart also includes two further pairs of horizontal lines: those located at $\mu_0 \pm 2\sigma/\surd n$ are called the upper/lower warning lines, whereas those drawn at $\mu_0 \pm 3\sigma/\surd n$ are called the upper/lower action lines.

Mean values obtained for the analyte under control are plotted in the chart as points.

The probability for a specific value to be found outside one of the action lines when the process is in control is known to be only 0.3%, i.e., once in 370 samples, so, in practice, the process is usually stopped and examined if that deviation occur for less samples.

On the other hand, there is a probability of ca. 5% (0.05) of a single point falling outside a warning line (but being still within the action lines) while the process remains in control. This outcome alone would not cause the process to be stopped, but if two successive values fell outside the same warning line, the probability ($0.025^2 \times 2 = 0.00125$, for both warning lines) would be so low that the process would be judged to be out of control.

Two possible explanations can be hypothesized if a Shewhart chart for mean values suggests that a process is out of control:

1) the process mean has changed
2) the process mean has remained unchanged but the variability in the process has increased

If the actual variability is increased, action and warning lines, based on the previous, lower, estimate of variability, are artificially located closer to the target line, thus significant changes in the mean value are deduced when in fact they do not occur.

Conversely, if the variability of the process is decreased (i.e., improved) with respect to the one used to draw action and warning lines, the latter become artificially more distant from the target line, thus allowing real changes in mean value to go undetected.

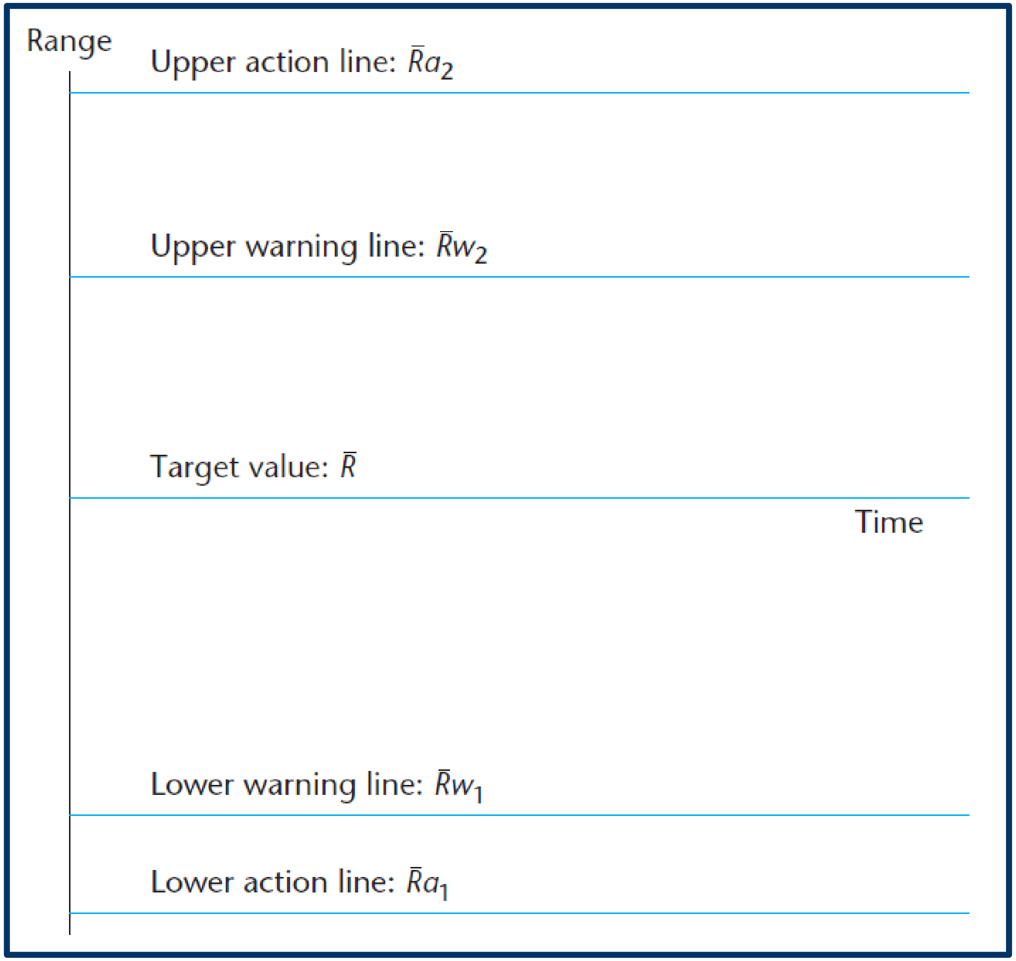
A careful monitoring of process variability must thus be planned.

This procedure also has its own intrinsic value: the variability of a process or analysis is one measure of its quality, and, in the laboratory context, it is directly linked to the repeatability (within-laboratory standard deviation) of the method.

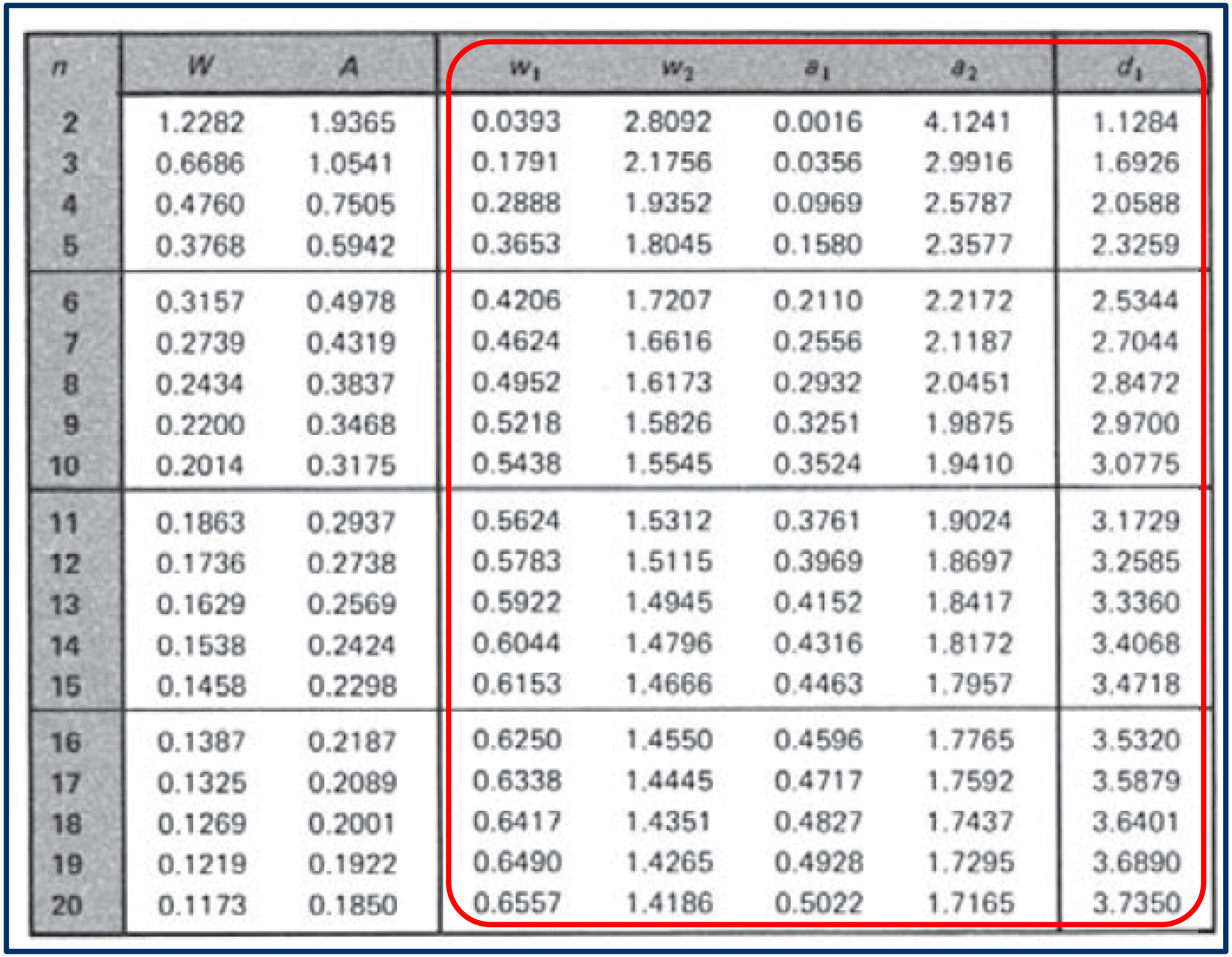
The variability of a process can be displayed by plotting another Shewhart chart, involving the range, *R* (highest value - lowest value), of each of the samples taken.

A typical Shewhart control chart for the range is shown in the figure on the right.

The general format of the chart is the same used in plotting mean values, with a line representing the target value, and pairs of action and warning lines.


Range
Upper action line: $\bar{R}a_2$
Upper warning line: $\bar{R}w_2$
Target value: $\bar{R}$
Time
Lower warning line: $\bar{R}w_1$
Lower action line: $\bar{R}a_1$

The most striking difference between the two charts is that the pairs of lines are not symmetrical with respect to the target value for the range, $\bar{R}$. The target value of *R* can be calculated as the product between σ and a constant, d_1, that can be found in appropriate statistic tables, like the one reported in the next slide, including also values of constants a_1, a_2, w_1 and w_2, required for the calculation of action and warning lines, once $\bar{R}$ is known.



n	W	A	w_1	w_2	a_1	a_2	d_1
2	1.2282	1.9365	0.0393	2.8092	0.0016	4.1241	1.1284
3	0.6686	1.0541	0.1791	2.1756	0.0356	2.9916	1.6926
4	0.4760	0.7505	0.2888	1.9352	0.0969	2.5787	2.0588
5	0.3768	0.5942	0.3653	1.8045	0.1580	2.3577	2.3259
6	0.3157	0.4978	0.4206	1.7207	0.2110	2.2172	2.5344
7	0.2739	0.4319	0.4624	1.6616	0.2556	2.1187	2.7044
8	0.2434	0.3837	0.4952	1.6173	0.2932	2.0451	2.8472
9	0.2200	0.3468	0.5218	1.5826	0.3251	1.9875	2.9700
10	0.2014	0.3175	0.5438	1.5545	0.3524	1.9410	3.0775
11	0.1863	0.2937	0.5624	1.5312	0.3761	1.9024	3.1729
12	0.1736	0.2738	0.5783	1.5115	0.3969	1.8697	3.2585
13	0.1629	0.2569	0.5922	1.4945	0.4152	1.8417	3.3360
14	0.1538	0.2424	0.6044	1.4796	0.4316	1.8172	3.4068
15	0.1458	0.2298	0.6153	1.4666	0.4463	1.7957	3.4718
16	0.1387	0.2187	0.6250	1.4550	0.4596	1.7765	3.5320
17	0.1325	0.2089	0.6338	1.4445	0.4717	1.7592	3.5879
18	0.1269	0.2001	0.6417	1.4351	0.4827	1.7437	3.6401
19	0.1219	0.1922	0.6490	1.4265	0.4928	1.7295	3.6890
20	0.1173	0.1850	0.6557	1.4186	0.5022	1.7165	3.7350

Adapted from: H.R. Neave, *Elementary Statistics tables*, Routledge, 1979, pp. 55-57

It is worth noting that the establishment of a proper value for σ is very important for a correct drawing of Shewhart charts, thus its estimate should be based on a substantial number of measurements.

However, in making such measurements, the same problem, i.e., distinguishing a change in the process mean from a change in the process variability, must be faced.
If σ is calculated directly from a long sequence of measurements, its value may be overestimated by any changes in the mean that occur during that sequence.

The solution to this problem is to take a large number of small samples, measure the range, R, for each of them, and then determine its mean value.
This procedure ensures that only the inherent variability of the process is measured, with any drift in the mean values eliminated. The mean value of R can subsequently be used to determine the action and warning lines for the range control chart.

Moreover, the warning and action lines for the control chart for the mean can be determined, respectively, from the mean value of x by adding/subtracting the products between the mean value of R and constants W and A reported in the table shown in the previous slide:

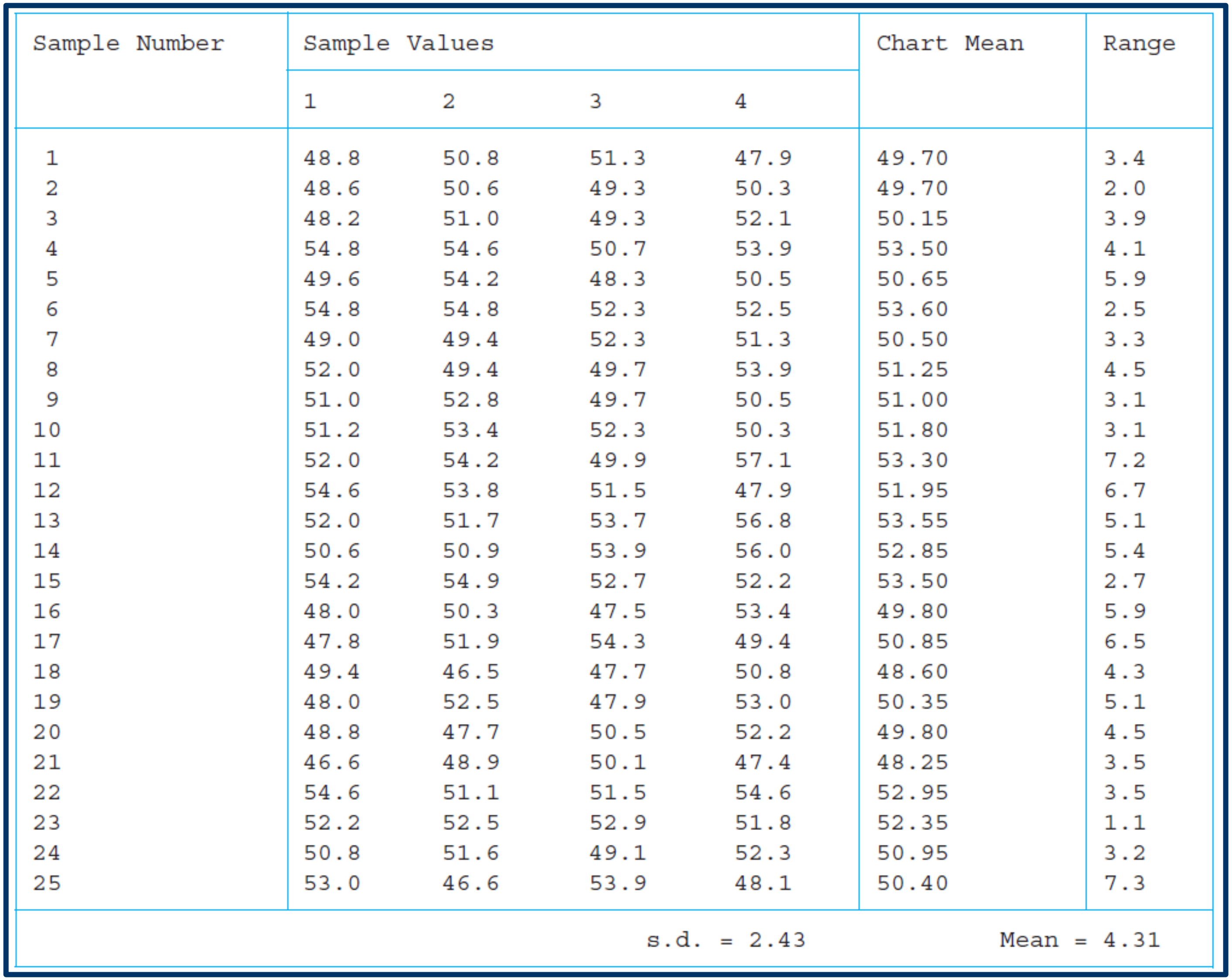
$$\text{Warning lines at } \overline{x} \pm W\overline{R}$$

$$\text{Action lines at } \overline{x} \pm A\overline{R}$$

A numerical example

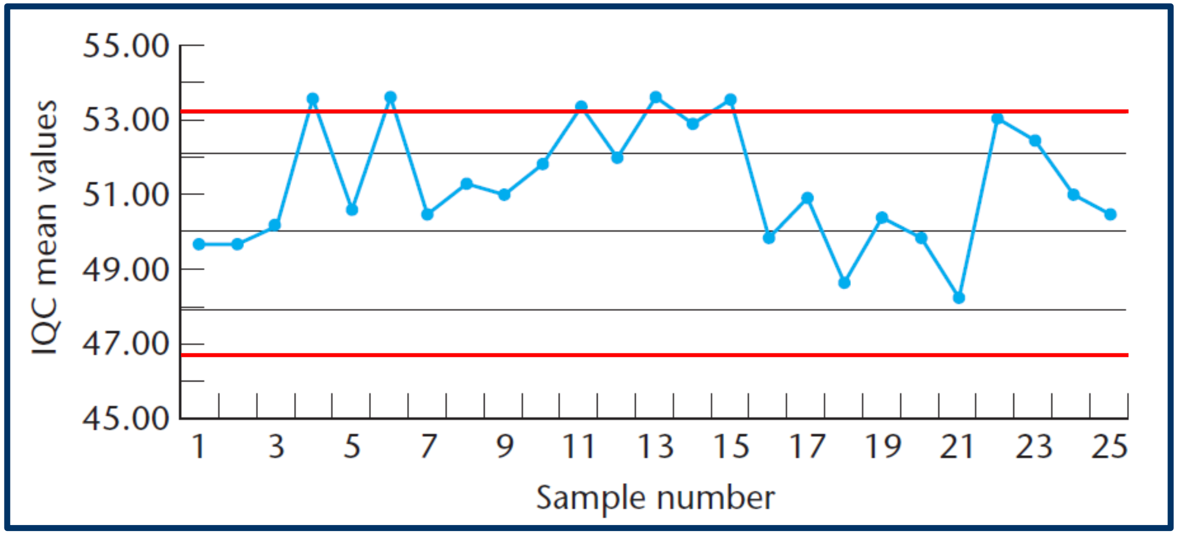
An internal quality control standard with an analyte concentration of 50 mg kg^{-1} was analyzed in a laboratory for 25 consecutive days, the sample size being four on each day.

The results are reported in the table on the right:



Sample Number	Sample Values				Chart Mean	Range
	1	2	3	4		
1	48.8	50.8	51.3	47.9	49.70	3.4
2	48.6	50.6	49.3	50.3	49.70	2.0
3	48.2	51.0	49.3	52.1	50.15	3.9
4	54.8	54.6	50.7	53.9	53.50	4.1
5	49.6	54.2	48.3	50.5	50.65	5.9
6	54.8	54.8	52.3	52.5	53.60	2.5
7	49.0	49.4	52.3	51.3	50.50	3.3
8	52.0	49.4	49.7	53.9	51.25	4.5
9	51.0	52.8	49.7	50.5	51.00	3.1
10	51.2	53.4	52.3	50.3	51.80	3.1
11	52.0	54.2	49.9	57.1	53.30	7.2
12	54.6	53.8	51.5	47.9	51.95	6.7
13	52.0	51.7	53.7	56.8	53.55	5.1
14	50.6	50.9	53.9	56.0	52.85	5.4
15	54.2	54.9	52.7	52.2	53.50	2.7
16	48.0	50.3	47.5	53.4	49.80	5.9
17	47.8	51.9	54.3	49.4	50.85	6.5
18	49.4	46.5	47.7	50.8	48.60	4.3
19	48.0	52.5	47.9	53.0	50.35	5.1
20	48.8	47.7	50.5	52.2	49.80	4.5
21	46.6	48.9	50.1	47.4	48.25	3.5
22	54.6	51.1	51.5	54.6	52.95	3.5
23	52.2	52.5	52.9	51.8	52.35	1.1
24	50.8	51.6	49.1	52.3	50.95	3.2
25	53.0	46.6	53.9	48.1	50.40	7.3
			s.d. = 2.43		Mean = 4.31	

The resulting Shewhart chart for the IQC mean value is the following:



A first examination of results indicates that, over the 25-day period of the analyses, the sample means are drifting up and down. Indeed, sample means from days 3–15 are greater than the target value of 50, whereas four of the next six means are below the target value, and the last four are all above it again.

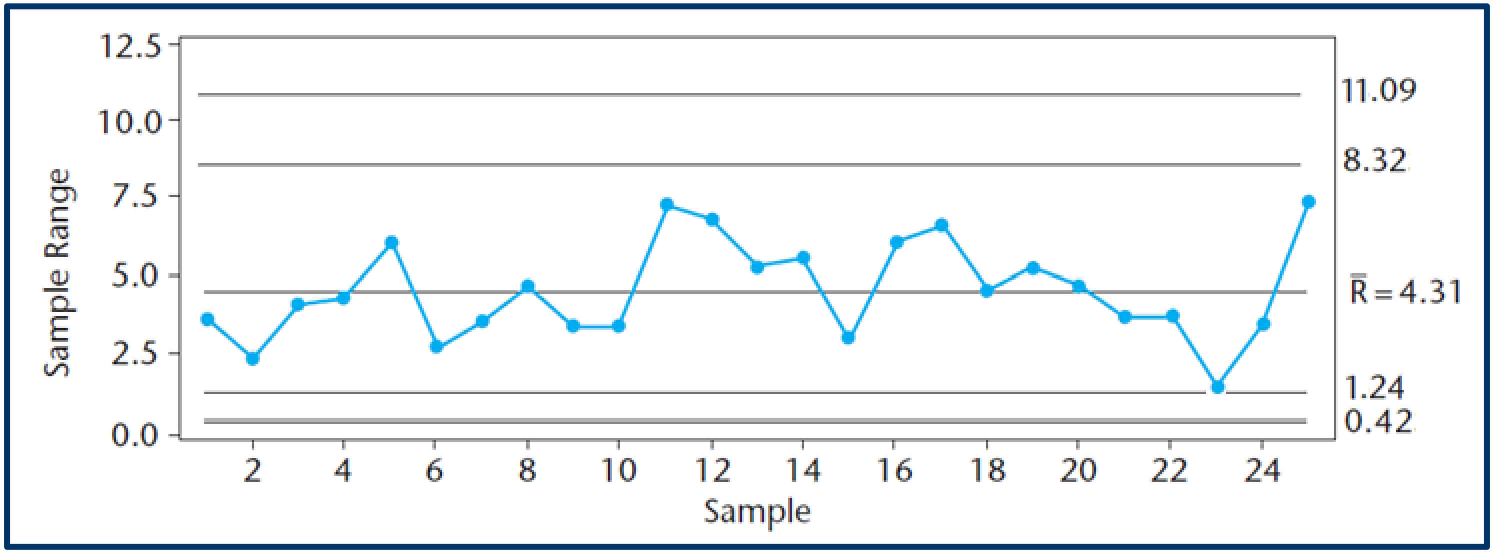
These are the circumstances in which it is important to estimate σ using the method described above. Based on *R*-values reported in the last column of data table, the mean range is found to be 4.31, whereas d_1, for n = 4, is equal to 2.059. Consequently, σ = 4.31/2.059 = 2.09.

Interestingly, the standard deviation of all the 100 measurements, treated as a single sample, is 2.43. If this value was used for σ, an overestimation would occur because of the drifts in the mean.

The control chart for the mean has then been plotted with the aid of equations shown before, using constants $W = 0.4760$ and $A = 0.7505$, leading to warning and action lines located at 50.00 ± 2.05 and 50.00 ± 3.23, respectively.

The chart shows that the process mean is not under control since several of the points fall outside the upper action line.

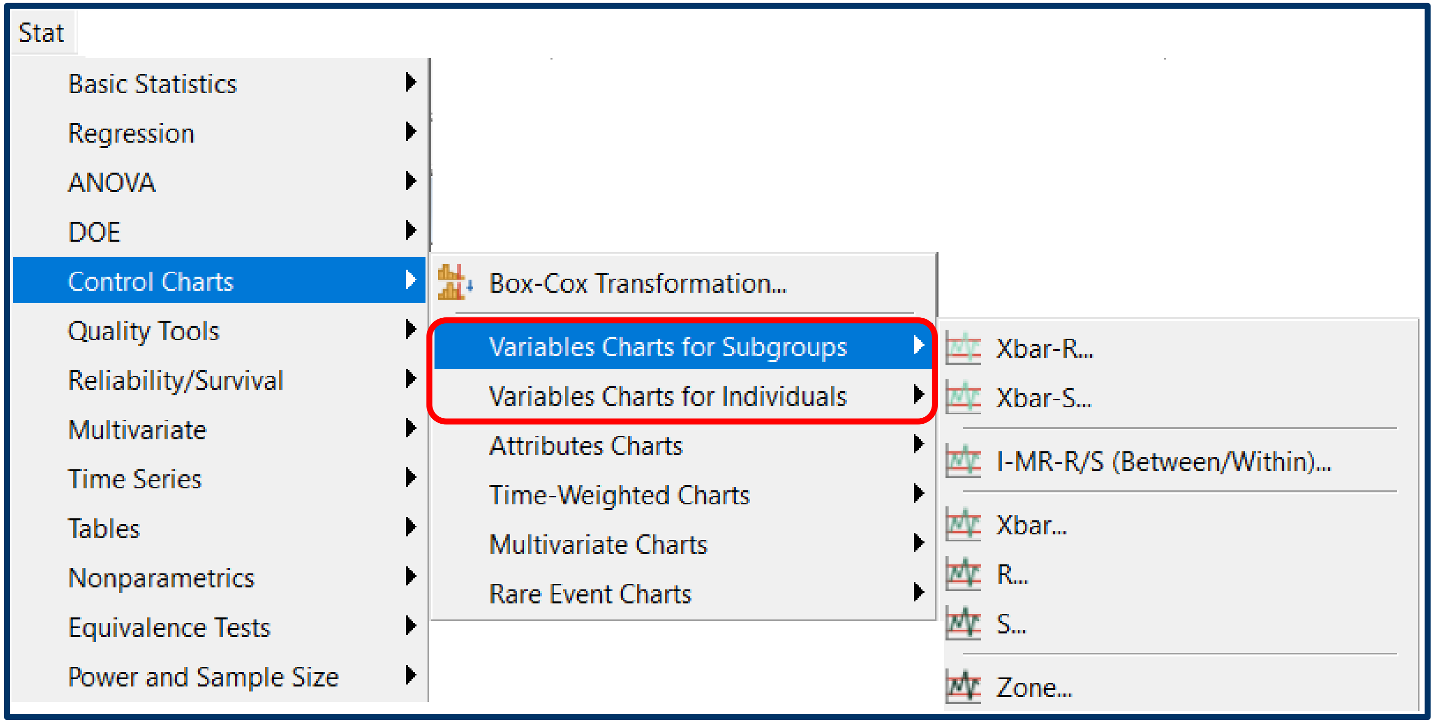
Using equations shown before, the Shewhart chart for the range can also be generated:



As apparent, with one exception, all the values of ranges lie well within the warning lines, thus indicating that the process variability is under control.

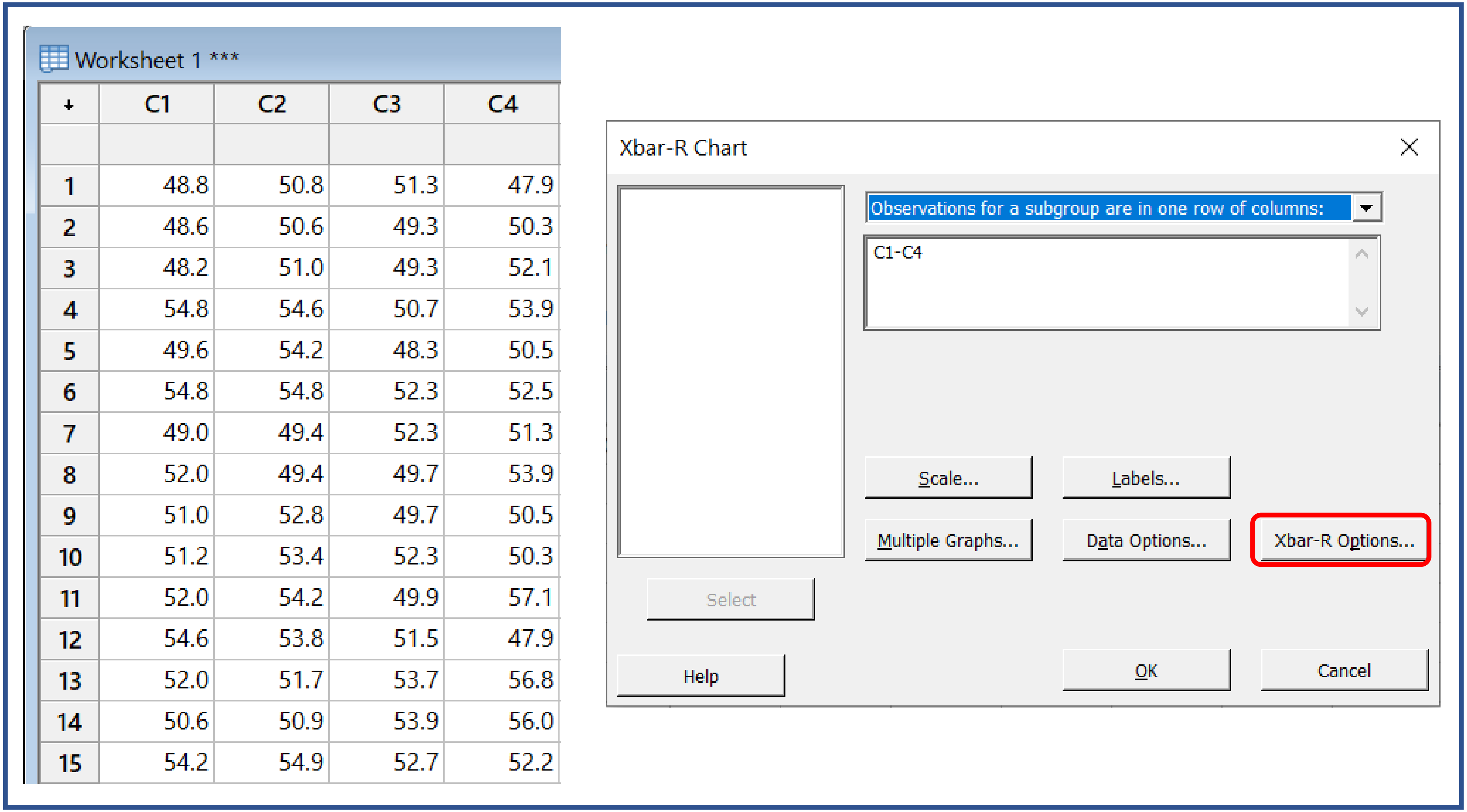
Using Minitab 18 to draw Shewhart control charts

Different types of control charts can be drawn by accessing the Stat > Control charts pathway in Minitab 18. In particular, Shewhart charts can be drawn by accessing sub-menus Variable charts for subgroups/individuals. When several replicates are available for each determination, as in the example shown before, the Variable charts for subgroups menu has to be used:



If control charts for mean and range are required, the Xbar-R… option has to be used. The option Xbar-S… is used to obtain charts for mean and standard deviation, whereas Xbar…, R… and S… are used for charts referred singularly to mean, range and standard deviation.

When subgroups are present, each row in the worksheet represents a group of replicates and columns including replicates have to be indicated, along with the specification "Observation for a subgroup are in one row of columns", in the window selected for control chart drawing:



Worksheet 1 ***

↓	C1	C2	C3	C4
1	48.8	50.8	51.3	47.9
2	48.6	50.6	49.3	50.3
3	48.2	51.0	49.3	52.1
4	54.8	54.6	50.7	53.9
5	49.6	54.2	48.3	50.5
6	54.8	54.8	52.3	52.5
7	49.0	49.4	52.3	51.3
8	52.0	49.4	49.7	53.9
9	51.0	52.8	49.7	50.5
10	51.2	53.4	52.3	50.3
11	52.0	54.2	49.9	57.1
12	54.6	53.8	51.5	47.9
13	52.0	51.7	53.7	56.8
14	50.6	50.9	53.9	56.0
15	54.2	54.9	52.7	52.2

Xbar-R Chart

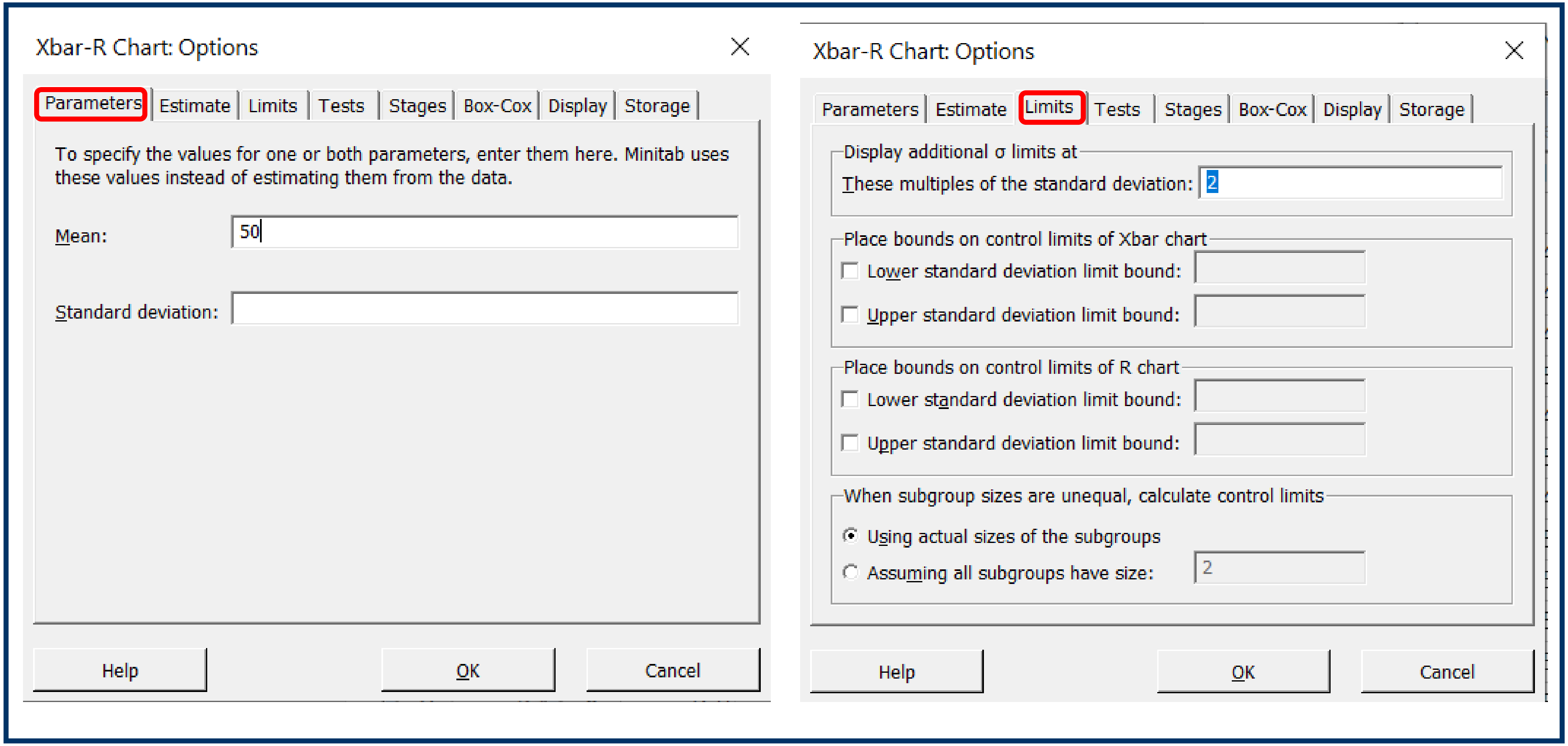
Observations for a subgroup are in one row of columns:

C1-C4

The Xbar-R Options... option has to be selected to indicate specific settings for the control chart.

In particular, the Parameters sub-menu can be used to set mean and standard deviation values, if the user does not want Minitab to estimate them.

In the present example, a mean equal to 50 was indicated, as in the calculations shown before, whereas the standard deviation box was left blank, thus it was calculated by Minitab:



Xbar-R Chart: Options

To specify the values for one or both parameters, enter them here. Minitab uses these values instead of estimating them from the data.

Mean: 50

Standard deviation:

Xbar-R Chart: Options

Display additional σ limits at

These multiples of the standard deviation: 2

Place bounds on control limits of Xbar chart

□ Lower standard deviation limit bound:

□ Upper standard deviation limit bound:

Place bounds on control limits of R chart

□ Lower standard deviation limit bound:

□ Upper standard deviation limit bound:

When subgroup sizes are unequal, calculate control limits

● Using actual sizes of the subgroups

○ Assuming all subgroups have size: 2

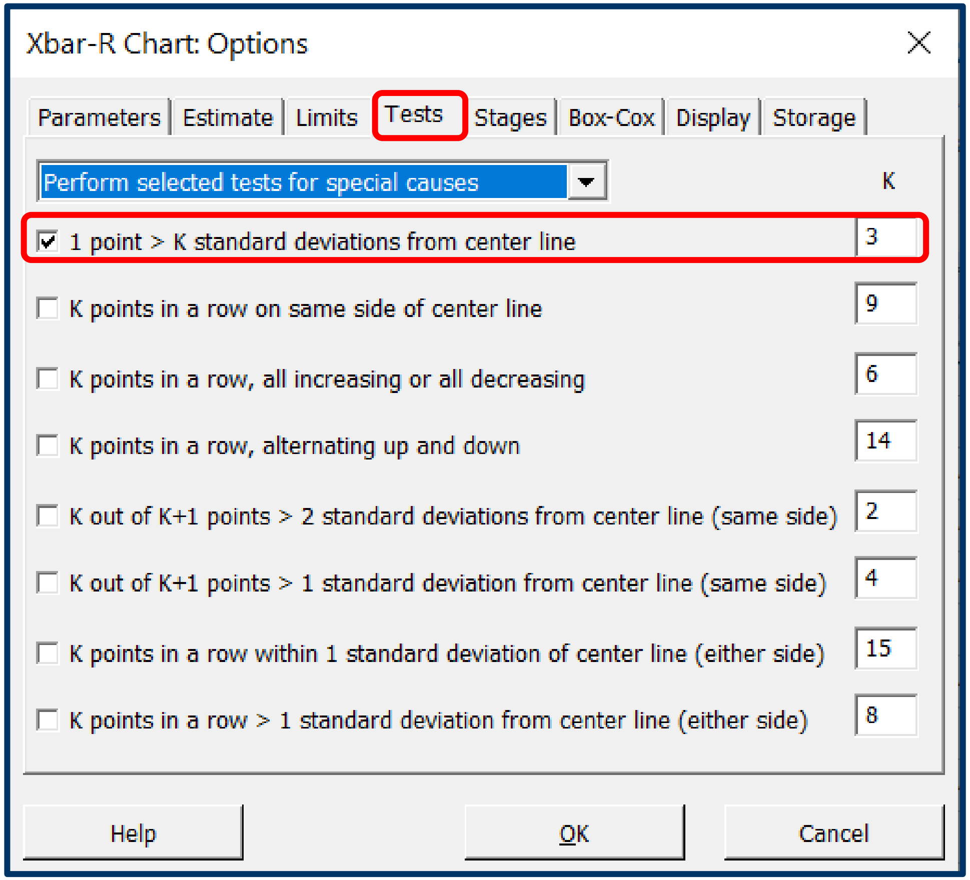
In the Limits sub-menu the display of additional limits in the control chart can be selected.

As a default setting, action lines are drawn by Minitab 18 by considering a difference from the mean value of ± 3 standard deviations. However, as shown in the figure, additional lines, namely warning lines, can be drawn at a ± 2 standard deviations distance from the mean.

In the Tests sub-menu several different types of tests on data can be selected.

As a default setting, a test evaluating if the response at one point goes beyond the limit represented by a K standard deviations distance from center line is performed by Minitab 18.

As shown in the figure, the default value for K is 3, thus meaning that the upper or lower action lines are considered as limits for the test.

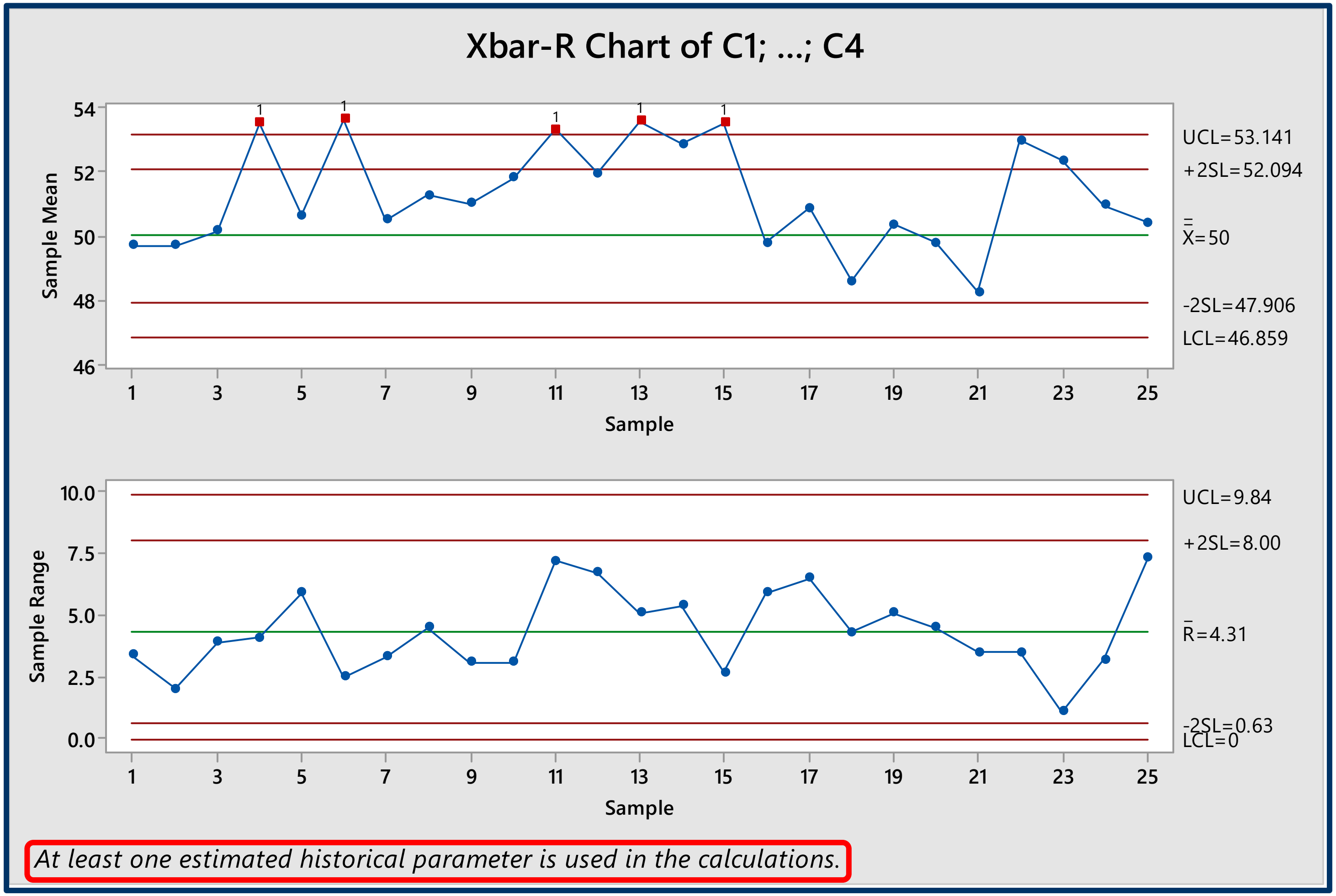


Xbar-R Chart: Options

Perform selected tests for special causes

	K
☑ 1 point > K standard deviations from center line	3
□ K points in a row on same side of center line	9
□ K points in a row, all increasing or all decreasing	6
□ K points in a row, alternating up and down	14
□ K out of K+1 points > 2 standard deviations from center line (same side)	2
□ K out of K+1 points > 1 standard deviation from center line (same side)	4
□ K points in a row within 1 standard deviation of center line (either side)	15
□ K points in a row > 1 standard deviation from center line (either side)	8

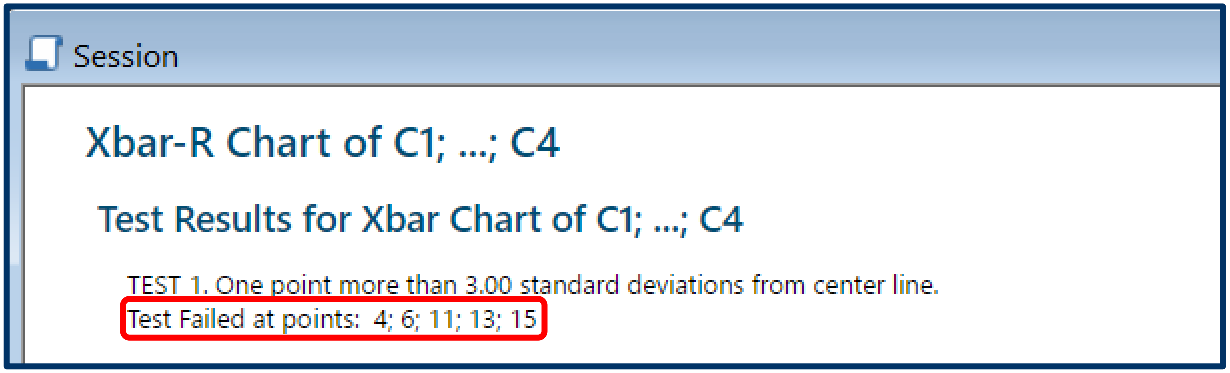
After calculations are performed, two graphs, showing Shewhart control charts for the sample mean and range, respectively, are obtained:



At least one estimated historical parameter is used in the calculations.

The note reported below the two graphs reminds the user that an estimated historical parameter, i.e., the mean, in the specific case (50), was adopted.

Critical points, located outside action lines (indicated as UCL and LCL by Minitab 18), are drawn in red color. Their number in the list of subgroups is indicated in the summary reported in the Session window:



Xbar-R Chart of C1; ...; C4

Test Results for Xbar Chart of C1; ...; C4

TEST 1. One point more than 3.00 standard deviations from center line.
Test Failed at points: 4; 6; 11; 13; 15

It is worth noting that Minitab calculates the warning and action lines for the range by approximating the asymmetrical distribution of the mean range by a normal distribution, thus a slight difference can be observed with calculations based on tables shown before.

Average run length (ARL) and Cumulative Sum (CUSUM) charts

An important property of a control chart is the ability to detect any change in the process mean as soon as possible. The average number of measurements required to detect any particular change in the process mean is called average run length (ARL).

Notably, since the positions of the action and warning lines on a Shewhart chart for the process mean depend on the value of sampling standard deviation, **σ**/√n, the ARL for that chart will depend on the size of the change in the mean compared with the latter.

A larger change will thus be detected more rapidly than a smaller one and the ARL will be reduced by using a larger sample size, n.

As calculated before, the average run length for a process under control would be *ca.* 370 if action lines are located at a 3 **σ**/√n distance from the target line. This value is indicated with the symbol ARL_0.

The calculation of the expected ARL when the system goes out of control, that is indicated as ARL_Δ, is more complicated.

First, let us define β the probability that the chart does not indicate a deviation despite the fact that the process is out of control.

Consequently, 1-β is the probability that a deviation is indicated by the control chart when the process is out of control. This probability corresponds to the probability of discovering the deviation in just one measurement, the first of the run.

If two measurements were required, the first of them would be unable to discover the deviation, then the second would discover it, thus the probability would be: $\beta * (1-\beta)$.
By analogy, if k measurements were required, the probability would be $\beta^{k-1} * (1-\beta)$.

In other words, a probability can be associated to the number of measurements, i.e., the run length (RL), required to discover the deviation (k), thus the expectation for the RL can be calculated, using the formula of expectation for a discrete random variable:

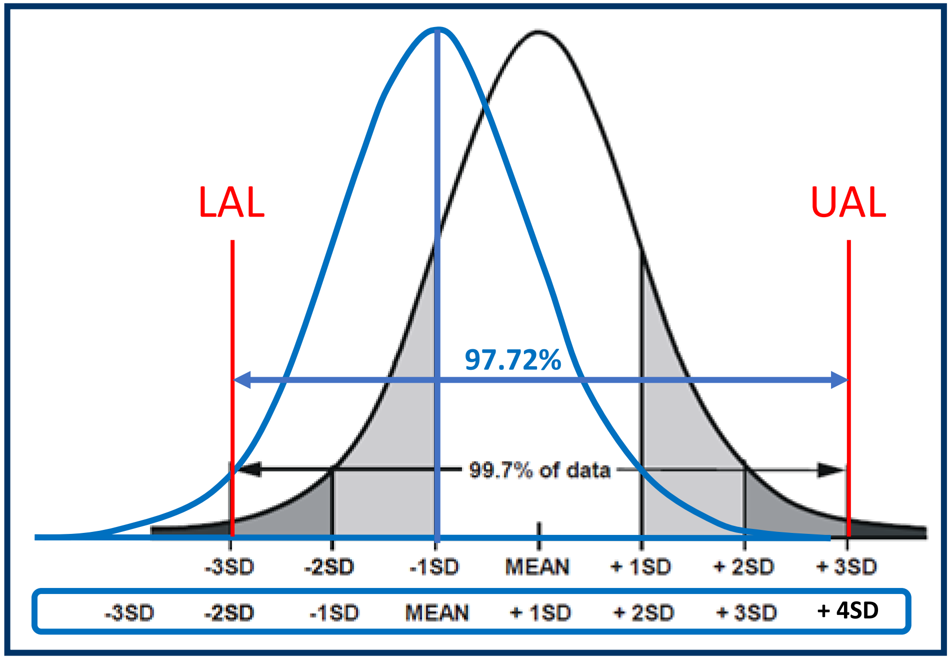
$$E(RL) = \sum_{k=1}^{\infty} k\,\beta^{k-1}\,(1-\beta) = (1-\beta)\sum_{k=1}^{\infty} k\,\beta^{k-1}$$

The sum reported in the equation corresponds to the series $1 + 2\beta + 3\beta^2 + \ldots$, which is the MacLaurin series expansion of function $1/(1-\beta)^2$.

Consequently: $E(RL) = 1/(1-\beta)$.

Let now suppose that a negative deviation corresponding to 1 $\boldsymbol{\sigma}/\sqrt{n}$ (-1SD) occurs for the mean with respect to the value found when the process is in control:

The calculation for β, considering the previously set action lines (LAL/UAL), which are constant, but the new mean, can be described graphically, as shown in the figure on the right.



If the process was in control, the gaussian PDF would be centered on the correct mean value and would thus correspond to the black curve, with a 99.7% of probability of finding a measurement comprised between the action lines. However, the actual gaussian PDF is now the light blue one, thus the calculation of β is the following:

$$\beta \;=\; \Phi(4) - \Phi(-\,2) \;\cong\; 1 - 0.0228 = 0.9772$$

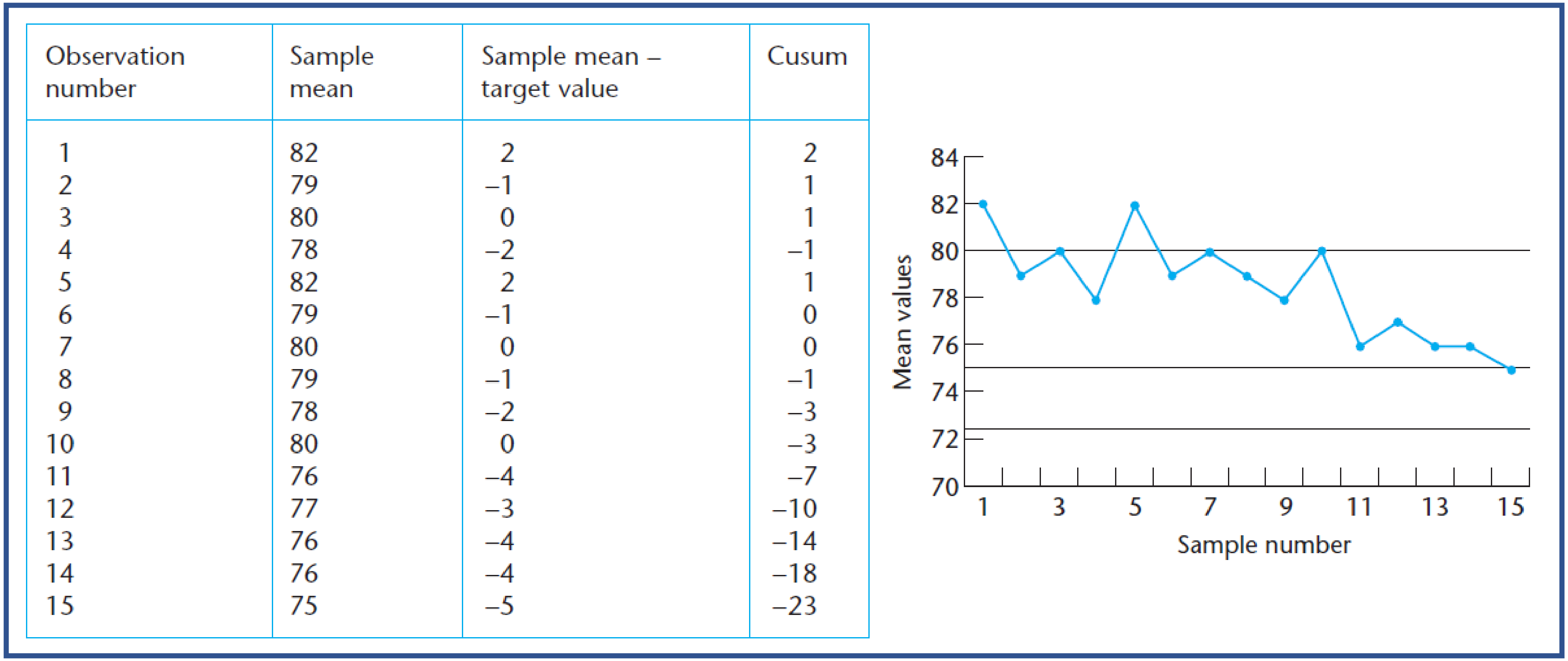
Consequently: $E(RL) = 1/(1-0.9772) = 43.86$.

Almost 44 runs would then be required, as an average, to discover a deviation of the mean by 1 $\sigma/\sqrt{n}$.

This delay in the recognition of a deviation from a target value inherent to Shewhart charts can be quite problematic if the target value is referred, for example, to a potentially toxic analyte analyzed in a food matrix or to an impurity in a drug.

A possible solution is represented by the use of a different control chart, the so-called cusum (cumulative sum) chart, in which the cumulation of deviations is considered.

As an example, let us consider the following data table and the corresponding Shewhart control chart for the mean:

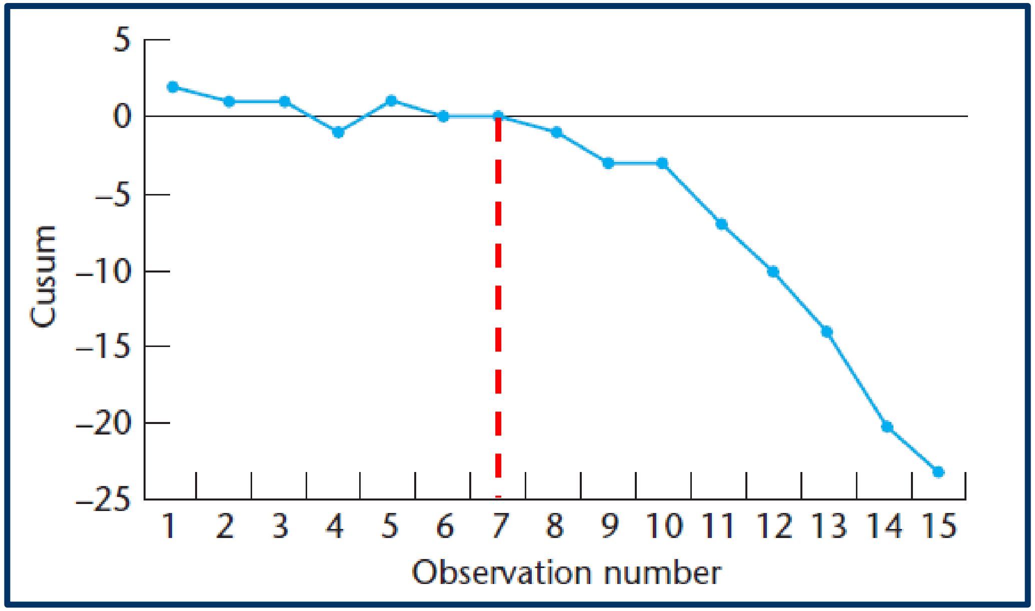


Observation number	Sample mean	Sample mean – target value	Cusum
1	82	2	2
2	79	–1	1
3	80	0	1
4	78	–2	–1
5	82	2	1
6	79	–1	0
7	80	0	0
8	79	–1	–1
9	78	–2	–3
10	80	0	–3
11	76	–4	–7
12	77	–3	–10
13	76	–4	–14
14	76	–4	–18
15	75	–5	–23

A progressive deviation from the target value (80) is clearly observed in the Shewhart chart, yet the observed value reaches the lower warning line only after 15 measurements (which could correspond to as many days).

On the other hand, if the difference between sample mean and target value is calculated and cumulatively summed, thus obtaining the so-called cusum, as indicated in the last column of the table, the presence of a steady trend downwards becomes apparent quite soon.

A plot of Cusum vs observation number in this case would be the one shown in the figure on the right:



If a manufacturing or analytical process is under control, positive and negative deviations from the target value are equally likely and the cusum should oscillate about zero.

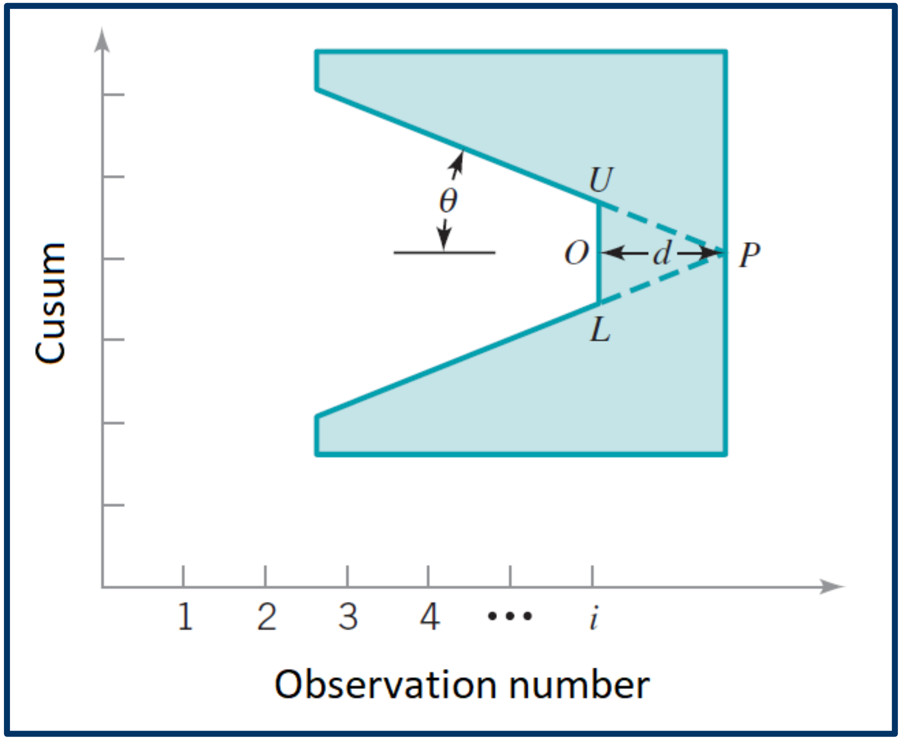
Conversely, if the process mean changes, the cusum will clearly move away from zero. In the example given, the process mean seems to start deviating systematically after the seventh observation, since the cusum becomes increasingly negative.

Cusum charts, proposed by the English statistician Ewan S. Page in 1954, exploit this concept. For a proper interpretation of cusum charts, i.e., to infer that a genuine change in the process mean has occurred, the so-called V-mask, proposed by the English statistician George A. Barnard in 1959, can be adopted.

In the original version, the V-mask was drawn on a transparent plastic sheet and placed over the control chart, drawn on paper, with its axis of symmetry positioned horizontally and its apex located at an appropriate distance *d* to the right of the last observation.

A typical V-mask is shown in the figure on the right, in which the two fundamental parameters required for its construction, i.e., the lead distance, *d*, and the semi-angle, θ, are evidenced.

The V-mask is placed over the cusum plot with its axis parallel to the horizontal axis and by aligning its origin O with the last point in the plot.
The distance between this point and the mask vertex corresponds to *d*.



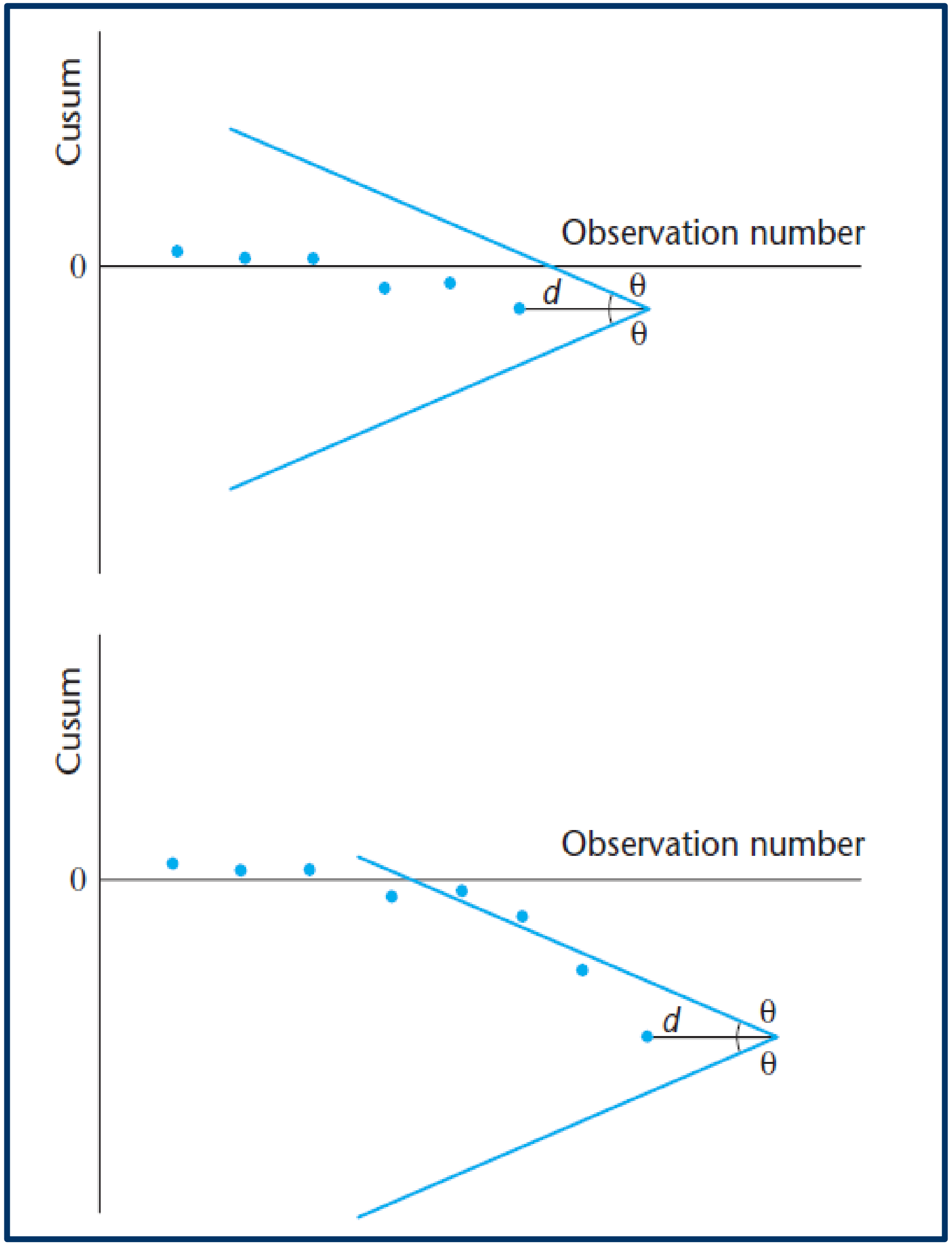
Values of *d* and tan θ are chosen so that significant changes in the process mean are detected quickly but, at the same time, false alarms are few.

In particular, the following equations are used:

$\tan\theta = \delta\,\sigma_x/2$ where $\delta\sigma_x$ is the shift in the mean value that the user wants to detect quickly

$d = (2/\delta^2)\ \ln[(1-\beta)/\alpha]$ where α is the probability of a false alarm (concluding that a shift has occurred while it has not) and β is the probability of not detecting a shift that has occurred.

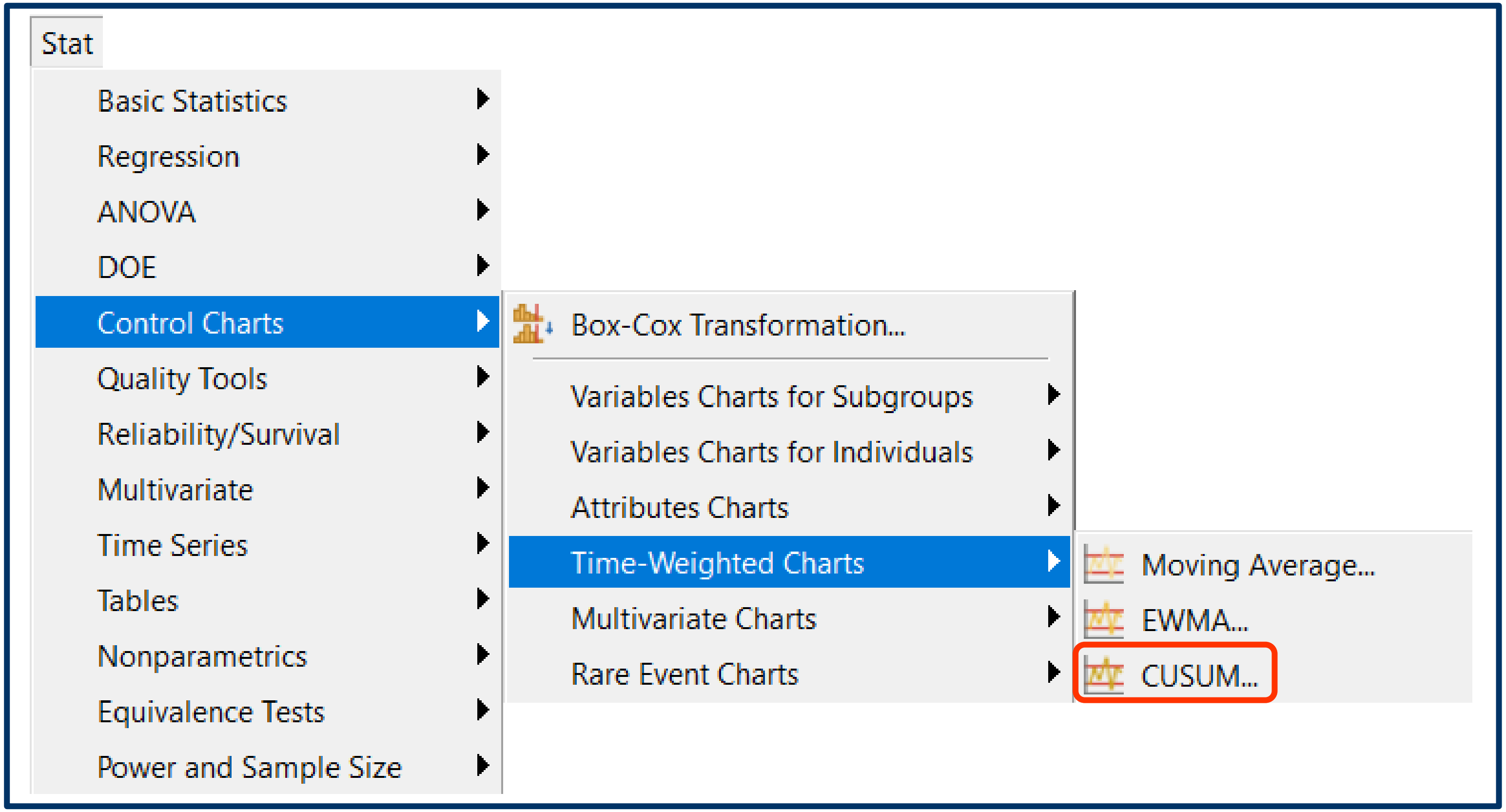
Notably, if all the points on the chart lie within the arms of the V mask, as shown in the upper panel of the figure on the right, then the process is in control.



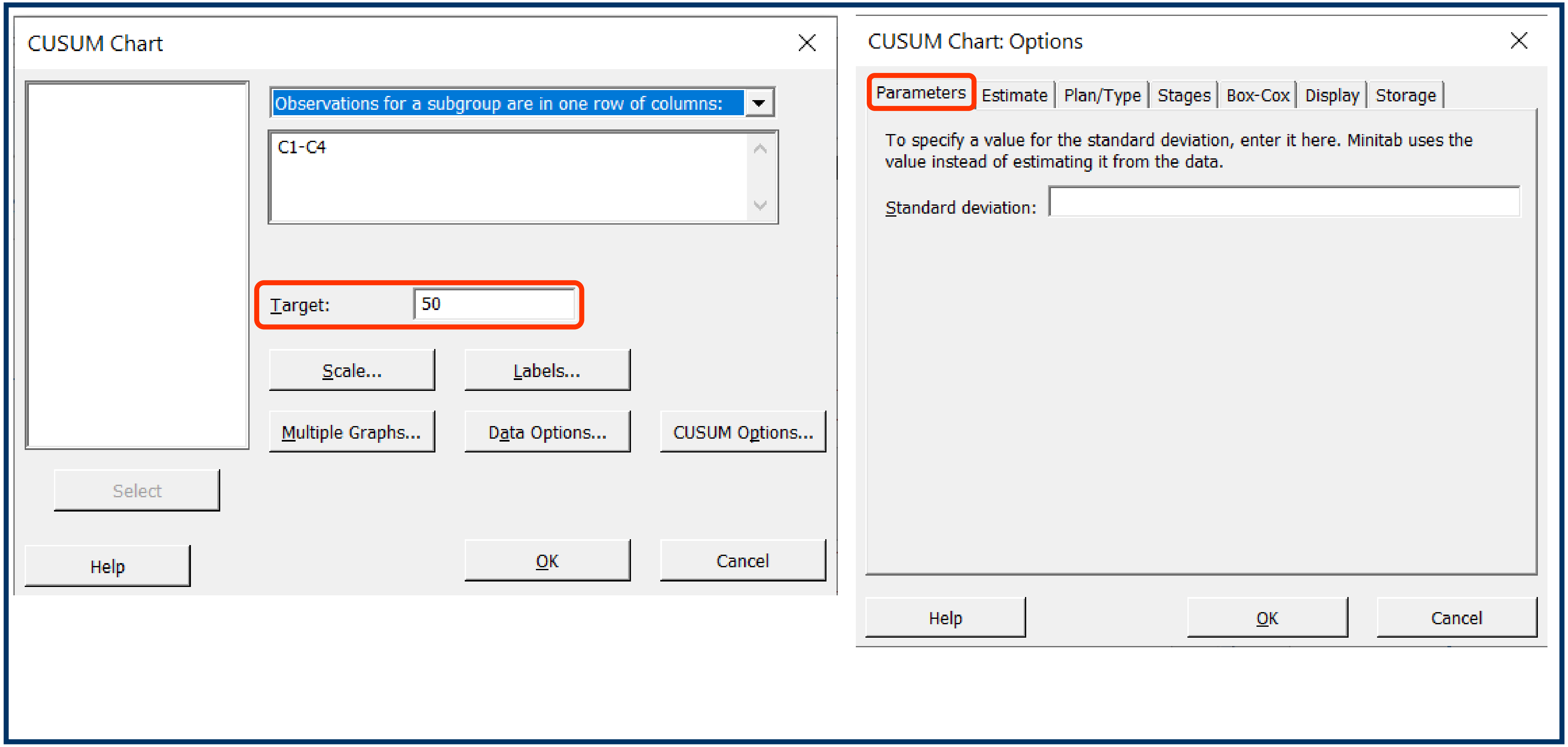
In the lower panel of the figure two points lie outside the upper arm of the V-mask, thus the process is considered out of control.

Using Minitab 18 to draw Cusum control charts with V-masks

Cusum charts can be drawn using Minitab 18 by accessing the Stat > Control charts > Time-Weighted Charts > CUSUM… pathway:



The CUSUM chart window is quite similar to the one used to choose settings for the Shewhart chart drawing, the main difference being the Target box, in which the target mean value has to be entered:



CUSUM Chart

Observations for a subgroup are in one row of columns:

C1-C4

Target: 50

CUSUM Chart: Options

To specify a value for the standard deviation, enter it here. Minitab uses the value instead of estimating it from the data.

Standard deviation:

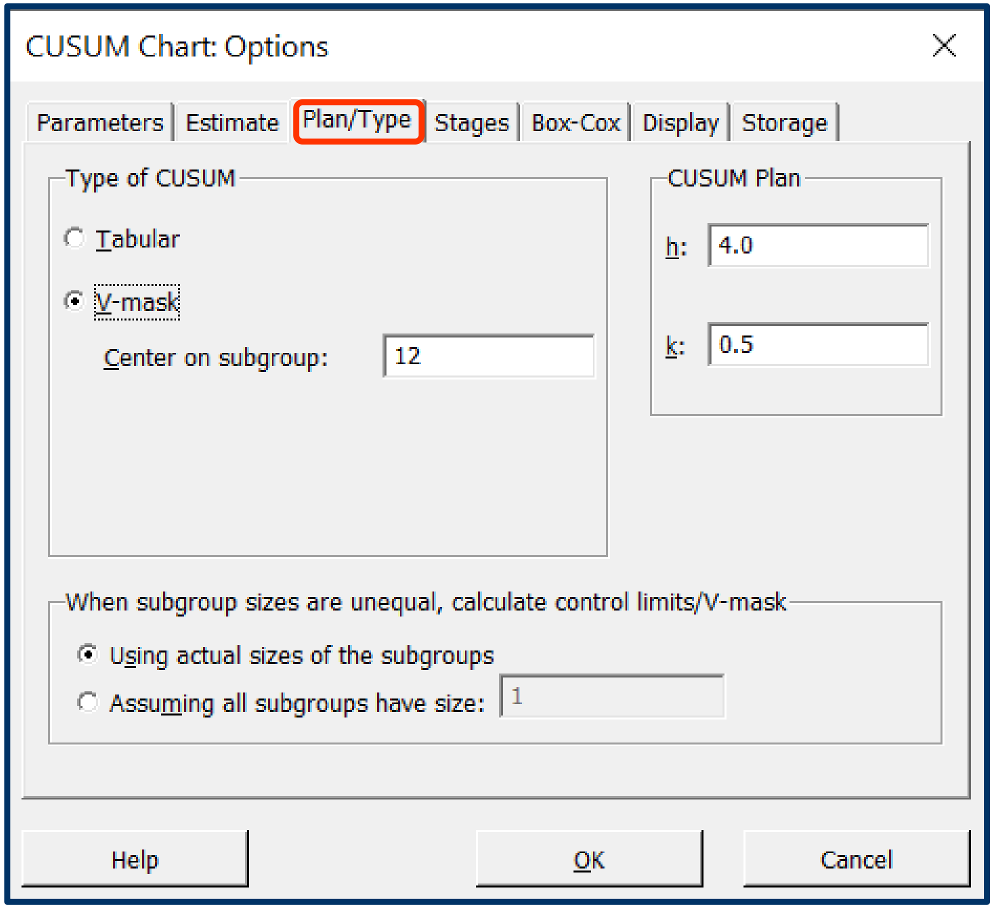
For this reason, only a box referred to Standard deviation is available in the Parameters sub-menu of the CUSUM Options… window. Different types of estimate for standard deviation can be selected in the Estimate sub-menu.

The most important sub-menu is the Plan/Type one.

Indeed, an alternative type of CUSUM chart, called *tabular*, can be selected in this sub-menu.

As for the V-mask type, the point in the cusum plot on which the mask has to be placed can be selected in the “Center on subgroup:” box.

In the example, the point corresponds to the 12th determination.



Type of CUSUM

○ Tabular

● V-mask

Center on subgroup: 12

CUSUM Plan

h: 4.0

k: 0.5

When subgroup sizes are unequal, calculate control limits/V-mask

● Using actual sizes of the subgroups

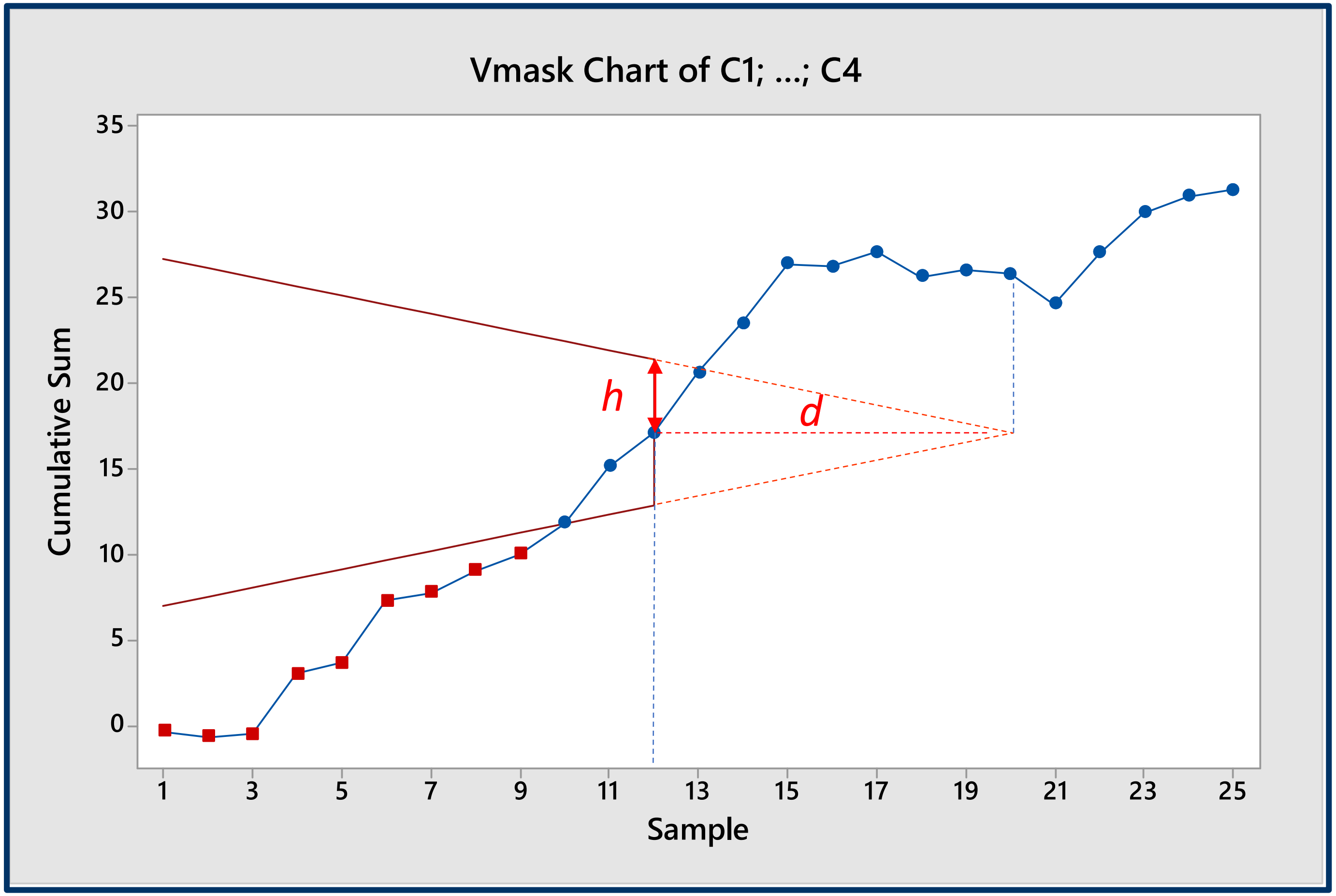
○ Assuming all subgroups have size: 1

Notably, Minitab uses a different notation to specify the V-mask parameters in the CUSUM Plan section, namely h and k.

Specifically:

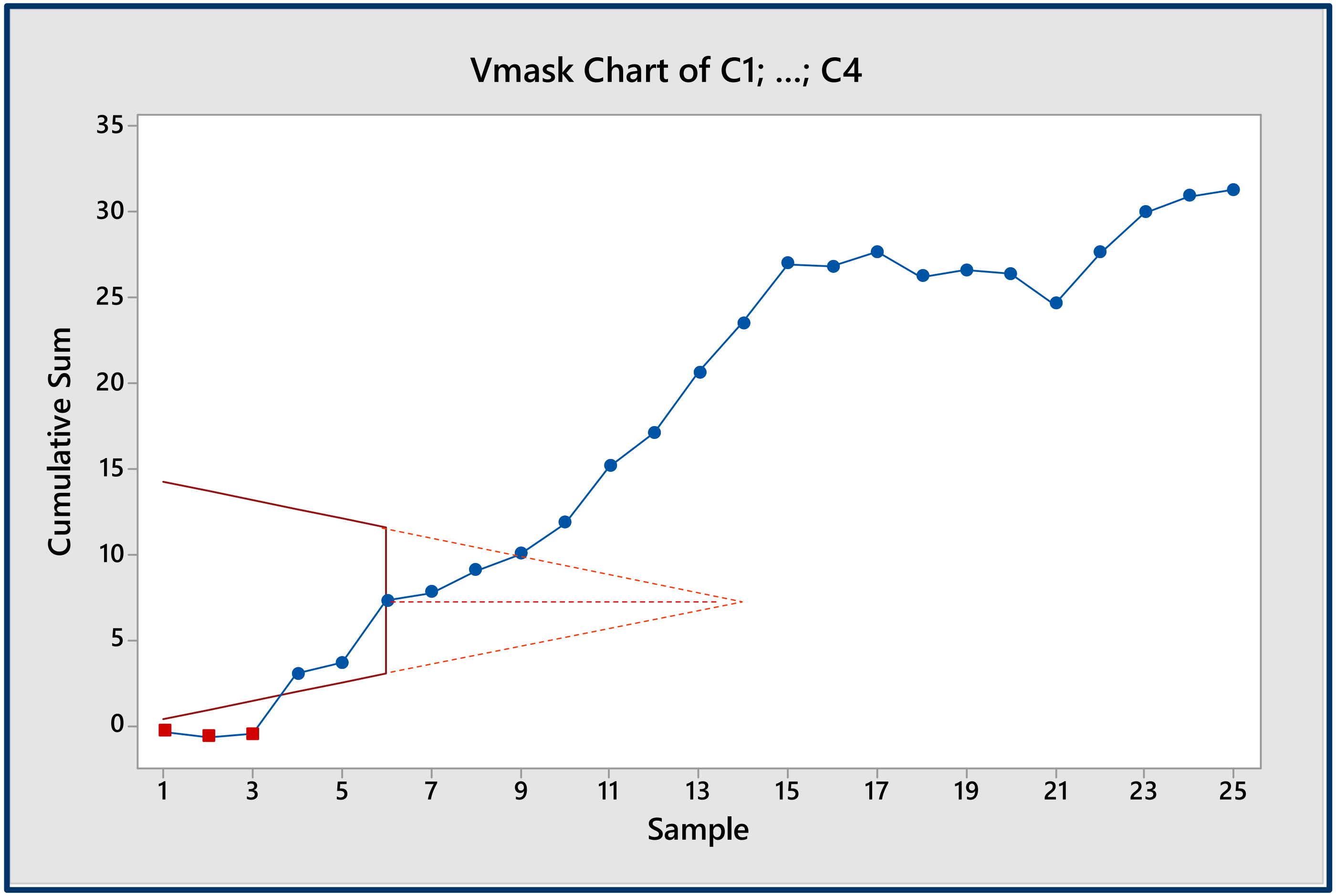
$k = \tan\theta = \delta\,\sigma_x/2$ and $h = d\tan\theta = (2/\delta^2)\ln[(1-\beta)/\alpha]\tan\theta$,
thus $k = 0.5$ corresponds to $\theta \cong 27°$ and $h = 4$ implies $d = 8$, that, in turn, leads to $\delta = 1$ for both α and β equal to 0.01.

The resulting cusum plot with V-mask superimposed on the point corresponding to the subgroup of responses #12 is shown in the following figure:



In the Minitab's representation the tip of the V-mask is not drawn, thus it was drawn in the figure (with a dotted line) to emphasize the geometric interpretation of parameter h (note that tan θ = h/d) and show that d = 8 (meant as subsequent determinations). Deviations soon occurring in the cumulative sum of the data series are apparent from the figure.

Interestingly, deviations occurring at the beginning of the data series, already shown by the Shewhart charts, are rapidly evidenced also by the Cusum chart when the mask is located on the sixth point:



Actually, estimates made when deviations from the target value are less pronounced indicate that Cusum charts are able to detect them well before Shewhart charts, although geometric parameters of the V-mask play a key role in determining their performance.

Proficiency testing

The quality of analytical measurements is enhanced by two types of testing scheme, in each of which a number of laboratories participate simultaneously:

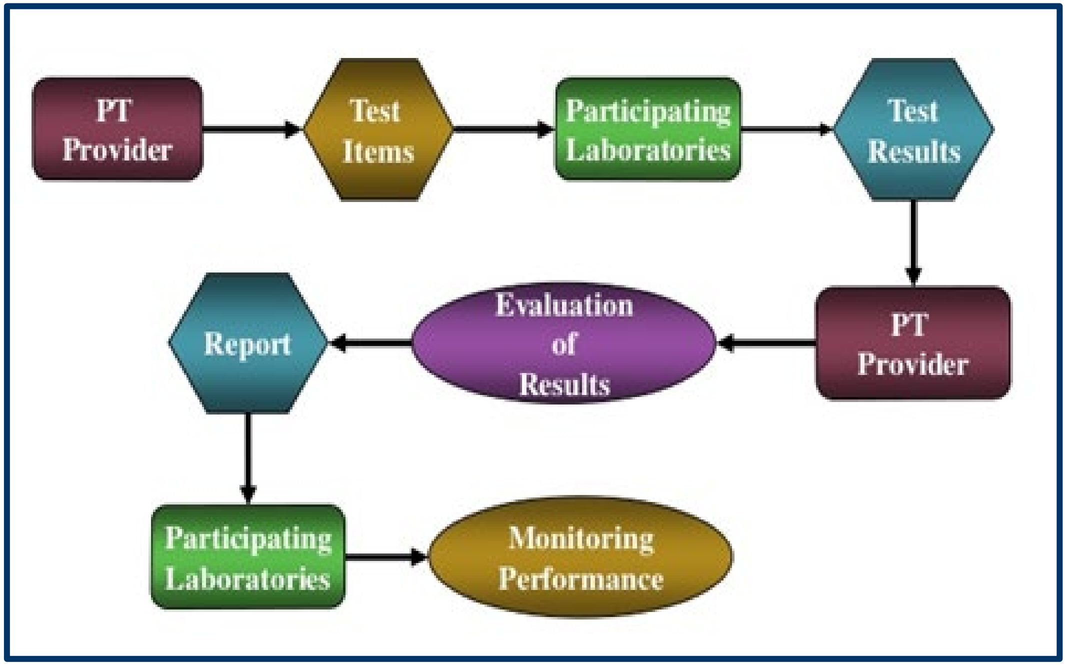
1) Proficiency testing (PT)
2) Collaborative trials (CT)

In proficiency testing schemes, aliquots from homogeneous materials are circulated to a number of laboratories for analysis, using their methods, at regular intervals (every few weeks or months), and the resulting data are reported to a central organizer.

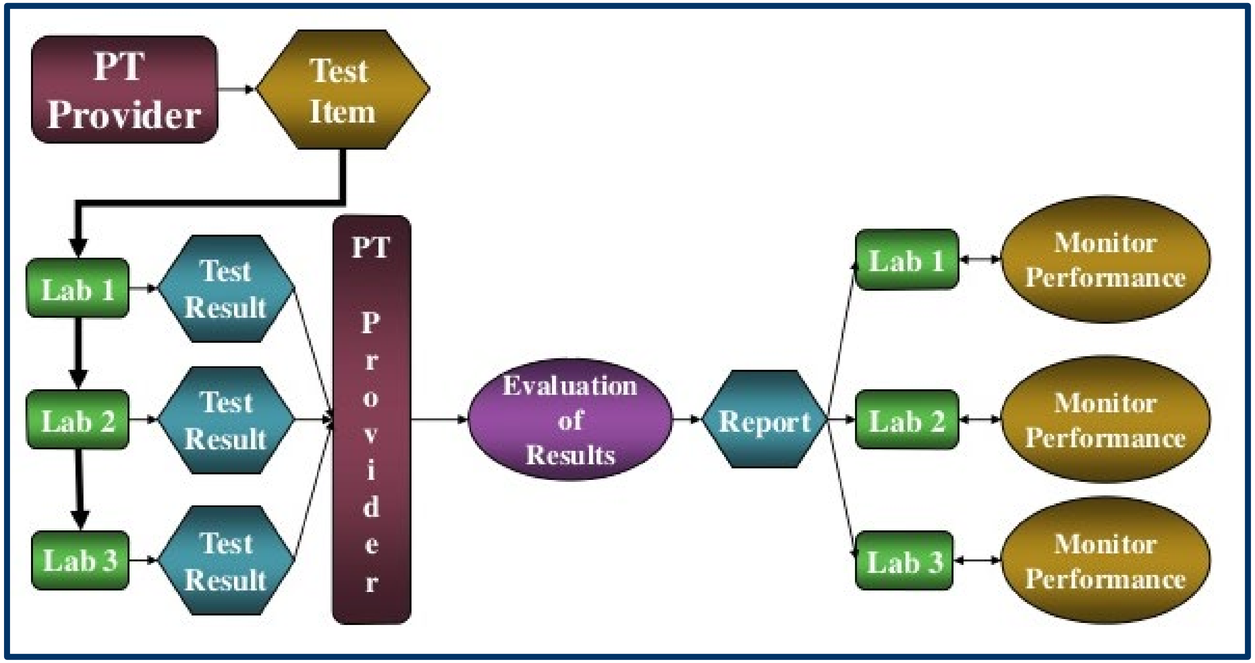
Two types of PT test are usually performed.

In the simultaneous scheme each participating laboratory makes its analysis independently and sends test results to the PT provider.

After evaluation of results, a report is sent to each laboratory by the provider.



In the sequential scheme, participating laboratories work one after the other and, once the analysis is done, each of them sends information to the next laboratory.



It is worth noting that each laboratory analyzes its portion of sample using its own usual method, and the material circulated is designed to resemble as closely as possible the samples normally submitted for analysis in the relevant field of application.

The results of all the analyses are circulated to all the participants, who thus gain information on how their measurements compare with those of others, how their own measurements improve or deteriorate with time, and how their own measurements compare with an external quality standard.

Actually, it is not uncommon that relevant differences are observed between different laboratories, even if they are well equipped and well staffed.

In one of the most common clinical analyses, the determination of blood glucose at mM levels, most of the results obtained during a proficiency test for a single blood sample approximated to a normal distribution with values between 9.5 and 12.5 mM, in itself a not negligible range.

However, the complete range of results was from 6.0 to 14.5 mM, i.e., some laboratories obtained values almost 2.5 times higher than those of others! The worrying implications of this discrepancy in clinical diagnosis are obvious.

In more difficult areas of analysis the results can be so divergent that there is no real consensus between different laboratories.

The importance of PT schemes in highlighting such alarming differences, and in helping to minimize them by encouraging laboratories to compare their performance, is very clear, and they have unquestionably helped to improve the quality of analytical results in many fields.

From a statistical point of view, particularly important aspects of proficiency testing are the methods of assessing participants’ performance and the need to ensure that the bulk sample from which aliquots are circulated is homogeneous.

The recommended method for verifying homogeneity of the sample involves taking n (≥ 10) portions of the test material at random, separately homogenizing them, if necessary, taking two test samples from each portion, and analyzing the 2n samples in a random order by a method whose standard deviation under repeatability conditions is (for example) not more than 30% of the target standard deviation (i.e., the expected reproducibility) of the proficiency test.

If the homogeneity is satisfactory, one-way ANOVA should then show that the between-sample mean square is not significantly greater than the within-sample mean square.

The results obtained by the laboratories participating in a PT scheme are most commonly expressed as z-scores, where z is given by:

$$z = \frac{x - x_a}{\sigma}$$

with x representing the result obtained by a single laboratory for a given analysis, x_a the assigned value for the concentration of the analyte, and σ the target value for the standard deviation of the test results.

The assigned value x_a can be obtained by using a certified reference material (CRM), if one is available and suitable for distribution to the participants.

In some cases, this approach is not feasible, and the relevant ISO (International Standard Organization) standard recommends three other possible approaches, in order of decreasing statistical rigor:

(i) a reference value obtained from one laboratory by comparing random samples of the test material against a CRM;
(ii) a consensus value obtained by the analysis of random samples of the test material by expert laboratories;
(iii) a consensus value obtained from all the participants in a given round of the scheme.

In the third case the median, or the mean of the interquartile range, can be finally adopted as the most reliable value, overcoming the eventual presence of outliers.

The target value for the standard deviation, σ, should be circulated in advance to the PT participants along with a summary of the method by which it has been established.

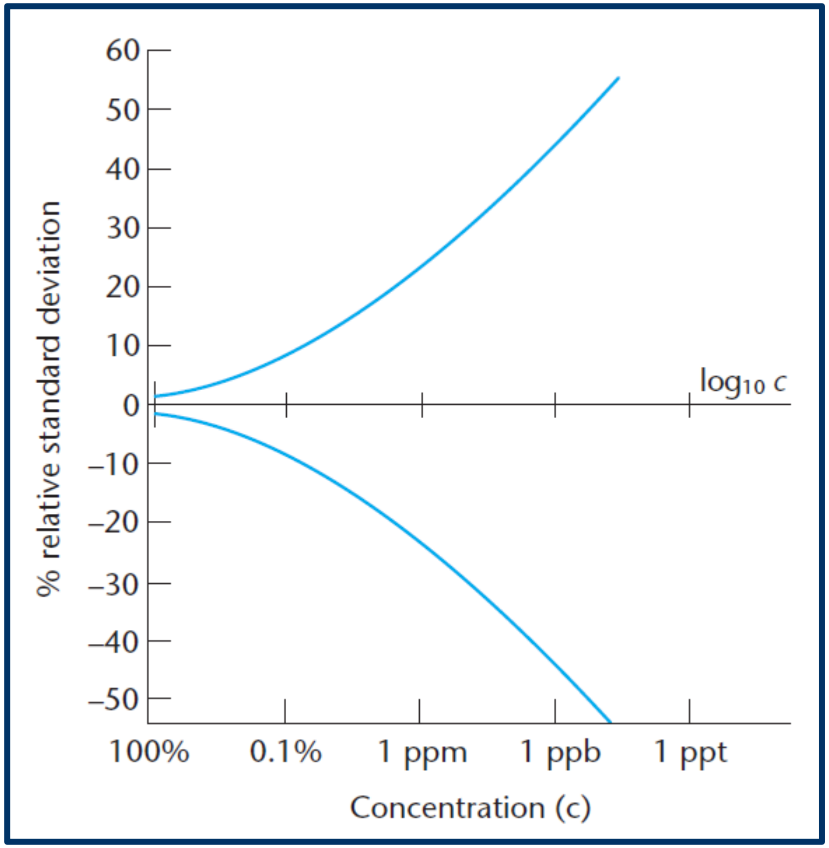
Since σ may vary with analyte concentration, a possible approach to its estimate is based on a functional relationship between concentration and standard deviation.

One of the most known relationships of this type is the Horwitz "trumpet", developed in 1982, so called because of the shape of its graphical representation.

Using many results from collaborative trials, the American statistician William Horwitz showed that the relative standard deviation of a method varied with the concentration, *c*, according to the approximate and empirical equation:

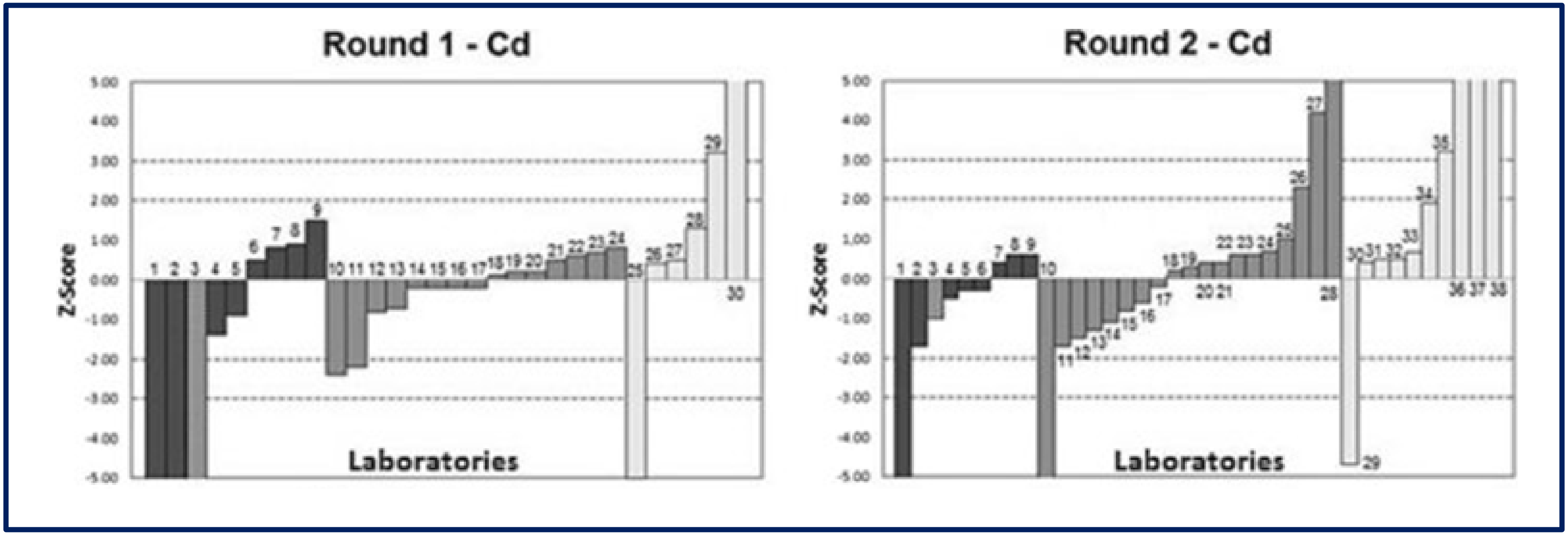
$$\mathrm{RSD} = \pm 2^{(1 - 0.5 \log c)}$$

A graphical representation of this relationship is shown in the figure on the right.



Another approach for the estimate of σ uses fitness for purpose criteria: if the results of the analysis, used routinely, require a certain precision for the data to be interpreted properly and usefully or to fulfil a legislative requirement, that precision provides the largest acceptable value of σ.

The results of a single round of a PT scheme, i.e., the z-scores obtained by different laboratories, are frequently summarized as shown in the following figure, in which values obtained during two PT rounds for the determination of Cadmium content in animal feed, based on a different extraction method (indicated by a different colour of bars) and different analytical techniques (ETAAS, Flame-AAS, ICP-AES and ICP-MS) are reported:



If the results follow a normal distribution with mean x_a and standard deviation σ, the *z*-scores will be a sample from the standard normal distribution, i.e., a normal distribution with mean zero and variance 1.

Consequently, a laboratory with a z-score value lower than 2 (in absolute value) is generally regarded as having performed satisfactorily; a z-score value between 2 and 3 is questionable (two successive values in this range from one laboratory would be regarded as unsatisfactory), and z-score values greater than 3 are unacceptable.

If an unsatisfactory score is obtained several issues should be considered:

✓ The overall standard of performance for the round:

Did a large number of participants obtain unsatisfactory results? If so, the problem may not lie within a specific laboratory.

✓ Test method performance:

- Which test methods were used by the other participants in the round? Did other laboratories use methods with very different performance characteristics?
- How was the standard deviation for proficiency assessment established? Was it appropriate for the laboratory's own needs?

- ✓ Test sample factors:

  Was the material for that round within the scope of the laboratory’s normal operations? Proficiency testing schemes often cover a range of materials appropriate to the scheme, but individual laboratories may receive materials that differ in composition from their routine test samples.

- ✓ Proficiency testing scheme factors:

  - How many results were submitted? Small numbers of results can make it difficult to establish the assigned value if the consensus approach is used.
  - Were there any problems with the organization of that particular round? Occasionally there may be problems such as unexpected test material behaviour, data entry or reporting errors, software problems or unsuitable evaluation criteria (for example, choice of assigned value or standard deviation for proficiency assessment).

If none of the above applies, the laboratory should investigate further to try to identify the cause of the unsatisfactory result and implement and document any appropriate corrective actions. There are many possible causes of unsatisfactory performance. These can be categorized as analytical errors or non-analytical errors; both are equally important.

- ✓ Examples of analytical errors

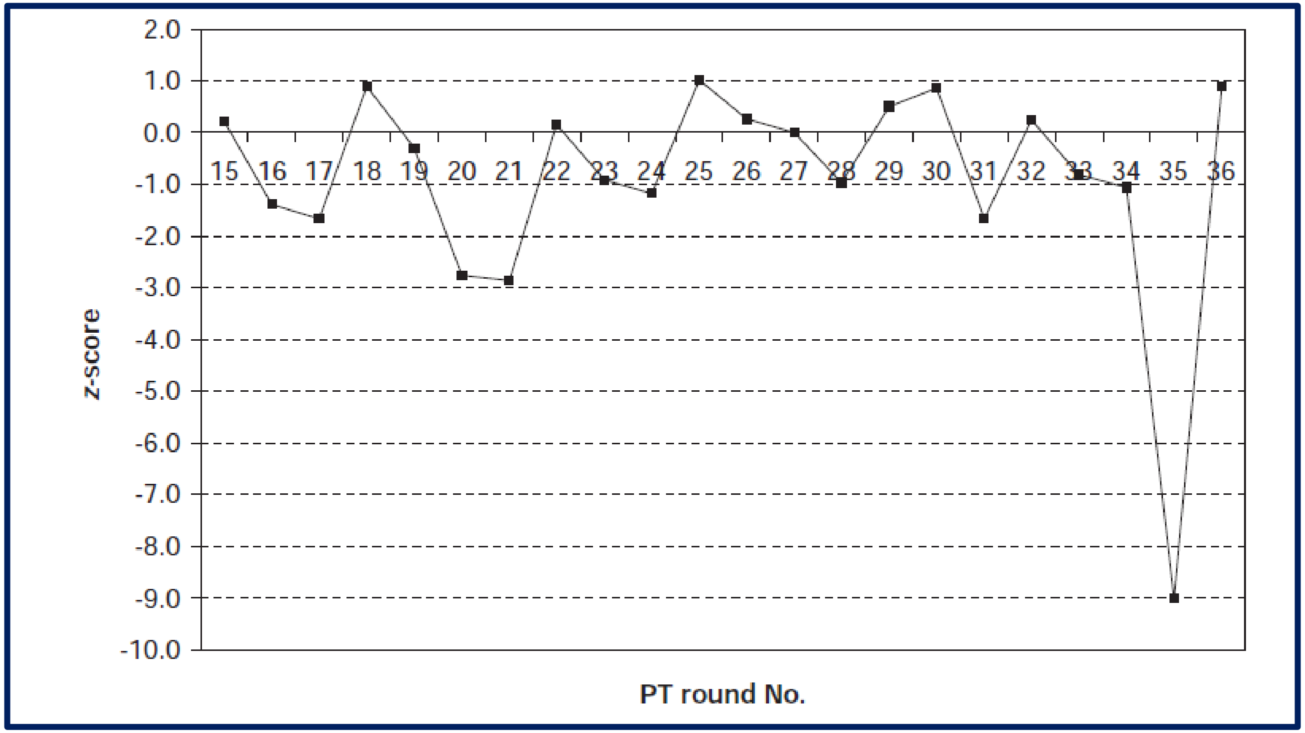
  - incorrect calibration of equipment
  - analyst error such as incorrect dilution of samples or standards
  - problems with extraction and clean-up of samples, such as incomplete extraction of the analyte from the sample matrix
  - interferences
  - performance characteristics of the chosen test method not fit for purpose
  - instrument performance not optimized.

- ✓ Examples of non-analytical errors

  - calculation errors
  - transcription errors
  - results reported in incorrect units or incorrect format

The performance of a laboratory can be monitored over time, through participation to periodic proficiency tests.

In this case, simple graphic representations based on z-scores can be adopted to visualize how the laboratory performance is changing over time, as shown in the following figure:



This graph, reporting z-scores obtained by a laboratory during 22 consecutive rounds for the determination of nickel in soil by aqua regia extraction, shows that only a specific round led to very poor results, with z-score being much higher than 3, in absolute value.

Collaborative trials

While proficiency testing schemes allow the competence of laboratories to be monitored, compared and, perhaps, improved, a collaborative trial (CT) aims to evaluate the precision of an analytical method, and sometimes its ability to provide results free from bias.

It is normally a one-off experiment involving competent laboratories, all of which, by definition, use the same technique.

A crucial preliminary experiment to be performed is a ruggedness test, i.e., an evaluation of how some experimental factors related to the analytical method (e.g., temperature, solvent composition, pH, humidity, reagent purity, concentration, etc.) will affect the results.

In some cases, a method is found to be so sensitive to small changes in one factor that it is in practice very difficult to control (e.g., it requires a very high reagent purity), thus the method is rejected as impracticable before a CT takes place.
In other instances, the trial will continue, but the collaborators will be warned of the factors to be most carefully controlled.

Full or fractional factorial designs can be often exploited for the preliminary evaluation of the method ruggedness.

In recent years international bodies have moved towards an agreement on how CTs should be performed. First, at least eight laboratories ($k \geq 8$) should be involved.

Since the precision of a method usually depends on the analyte concentration, it should be applied to at least five different levels of analyte in the same sample matrix, with duplicate measurements ($n = 2$) at each level.

A crucial requirement of a CT is that it should distinguish between the repeatability standard deviation, s_r, and the reproducibility standard deviation, s_R.
At each analyte level these are related by the equation:

$$s_R^2 = s_r^2 + s_L^2$$

where s^2_L is the variance due to inter-laboratory differences, which reflect different degrees of bias in different laboratories.

Note that, in this particular context, reproducibility refers to errors arising in different laboratories and equipment but using the *same* analytical method: this is a more restricted definition of reproducibility than that used in other cases.

The separation of variances in the equation reported above can be performed using one-way ANOVA, if the mean of responses obtained in different laboratories is normally distributed and the repeatability variance among laboratories is equal.

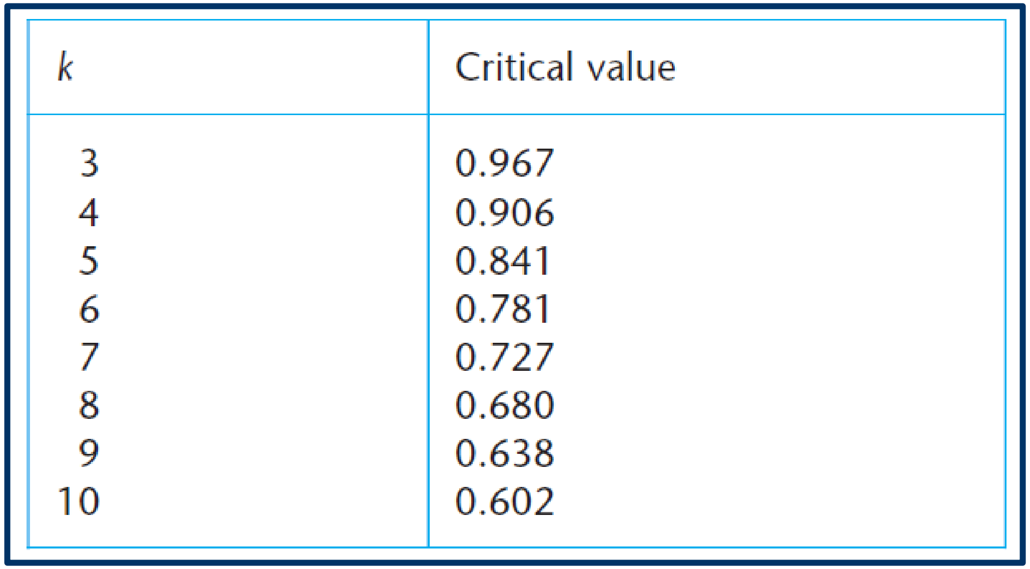
The homogeneity of variance can be tested first, using, for example, the Cochran's method, usually adopted to check if a specific variance in a set of variances can be considered an outlier. The test is based on the calculation of the following statistic:

$$C = \frac{w_{\max}^2}{\sum_j w_j^2}$$

where w_{max} is the largest of ranges w_j, i.e., of the differences between the two results obtained by each laboratory. Obviously, ranges are replaced by variances if more than 2 analyses are performed by each laboratory.

The realization of statistic C is compared with the appropriate critical value among those reported in the table on the right:

If C is greater than the critical value, the null hypothesis is rejected and the results from the laboratory providing w_{max} as a range are discarded.



k	Critical value
3	0.967
4	0.906
5	0.841
6	0.781
7	0.727
8	0.680
9	0.638
10	0.602

The absence of outlying results is evaluated using the Grubbs' test , which is applied first as a test for single outliers, and then (since each laboratory makes duplicate measurements) as a test for paired outliers.

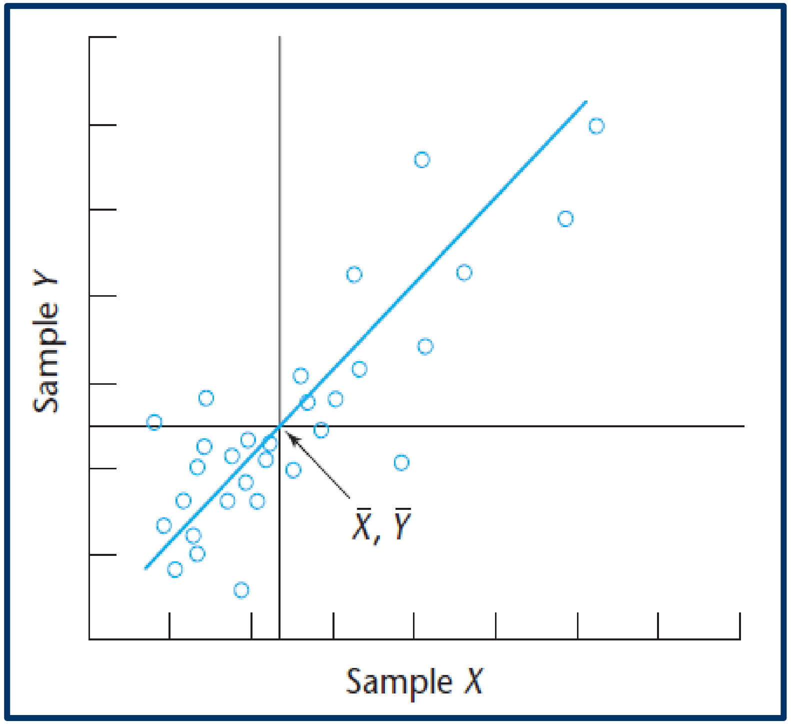
Once again, all the results from laboratories producing outlying results are eliminated from the trial unless this would result in the loss of too many data.

In many circumstances it is not possible to carry out a full CT as described so far, for example when the test materials are not available with a suitable range of analyte concentrations. In such cases, the Youden matched pairs or two-sample method, first described by the Australian statistician William John Youden in 1959, is adopted.

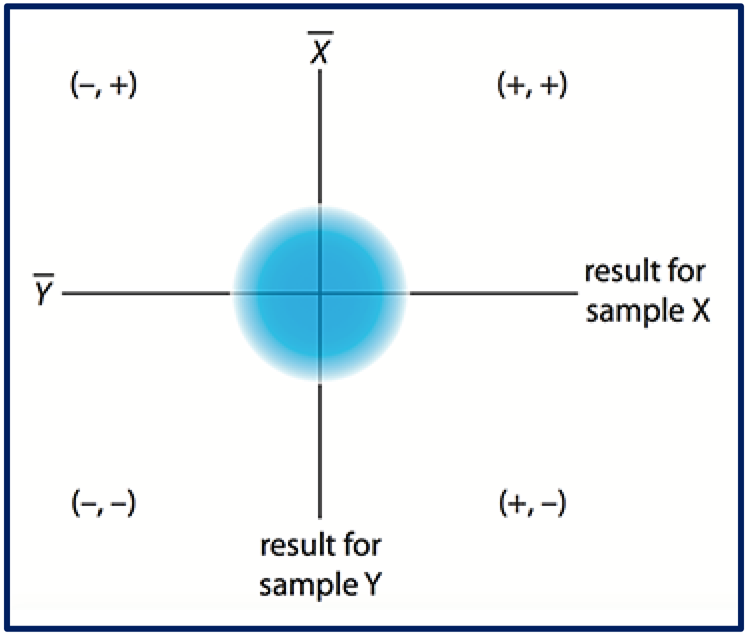
According to this method, each participating laboratory receives two materials of nearly identical composition, *X* and *Y*, and is asked to make one determination on each material.

The results are plotted as shown in the figure on the right, in which each point represents a pair of results from one laboratory.

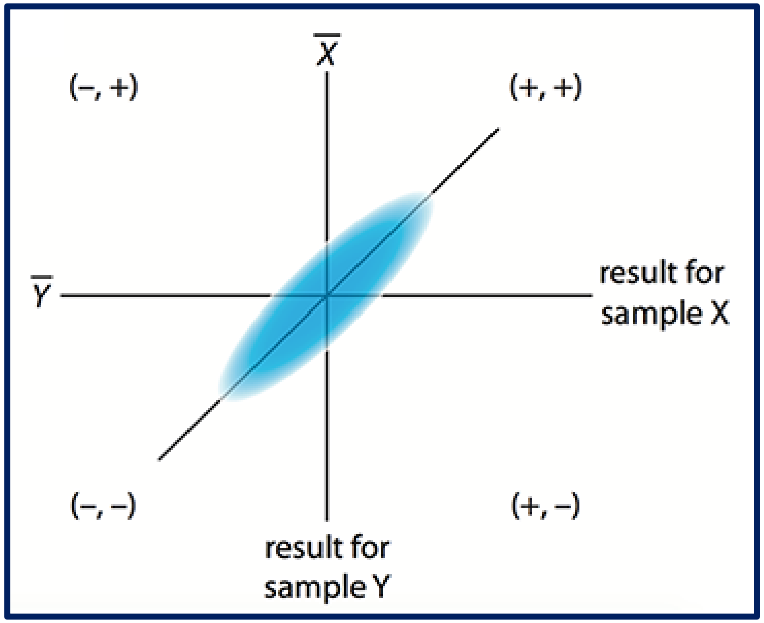
The mean values obtained for the two materials are also determined, and vertical and horizontal lines are drawn through the point ($\bar{X}$,$\bar{Y}$), thus dividing the chart into four quadrants.



If only random errors occur, the *X* and *Y* determinations may give results which are both high (+,+), both low (-,-), or one high and one low or viceversa (+,- or -,+).
These four outcomes are equally likely, thus the number of points in each of the quadrants should be roughly equal.

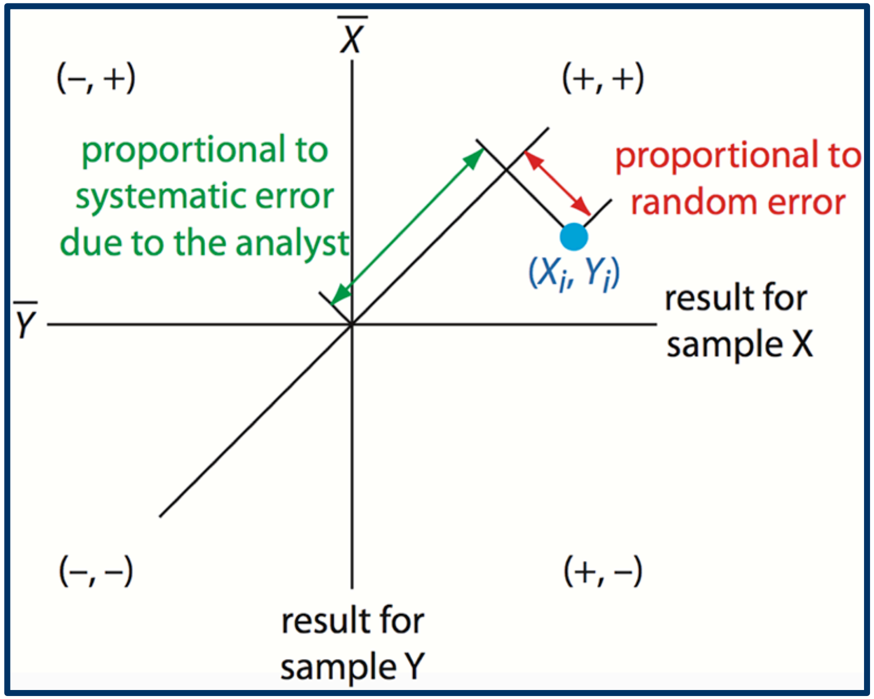


If a systematic error occurs in a laboratory, it is likely that its results for *X* and *Y* will be both high or low. Consequently, if systematic errors dominate, most of the points will be in the top-right and bottom-left quadrants. This is indeed the result obtained in most cases.



Notably, if random errors were absent all the results would lie on a line forming an angle of 45° with the axes of the plot. Consequently, when, in practice, such errors do occur, the perpendicular distance of a point from that line is a measure of the random error of the laboratory.

On the other hand, the distance from the intersection of that perpendicular with the 45° line to the point ($\bar{X}$, $\bar{Y}$) measures the systematic error due to the analyst or the laboratory.



The Youden method is a very effective approach to a collaborative trial, being capable of yielding a good deal of information in a simple form.

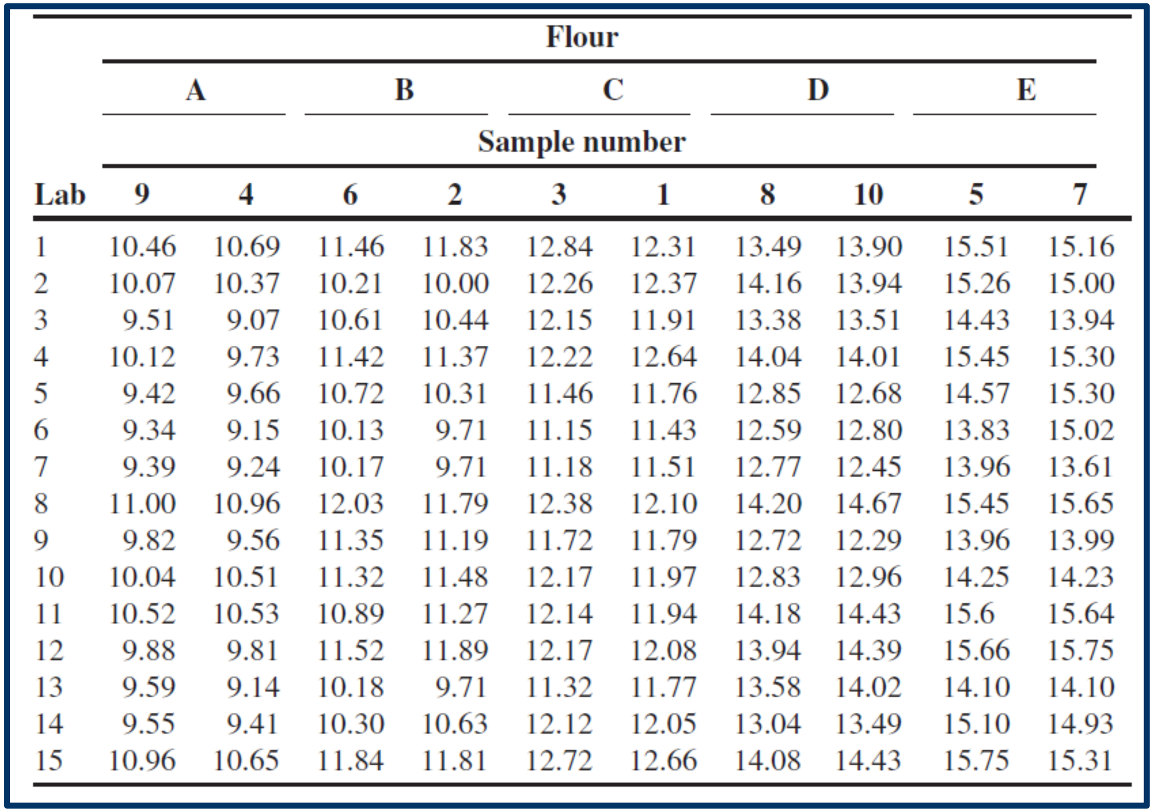
The Youden approach has the further advantage that participating laboratories are not tempted to censor one or more replicate determinations, and that more materials can be studied without large numbers of experiments.

An example of collaborative trial

A collaborative trial was devised to evaluate a new method (Q1) developed to measure quickly wheat flour protein concentrations.

Five different wheat flours with different protein levels (A-E) were prepared and 15 different laboratories around the world were contacted and agreed to participate, receiving 10 flour samples each, i.e., five flours prepared in duplicate and assigned a number by a random process.

Each laboratory was asked to analyze each sample by the Q1 method and report the results as percentages estimated to two decimal places, as shown in the following table:

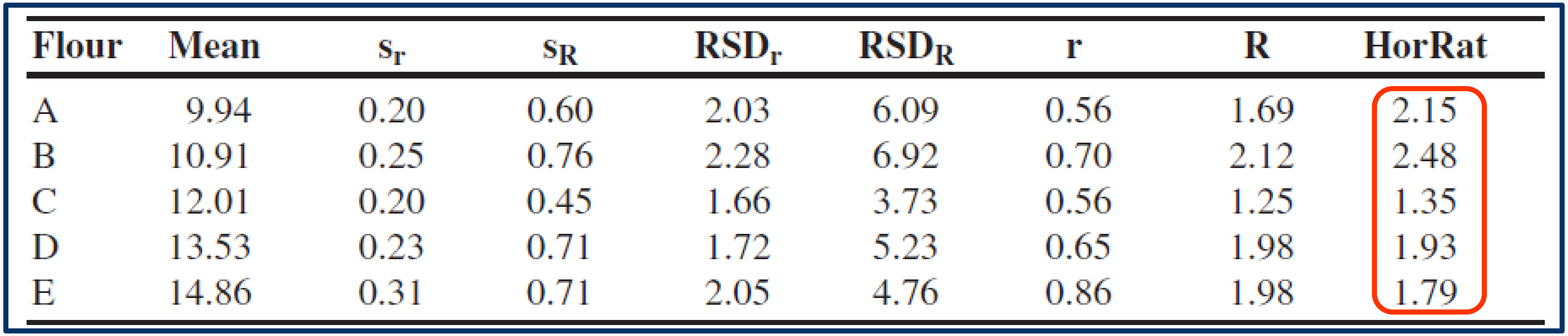


	Flour									
	A		B		C		D		E	
	Sample number									
Lab	9	4	6	2	3	1	8	10	5	7
1	10.46	10.69	11.46	11.83	12.84	12.31	13.49	13.90	15.51	15.16
2	10.07	10.37	10.21	10.00	12.26	12.37	14.16	13.94	15.26	15.00
3	9.51	9.07	10.61	10.44	12.15	11.91	13.38	13.51	14.43	13.94
4	10.12	9.73	11.42	11.37	12.22	12.64	14.04	14.01	15.45	15.30
5	9.42	9.66	10.72	10.31	11.46	11.76	12.85	12.68	14.57	15.30
6	9.34	9.15	10.13	9.71	11.15	11.43	12.59	12.80	13.83	15.02
7	9.39	9.24	10.17	9.71	11.18	11.51	12.77	12.45	13.96	13.61
8	11.00	10.96	12.03	11.79	12.38	12.10	14.20	14.67	15.45	15.65
9	9.82	9.56	11.35	11.19	11.72	11.79	12.72	12.29	13.96	13.99
10	10.04	10.51	11.32	11.48	12.17	11.97	12.83	12.96	14.25	14.23
11	10.52	10.53	10.89	11.27	12.14	11.94	14.18	14.43	15.6	15.64
12	9.88	9.81	11.52	11.89	12.17	12.08	13.94	14.39	15.66	15.75
13	9.59	9.14	10.18	9.71	11.32	11.77	13.58	14.02	14.10	14.10
14	9.55	9.41	10.30	10.63	12.12	12.05	13.04	13.49	15.10	14.93
15	10.96	10.65	11.84	11.81	12.72	12.66	14.08	14.43	15.75	15.31

T.C. Nelsen, P.W. Wehling, *Cereal Foods World*, 53, 2008, 285-288.

Cochran’s and Grubbs’ tests were performed preliminarily to evaluate if outlying ranges (i.e., variations between the two replicated results obtained by a single laboratory) or data were present, respectively. No deviation was observed.

Further method performance statistics were thus considered subsequently:



Flour	Mean	s_r	s_R	RSD_r	RSD_R	r	R	HorRat
A	9.94	0.20	0.60	2.03	6.09	0.56	1.69	2.15
B	10.91	0.25	0.76	2.28	6.92	0.70	2.12	2.48
C	12.01	0.20	0.45	1.66	3.73	0.56	1.25	1.35
D	13.53	0.23	0.71	1.72	5.23	0.65	1.98	1.93
E	14.86	0.31	0.71	2.05	4.76	0.86	1.98	1.79

In this table:

s_r and s_R correspond to repeatability (within-laboratory) and reproducibility (within + between laboratory) standard deviations, respectively;

RSD_r and RSD_R correspond to their ratios with mean, expressed as percentages;

r and R parameters correspond to, respectively, repeatability and reproducibility as expressed by the ISO standard 5725, i.e.: $r = 2.8 \times s_r$ and $R = 2.8 \times s_R$.

Finally, HorRat values correspond to ratios between experimental RSD_R values and those predicted from the Horwitz's trumpet.

The HorRat value should be comprised between 0.5 and 2.0; values lower than 0.5 can be suspected of being too good to be true, whereas values comprised between 1.5 and 2.0 can be an indication of material instability, among other possible problems. HorRat values greater than 2 can cause the method to be judged unreliable and thus unacceptable.

In the specific case HorRat values were generally quite high, and, additionally, RSD_r values were much lower than RSD_R, which is another indication of unreliability.

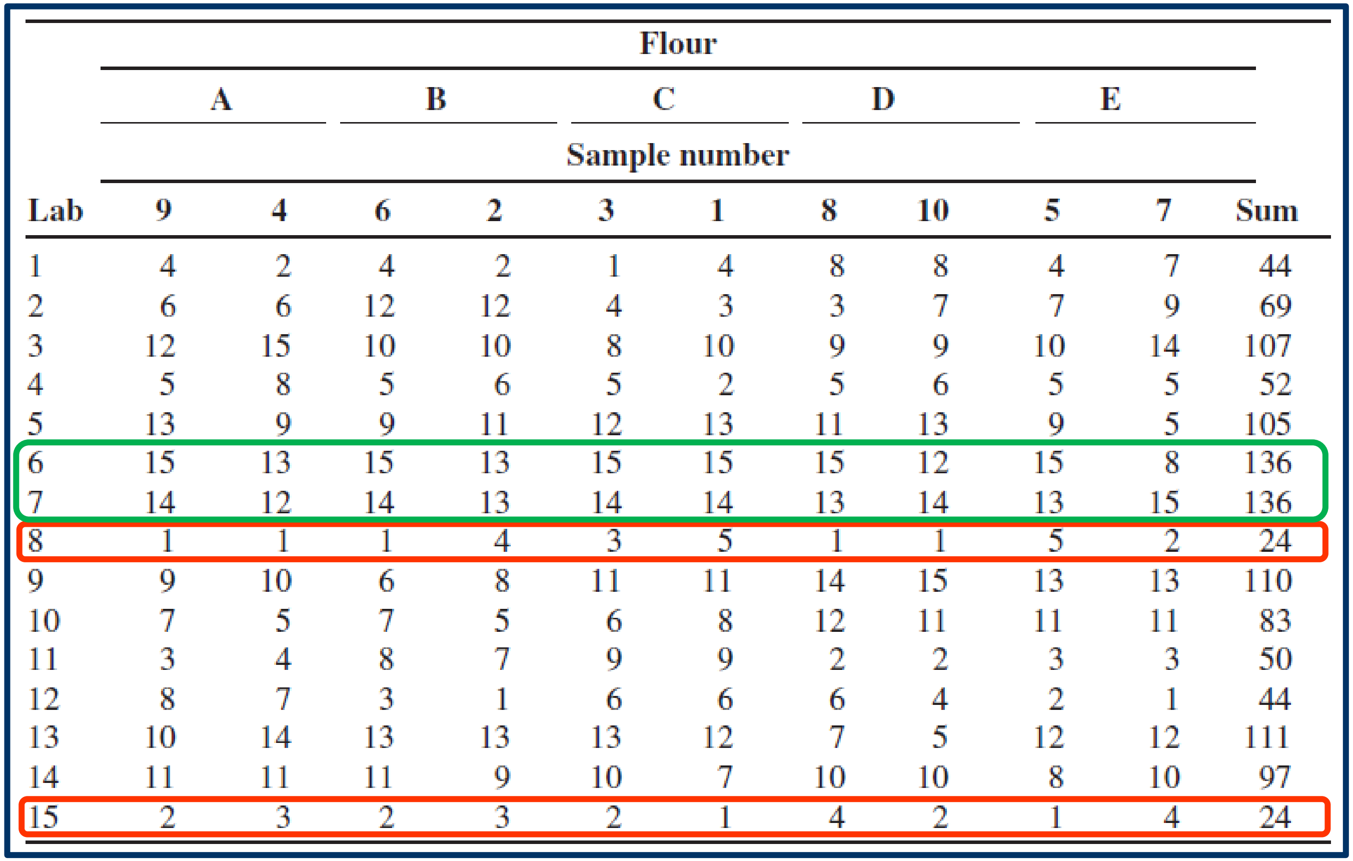
In order to make a final decision on the method, a ranks test was performed.

First, protein values for each sample were ranked from largest to smallest. Protein values obtained by each laboratory were then replaced by the corresponding rank and finally the 10 ranks were summed for each laboratory.

The rank sum can be tested for significance. Indeed, if the method is truly independent on a specific laboratory, then no lab should provide a sum consistently higher or lower than the other labs.

In the case of 15 labs and 10 samples no rank sum should be lower than 41 or greater than 119 ($P < 0.05$).

Ranks observed in the specific case are shown in the following table:



	Flour										
	A		B		C		D		E		
	Sample number										
Lab	9	4	6	2	3	1	8	10	5	7	Sum
1	4	2	4	2	1	4	8	8	4	7	44
2	6	6	12	12	4	3	3	7	7	9	69
3	12	15	10	10	8	10	9	9	10	14	107
4	5	8	5	6	5	2	5	6	5	5	52
5	13	9	9	11	12	13	11	13	9	5	105
6	15	13	15	13	15	15	15	12	15	8	136
7	14	12	14	13	14	14	13	14	13	15	136
8	1	1	1	4	3	5	1	1	5	2	24
9	9	10	6	8	11	11	14	15	13	13	110
10	7	5	7	5	6	8	12	11	11	11	83
11	3	4	8	7	9	9	2	2	3	3	50
12	8	7	3	1	6	6	6	4	2	1	44
13	10	14	13	13	13	12	7	5	12	12	111
14	11	11	11	9	10	7	10	10	8	10	97
15	2	3	2	3	2	1	4	2	1	4	24

In this case rank sums for labs 8 and 15 were consistently low and were lower than 41 and those for labs 6 and 7 were consistently high and were higher than 119.

This outcome can be interpreted with the presence of a systematic bias for the method and the method developer should try to find the source of this bias and correct it.