TECHNICAL REPORT

ISO/TR 22971

First edition
2005-01-15

Accuracy (trueness and precision) of measurement methods and results — Practical guidance for the use of ISO 5725-2:1994 in designing, implementing and statistically analysing interlaboratory repeatability and reproducibility results

Exactitude (justesse et fidélité) des résultats et méthodes de mesure — Lignes directrices pratiques pour l'utilisation de l'ISO 5725-2:1994 pour la conception, la mise en œuvre et l'analyse statistique des résultats de répétabilité et de reproductibilité interlaboratoires

Reference number
ISO/TR 22971:2005(E)

PDF disclaimer

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

Adobe is a trademark of Adobe Systems Incorporated.

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

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

Published in Switzerland

Contents

Page

Foreword

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

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

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

In exceptional circumstances, when a technical committee has collected data of a different kind from that which is normally published as an International Standard ("state of the art", for example), it may decide by a simple majority vote of its participating members to publish a Technical Report. A Technical Report is entirely informative in nature and does not have to be reviewed until the data it provides are considered to be no longer valid or useful.

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/TR 22971 was prepared by Technical Committee ISO/TC 69, *Applications of statistical methods*, Subcommittee SC 6, *Measurement methods and results*.

Introduction

ISO 5725 consists of six parts, the general structure of which is shown in Figure 1.

ISO 5725-2 was developed as a guidance document for ISO Technical Committees and other organizations responsible for undertaking inter-laboratory studies for characterizing the variability of standard measurement methods. Two measures of variability, repeatability and reproducibility, are accepted in many disciplines as representative of data encountered in measurement processes.

Repeatability refers to the variability among measurements made on nominally identical samples or materials under identical circumstances. It is recognized that, because of unknown or uncontrollable factors which influence the measurement process, repeated measurements will usually not agree. The extent of this variability can be expressed by a standard deviation, called the repeatability standard deviation, of the results of within-laboratory comparisons.

Reproducibility refers to the variability among measurements made on identical samples or materials under differing conditions by different laboratories following the same standard measurement method. Reproducibility includes effects caused by differences among instruments, reagents, operators, laboratories, and environmental conditions. The variability of results under these conditions may be described by a standard deviation called the reproducibility standard deviation.

This guidance document is divided into four clauses in addition to the Scope (Clause 1):

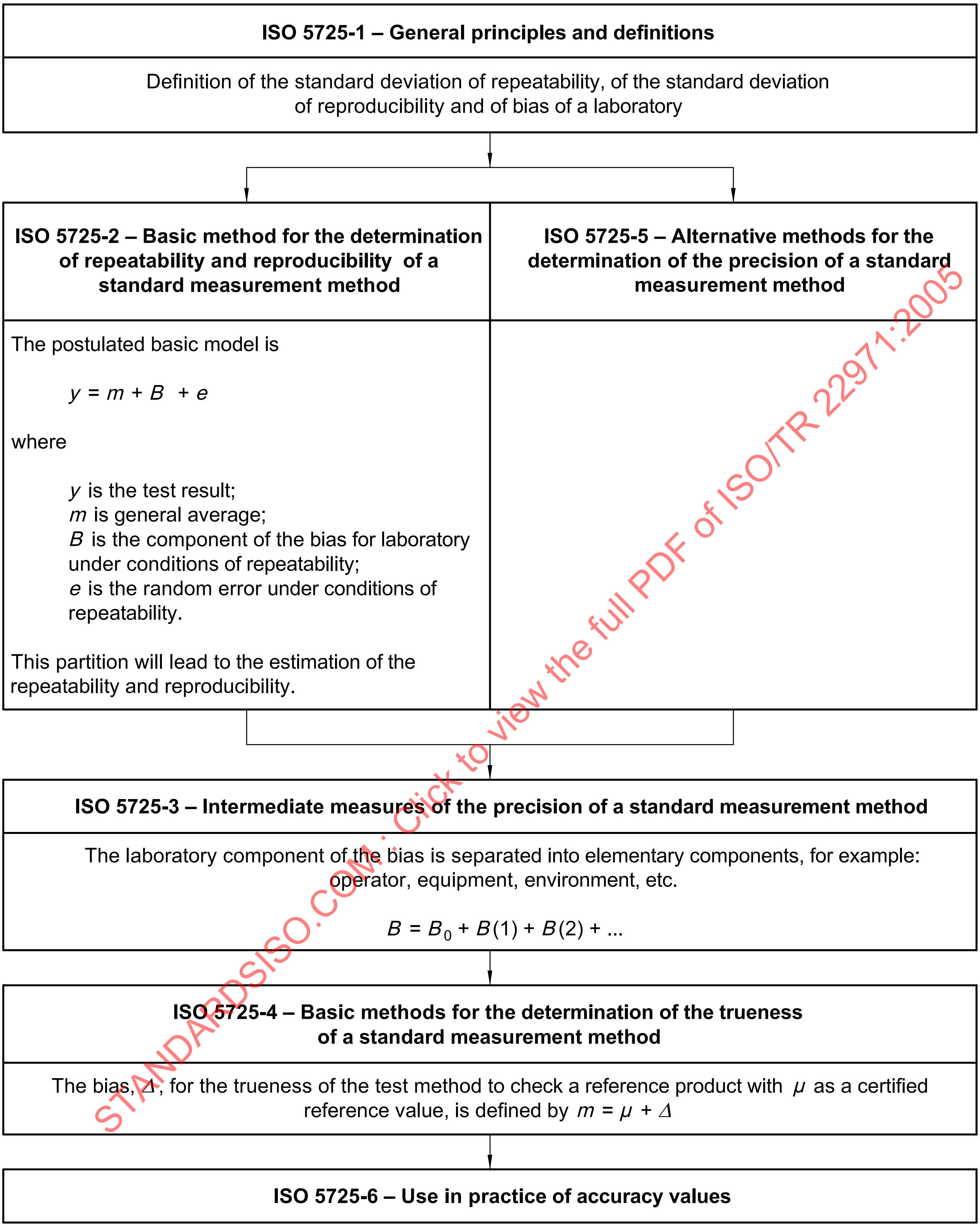
- Clause 2, Organization of an inter-laboratory programme, deals with the organization of the inter-laboratory test and covers the roles of the executive officer, laboratory personnel, and statistician in preparing for and administering the test; the choice of materials and levels of interest for the test; and the selection of laboratories. It also describes how the number of replicate measurements (to be made on each sample) is to be statistically treated and the manner in which the resulting data are to be reported.

- Clause 3, Critical examination of the data, deals with data using graphical and numerical procedures. Guidance is given as to when data are anomalous, i.e. if they are inconsistent with other data from the study, and for outlier tests that are used to identify the presence or absence of anomalous data.

- Clause 4, Estimation of repeatability and reproducibility standard deviations, deals with the estimation and interpretation of repeatability and reproducibility standard deviations. Also included is a comparison of the relative contributions of the repeatability and reproducibility standard deviations to the total variability of the test method.

- Clause 5, Worked examples using statistical software, deals with worked examples that highlight various techniques that can be used.

It is recommended that this guidance document be read in conjunction with ISO 5725-2 and should not be used as a replacement for ISO 5725-2.



ISO 5725-1 – General principles and definitions

Definition of the standard deviation of repeatability, of the standard deviation of reproducibility and of bias of a laboratory

ISO 5725-2 – Basic method for the determination of repeatability and reproducibility of a standard measurement method

The postulated basic model is

$$y = m + B + e$$

where

y is the test result;
m is general average;
B is the component of the bias for laboratory under conditions of repeatability;
e is the random error under conditions of repeatability.

This partition will lead to the estimation of the repeatability and reproducibility.

ISO 5725-5 – Alternative methods for the determination of the precision of a standard measurement method

ISO 5725-3 – Intermediate measures of the precision of a standard measurement method

The laboratory component of the bias is separated into elementary components, for example: operator, equipment, environment, etc.

$$B = B_0 + B(1) + B(2) + \ldots$$

ISO 5725-4 – Basic methods for the determination of the trueness of a standard measurement method

The bias, Δ, for the trueness of the test method to check a reference product with μ as a certified reference value, is defined by $m = \mu + \Delta$

ISO 5725-6 – Use in practice of accuracy values

Figure 1 — Structure of ISO 5725 — Application of a standardized test method to the analyses of a sample or product in different laboratories

Accuracy (trueness and precision) of measurement methods and results — Practical guidance for the use of ISO 5725-2:1994 in designing, implementing and statistically analysing interlaboratory repeatability and reproducibility results

1 Scope

This Technical Report provides users with practical guidance to the use of ISO 5725-2:1994 and presents simplified step-by-step procedures for the design, implementation, and statistical analysis of inter-laboratory studies for assessing the variability of a standard measurement method and on the determination of repeatability and reproducibility of data obtained in inter-laboratory testing.

2 Organization of an inter-laboratory programme

2.1 Requirements for a precision experiment

The whole experiment is organized

a) to provide a complete set of results for:

— p_{Lab} number of laboratories demonstrating that the test procedure is well controlled and for quantifying the observed scatter, estimated by the reproducibility;

NOTE The symbol p_{Lab} used in this Technical Report has the same meaning as the symbol p used in ISO 5725-2:1994. The change was made to clearly distinguish this symbol from the symbol, P, used for “probability”. The lowercase and uppercase P's are sometimes difficult to distinguish, particularly as subscripts.

— q number of samples or products representing different levels of results or performance. A minimum value for q is two, but from five to ten is more appropriate for demonstrating that the test procedure is able to discriminate correctly between levels;

— n number of replications cell demonstrating that the test procedure is well controlled within a single laboratory. When the number of laboratories and of levels is sufficient, at least two determinations are required;

b) to analyse statistically (see Clauses 2 and 3) a table of results reported by p_{Lab} laboratories analysing q samples, tested n times under conditions of repeatability.

The table of the results submitted to the executive officer is shown in Table 1 (see ISO 5725-2:1994, 7.2.8).

2.2 The responsibilities of the personnel involved in a precision experiment

2.2.1 General

An inter-laboratory programme is very expensive, both in terms of its co-ordination and its participation. Hence, the performance testing should be well co-ordinated and planned. In any inter-laboratory programme, it is necessary to consider three types of activity as shown in Figure 2.

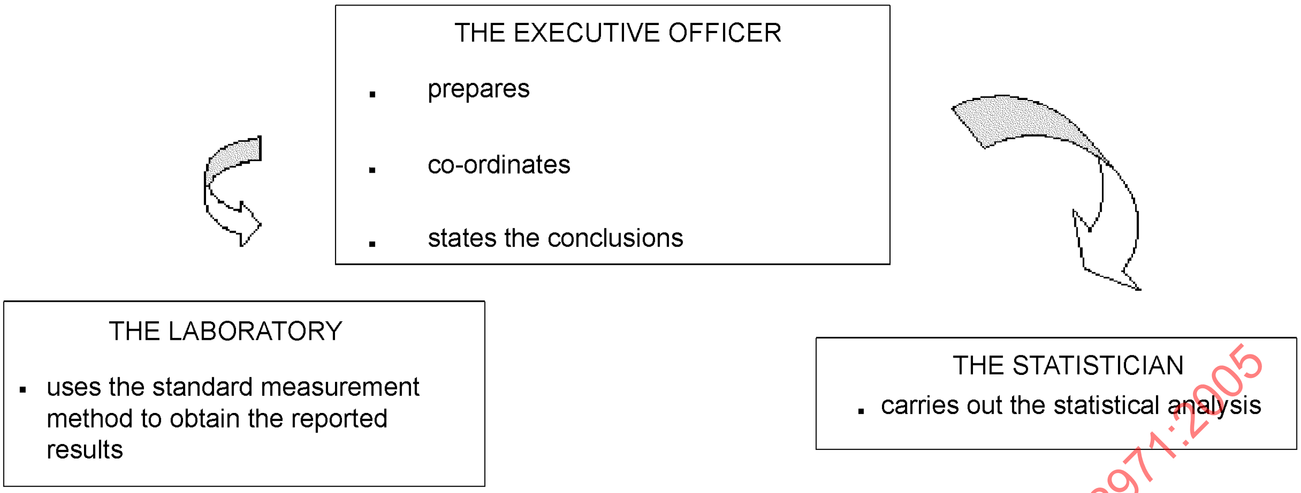

THE EXECUTIVE OFFICER
▪ prepares
▪ co-ordinates
▪ states the conclusions
THE LABORATORY
▪ uses the standard measurement method to obtain the reported results
THE STATISTICIAN
▪ carries out the statistical analysis

Figure 2 — The responsibilities of executive functions

2.2.2 Executive officer

The executive officer's main tasks are

- to organize the inter-laboratory programme, with the advice of the statistician for the construction of the experimental design;
- to co-ordinate the progress;
- to state the conclusions.

The duties of the executive officer may be undertaken by more than one person. However, only one person should be responsible for the entire programme. The executive officer should be familiar with the standard method, but should not participate in the measurement process.

2.2.3 Laboratory

The laboratory personnel should be fully experienced with the test measurement method.

The laboratory shall undertake the analysis, adhering to the test procedures received from the executive officer. Any comments by the laboratory on the use of the test method should be reported to the executive officer. However, the procedures carried out by the laboratory should be those provided by the executive officer.

The laboratory shall comply with any requirements prescribed by the executive officer, including

- storage of the samples;
- date and order of carrying out the analysis.

The laboratory shall provide the executive officer with the results of analysis in a manner prescribed by the executive officer.

2.2.4 Statistician

The statistician shall receive from the executive officer the raw data obtained using the stated method and as reported in Table 1.

The statistician shall examine the data and apply any statistical test, preferentially the tests described in ISO 5725-2, to identify potential outliers. Statistical outliers shall be brought to the attention of the executive officer. The executive officer shall undertake an appropriate investigation to ascertain whether to retain, reject or modify any data.

The statistician shall carry out the statistical analyses, prepare graphical plots, and provide estimates of the means and variances (ISO 5725-2:1994, 7.1.2). The statistician shall summarize all the results of the statistical analyses in a report that shall be sent to the executive officer.

3 Critical examination of the data

3.1 Description of the data

3.1.1 Raw data

The data are presented as shown in Table 1. Tables 2 and 3 are derived from Table 1. Some validated statistical software packages may provide different presentations of the same information.

Table 1 — Collation of all raw data

Laboratory	Level						
	1	**2**	…	j	…	$q-1$	q
1							
2							
…							
i				y_{ij1} y_{ij2} y_{ijn}			
…							
p_{Lab}							

Table 2 — Collation of the mean values for each cell in Table 1

Laboratory	Level						
	1	**2**	…	j	…	$q-1$	q
1							
2							
…							
i				m_{ij}			
…							
p_{Lab}							

Table 3 — Collation of values indicating the spread [a] of the values for each cell in Table 1

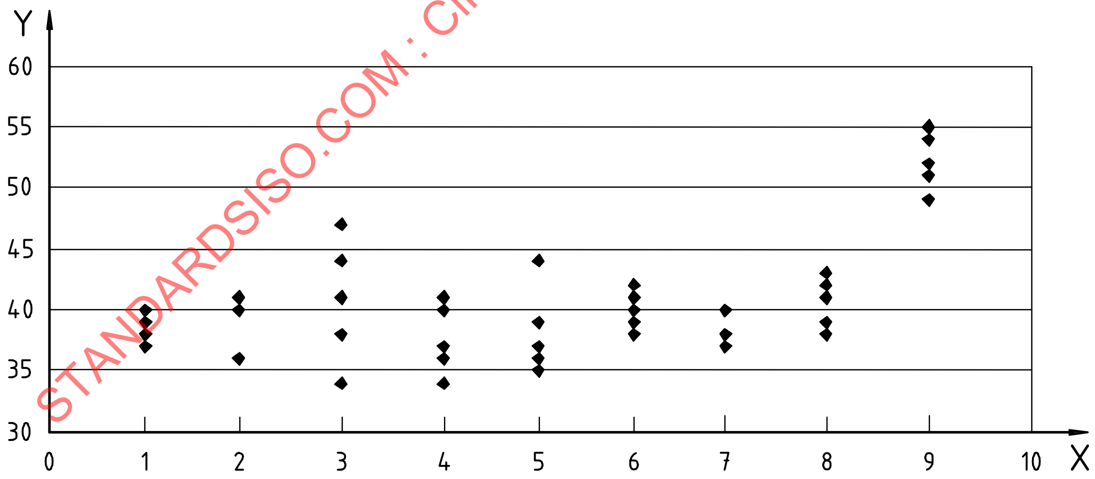
Laboratory	Level						
	1	**2**	…	j	…	$q-1$	q
1							
2							
…							
i				s_{ij}			
…							
p_{Lab}							

[a] The most common measure of spread is the standard deviation.

3.1.2 Graphical representation of the data

3.1.2.1 Results plotted versus laboratory number (raw data plot)

Before carrying out any tests to determine potential outliers, it is recommended that graphical plots of the raw data be made. In this way, an instant "picture" of the results can be depicted, for example, as shown in Figure 3 (which is based on ISO 5725-2:1994, Figures B.1 to B.4). A great deal of information can be obtained by a visual inspection of a graphical plot of the raw data, and an instant appraisal of the spread of data ascertained. Hence, an indication of the presence of outliers might be suggested, or unusual differences might become apparent, at particular levels of interest, simply by a visual inspection of the appropriate plot of data. For example, in Figure 3, the plot of results for laboratory 3 might suggest a larger-than-expected spread of results compared to all the other laboratories; hence, the overall repeatability will be affected. This possibility can be confirmed by Cochran's test. In addition, the results for laboratory 9 might suggest an outlier with respect to the mean value of the laboratory when compared to the other mean values for all the other laboratories. Hence, reproducibility might be affected and this can be confirmed by Grubbs' test.



Key

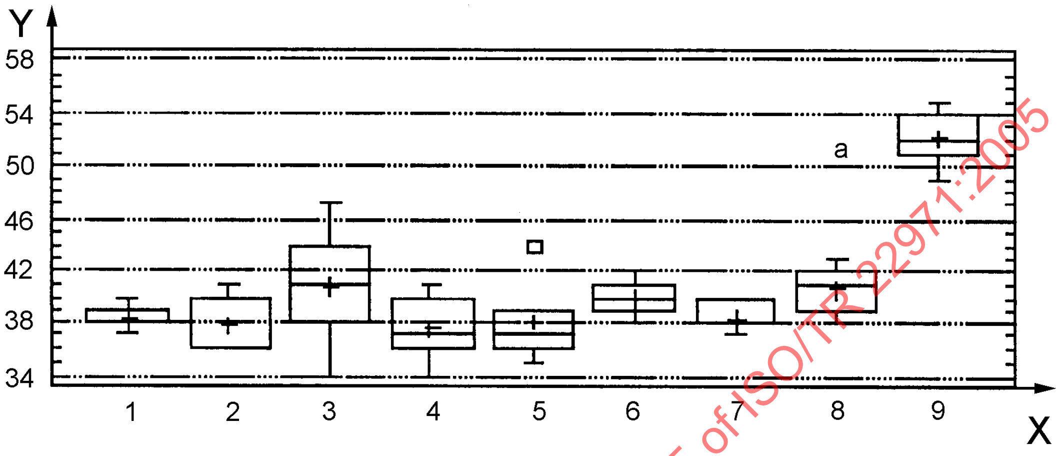
X Laboratory identification

Y Results per laboratory

Figure 3 — Graphical representation of raw data for a particular level of interest

3.1.2.2 Boxplot (“box-and-whiskers” plot)

Where many results are reported, especially for a particular level of interest, a “box-and-whiskers” plot can reveal information similar to that in 3.1.2.1; for an example, see Figure 4. However, this type of plot, which is based on robust statistics including the determination of the median value, is not described in ISO 5725-2. It is, however, defined and illustrated in the examples in Clause 4, since these graphs are available in most statistical software packages.



Key

X Laboratory identification

Y Results per laboratory

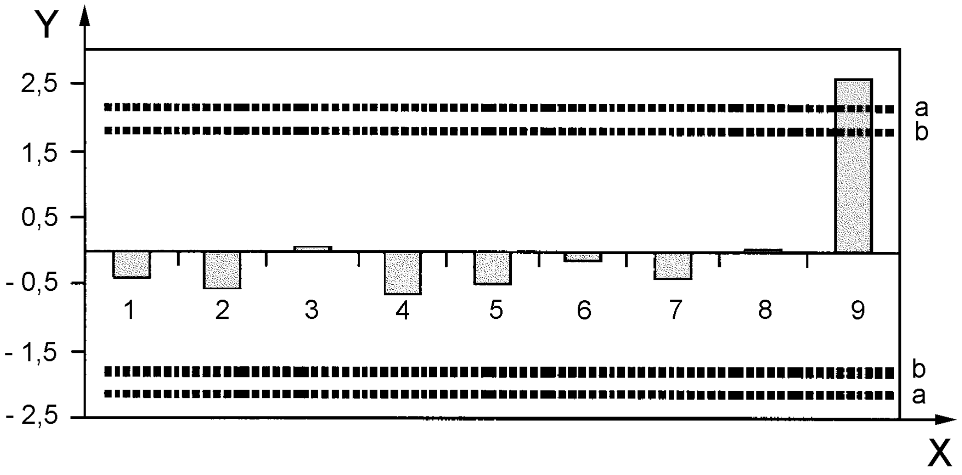
[a] “+” indicates the average.

Figure 4 — “Box-and-whiskers” plot

3.1.2.3 Mandel's h and k tests statistics

3.1.2.3.1 Mandel's h plot

For a particular level of interest, the mean values obtained for all the laboratories are used to calculate a single overall mean value. This value is then used to calculate Mandel's h statistic for all the laboratories for this level. This statistic is defined in ISO 5725-2:1994, Equation (6). This statistic is the ratio of the difference between the mean for a particular set of data and the mean of all sets of data, and the standard deviation of the means from all the sets of data. This quotient value is then plotted and compared with computed or tabulated ratio values obtained for 95 % and 99 % confidence levels. The same procedure is then used to calculate Mandel's h statistic for all the laboratories for all the other levels of interest (see Figure 5). It should be noted that both positive and negative values can be plotted.



Key

X Laboratory identification

Y Mendel's h

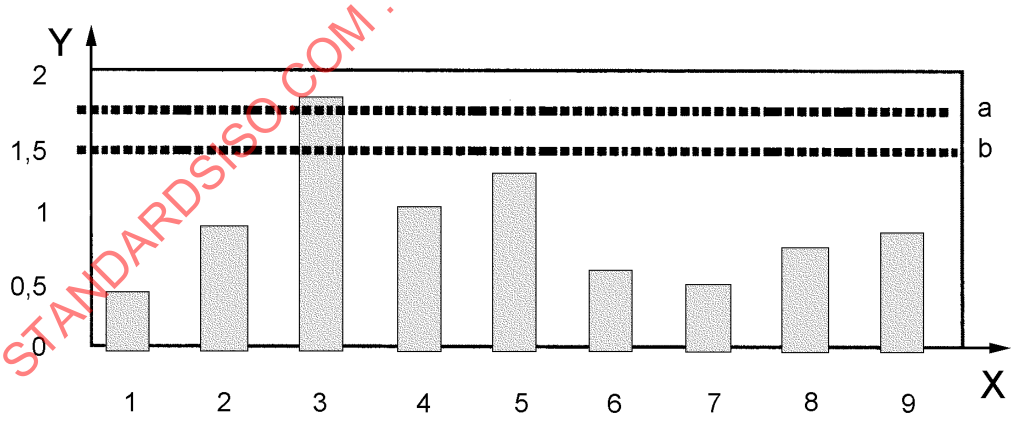
[a] 99 % confidence level

[b] 95 % confidence level

Figure 5 — Mandel's *h* plot

3.1.2.3.2 Mandel's k plot

For a particular level of interest, the standard deviations obtained for all the laboratories are used to calculate a mean standard deviation or pooled single standard deviation. This value is then used to calculate Mandel's k statistic for all the laboratories for this level. This statistic is defined in ISO 5725-2:1994, Equation (7). It is the quotient of the standard deviation of results and the mean or pooled standard deviation. This quotient value is then plotted and compared with computed or tabulated ratio values obtained for 95 % and 99 % confidence levels. The same procedure is then used to calculate Mandel's k statistic for all the laboratories for all the other levels of interest (see Figure 6). It should be noted that only positive values are plotted.



Key

X Laboratory identification

Y Mendel's k

[a] 99 % confidence level

[b] 95 % confidence level

Figure 6 — Mandel's *k* plot

3.1.2.3.3 Graphical inspection

From the plots, individual results can be identified for each laboratory that might be considered different from the expected distribution of results. For example, the *h* plot for particular levels of interest for each laboratory might approach or exceed the computed Mandel's *h* statistic value at the 95 % or 99 % confidence level if the Grubbs' test shows outliers to be present. In addition, the *k* plot for particular levels of interest for each laboratory might approach or exceed the computed Mandel's *k* statistic value at the 95 % or 99 % confidence level if Cochran's test shows outliers to be present.

3.2 Tests for outliers

3.2.1 General points

3.2.1.1 Level of confidence

The treatment of outliers is dealt with in ISO 5725-2:1994, Clause 7, particularly 7.1 to 7.3. An outlier can be considered as a result which is sufficiently different from all other results to warrant further investigation. Depending on the type of distribution into which the results fit, a result that appears to be an outlier could, in reality, be a valid result. ISO 5725-2:1994, 7.3.2.1 and 7.3.3.2, recommends confidence levels of 95 % for outliers termed “stragglers”, and 99 % for outliers termed “statistical outliers”. For individual circumstances, the selection of 95 % and 99 % confidence levels means that one result in 20, and one result in 100, respectively, might be erroneously misinterpreted. Hence, this one result could occur by chance and the degree of confidence stated in ISO 5725-2 might not be appropriate for individual needs. This might represent a degree of acceptability that is not sufficient for certain purposes. This would mean that individual circumstances would merit individual consideration as to whether ISO 5725-2, in terms of the confidence levels used, should be applied.

3.2.1.2 Basic assumptions

In the tests used to determine the presence or absence of outliers, it is assumed that the results are distributed in a Gaussian manner (commonly referred to as a normal distribution; ISO 5725-2:1994, 1.4) or at least a single unimodal distribution (ISO 5725-2:1994, 7.3.1.7). Hence, before undertaking any test, especially one that involves a large number of results, a check to confirm this assumption should be made. It is also assumed (ISO 5725-2:1994, 1.3 and 5.1.1) that the number of results within each set of data (from each laboratory) is the same and that the number of results for each level of interest, or number of different samples, is the same. Thus, the results are “balanced”. If the results are not “balanced”, then it is recommended (ISO 5725-2:1994, 7.2.2) that results from appropriate sets of data be randomly discarded until a “balanced” situation is created. Although a “balanced” situation is preferred, it is recognized (even within the examples illustrated in ISO 5725-2) that “unbalanced” situations can be accommodated. It is further assumed (ISO 5725-1:1994, 4.4, and ISO 5725-2:1994, 7.3.3.3) that results are obtained under repeatability conditions. Hence, it can be assumed that the samples for a specific level of interest are homogeneous, identical in all respects, and analysed within a short period of time using the same reagents and calibration solutions. In theory, these criteria have to be satisfied before any tests can be used to establish the presence or absence of outliers.

3.2.1.3 Declaration of outliers

When carrying out the outlier tests, it should be understood that outliers should not be discarded or rejected purely from a statistical point of view. For each sample, the reason why the result is different from all the others should be investigated and identified. Outlier tests (based on the assumptions used) indicate whether there is sufficient statistical cause for an outlier; it will not indicate why it has occurred. It is only after thorough investigations have been undertaken to identify likely causes that data should be declared outliers and discarded.

When a particular level of interest has been analysed by Cochran's test, Grubbs' test or some other test and no outliers or further outliers are identified, then other levels of interest are similarly tested. If several outliers are identified in different levels of interest for a given set of data produced by a single laboratory, it may be necessary to consider whether all the data sets for all the levels of interest should be further investigated.

Furthermore, it should be considered whether to discard only the outliers identified for that laboratory or whether to reject the entire set of data for that laboratory. Experience in this matter will dictate the necessary course of action and should be undertaken on an individual basis and with knowledge of the investigations carried out to identify the possible causes.

Inspection of graphical outlier plots may provide additional evidence to that provided by the (numerical) outlier tests. Hence, Mandel's h and k plots can be used to facilitate these decisions. Consideration should be given to rejecting all results from a laboratory if a particular set of data (for example, for laboratory 9 in Figure 5 or for laboratory 3 in Figure 6) shows that all the calculated values are positive and approach or exceed the tabulated values at the 95 % and 99 % confidence levels. It might be that, in the example chosen, all mean values for this laboratory are greater than the corresponding values for all the other laboratories. This fact may be a cause for concern. As before, the decision to reject or discard data should be taken only in light of appropriate investigations undertaken to ascertain the likely cause of the outliers.

3.2.1.4 Choice of tests

For each laboratory or level of interest or particular sample, most outlier tests compare some measure of the relative distance of a suspect result to the mean of all the results, and assess this comparison to ascertain if the result could have occurred by chance. Many tests are available but, for practical purposes, not all are fully described in ISO 5725-2:1994, as stated in ISO 5725-2:1994, 7.1.3. Thus, in addition to those tests referred to in ISO 5725-2:1994, 7.1 to 7.3, it should be understood that other tests may also be used for determining potential outliers. Hence, whether to use outlier tests that are not described in ISO 5725-2 is left to the individual statistician to decide. Full details of the test used must be recorded with the results.

3.2.2 Cochran's test

3.2.2.1 Principle

Cochran's test is used to check if there are standard deviations for cells in Table 3 that are exceptionally large and that would inflate the estimate of the repeatability standard deviation if retained. The statistic used in Cochran's test is closely related to Mandel's k statistic.

3.2.2.2 Interpretation

From Table 3, Cochran's test can be used to identify whether the largest variation for a given set of data is consistent with the variations in data from all the sets of data, for that level of interest. This test enables the quotient of the largest variance obtained for a particular laboratory and the sum of all the variances obtained for all laboratories to be calculated. This calculated value is then compared with the computed or tabulated (critical) ratio and the presence of stragglers or outliers is assessed.

Cochran's test will identify those variances that are greater than the expected variances for that level of interest. In this respect, the test is a one-sided test, as that laboratory with the smallest variance (in relation to the other laboratories) is not subjected to this test. The decision to repeat Cochran's test will depend on whether an outlier is identified and on the number of data sets to be tested for that particular level of interest. If an outlier is not indicated, then the test is not repeated. If an outlier is indicated, then caution should be exercised whether to repeat Cochran's test on the remaining sets of data for that level of interest. This caution is particularly relevant for small quantities of data, especially if approximately 20 % of the data are eventually rejected as outliers.

3.2.2.3 Alternative tests

Alternative to Cochran's test are Bartlett's test, Levene's test and Hartley's test. However, there may be occasions, particularly with borderline cases, when outliers are identified using one test but are not identified using another test. Hence, it is important for the statistician to report which tests have been used and the conclusions reached.

3.2.3 Grubbs' test

3.2.3.1 Principle

Grubbs' test is used to check if there are cell means in Table 2 that are exceptionally high or low and would inflate the estimate of the reproducibility standard deviation if retained. The statistic used in Grubbs' test is closely related to Mandel's h statistic.

3.2.3.2 Interpretation

After Cochran's test has been carried out, the tabulated mean values for each particular level of interest shown in Table 2 are then arranged in non-decreasing order. Several Grubbs' tests are then carried out. Firstly, the test is carried out to establish whether the highest or lowest mean value can be identified as a single outlier. If an outlier is indicated, it is discarded and the test is repeated for the other extreme value. For a particular level of interest, a Grubbs' test for one outlier enables the calculation of the quotient of the difference between the suspect value and the mean of all the values for that level, and the standard deviation of all values. This ratio is then compared with computed or tabulated (critical) ratio values at 95 % and 99 % confidence levels.

If no single outlier is identified, a further Grubbs' test is carried out to establish the presence (or absence) of two extreme outliers. For example, the two lowest mean values are tested and if no outlier is shown, the two highest mean values are tested. In this test, if the calculated quotient is more than the computed ratio at the stated level of confidence, the means can be regarded as satisfactory.

3.2.3.3 Alternative tests

An alternative to Grubbs' test is Dixon's test. Again, however, there may be occasions, particularly with borderline cases, when an outlier is identified using one test but is not identified using the other test. Hence, it is important for the statistician to report which tests have been used.

3.3 Conclusions

The flow diagram in Figure 7 highlights the main procedures that need to be undertaken.

The data that have been identified as statistical outliers are reported to the executive officer. The executive officer shall undertake an appropriate investigation to ascertain whether to retain, reject or modify any data. When this investigation is completed and depending on its outcome, the statistician may receive a revised set of data in the same format as Table 1. If appropriate, this enables additional tables similar to Tables 2 and 3 to be recalculated.

The statistician can then estimate, for each level of interest, the repeatability and reproducibility standard deviations.

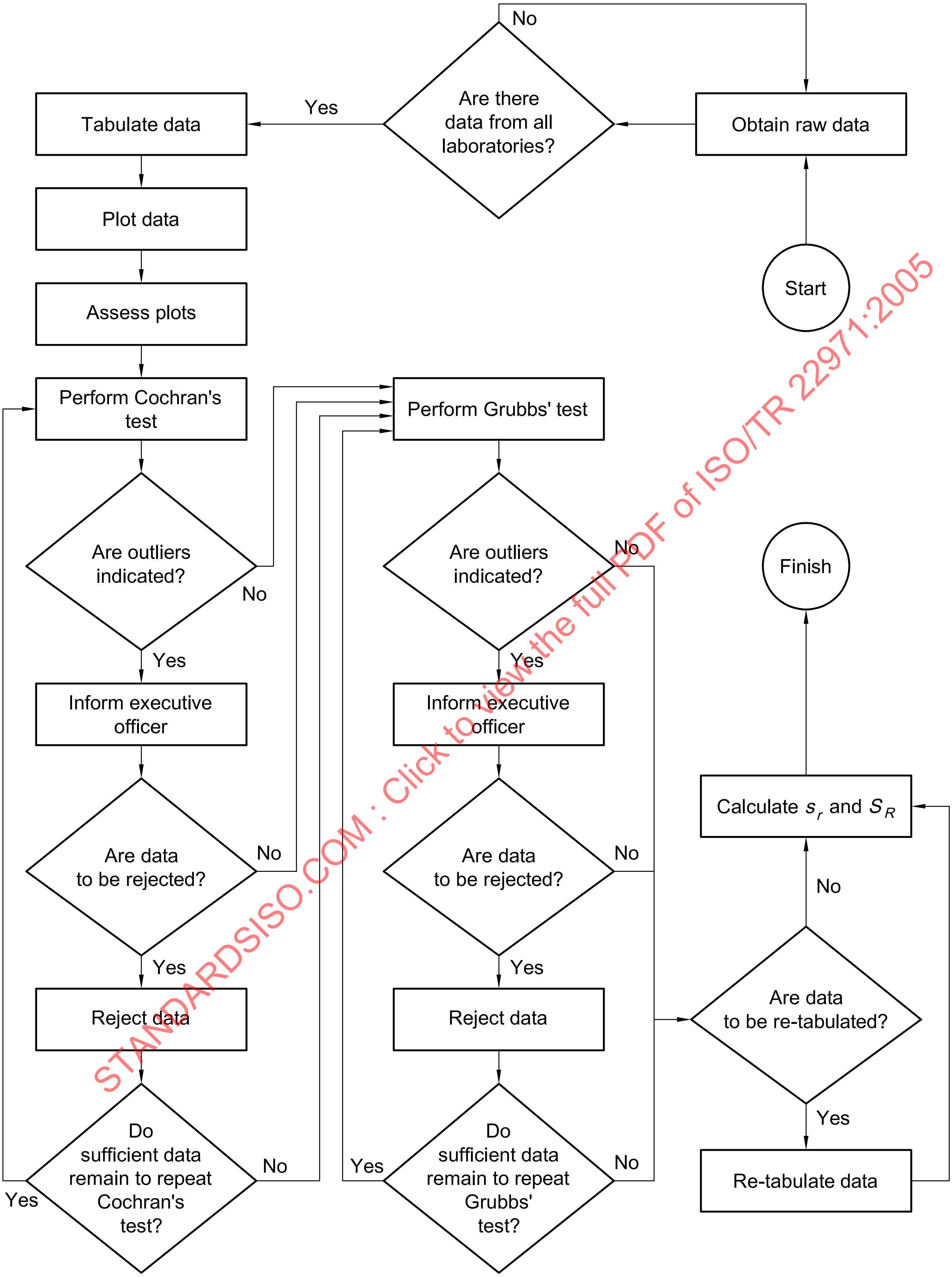


Figure 7 — Schematic flow chart for statistical treatment of outliers

4 Estimation of repeatability and reproducibility standard deviations

4.1 Analysis of variance

In an inter-laboratory programme, a large number of laboratories perform repeated tests on the same sample of material. The design can be depicted as shown in Figure 8.

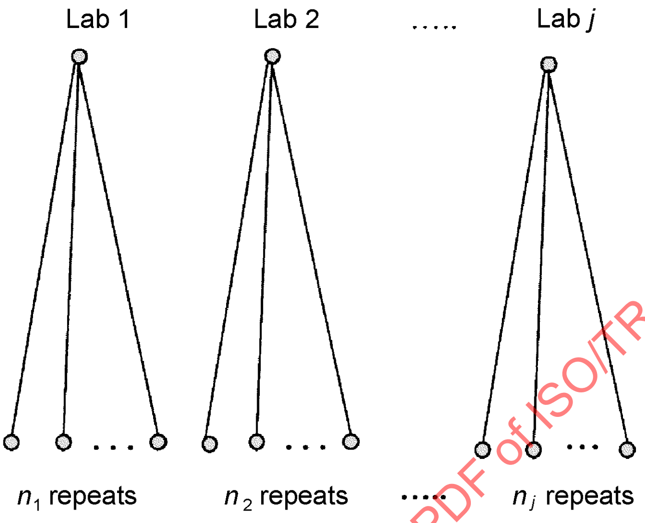


Figure 8 — Design of the inter-laboratory programme

If the sample is tested once in different laboratories, the variation of the resulting characteristic, y, will reflect some combination of the variability arising from within-laboratory variation and of the variability arising from between-laboratory variation. The model described in 4.2 is used to explain the generation of the individual results, y.

4.2 Description of the model

The precision (repeatability and reproducibility) of a measurement method can be estimated from the analysis of data from a group of laboratories selected from a population of laboratories using the same method. Each laboratory provides a set of results that is statistically compared with other sets of results.

NOTE These procedures can also be used to estimate the parameters of a group of analysts or operators in place of a group of laboratories.

The statistical treatment is based on the following:

- several samples, at different levels of results or performances, are analysed by several laboratories;
- each laboratory carries out the test method and reports replicate results;
- at each level, the total scatter of the results is broken down into random variations.

This process is defined in ISO 5725-2:1994, 4.1 by the equation:

$$y = m + B + e$$

where

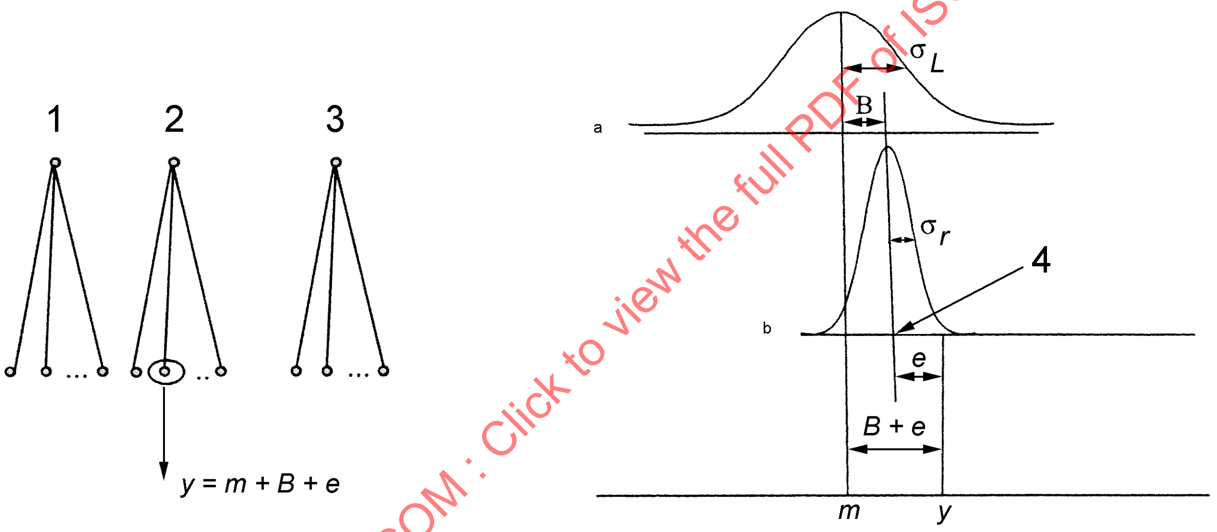
m is the mean of the results;

B is the laboratory component of the bias under repeatability conditions;

e is the random variability occurring during any measurement, under repeatability conditions.

The best available estimate of the true value of the characteristic under study is m, the overall mean of all results. A single reported value, y, will not, in general, be equal to m.

The overall error $(y - m)$ contains two parts: e and B, where e is the deviation of y from the mean of a large number of results that could come from the particular laboratory where y was generated, and B is the deviation (or difference) between this conceptual mean and the overall mean. This partition, $y = m + B + e$, is depicted in Figure 9.



Key

1 lab 1

2 lab 2

3 lab j

4 lab 2 mean

[a] Variation of the laboratory component of the bias.

[b] Variation of test results within one laboratory.

Figure 9 — Partition of the overall error

It is expected that e and B are close to zero and it is the purpose of the ISO 5725-2 to specify how to estimate their variances. The variance of e is the within-laboratory variance, σ_r^2, and the variance of B is the between-laboratory variance, σ_L^2.

The within-laboratory variance is known as the repeatability variance.

The sum of the between-laboratory variance and the within-laboratory variance $\sigma_R^2 = \sigma_L^2 + \sigma_r^2$ represents the variance of results obtained under conditions that are appreciably different (such as a change of laboratories) and is known as the reproducibility variance, σ_R^2.

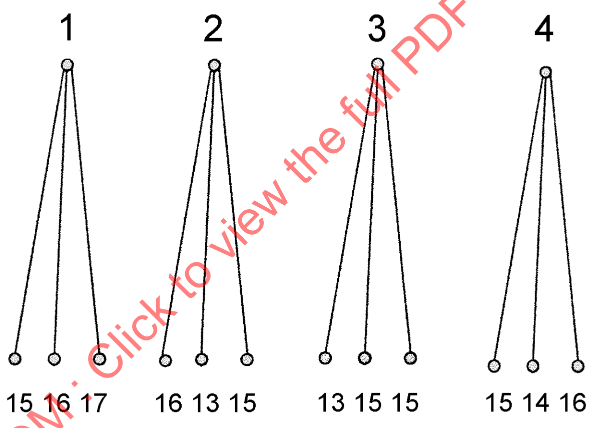
The structure described in Figure 8 is called a nested or hierarchical design. It is, therefore, necessary to estimate σ_r^2 and σ_L^2 in order to assess the variance of e and B.

Statistically speaking, the laboratory effect, B, is said to be random. As such, this means that the set of laboratories that participated to the inter-laboratory programme is a random choice from a conceptual infinite by number of laboratories that might use the test method. It is necessary to estimate how consistent these laboratories are with one another.

4.3 Examples

4.3.1 Example 1

To explain how both variance components, σ_r^2 and σ_L^2, can be estimated, suppose there are four participating laboratories, each providing three repeats of the same sample, with no missing data and no outliers. For example, the reported data are as shown in Figure 10.



Key

1 lab 1

2 lab 2

3 lab 3

4 lab 4

Figure 10 — Design of example 1

Within Lab 1, an estimate, $s_{r(\text{Lab1})}^2$, of σ_r^2 is given by the variance of the three results, hence:

$s_{r(\text{Lab1})}^2$ = var (15, 16, 17) = 1,00 with two degrees of freedom.

Within Lab 2, an estimate, $s_{r(\text{Lab2})}^2$, of σ_r^2 is given by the variance of the three results, hence

$s_{r(\text{Lab2})}^2$ = var (16, 13, 15) = 2,33 with two degrees of freedom.

Within Lab 3, an estimate, $s^2_{r(\text{Lab3})}$, of σ^2_r is given by the variance of the three results, hence:

$s^2_{r(\text{Lab3})}$ = var (13, 15, 15) = 1,33 with two degrees of freedom.

Within Lab 4, an estimate, $s^2_{r(\text{Lab4})}$, of σ^2_r is given by the variance of the three results, hence:

$s^2_{r(\text{Lab4})}$ = var (15, 14, 16) = 1,00 with two degrees of freedom.

A better estimate, s^2_r, of the true unknown repeatability variance, σ^2_r, is obtained by averaging $s^2_{r(\text{Lab1})}$, $s^2_{r(\text{Lab2})}$, $s^2_{r\,(\text{Lab3})}$ and $s^2_{r\,(\text{Lab4})}$ under the hypothesis of equality of the four variances. This hypothesis can be tested by Cochran's test, for instance as in Clause 2. For a maximum variance of 2,33 and a sum of all variances of 5,66, the Cochran's statistic is (2,33/5,66) = 0,41. Compared to the critical value 0,768 at 5 % for the four laboratories and three replicates, the equality of the variance cannot be rejected. Hence, s^2_r = 1,42 with eight degrees of freedom.

There are four means (16,00, 14,67, 14,33 and 15,00), one for each laboratory, each mean being an estimate of the measured characteristic. The overall average, m, is 15,00 and the variance is 0,52 with three degrees of freedom. Because in this case each laboratory mean is an average of three single results, the variance between means is not an estimate of σ^2_L alone but contains in addition part of the repeatability variance (one third, in this example).

This information can be summarized as in Table 4:

Table 4 — Estimation of variance components

Quantity estimated	Estimate
$\sigma^2_L + (\sigma^2_r)/3$	0,52
σ^2_r	1,42

By equating the estimates to the corresponding quantities estimated, s^2_r and s^2_L are derived as estimates of σ^2_r and σ^2_L

where

s^2_r is equal to 1,42;

s^2_L is equal to 0,05 (e.g. 0,52 – 1,42/3).

The estimate, s^2_R, of the reproducibility variance is given by $s^2_L + s^2_r$, and is equal to 1,47 (e.g. 0,05 + 1,42).

Most software packages will report the following equivalent table (Table 5):

Table 5 — Estimation of variance components

Quantity estimated	Estimate
$3\sigma^2_L + \sigma^2_r$	1,56 (= 3 × 0,52)
σ^2_r	1,42

The calculations become more complicated when the numbers of results per laboratory are not equal and the use of a software package is recommended in this case. However, the principle remains the same. It will be presented in the worked examples.

4.3.2 Example 2

Table 6 shows the raw data for different laboratories and the associated means, variances.

Table 6 — Example 1 means and variances

Parameter	Lab 1	Lab 2	Lab 3	Lab 4
Results	63	44	50	53
	57	51	40	57
	54	43	42	46
Mean	58	46	44	52
$s_{r,i}^2$ [a]	21	19	28	31
Degrees of freedom	2	2	2	2

[a] Subscript i indicates the ith laboratory.

From these results, it follows:

- overall mean, m = 50,00 [i.e. (58 + 46 + 44 + 52) / 4];
- variance of the means = 40,00;
- repeatability variance, s_r^2 = 24,75 [i.e. (21 + 19 + 28 + 31) / 4].

Hence,

- between-laboratory variance, s_L^2 = 31,75 (i.e. 40 – 24,75 / 3);
- reproducibility variance, $s_R^2 = s_L^2 + s_r^2$ = 56,50 (i.e. 31,75 + 24,75).

4.4 Use of repeatability and reproducibility limits

Estimation of the repeatability and reproducibility standard deviations can be used to derive values for the repeatability and reproducibility limits. These limits are used in practice to judge whether the measurement method is likely to have been carried out without significant adverse effects and/or to assess if two products differ significantly.

The repeatability and reproducibility limits represent the absolute difference between two single results obtained under repeatability and reproducibility conditions, respectively, that will not be exceeded more than 95 % of the time. A rough estimation of the repeatability and reproducibility limits can be obtained by multiplying the repeatability and reproducibility standard deviations, respectively, by 2,8 (ISO 5725-6:1994, 4.1) and applying the following rules.

a) If two results obtained from the same product under repeatability conditions (or reproducibility conditions) differ by more than the repeatability (or reproducibility) limit, it is likely that a problem occurred during the application of the method and/or the sampling. Further investigation is recommended. Possibly more results will be needed.

b) if two results obtained each from two different products under repeatability conditions (or reproducibility conditions) differ by more than the repeatability (or reproducibility) limit, it is reasonable to question whether these two products are of different quality.

NOTE The repeatability and reproducibility standard deviations and limits are precision estimates, and therefore are subject to estimation errors. They are meant to be used as guides to assess the validity or relevance of results produced by the measurement method and not as rigid numerical criteria to discard or validate results. In particular, the value of the factor 2,8 used in the derivation of the repeatability and reproducibility limits is based on the assumption of a normal distribution, entirely determined before any result is obtained, which rarely happens in practice. Many practitioners have found it adequate to use the value of three for this constant, without loss of information. Common sense should prevail in all cases.

The results of the application of this evaluation estimate to example 2 are as follows:

— repeatability limit: $r = 2{,}8 \times s_r = 13{,}93$;

— reproducibility limit: $R = 2{,}8 \times s_R = 21{,}05$.

The difference between the two results obtained from the same product under repeatability (or reproducibility) conditions will not exceed 13,93 (or 21,05) in 95 cases out of 100.

5 Worked examples using statistical software

5.1 General

This clause illustrates the extent to which graphs and statistics can be derived automatically using statistical software programs. The objective is not to advertise or endorse specific packages, but to emphasise the major advantages of automated computation, namely accuracy, speed and the ability to use procedures that might be too complicated for the use of a pocket calculator alone.

There are certain limitations, however, to using software packages. Not all packages possess the ability to calculate all the statistical parameters required, for example, Mandel's h and k statistic values or Grubbs' statistic values.

Two of the examples from ISO 5725-2:1994, Annex B, are presented here: ISO 5725-2:1994, B.1, example 1, “Determination of sulfur content in coal”, and ISO 5725-2:1994, B.3, example 3, “Thermometric titration of creosote oil”. When applicable, reference to specific sections of this annex will be highlighted.

5.2 Determination of sulfur content in coal

5.2.1 Original data

In general, the data can be entered following the layout of Table B.1 of ISO 5725-2:1994. One column contains the laboratory number and, in this case, four columns, one per level, contain the data. See Table 7.

Table 7 — Presentation of data (representing the sulfur content of coal [a,b])

Laboratory	Level			
	1	2	3	4
1	0,71	1,20	1,68	3,26
1	0,71	1,18	1,70	3,26
1	0,70	1,23	1,68	3,20
1	0,71	1,21	1,69	3,24
2	0,69	1,22	1,64	3,20
2	0,67	1,21	1,64	3,20
2	0,68	1,22	1,65	3,20
3	0,66	1,28	1,61	3,37
3	0,65	1,31	1,61	3,36
3	0,69	1,30	1,62	3,38
4	0,67	1,23	1,68	3,16
4	0,65	1,18	1,66	3,22
4	0,66	1,20	1,66	3,23
5	0,70	1,31	1,64	3,20
5	0,69	1,22	1,67	3,19
5	0,66	1,22	1,60	3,18
5	0,71	1,24	1,66	3,27
5	0,69	—	1,68	3,24
6	0,73	1,39	1,70	3,27
6	0,74	1,36	1,73	3,31
6	0,73	1,37	1,73	3,29
7	0,71	1,20	1,69	3,27
7	0,71	1,26	1,70	3,24
7	0,69	1,26	1,68	3,23
8	0,70	1,24	1,67	3,25
8	0,65	1,22	1,68	3,26
8	0,68	1,30	1,67	3,26

[a] Original data from ISO 5725-2:1994, B.1.2.

[b] Concentration units are % mass fraction.

Alternatively, the data can be entered using three columns, one for laboratory, one for level, one for the data, in this case the sulfur content, expressed in % mass fraction. An example of this layout is given in Table 8 for laboratories 1 and 2 only. The whole file continues in a similar fashion to include data from laboratories 3 to 8.

Table 8 — Alternate presentation of data (representing the sulfur content of coal [a,b]) for laboratories 1 and 2

Laboratory	Level	Sulfur
1	1	0,71
1	1	0,71
1	1	0,70
1	1	0,71
1	2	1,20
1	2	1,18
1	2	1,23
1	2	1,21
1	3	1,68
1	3	1,70
1	3	1,68
1	3	1,69
1	4	3,26
1	4	3,26
1	4	3,20
1	4	3,24
2	1	0,69
2	1	0,67
2	1	0,68
2	2	1,22
2	2	1,21
2	2	1,22
2	3	1,64
2	3	1,64
2	3	1,65
2	4	3,20
2	4	3,20
2	4	3,20

[a] Original data from ISO 5725-2:1994, B.1.2.

[b] Concentration units are % mass fraction.

As already stated, it is recommended to plot the data before performing tests to detect potential outliers. Below are two examples of such plots. The scatter of the raw data for level 1 is demonstrated in Figure 11, which is similar to ISO 5725-2:1994, Figures B.1 to B.4. These plots show the results by laboratory, with a horizontal line drawn at the value of the overall mean. The “box-and-whiskers” plot of Figure 12, consisting of a box, “whiskers” and outliers, summarizes results by laboratory for level 1 and is useful for obtaining an initial assessment of potential outliers. It is an interesting and very informative display of the data.

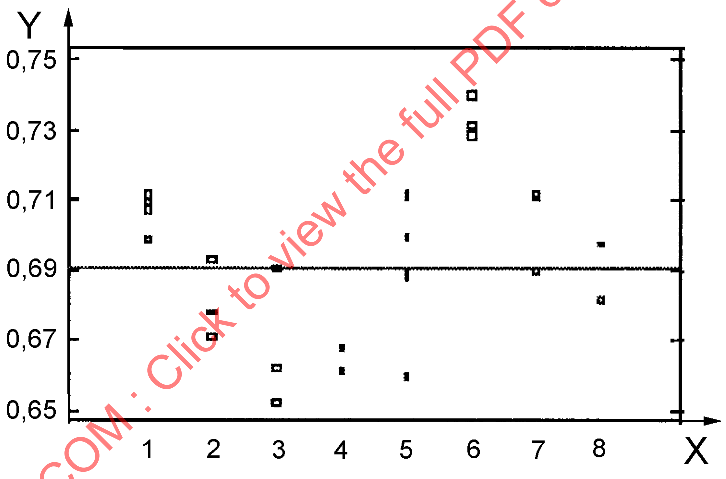
The average and the median for each laboratory are represented by a cross and a darker line, respectively. The bottom of the box is at the first quartile (Q1), the top is at the third quartile (Q3). The approximate range of the data is indicated by the lines from the top and the bottom of the box, called whiskers, which extend from the top and bottom of the box to the the highest and the lowest observations, respectively, that are still inside the region defined by the following limits:

— lower limit: Q1 − 1,5 (Q3 − Q1);

— upper limit: Q3 + 1,5 (Q3 − Q1).

For a Gaussian or normal distribution, this corresponds to about 99,30 % of the area under the probability curve. John Tukey, inventor of the “box-and-whiskers” plot, called these limits the “inner fences”. He further defined the “outer fences” obtained by using a multiplier of 3 instead of 1,5. Data points that are between the “inner” and “outer” fences are defined as possible outliers. Data that fall beyond the “outer” fence are highly suspicious due to the fact that the area within the outer fences encompasses 99,99 % of the area under the curve. Some statistical software packages use different symbols for these two classes of outliers, such as “*” for the data between the inner and outer fences and “0” for data beyond the outer fences.

Both graphs show that all the results for laboratory 6 are higher than corresponding values for the other laboratories. In addition, there is a value for laboratory 5, easily identified on Figure 12, that might merit further investigation.



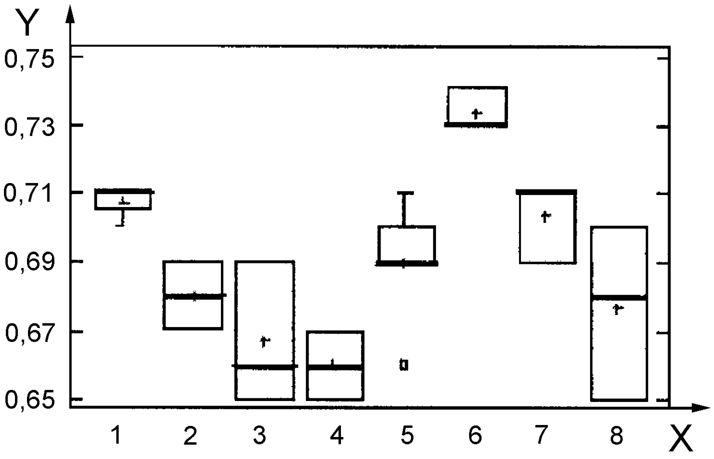
Key

X Laboratory number

Y Concentration, percent mass fraction

— $m_1 = 0{,}69$.

Figure 11 — Scatter plot of the level-1 data



Key

X Laboratory number

Y Concentration, percent mass fraction

▬ is the median

+ is the average

⊥ is a “whisker”

Figure 12 — “Box and whiskers” plot of the level-1 data

5.2.2 Table of cell means and standard deviations

Table 9 shows results from ISO 5725-2:1994, B.1.3 and B.1.4, calculated for an inter-laboratory trial as for a typical software package.

Table 9 — Table of cell means and standard deviations

Laboratory number	Count	Mean	Standard deviation
Overall			
—	27	0,690 37	0,025 49
For individual laboratories			
1	4	0,707 50	0,005 00
2	3	0,680 00	0,010 00
3	3	0,666 67	0,020 82
4	3	0,660 00	0,010 00
5	5	0,690 00	0,018 71
6	3	0,733 33	0,005 77
7	3	0,703 33	0,011 55
8	3	0,676 67	0,025 17

As shown in ISO 5725-2:1994, Tables 2 and 3, these values are required to calculate Mandel's h and k statistics, as well as the Grubbs' test, statistics which are not available in all software packages.

5.2.3 Scrutiny for consistency and outliers

In many packages, several tests for assessing the homogeneity of variances are available, including those of Cochran, Bartlett, Hartley and Levene. The results of these tests on the data for the sulfur content are shown in Table 10.

Table 10 — Scrutiny for consistency and outliers

Test	Statistic	P-value
Cochran's C	0,350	0,308
Bartlett's	1,679	0,296
Levene's	1,679	0,332
Hartley's	25,333	—

The P-value represents the probability of observing a statistic as large as that which would be computed if the variances are homogeneous. A low P-value, 0,05 or below, is sufficient to conclude that the within-laboratory variances differ. Hence, for Cochran's test, there is insufficient evidence to conclude that the intra-laboratory variances differ.

5.2.4 Computation of the general mean and of the repeatability and reproducibility standard deviations

As described in ISO 5725-2 and in ISO 5725-3, the analysis of variance technique can be used to compute s_{rj} and s_{Rj}, the repeatability and the reproducibility standard deviations, respectively, for level j.

Most packages have the ability to compute the variance components directly. This ability can usually be found among other statistical tools by looking for an option entitled “Variance Components” or equivalently “Hierarchical Designs” or “Nested Designs”.

The nesting occurs in that the samples analysed differ from laboratory to laboratory, even though they originate from one single lot or batch of product. Therefore, samples are nested within “laboratory”, or equivalently, laboratories are hierarchically above samples.

If no such option exists, another choice is to use a “one-way” analysis of variance where the factor “laboratory” (the “one-way” factor) is declared random.

The variance components refer to partitioning the total variance into two parts, the between- and the within-laboratory variances, s_{Lj}^2 and s_{rj}^2, respectively. An example of the resulting one-way analysis of variance is given in Table 11, for level 1 ($j = 1$).

Table 11 — Variance components analysis — Analysis of variance for sulfur content — Level 1

Source	Sum of squares	Degrees of freedom	Mean square	Variance component	Percent
laboratory	0,012 554 6	7	0,001 793 5	0,000 466 5	67,1
error	0,004 341 7	19	0,000 228 5	0,000 228 5	32,8
total (corrected)	0,016 896 3	26	—	—	—

The repeatability standard deviation, s_{r1}, is obtained by taking the square root of the variance component “error”. Hence:

$$s_{r1} = \sqrt{0{,}000\,228\,5} = 0{,}015\,.$$

The reproducibility standard deviation, s_{R1}, is obtained by taking the square root of the sum of the variance component "error" and variance component "laboratory". Hence:

$$s_{R1} = \sqrt{0{,}000\ 228\ 5 + 0{,}000\ 466\ 5} = 0{,}026\,.$$

The relative weight of the variance component "laboratory" and of the variance component "error" for the reproducibility variance is indicated in the column "Percent".

Should this option not be available, then the analysis of variance table might not display the variance components directly and results might appear as shown in Table 12. These values can be derived by manual calculation, using the formula given in ISO 5725-2:1994, 7.4.5.2, as shown below:

Table 12 — ANOVA table for sulfur content — Level 1 — Analysis of variance

Source	Sum of squares	Degrees of freedom	Mean square	F-ratio	P-value
between-laboratory	0,012 554 6	7	0,001 793 5	7,85	0,000 2
within-laboratory (error)	0,004 341 7	19	0,000 228 5	—	—
total (corrected)	0,016 896 3	26	—	—	—

The parameter s_{r1}^2 is given by the value of the within-laboratory mean squared: $s_{r1}^2 = 0{,}000\ 228\ 5$.

As is shown in ISO 5725-2, 7.4.5.2, Equation (21):

$$S_{L1}^2 = \frac{s_{d1}^2 - s_{r1}^2}{\bar{\bar{n}}_1}$$

where

s_{d1}^2 is the between-laboratory mean squared equal to 0,001 793 52;

s_{r1}^2 $= 0{,}000\ 228\ 5$ (see above);

$$\bar{\bar{n}}_1 = \frac{1}{p_{\text{Lab}} - 1}\left(\sum n_{i1} - \frac{\sum n_{i1}^2}{\sum n_{i1}}\right)$$ [ISO 5725-2:1994, 7.4.5.2, Equation (23)]

$$= \frac{1}{7}\left(27 - \frac{\left(4^2 + 3^2 + 3^2 + 3^2 + 5^2 + 3^2 + 3^2 + 3^2\right)}{27}\right) = 3{,}35$$

NOTE An approximation of $\bar{\bar{n}}_j$ is given by the mean number of results per laboratory (i.e. 27/8 = 3,375).

Hence:

$$s_{L1}^2 = \frac{0{,}001\,793\ 5 - 0{,}000\ 228\ 5}{3{,}35} = 0{,}000\ 467\ 2$$

Since $s_{R1}^2 = s_{L1}^2 + s_{r1}^2$, then:

$$s_{R1} = \sqrt{0{,}000\ 228\ 59 + 0{,}000\ 467\ 2} = 0{,}026.$$

5.2.5 Dependence of precision on m

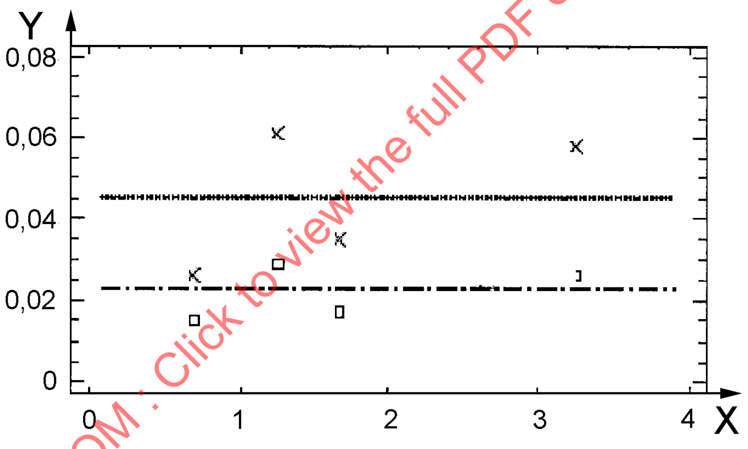
When the statistical analysis is repeated for all other levels, analogous repeatability and reproducibility standard deviations are determined.

The data can be presented as in Table 13:

Table 13 — Repeatability and reproducibility standard deviations for multiple levels

Level	Overall mean	Repeatability standard deviation	Reproducibility standard deviation
1	0,690	0,015	0,026
2	1,252	0,029	0,061
3	1,667	0,017	0,035
4	3,250	0,026	0,058

The dependence of precision on the level overall means (see also ISO 5725-2, B.1.7) can be assessed by observing the scatter plot below. The averaged repeatability and reproducibility standard deviations, 0,022 and 0,045, respectively, are calculated and can be plotted as shown in Figure 13.



Key

X Overall mean, in percent mass fraction

Y Percent mass fraction

□ repeatability

x reproducibility

… average s_R

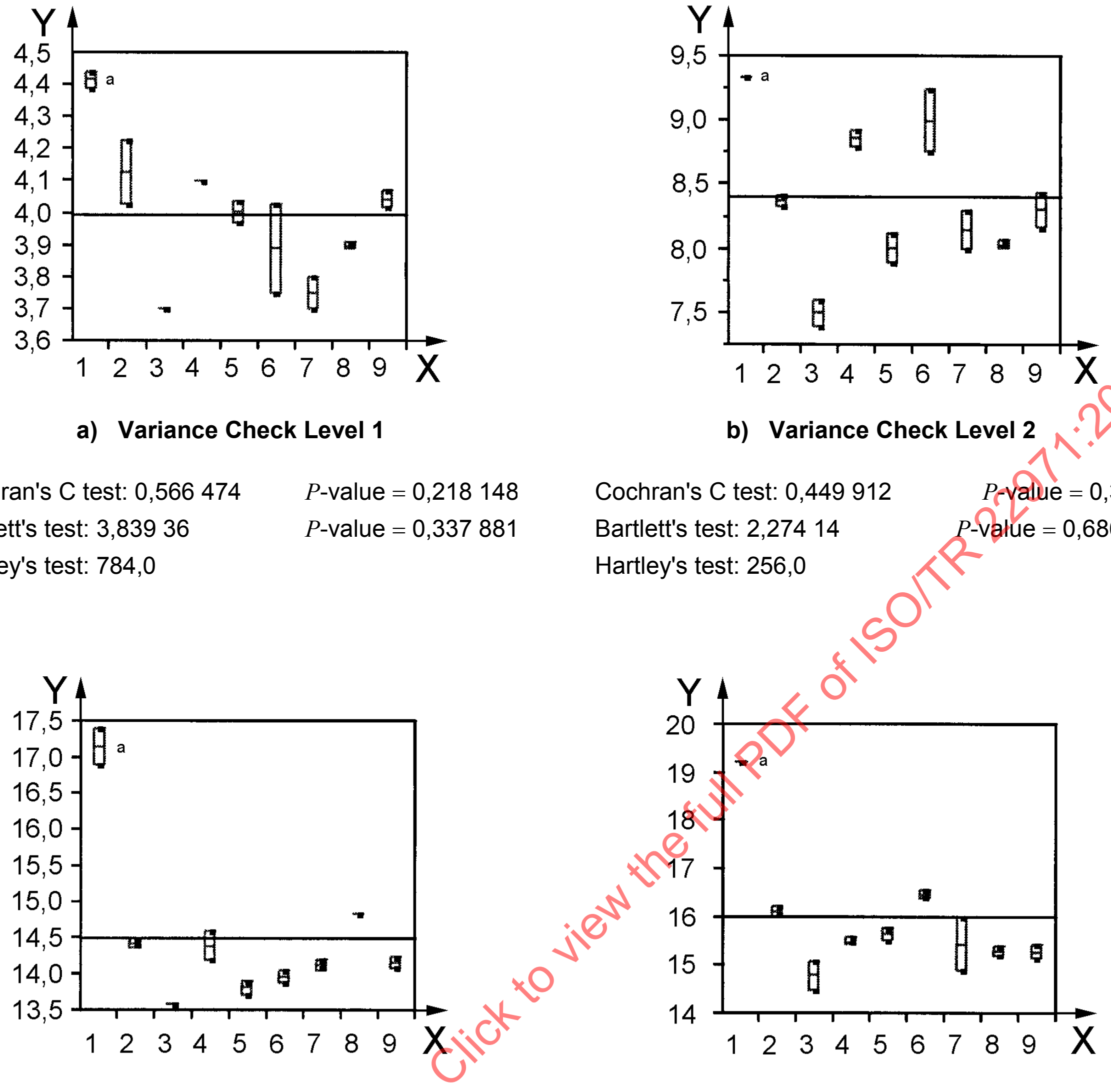
—·—·— average s_r

Figure 13 — Plot of s_r and s_R versus overall mean

5.3 Thermometric titration of creosote oil

5.3.1 Original data

Plots of the raw data from ISO 5725-2:1994, B.3.2, for the thermometric titration of creosote oil show that laboratory 1 always produces data that are consistently higher than all the other laboratories. This fact might invite further investigation. In addition, outlier tests might suggest whether these data should be considered for rejection.



a) Variance Check Level 1

Cochran's C test: 0,566 474 *P*-value = 0,218 148
Bartlett's test: 3,839 36 *P*-value = 0,337 881
Hartley's test: 784,0

b) Variance Check Level 2

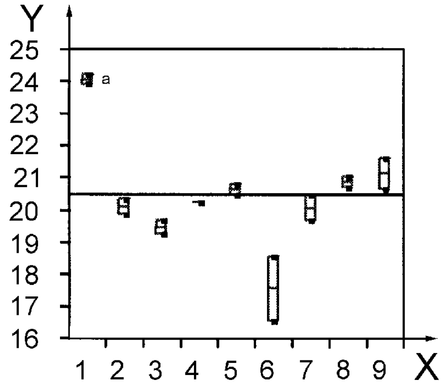
Cochran's C test: 0,449 912 *P*-value = 0,383 8
Bartlett's test: 2,274 14 *P*-value = 0,686 763
Hartley's test: 256,0

c) Variance Check Level 3

Cochran's C test: 0,492 417 *P*-value = 0,366 708
Bartlett's test: 2,171 5 *P*-value = 0,686 072
Hartley's test: 39,062 5

d) Variance Check Level 4

Cochran's C test: 0,666 703 *P*-value = 0,057 952 4
Bartlett's test: 3,271 86 *P*-value = 0,439 721
Hartley's test: 121,0



e) Variance Check Level 5

Cochran's C test: 0,635 778 P-value = 0,080 422 1

Bartlett's test: 2,485 46 P-value = 0,623 743

Hartley's test: 50,005 1

Key

X Laboratory number

Y Creosote, percent mass fraction

— average of all laboratory averages

[a] Systematically high value.

Figure 14 — “Box-and-whiskers” plots and variance tests

The “box and whiskers” plots clearly show that laboratory 1 exhibits a systematic higher value than the other laboratories, although laboratories 6 and 7 show a higher variability at some specific levels.

5.3.2 Scrutiny for consistency and outliers

The data can be presented in the form of Table 14 in order to assess the consistency and the identification of outliers; see also ISO 5725-2, B.3.2 and B.3.5.

Table 14 — Scrutiny for consistency and outliers by laboratory and level

Lab	Observation	Mean by level				
		Level 1	Level 2	Level 3	Level 4	Level 5
1	high ?	4,415	9,340	17,150	19,230	24,140
2	—	4,130	8,375	14,460	16,140	20,155
3	—	3,700	7,500	13,600	14,800	19,500
4	—	4,100	8,865	14,400	15,550	20,300
5	—	4,005	8,005	13,825	15,660	20,700
6	—	3,890	9,000	13,980	16,500	17,570
7	—	3,750	8,150	14,150	15,450	20,100
8	—	3,905	8,055	14,840	15,315	20,940
9	—	4,045	8,305	14,170	15,290	21,185
Parameter		Value by level				
Mean		3,993	8,399	14,508	15,993	20,511
Std Dev		0,216	0,572	1,056	1,310	1,727
Sdt Error Mean		0,072	0,191	0,352	0,437	0,576

EXAMPLE For level 3 and laboratory 1, the single high Grubbs' statistic, G_b, is given by

$$G_b = \frac{17,15 - 14,508}{1,056} = 2,50$$

The critical values at the 5 % and 1 % significance level are 2,215 and 2,387, respectively, for the nine laboratories. Hence, with 99 % confidence, the data for level 3 and laboratory 1 can be considered as outliers.

5.3.3 Computation of the overall mean and of the repeatability and reproducibility standard deviations

On this basis, all data from laboratory 1, together with data from laboratory 6 at level 5 are rejected. Following this, the analysis of variance for level 5 ($j = 5$) and the calculations for the repeatability and reproducibility standard deviations, s_r and s_R, respectively, are given as shown in Table 15 and the following equations; see also ISO 5725-2, B.3.6.