International Standard 5725

INTERNATIONAL ORGANIZATION FOR STANDARDIZATION•МЕЖДУНАРОДНАЯ ОРГАНИЗАЦИЯ ПО СТАНДАРТИЗАЦИИ•ORGANISATION INTERNATIONALE DE NORMALISATION

Precision of test methods — Determination of repeatability and reproducibility for a standard test method by inter-laboratory tests

Fidélité des méthodes d'essai — Détermination de la répétabilité et de la reproductibilité d'une méthode d'essai normalisée par essais interlaboratoires

Second edition — 1986-09-15

UDC 519.248 : 620.1 **Ref. No. ISO 5725-1986 (E)**

Descriptors : tests, reproducibility, statistical analysis.

Price based on 49 pages

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.

Draft International Standards adopted by the technical committees are circulated to the member bodies for approval before their acceptance as International Standards by the ISO Council. They are approved in accordance with ISO procedures requiring at least 75 % approval by the member bodies voting.

International Standard ISO 5725 was prepared by Technical Committee ISO/TC 69, *Applications of statistical methods.*

This second edition cancels and replaces the first edition (ISO 5725-1981), of which it constitutes a minor revision.

Users should note that all International Standards undergo revision from time to time and that any reference made herein to any other International Standard implies its latest edition, unless otherwise stated.

Printed in Switzerland

Contents

Precision of test methods — Determination of repeatability and reproducibility for a standard test method by inter-laboratory tests

0 Introduction

0.1 Tests performed on presumably identical material (see 4.2) in presumably identical circumstances do not, in general, yield identical results. This is attributed to unavoidable random errors inherent in every test procedure; the factors that may influence the outcome of a test cannot all be completely controlled. In the practical interpretation of test data, this variability has to be taken into account. For instance, the difference between a test result and some specified value may be within the scope of unavoidable random errors, in which case a real deviation from such a specified value has not been established. Similarly, comparing test results from two batches of material will not indicate a fundamental quality difference if the difference between them can be attributed to inherent variation in the test procedure.

0.2 Many different factors (apart from variations between supposedly identical specimens) may contribute to the variability of a test procedure, including the following :

a) the operator;

b) the equipment used;

c) the calibration of the equipment;

d) the environment (temperature, humidity, air pollution, etc.).

The variability between tests performed by different operators and/or with different equipment will usually be greater than between tests carried out by a single operator using the same equipment.

0.3 Precision is a general term for the variability between repeated tests. Two measures of precision, termed repeatability and reproducibility, have been found necessary and, for many practical cases, sufficient for describing the variability of a test method. Repeatability refers to tests performed under conditions that are as constant as possible, with the tests performed during a short interval of time (see 4.3) in one laboratory by one operator using the same equipment. On the other hand, reproducibility refers to tests performed in widely varying conditions, in different laboratories with different operators and different equipment. Under repeatability conditions factors a) to d) listed in 0.2 are considered constants and do not contribute to the variability, while under reproducibility conditions they vary and contribute to the variability of the test results. Thus repeatability and reproducibility are two extremes, the first measuring the minimum and the second the maximum variability in results. Other intermediate measures of variability between these two extremes are conceivable, such as repetition of tests within a laboratory at longer time intervals, or by different operators, or including the effects of recalibration but these are not considered in this International Standard. If, in a particular situation, some intermediate measure should be needed, it must be clearly defined by some responsible authority, together with the circumstances under which it applies and the method by which it should be determined.

0.4 The definitions used in this International Standard are given in clause 3 and the symbols and subscripts used are given in annex C.

A bibliography of the publications referred to in this International Standard is appended.

1 Scope

This International Standard establishes practical definitions of repeatability r and reproducibility R which lend themselves to numerical estimation by experiment (see clause 3). It does not provide any measure of the errors in estimating the values of r and R. It discusses the implications of these definitions of r and R.

This International Standard establishes basic principles for the layout, organization and analysis of experiments designed for estimating r and R (see clauses 6 to 17). Experiments for this purpose will be referred to as precision experiments. Only the simplest type of experiment for the estimation of r and R is considered, which consists of tests on samples of identical material by several laboratories.

This International Standard also presents rules for the interpretation and application of these estimates of r and R in practical situations (see clauses 18 to 20).

This International Standard does not deal with determining the accuracy of the test method, as measured by the difference between the overall mean value and the true value or conventional true value.

2 Field of application

This International Standard deals exclusively with test methods which yield a single numerical figure as the test result, although this single figure may be the outcome of a calculation from a set of observations.

The essence of the determination of precision values is that they measure the ability of a test method to repeat a given determination. Thus the implication is that exactly the same thing is being measured in exactly the same way.

In order that the measurements are made in the same way, the test method shall have been standardized and in use in a number of different laboratories. All tests forming part of a precision experiment shall be carried out in accordance with that standard.

Ideally, the various tests should be carried out using the same specimen. Unfortunately many tests are destructive in nature (chemical tests, strength tests of materials) so that the same specimen is not available for further determinations. Under such circumstances, different specimens shall be used, but to conform to the basic principle every effort shall be made to ensure that the specimens are as nearly identical as is possible. Care shall also be taken that the specimens are not just identical when the samples are prepared, but that they should also be identical at the time of testing.

Because of the above principles, precision should not be determined using specimens which are known not to be, or are even suspected of not being identical. Thus the specimens for test should be taken as similar sub-samples of one bulk sample, and shall never be drawn from different lots or different consignments. These points are discussed further in 4.2.

In practice, where destructive testing is involved, the contribution to the variability in the test results arising from differences between the specimens on which the tests are performed shall either be negligible compared to the variability of the test method itself or else form an inherent part of the test method, and thus truly a component of precision (see 4.2).

The statistical model described in clause 5 is accepted as a suitable basis for the interpretation and analysis of the test results given by a precision experiment which conforms to the principles stated above.

Section one : General principles

3 Quantitative definitions of repeatability and reproducibility of a standard test method

3.1 For practical purposes quantitative definitions are needed; the following definitions conform with ISO 3534[1].

3.1.1 observed value : The value of a characteristic determined as the result of an observation.

NOTE — This is a single value obtained from a single observation.

3.1.2 test result : The value of a characteristic determined by carrying out a specified test method.

NOTE — The test method may specify that a number of individual observations be made, and their average reported as the test result. It may also require standard corrections to be applied, such as correction of gas volumes to standard temperature and pressure. Thus a single test result can be a result calculated from several observed values.

3.1.3 level of the test : The general average of the test results from all laboratories for one particular material or specimen tested.

3.1.4 cell : The test results at a single level obtained by one laboratory.

3.1.5 precision : The closeness of agreement between mutually independent test results obtained under stipulated conditions.

NOTES

1 Precision depends only on the distribution of random errors and does not relate to the true value, conventional true value or specified value.

2 Repeatability and reproducibility are concepts of precision.

3.1.6 repeatability : The closeness of agreement between mutually independent test results obtained under repeatability conditions.

3.1.7 repeatability conditions : Conditions where mutually independent test results are obtained with the same method on identical test material in the same laboratory by the same operator using the same equipment within short intervals of time.

3.1.8 repeatability standard deviation : The standard deviation of test results obtained under repeatability conditions. It is a parameter of dispersion of the distribution of test results under repeatability conditions.

NOTE — Similarly, repeatability variance and repeatability coefficient of variation could be defined and used as parameters of dispersion of test results under repeatability conditions.

3.1.9 repeatability value, r : The value below which the absolute difference between two single test results obtained under repeatability conditions may be expected to lie with a probability of 95 %.

NOTE — For brevity, in the remainder of this International Standard, "repeatability value" is shortened to "repeatability" or just "r" where the context makes it clear that it is the values that are referred to.

3.1.10 repeatability critical difference : The value below which the absolute difference between two single test results obtained under repeatability conditions may be expected to lie with a specified probability.

NOTES

1 The specified probability has to be attached as a subscript to the symbol r of the repeatability critical difference, for example r_{90} is the repeatability critical difference for a probability of 90 %.

2 The repeatability value r is the repeatability critical difference for a probability of 95 %, the subscript being omitted in this special case.

3.1.11 reproducibility : The closeness of agreement between test results obtained under reproducibility conditions.

3.1.12 reproducibility conditions : Conditions where test results are obtained with the same method on identical test material in different laboratories with different operators using different equipment.

3.1.13 reproducibility standard deviation : The standard deviation of test results obtained under reproducibility conditions. It is a parameter of dispersion of the distribution of test results under reproducibility conditions.

NOTE — Similarly, reproducibility variance and reproducibility coefficient of variation could be defined and used as parameters of dispersion of test results under reproducibility conditions.

3.1.14 reproducibility value, R : The value below which the absolute difference between two single test results obtained under reproducibility conditions may be expected to lie with a probability of 95 %.

NOTE — For brevity, in the remainder of this International Standard, "reproducibility value" is shortened to "reproducibility" or just "R" where the context makes it clear that it is the values that are referred to.

3.1.15 reproducibility critical difference : The value below which the absolute difference between two single test results obtained under reproducibility conditions may be expected to lie with a specified probability.

NOTES

1 The specified probability has to be attached as a subscript to the symbol R of the reproducibility critical difference, for example R_{90} is the reproducibility critical difference for a probability of 90 %.

2 The reproducibility value R is the reproducibility critical difference for a probability of 95 %, the subscript being omitted in this special case.

3.2 The definitions given in 3.1 apply to results variable on a continuous scale. If the test result is discrete or rounded off, r and R are each the minimum value equal to or below which the absolute difference between two single test results is expected to lie with a probability of not less than the specified value.

3.3 The terms "repeatability" and "reproducibility" have been adopted because they have been in common use for several years. The symbols r and R are already in general use for other purposes; in ISO 3534, r is recommended for the correlation coefficient and R (or w) for the range of a single series of observations. There should, however, be no confusion if the full wordings "repeatability r" and "reproducibility R" are used whenever there is a possibility of misunderstanding; particularly when r and R are quoted in standards.

3.4 R and r as defined in this International Standard are meant in the first place as criteria by which to judge how far a difference between two single test results can be ascribed to random fluctuations; a difference larger than r or R is suspect and may justify the conclusion that there exists a systematic difference, or lead to some additional investigation. In this sense, r and R can be termed critical differences, to be applied to a pair of test results obtained under repeatability and reproducibility conditions respectively.

3.4.1 It is sometimes required to compare the averages of two or more tests or to compare the average of a series with a specified value. Critical differences for such purposes can be derived from r and R as explained in 19.2.1 to 19.2.4.

3.4.2 As defined, r and R are associated with a probability level of 95 %. Sometimes critical differences with a probability level other than 95 % may be preferred; these can be computed as explained in 19.1.1. In such cases, to avoid misinterpretations, the probability level should then be attached as a subscript; for example r_{99} or R_{90}.

3.4.3 The definitions in 3.1.9 and 3.1.14 refer to theoretical constants which in reality remain unknown. The values of r and R actually determined from a precision experiment as described in this International Standard are, in statistical terms, estimates of these constants, and as such are subject to errors. Consequently, the probability levels associated with r and R will not be exactly 95 % but only of the order of 95 %, and this will also be true for other critical differences derived from them. This is unavoidable but does not seriously detract from their practical value as they are primarily designed to serve as tools for judging whether the difference between results could be ascribed to random uncertainties inherent in the test method or not. Differences larger than r or R are suspect.

3.5 If the requirements of this International Standard concerning the number of laboratories to be included in a precision experiment and the number of tests they should each carry out are followed (see 10.1), the resulting estimates of r and R will be sufficiently precise for practical purposes, with the proviso that the laboratories participating are truly representative of all laboratories using the standard method. This hypothetical population is defined by requirements similar to those given in 10.2. If at some future date it should become evident that this condition was not or is no longer satisfied by the original precision experiment, then a fresh precision experiment may be required, unless it should be possible to re-estimate r and R to conform to the altered conditions.

3.6 In principle, repeatability r, as defined in 3.1.9, can be applied to any test method within any single laboratory. A basic assumption underlying this International Standard is that, for a standardized test method, repeatability will be, at least approximately, the same for all laboratories applying the standard method, so that it is possible to establish one common average repeatability applicable to any laboratory. However, any laboratory can, by carrying out a series of tests under repeatability conditions, arrive at an estimate of its own particular repeatability for the test method, and check it against the common standard value. Such a procedure has not been worked out in detail in this International Standard.

3.7 When the reproducibility is to be used as a critical difference, the pair of test results to be compared shall have been obtained from two laboratories selected at random from the total population of laboratories using the standard test method. Where test results are always compared between the same two laboratories, caution is needed, because the probability level associated with R may then no longer hold true owing to a possible systematic difference between the results from these two particular laboratories. If it is thought that this may be the case, the two laboratories in question should organize a precision experiment between themselves in order to determine the magnitude of this systematic difference.

3.8 Although throughout this International Standard repeatability and reproducibility are considered in terms of critical differences, there is no reason for preventing the expression of precision results in terms of standard deviations or coefficients of variation if, for any particular application, this would be more appropriate.

3.9 The values of r and R, once determined, can be used in a variety of ways. For example, they can serve

— to verify that the experimental technique of a laboratory is up to standard (see 3.6);

— in designing quality control procedures;

— in comparing test results from a batch of material with a product specification;

— in designing the specifications in the first place to ensure that conformity is verifiable by the test method;

— in comparing test results on the same batch of material obtained by a supplier and a consumer;

— in assessing the suitability of rival test methods.

In some applications, various other factors may have to be taken into consideration, for example see 4.2.6.

4 Practical implications of the definitions

4.1 Standard test method

4.1.1 As stated in clause 2, the test method under investigation has to be one that has been standardized. This means that

there has to be a standard, i.e. a written document that lays down in full detail how the test should be carried out, preferably including a description as to how the test specimen should be obtained and prepared. The estimates of r and R derived from such an experiment should always be quoted as valid only for tests carried out in accordance with the standard method.

4.1.2 The existence of a standard for the test method implies the existence of an organization responsible for the establishment of the standard under study.

4.1.3 Preparing a standard for a test method requires a careful evaluation of the method (or possibly several alternative methods) by means of experiments in which a number of laboratories take part. Such a standardization experiment will provide some preliminary information concerning the values of r and R. The essential points underlying a precision experiment to determine r and R is that it will usually require the cooperation of a larger number of laboratories than for a standardization experiment, and that these laboratories shall be recruited from among all those using, or likely to use, the standard in normal operations and not exclusively consist of laboratories that have gained special experience during the process of standardizing the method. Thus a precision experiment arranged for the determination of r and R should not as a rule be organized until after the standard for the test method has been issued and is in general use. This does not mean, however, that any information regarding the possible values of r and R obtained from a standardization experiment is of no value, as they can be taken into consideration when designing the precision experiment.

4.1.4 A precision experiment can also be considered as a practical test of the adequacy of the standard. One of the main purposes of standardization is to eliminate differences between users (laboratories) as far as possible, and the data provided by a precision experiment will reveal how far this purpose has been achieved. Pronounced differences may indicate that the standard is not yet sufficiently detailed and can possibly be improved. If so, this should be reported to the standards panel with a request for further investigation. [See 9.6 c), 17.2 b) and c) and 17.3.]

4.2 Identical material

4.2.1 In a precision experiment, samples of a specific material or specimens of a specific product are sent from a central point to a number of laboratories in different places, different countries, or even in different continents. The requirement that the tests in these laboratories shall be made on identical material refers to the moment when these tests are actually carried out, and in order to achieve this the following two different conditions have to be satisfied :

a) the samples have to be identical when despatched to the laboratories, and

b) they have to remain identical during transport and during the different time intervals that may elapse before the tests are actually performed in the participating laboratories.

In organizing precision experiments, both conditions shall be carefully observed.

4.2.2 A fluid or fine powder can be homogenized by stirring, and samples drawn from such batches can then be considered as identical at the moment they are prepared. Additional precautions may be needed to ensure that they remain identical up to the time the tests are carried out. If the material to be tested consists of a mixture of powders of different relative density or of different grain size, some care is needed because segregation may result from shaking, for example during transport. When reaction with the atmosphere may be expected, the specimens may be sealed into ampoules, either evacuated or filled with an inert gas. For perishable materials, such as foodstuffs or blood samples, it may be necessary to send them to the participating laboratories in a deep-frozen state with detailed instructions of the procedure for thawing. Each case has to be judged on its merits.

4.2.3 When the tests have to be performed on discrete objects that are not altered by testing, they could, in principle at least, be carried out using the same set of objects in different laboratories. This, however, would necessitate circulating the same set of objects around many laboratories often situated far apart, in different countries or continents, with a considerable risk of loss or damage during transport.

4.2.4 When tests have to be performed on solid materials that cannot be homogenized (such as metals, rubber or textile fabrics) and when the tests cannot be repeated on the same test piece, inhomogeneity in the test material will form an essential component of the precision of the measurement and the idea of identical material no longer holds good. Precision experiments can still be carried out, but the values of r and R may only be valid for the particular material used and should be quoted as such. A more universal use of r and R will be acceptable only if it can be demonstrated that the values do not differ significantly between material produced at different times or by different producers. This would require a more elaborate experiment than has been considered in this International Standard.

4.2.5 In 4.2.1 to 4.2.4, reference is made to testing in different laboratories, with the implication of transportation of the test specimens to the laboratory, but some test specimens are not transportable, such as an oil storage tank. In such cases, testing by different laboratories means that different operators are sent with their equipment to the test site. In other cases, the quantity being measured may be transitory or variable, such as water flow in a river, when care shall be taken that the different measurements are made under as near as possible the same conditions. The guiding principle shall always be that the objective is to determine the ability to repeat the same measurement.

4.2.6 In practice, r and R, or other critical differences derived from them using the methods specified in 19.1.1 and/or 19.1.2, are often used in order to compare batches of commercial material with a specification or to compare two batches with each other. A difference larger than that critical difference can then, *inter alia*, be explained by the normal commercial inhomogeneity in the batches of material unless it has been possible to include this lack of homogeneity in the determination of r and R. However, in that case, the difficulties will be the same as those mentioned in 4.2.4.

4.3 Short intervals of time

According to the definition in 3.1.7, tests for the determination of repeatability have to be made under constant operating conditions, i.e. during the time covered by the tests, factors such as those in 0.2 should be constant. In particular, the equipment should not be recalibrated between the tests unless this is an essential part of every single determination. In practice, tests under repeatability conditions should be conducted in as short a time as possible in order to minimize changes in those factors, such as environmental ones, which cannot always be guaranteed constant. [See 10.4.1 c).]

5 Statistical model

5.1 Basic model

For estimating the precision of a test method, it is useful to assume that every single test result, y, is the sum of three components :

$$y = m + B + e \qquad \ldots(1)$$

where, for the particular material tested,

m is the general average;

B is the between-laboratory variation;

e is the random error occurring in every test.

Other models are sometimes used, but the above will cover the majority of practical cases. (See 5.6.)

5.2 General average, m

5.2.1 The general average, m, of the material tested is called the "level of the test property"; specimens of different purities of a chemical or different materials (e.g. different types of steel) will correspond to different levels. In many technical situations, the level of the test property is exclusively defined by the test method, and the notion of an independent true value does not apply. However, in some situations, the concept of a true value μ of the test property may hold good, such as the true concentration of a solution that is being titrated. The level m is not necessarily equal to the true value μ; the difference $(m - \mu)$, when it exists, is called the "bias of the test method".

5.2.2 When r and R are used to test the difference between test results, a bias will have no influence and can be ignored. But when these criteria are used to compare test results with a value specified in a contract or in a standard, a bias will have to be taken into account where the contract or specification refers to the true value, μ, and not to the test level, m. If a true value exists and is known, the analysis of a precision experiment can indicate that there is a bias. (See note to 19.2.4.)

5.3 Term B in the basic model (see 5.1)

5.3.1 The term B is considered to be constant during any series of tests performed under repeatability conditions, but it is considered to behave as a random variable in a series of tests performed under reproducibility conditions. The procedures given in this International Standard were developed assuming that the distribution of this error variable was approximately normal, but, in practice, they work for most distributions provided that they are unimodal and that the critical differences are for the 95 % level. Its variance is called the between-laboratory variance and is expressed as

$$\text{var}(B) = \sigma_L^2$$

where σ_L^2 includes the between-operator and the between-equipment variabilities.

5.3.2 In general, B can be considered as the sum of both random and systematic components, but they are not separated in this analysis. No attempt has been made in this International Standard to give an exhaustive list of the factors that contribute to B, but they include different climatic conditions, variations of equipment within the manufacturer's tolerances, and even the techniques in which operators are trained in different places.

5.3.3 If there are a limited number of laboratories likely to use the method at any time, B can only take a limited number of values, and to be of practical use, reproducibility shall be determined from a set of laboratories which can be considered as selected at random from all those likely to use the method. Some caution is needed when the test results to be compared are always performed by the same laboratories. An example of the sort of problem that can arise in this situation is given in clause 23, in which the results from two (11 and 1) of the laboratories are shown to differ consistently by about 4 °C. Where only two laboratories are regularly concerned, reproducibility as such should not be used, but a cooperative experiment between the two laboratories to determine their relative bias, and thus their own specific reproducibility, should be carried out.

5.4 Error term e in the basic model (see 5.1)

5.4.1 The term e represents a random error occurring in every single test result and the procedures given in this International Standard were developed assuming that the distribution of this error variable was approximately normal, but, in practice, they work for most distributions provided that they are unimodal and that the critical differences are for the 95 % level. Within a single laboratory its variance is called the within-laboratory variance and is expressed as

$$\text{var}(e) = \sigma_w^2$$

5.4.2 It may be expected that σ_w^2 will have different values in different laboratories due to differences such as in the skills of the operators, but, in this International Standard, it is assumed that for a properly standardized test method such differences between laboratories should be small and that it is justifiable to establish a common value of within-laboratory variance for all the laboratories using the test method. This common value, which is the average of all the within-laboratory variances taken over all the labaoratories taking part in the precision experiment, is called the repeatability variance and is expressed as

$$\overline{\text{var}}(e) = \sigma_r^2$$

5.5 Relation between the basic model, and r and R

When the basic model (see 5.1) is adopted, the repeatability value r depends solely on the repeatability variance (5.4.2), while the reproducibility value R depends on the sum of the repeatability variance and the between-laboratory variance (see 5.3.1). Thus, there are two quantities, called the repeatability standard deviation, expressed as

$$\sigma_r = \sqrt{\overline{\text{var}(e)}}$$

and the reproducibility standard deviation, expressed as

$$\sigma_R = \sqrt{\sigma_L^2 + \sigma_r^2}$$

Hence

repeatability value $r = f\sqrt{2}\,\sigma_r$, and ...(2)

reproducibility value $R = f\sqrt{2}\,\sigma_R$ ...(3)

where

the coefficient $\sqrt{2}$ is derived from the fact that r and R refer to the difference between two single test results;

f is a factor the value of which depends both on the number of test results available for estimating each of the variances and on the shape of the distributions of the components B and e (see 5.1).

However, if these distributions are approximately normal and the number of test results is not too small, then for a probability level of 95 % the factor f will never differ much from the value 2 and the use of this value is therefore recommended in this International Standard, with the value of $f\sqrt{2}$ being rounded to be 2,8. (Taking into account variations in the factor f would lead to considerable complications and would not effectively contribute to the practical value of r and R.)

In practice, as the exact values of the repeatability standard deviation and the reproducibility standard deviation are not known, they are replaced by their estimates s, leading to

$$r = 2{,}8\,s_r \quad ...(4)$$

$$R = 2{,}8\,s_R \quad ...(5)$$

5.6 Suitability of the basic model

It is clear that the basic model presented in 5.1 is an approximation that, by extensive experience, is known to satisfy practical requirements as a working hypothesis for designing the experiments and analysing the data. For the purposes of this International Standard, the model is an acceptable approximation as long as the experimental requirements laid down in section two are met and the statistical tests specified in section three do not yield significant results that indicate its unsuitability. The action that should be taken when these statistical tests indicate that the model is unsuitable are discussed in clauses 16 and 17.

6 Design of a precision experiment

6.1 In one layout, samples from q batches of material, representing q different levels of the test property, are sent to p laboratories which each perform n tests under repeatability conditions at each level. These n tests are thus made on identical material. This type of experiment is called a uniform-level experiment.

6.2 An alternative preferred in certain cases (see 10.4.2) is the split-level experiment. Each level is split into two sub-levels, a and b, which are only slightly different. Each laboratory receives one sample from each of these sub-levels for testing.

6.3 Full examples of both layouts are given in the case studies in section five. Practical considerations in planning and execution are given in section two.

7 Analysis of the data

7.1 The analysis of the data produced by a precision experiment, which should be considered as a statistical problem to be entrusted to a statistical expert (see 8.2 and 9.2), involves the following three successive stages :

a) critical examination of the data in order to identify and treat outliers or other irregularities and to test the suitability of the model;

b) computation of preliminary values of r and R for each level separately;

c) establishment of final values of r and R, including the establishment of a relation between r, R and m when the analysis indicates that either of the first two depend on the level m.

7.2 As detailed in 14.7 to 14.10, the analysis of a precision experiment first computes, for each level separately, estimates of the repeatability variance s_r^2, the between-laboratory variance s_L^2 and the reproducibility variance s_R^2, as defined in 5.3, 5.4 and 5.5, and then the values of repeatability r and the reproducibility R.

7.3 The analysis, especially 7.1 a), includes a systematic application of statistical tests, a great variety of which are available from the literature and which could be used for the purposes of this International Standard. For practical reasons, only a limited number of these tests, as explained in section three, have been incorporated in this International Standard.

Section two : Organization of an inter-laboratory precision experiment

NOTE — The methods of operation within different organizations are not expected to be identical. Therefore, the contents of this section are only intended as a guide to be modified as appropriate to cater for a particular situation.

8 Personnel requirements

8.1 Panel

The actual planning of the experiment should be the task of a panel of experts familiar with the test method and its application.

8.2 Statistical expert

At least one member of the panel should have experience in the statistical design and analysis of experiments.

8.3 Executive officer

The actual organization of the experiment should be entrusted to a single laboratory. A member of the staff of that laboratory shall take full responsibility; he is called the executive officer.

8.4 Supervisors

A staff member in each of the participating laboratories should be made responsible for organizing the actual performance of the tests in keeping with instructions received from the executive officer, and for reporting the test results.

8.5 Operators

In each laboratory, the tests shall be carried out by one operator selected as representative of those likely to perform the tests in normal operations. He should be instructed by the supervisor as to the dates on which, and the order in which, the tests have to be carried out, but the instructions should not amplify the test method itself.

9 Tasks and problems

9.1 The following questions should be discussed by the panel :

a) Is a satisfactory standard available for the test method ?

b) What is the range of levels encountered in practice ?

c) How many levels should be used in the experiment ? (See 10.1.)

d) What are suitable materials to represent these levels ?

e) Should the material be specially homogenized before preparing the samples or should the heterogeneity in the material be included in the values of r and R ? (See 10.3.)

f) What number n of replicates should be specified and what amount of material should be sent to the laboratories ? (See 10.1.)

g) Should each laboratory be sent n separate samples for each level or one sample for n replicate tests ? (See 10.3.) Or is a split-level experiment desirable ? (See 10.4.2.)

h) Should the laboratories be sent additional material for practical exercises before the official tests are performed ? (See 10.5.2.)

i) How many laboratories should be recruited to cooperate in the experiment ? (See 10.1.)

j) How should the laboratories be recruited and what requirements should they satisfy ? (See 10.2.)

k) What instructions should be issued to the supervisors concerning the execution of the tests, and to how many significant figures should the test results be reported ? (See 10.4.1 and 10.5.1.)

l) What information should be requested in addition to the numerical test results ? (See 10.6.)

m) Who should be appointed to be executive officer ?

9.2 The tasks of the statistical expert are

a) to contribute his specialized knowledge in designing the experiment;

b) to analyse the data;

c) to write a report for submission to the panel following the instructions contained in section three.

9.3 The task of the executive officer is to organize the experiment as planned by the panel, in particular

a) to enlist the cooperation of the requisite number of laboratories and to ensure that supervisors are appointed;

b) to organize and supervise the preparation of the materials and samples, and the despatch of the samples. For each level, a certain quantity of material should be set aside as a reserve stock;

c) to draft instructions and circulate them to the supervisors early enough in advance for them to raise comments or queries;

d) to design suitable forms for the operator to use as a working record and for the supervisor to report the test results;

e) to collect the test results and prepare a table suitable for the statistical analysis.

9.4 The tasks of the supervisor are

a) to hand out the samples to the operator(s) in keeping with the instructions of the executive officer;

b) to supervise the execution of the tests (the supervisor shall not take part in performing the tests);

c) to collect the test results, including any anomalies and difficulties experienced, and to report them to the executive officer.

9.5 The tasks of the operators are

a) to perform the tests in accordance with the standard test method;

b) to report any anomalies or difficulties experienced (see 10.5.1 and 10.5.3).

9.6 The final tasks of the panel are

a) to discuss the report by the statistical expert;

b) to establish final values for the repeatability and reproducibility;

c) to decide if further actions are required for improving the standard for the test method or with regard to laboratories whose results have been rejected as outliers [see 11.2.3 d)].

9.7 As 9.2 and 9.6 are considered during the final stages of the statistical analysis, they are discussed further in section three.

10 Comments on clauses 8 and 9

10.1 Number of laboratories and levels

No hard and fast rules can be laid down. The number of levels in a precision experiment should be chosen in relation to the range of levels to be covered, bearing in mind the cost of performing tests.

If the range of levels is very wide, r and R can be expected to depend on the level m. The use of at least six levels is desirable in order to establish the relationship between these quantities in a satisfactory manner. On the other hand, for the example on the determination of the softening point of a tar product given in clause 23 (with a range of levels from 88 to 102 °C), the use of four levels may be considered as more than adequate.

The number of laboratories should to some extent depend on the number of levels. It is recommended that the number of laboratories should never be fewer than eight; and if only one level is of interest, the number of laboratories should preferably be higher, for example 15 or more.

Regarding the value of n, the recommended figure is two except where it is customary to make a larger number of replicates, such as with certain simple physical tests.

10.2 Recruitment of participating laboratories

10.2.1 From a statistical point of view, the laboratories participating in a precision experiment should be chosen at random from all laboratories likely to use the test method. Volunteers may not represent a realistic cross-section of laboratories. However other practical considerations may intervene, for example, a requirement that participating laboratories be distributed over different continents or climatic regions may affect the pattern of representation. The panel should lay down the recruitment policy and the requirements for the participating laboratories.

10.2.2 In enlisting the cooperation of the requisite number of laboratories, their responsibilities should be clearly stated. An example of a suitable enlistment questionnaire that may be used for this purpose is given below :

Questionnaire on inter-laboratory study

Title of test method (copy attached) : .....
.....
.....

1 Our laboratory is willing to participate in the precision experiment for this standard test method.

YES ☐ NO ☐ (tick the appropriate box)

2 As a participant, we understand that

a) all essential apparatus, chemicals and other requirements specified in the method shall be available in our laboratory when the programme begins;

b) specified "timing" requirements, such as starting date, order of testing specimens and finishing date of the programme, shall be rigidly met;

c) the method shall be strictly adhered to;

d) samples shall be handled in accordance with instructions;

e) a qualified operator shall perform the test.

Having studied the method and having made a fair appraisal of our capabilities and facilities, we feel that we will be adequately prepared for cooperative testing of this method.

3 Comments : .....
.....
.....

Signature : .....

Company or laboratory : .....

10.3 Heterogeneity of the material

When the material to be tested is not homogeneous, it is important to prepare the samples in the manner stipulated by the method, preferably starting with one batch of commercial material for each level. Some modification may be necessary to ensure that a sufficient amount of material is available to cover the experiment and keep a certain stock in reserve. For the samples at each level, n separate containers should be used for each laboratory if there is any danger of the materials deteriorating once the container has been opened (e.g. by oxidation, by losing volatile components or in the case of hygroscopic material). In the case of unstable materials, special instructions on storage and treatment should be stipulated.

In general when publishing values of r and R, it is recommended that the material used in the precision experiment should be clearly specified along with the range of materials to which the values can be expected to apply.

10.4 Actual organization of the tests

10.4.1 With q levels and n replicates, each participating laboratory has to carry out qn tests. The performance of these tests should be organized and the operators instructed as follows :

a) All qn tests should be performed by one and the same operator using the same equipment throughout.

b) Each group of n tests belonging to one level shall be carried out under repeatability conditions, i.e. in a short interval of time and by the same operator, and without any intermediate recalibration of apparatus unless this is an integral part of making a determination.

c) It is not essential that the q groups of n tests each be performed strictly within a short interval; different groups of tests may be carried out on different days.

d) If in the course of the tests the operator should drop out, another operator may complete the tests, provided that the change does not occur within a group of n tests at one level but only occurs between two of the q groups. Any such change shall be reported with the test results.

e) It is essential that a group of n tests under repeatability conditions be performed independently as if they were n tests on different material. As a rule, however, the operator will know that he is testing identical material, but the point should be stressed to him in his instructions that the whole object of the experiment is to determine what differences in results can occur in actual testing. If it is feared that, despite this warning, previous results may influence succeeding test results and thus the repeatability variance, then a split-level experiment is considered the correct procedure (see 10.4.2).

10.4.2 An alternative procedure, sometimes adopted when n equals 2, is that of using split-level experiments. Adoption of this procedure may be considered when it is feared that the operator, when testing successive identical samples, may be influenced by the result of his first test. In this procedure, instead of using two samples that the operator has been told should be identical, or performing two tests on the same specimen of material, two series of p samples are prepared at slightly different levels m_a and m_b (where $m_a - m_b$ is small) and each of the p laboratories receives one sample from series **a** and one from series **b** for testing. It shall be distinguished clearly which test result belongs to series **a** and which to series **b**; they cannot be interchanged as can two test results on identical material. The values of r and R derived from a split-level experiment are valid for the mean level m equal to $(m_a + m_b)/2$.

The split-level design requires a slight modification in the statistical analysis, as discussed in section three.

10.4.3 Additional aspects of organizing the tests are as follows :

a) it may be necessary to limit the time that should be allowed to elapse between the day the samples are received and the day the tests are performed;

b) any preliminary checking of equipment should be as laid down in the standard method;

c) all samples should be clearly labelled with the name of the experiment and a sample identification.

10.5 Instructions to the operators

10.5.1 Before performing the tests the operators should receive no instructions that supplement those contained in the standard test method; these alone should suffice. The operators should, however, be encouraged to comment on the standard, in particular to state whether the instructions contained in it are sufficiently unambiguous and clear. For example, ambiguities may arise when a standard has been translated into different languages. However, it is desirable that all participating laboratories report their test results to the same number of decimal places, and the supervisors should be instructed accordingly. In commercial practice, the test results may be rounded rather crudely, and in a precision experiment it may be advisable to use one more decimal than is customary or laid down in the standard method. When r or R may depend on the level m, different rules for rounding may be needed for different levels.

10.5.2 An operator may not achieve normal precision when he carries out a test method for the first time or after a long interval. In such cases, subject to the decision of the panel or of the supervisors, the operators may be allowed to carry out a few unofficial tests in order to gain experience with the method before starting testing on the official samples of the precision experiment. Such preliminary familiarization tests shall never be performed on the official samples, and material for them should be supplied by the executive officer.

10.5.3 The operators should be told to report any occasions when they are not able to follow their instructions or when they accidentally fail to keep to the instructions. They should also be told that it is better to report a mistake than to adjust the results, because one or two missing results will not spoil the experiment and may indicate a deficiency in the standard.

10.6 Reporting the test results

The supervisor of each laboratory should write a full report on the tests which should contain the following information :

a) the final test results, taking particular care to avoid transcription and typing errors, e.g. by using photocopies of the operators' results;

b) the original observations or readings (if any) from which the final results were derived, possibly by photocopying the operators' workbook;

c) comments by the operators on the standard for the test;

d) information about irregularities or disturbances that may have occurred during the tests, including any change of operator that may have occurred along with a statement as to which tests were performed by which operator;

e) the date(s) on which the samples were received;

f) the date(s) on which each sample was tested;

g) information about the equipment used, if relevant;

h) any other relevant information.

Section three : Statistical analysis of the results of an inter-laboratory experiment

11 Preliminary considerations

11.1 Statistical expert

The analysis of the test results produced by a precision experiment is the task of the statistical expert who is a member of the panel and has taken part in planning the experiment. (See 8.2 and 9.2.)

11.2 Cells

Each combination of a laboratory and a level is called a cell of the precision experiment. In the ideal case, the results of an experiment with p laboratories and q levels consist of a table with pq cells each containing n replicate results that can all be used for computing the repeatability r and the reproducibility R. This ideal situation is not, however, always attained in practice. Departures occur due to redundant data, missing data and outliers.

11.2.1 Redundant data

Sometimes a laboratory may carry out and report more than the n replicates officially stipulated. In that case, the supervisor (see 8.4 and 9.4) reports, or is asked to report, why this was done and which are the correct test results. If the answer is that they are all equally valid, they can all be taken into account by using the computational procedure of 14.9.

11.2.2 Missing data

In other cases, some of the test results may be missing, e.g. due to the loss of a sample or a slip in performing the test. The analysis recommended in clause 16 is such that completely empty cells can simply be ignored, while partly empty cells can be taken into account by the computational procedure of 14.9. The reasons for the missing test results should be given in the supervisor's report.

NOTE — If one of the two test results in a cell of a split-level experiment (see 10.4.2) is missing, the single test result available has to be discarded and the cell treated as an empty one.

11.2.3 Outliers

These entries among the original test results, or in the tables derived from them, deviate so much from the comparable entries in the same table that they are considered as irreconcilable with the other data. Experience has taught that outliers cannot always be avoided and have to be taken into consideration.

The following practice is recommended for dealing with outliers :

a) Cochran's one-sided outlier test (see clause 12) and Dixon's outlier test (see clause 13) are recommended in combination with the following procedures :

$P > 5$ %, i.e. Cochran's or Dixon's test statistic is less than its 5 % critical value : the item tested is accepted as correct;

5 % $\geqslant P >$ 1 %, i.e. the test statistic lies between its 5 % and 1 % critical values : the item tested is called a straggler and is marked with a single asterisk;

$P \leqslant 1$ %, i.e. the test statistic is greater than its 1 % critical value : the item is called a statistical outlier and is marked with a double asterisk;

P is the probability of the observed value of the test statistic.

The 5 % and 1 % critical values for Cochran's and Dixon's tests are given in annexes A and B.

b) Sometimes the actual application of these statistical tests may be omitted or other statistical tests may be chosen because a statistical expert will see from a cursory examination of the data (for example from a graphical presentation) that the test will yield either a non-significant or a highly significant result. In case of any doubt, however, the test should always be applied.

c) It is next investigated whether the stragglers and/or statistical outliers can be explained by some technical error, e.g. a slip in performing the test, a computational error, a clerical error in transcribing a test result or the analysis of a wrong sample. Where the error is of the computation or transcription type, the suspect test result should be replaced by the correct value; where the error is in analysing the wrong sample, the test result should be placed in its correct cell. After such correction has been made, the examination for stragglers/outliers should be repeated. If the explanation of the technical error is such that it proves impossible to replace the suspect test result, it should be discarded as a real outlier that does not belong to the experiment proper.

d) When several unexplained stragglers and/or statistical outliers occur at different levels within the same laboratory, that laboratory may be considered as an outlier, having too high a within-laboratory variance, and/or too large a systematic error in the level of its test results. It may then be reasonable to discard some or all the data from such an outlying laboratory.

(This International Standard does not provide a statistical test by which suspected laboratories may be judged. The primary decision should be the responsibility of the statistical expert, but all rejected laboratories shall be reported to the panel for further action. Examples of outlying laboratories occur in the case study of clause 24.)

e) When any stragglers and/or statistical outliers remain that have not been explained or rejected as belonging to an outlying laboratory, the stragglers are retained as correct items, and the statistical outliers are discarded, unless the statistician for good reasons decides to retain them.

11.3 Computation of r and R

The computation of the repeatability r and the reproducibility R is carried out, for each level separately, from the data remaining after elimination or correction of the stragglers and/or outliers.

11.4 Functional relation between r, R and m

Provided that there are several levels and that a functional relation between r (and/or R) and m is expected (see 15.1), it is then investigated whether r (and/or R) depends on m and, if so, the relationship between these quantities is determined.

11.5 Notations used

As stated in 11.2, the ideal case is p laboratories L_i ($i = 1, 2, \ldots, p$), each testing q levels M_j ($j = 1, 2, \ldots, q$) and n replicates at each level (each $L_i M_j$ combination) giving a total pqn results of the tests. As a result of redundant (see 11.2.1), missing (see 11.2.2) or deviating (see 11.2.3) results, or deviating laboratories [see 11.2.3 d)], this ideal situation is not always attained. Under these conditions, the notations given in 11.5.1 to 11.5.3 will be used in the remainder of this International Standard. Specimen recommended tables for the statistical analysis are given in figure 1. For convenience they will be referred to simply as tables A, B and C rather than as figure 1.

11.5.1 Original results (tabler A)

11.5.1.1 Case of a uniform-level experiment

n_{ij} is the number of results in cell $L_i M_j$,

y_{ijk} is any one of these results ($k = 1, 2, \ldots, n_{ij}$),

p_j is the number of laboratories reporting at least one result (after any result designated as an outlier has been eliminated).

11.5.1.2 Case of a split-level experiment

y_{ija} and y_{ijb} are the results obtained, respectively at sub-levels a and b, level j, laboratory i. The notation p_j is applicable to this case only when both results for the two sub-levels exist.

11.5.2 Measures of cell spread (table B)

These are derived from table A (see 11.5.1) and table C (see 11.5.3) as described in 11.5.2.1 and 11.5.2.2.

11.5.2.1 Case of a uniform-level experiment

For the general case, use the intra-cell standard deviation

$$s_{ij} = \sqrt{\frac{1}{n_{ij} - 1} \sum_{k=1}^{n_{ij}} (y_{ijk} - \bar{y}_{ij})^2} \qquad \ldots (6)$$

$$s_{ij} = \sqrt{\frac{1}{n_{ij} - 1} \left[\sum_{k=1}^{n_{ij}} y_{ijk}^2 - \frac{1}{n_{ij}} \left(\sum_{k=1}^{n_{ij}} y_{ijk} \right)^2 \right]} \qquad \ldots (7)$$

Small, and in themselves unimportant, rounding errors in $\bar{y}_{ij}$ can produce considerable errors in s_{ij}. Hence formula (7) is preferred, and formula (6) should be used only with coded data. (See 14.11.)

The standard deviation should be expressed with one more decimal place than the results in table A.

For the particular cases where all $n_{ij} = n = 2$, use the cell range

$$w_{ij} = |y_{ij1} - y_{ij2}| \qquad \ldots (8)$$

that is, without regard for sign.

11.5.2.2 Case of a split-level experiment

$$d_{ij} = y_{ija} - y_{ijb} \qquad \ldots (9)$$

taking the sign into account.

11.5.3 Cell averages (for the two types of experiment) (table C)

These are derived from table A as follows :

$$\bar{y}_{ij} = \frac{1}{n_{ij}} \sum_{k=1}^{n_{ij}} y_{ijk} \qquad \ldots (10)$$

The cell averages should be recorded with one more decimal place than the test results in table A.

11.5.4 Simplified notations used in clauses 12, 13 and 14

Clauses 12 and 13 concern statistical tests and clause 14 relates to procedures for calculating r and R, which are applied separately at each level (fixed j). In these clauses, for clarity of layout, the subscript j will be omitted from the notations defined in 11.5.1 to 11.5.3, when this subscript is not indispensable.

Table A — Recommended form for the collation of the original data

Uniform-level experiment

Laboratory \ Level	1	2		j		q
1						
2						
i				⋯ ⋯ y_{ijk} ⋯		
p						

Split-level experiment

Laboratory \ Level	1		2		j		q	
	a	b	a	b	a	b	a	b
1								
2								
i					y_{ija}	y_{ijb}		
p								

Table B — Recommended form for the collation of measures of spread

Uniform-level experiment

Laboratory \ Level	1	2			j		q
1							
2							
i					s_{ij} or w_{ij}		
p							

Split-level experiment

Laboratory \ Level	1	2		j		q
1						
2						
i				d_{ij}		
p						

Table C — Recommended form for the collation of cell averages

Laboratory \ Level	1	2				j				q
1										
2										
i						$\bar{y}_{ij}$				
p										

Figure 1 — Recommended forms for the collation of test results for analysis

11.5.5 Corrected or rejected data

As some of the data may be corrected or rejected on the basis of the tests outlined in 11.2.3, the values of y_{ijk}, n_{ij} and p_j used for the final determinations of r and R may be different from the values referring to the original test results as recorded in tables A, B and C in figure 1. Hence, in reporting the final values of r and R, it should always be stated what data, if any, have been corrected or discarded.

11.6 Repeatability variance s_r^2 and between-laboratory variance s_L^2

The values of s_r^2 and s_L^2 are given by the following equations for a given level j. For convenience, the subscript j has been dropped.

11.6.1 Case of a uniform-level experiment

$$s_r^2 = \frac{\sum_{i=1}^{p} (n_i - 1)\, s_i^2}{\left(\sum_{i=1}^{p} n_i\right) - p} \qquad \ldots (11)$$

$$s_L^2 = \frac{\dfrac{1}{p-1}\left[\sum_{i=1}^{p} n_i(\bar{y}_i - \bar{\bar{y}})^2\right] - s_r^2}{\bar{\bar{n}}} \qquad \ldots (12)$$

where

$$\bar{\bar{y}} = \frac{\sum_{i=1}^{p} n_i \bar{y}_i}{\sum_{i=1}^{p} n_i} \qquad \ldots (13)$$

$$\bar{\bar{n}} = \frac{1}{p-1}\left[\sum_{i=1}^{p} n_i - \frac{\sum_{i=1}^{p} n_i^2}{\sum_{i=1}^{p} n_i}\right] \qquad \ldots (14)$$

The application of these formulae in an example is given in 14.8.2 and 14.9.2.

For the particular case where all $n_i = n = 2$, the cell range $w_i = \sqrt{2}\, s_i$ is used, giving

$$s_r^2 = \frac{1}{2p} \sum_{i=1}^{p} w_i^2$$

and

$$s_L^2 = \frac{1}{p-1}\left[\sum_{i=1}^{p} (\bar{y}_i - \bar{\bar{y}})^2\right] - \frac{s_r^2}{2}$$

These formulae are illustrated in an example given in 14.7.2.

11.6.2 Case of a split-level experiment

$$s_r^2 = \frac{1}{2(p-1)} \sum_{i=1}^{p} (d_i - \bar{d})^2 \qquad \ldots (15)$$

where

$$\bar{d} = \frac{1}{p} \sum_{i=1}^{p} d_i$$

$$s_L^2 = \frac{1}{p-1}\left[\sum_{i=1}^{p} (\bar{y}_i - \bar{\bar{y}})^2\right] - \frac{s_r^2}{2} \qquad \ldots (16)$$

where

$$\bar{\bar{y}} = \frac{1}{p} \sum_{i=1}^{p} \bar{y}_i$$

The application of these formulae in an example is given in 14.10.2.

12 Cochran's test

12.1 As explained in 5.4.2, this International Standard assumes that between laboratories only small differences exist in the within-laboratory variances. Experience, however, shows that this is not always the case, so that a test has been included to test the validity of this assumption.

Three tests[1)] could be used for this purpose, namely

a) Bartlett's (1937) variance homogeneity test;

b) Hartley's (1940) variance ratio test;

c) Cochran's (1941) one-sided outlier test.

1) All three tests are fully explained in PEARSON, E.S. and HARTLEY, H.O. Test for heterogeneity of variance, *Biometrika tables for statisticians*, 3rd ed., Cambridge University Press, 1976, Vol. 1, chapter 16[2] in which
- Bartlett's test is explained in 16.1 on page 63;
- Hartley's test is explained in 16.5 on page 67;
- Cochran's test is explained in 16.5 on page 67.

The first two tests, however, cannot be applied when one of the variances in a set is zero, which may easily happen as a result of rounding and of the small number of test results on which the variances are based. Moreover, these tests, even if no zeros occur, are very sensitive against the value of the smallest variance which, again due to rounding, is unreliable. For these reasons, only Cochran's test is given in detail in this International Standard.

Cochran's test applies only to uniform-level experiments, as it is a test based on homogeneity of variance. For a split-level experiment, Dixon's outlier test (see clause 13) is applied to the cell differences d_i.

12.2 Given a set of p standard deviations s_i, all computed from the same number n of replicate test results, Cochran's criterion C is given by

$$C = \frac{s_{max}^2}{\sum_{i=1}^{p} s_i^2} \qquad \ldots (17)$$

In the case of two replicates, the ranges w_i can be used instead of the standard deviations s_i, and Cochran's criterion then becomes

$$C = \frac{w_{max}^2}{\sum_{i=1}^{p} w_i^2}$$

In these expressions, s_{max} and w_{max} stand for the highest values in the set. If the test is significant, s_{max} (or w_{max}) is classified as a straggler or statistical outlier according to the procedure of 11.2.3 a). Critical values for Cochran's criterion at the 5 % and 1 % levels are given for $p = 2$ to 40 and $n = 2$ to 6 in annex A.

Cochran's test shall be applied to table B in figure 1 at each level separately.

12.3 As stated in 12.2, Cochran's criterion applies strictly only when all standard deviations are derived from the same number n of test results obtained under conditions of repeatability. In actual cases, this number may vary due to redundant, missing or discarded data. This International Standard assumes, however, that in a properly organized experiment, such variations in the number of test results per cell will be limited and can be ignored, and therefore Cochran's criterion is applied using for n the number of results occurring in the majority of cells.

Cochran's criterion tests only the highest value in a set of standard deviations or ranges and is therefore a one-sided outlier test. Variance heterogeneity may, of course, also manifest itself in some of the standard deviations being comparatively too low. However, small values of standard deviation or range may be very strongly influenced by the degree of rounding of the original test results and are for that reason not very reliable.

In addition, it seems unreasonable to reject the data from a laboratory because it has accomplished a higher precision in its test results than the other laboratories. Hence, Cochran's criterion is considered adequate.

12.4 A critical examination of table B of figure 1 may sometimes reveal that the standard deviations for a particular laboratory are at all or at most levels lower than those for other laboratories. This may indicate that the laboratory works with a lower repeatability than the other laboratories, which in turn may be caused either by a modified or incorrect application of the standard test method or by better technique and equipment.

If this occurs, it should be reported to the panel, which should decide whether the point is worthy of more detailed investigation. (An example of this is laboratory 2 in the experiment described in clause 22.)

12.5 If the highest standard deviation is classed as an outlier, then the value should be omitted and Cochran's test repeated on the remaining values. This process can be repeated but it may lead to excessive rejections when, as is sometimes the case, the underlying assumption of normality is not sufficiently well approximated. The repeated application of Cochran's test is proposed in this International Standard only as a helpful tool in view of the lack of a statistical test designed for testing several outliers together. Cochran's test is not designed for this purpose and great caution should be exercised in drawing conclusions, in particular when this technique reveals several statistical outliers in different laboratories within only one of the levels, some of these may not really be significant. The data have to be examined carefully to decide which outliers can be rejected and which can be retained. On the other hand, if several stragglers and/or statistical outliers are found at different levels within one laboratory, this may be a strong indication that the laboratory's within-laboratory variance is exceptionally high, and the whole of the data from that laboratory should be rejected.

13 Dixon's test

13.1 Given a set of data $z(h)$, $h = 1, 2, \ldots, H$, arranged in order of magnitude, then Dixon's test uses the following test statistics :

H	Test statistic
3 to 7	Q_{10} = the larger of $\frac{z(2) - z(1)}{z(H) - z(1)}$ and $\frac{z(H) - z(H-1)}{z(H) - z(1)}$
8 to 12	Q_{11} = the larger of $\frac{z(2) - z(1)}{z(H-1) - z(1)}$ and $\frac{z(H) - z(H-1)}{z(H) - z(2)}$
13 or more	Q_{22} = the larger of $\frac{z(3) - z(1)}{z(H-2) - z(1)}$ and $\frac{z(H) - z(H-2)}{z(H) - z(3)}$

Critical values of these test statistics at the 5 % and 1 % levels and for values of $H = 3$ to 40 are reproduced in annex B.

13.2 In analysing a precision experiment, Dixon's test should be applied to

a) the test results within a cell of table A in figure 1 when $n_{ij} \geqslant 3$, provided that Cochran's test has already indicated an anomaly. In this case, $h = k$, $H = n_{ij}$, and $z(h) = y_{ijk}$, i and j both being fixed;

b) the cell averages for a given level j in table C in figure 1, when in that case, $h = i$, $H = p_j$, and $z(h) = \bar{y}_{ij}$, j being fixed;

c) the cell differences, $d_{ij} = y_{ija} - y_{ijb}$, for a given level of a split-level experiment given in table B in figure 1, when in that case, $h = i$, $H = p_j$, and $z(h) = d_{ij}$, j being fixed.

13.3 If Dixon's test reveals one of the extreme values in a series (the highest or the lowest) as a straggler or statistical outlier, the test should again be applied to the remaining $H - 1$ values; and if this once more proves one of the extremes as suspect, the test should be applied afresh to the remaining set of $H - 2$ values. However, as explained in Cochran's test in 12.5, great caution should be exercised in drawing conclusions from the result of repeated applications of Dixon's test, and the comments in 12.5 also apply to this test.

14 Computation of the mean level m, the repeatability r and the reproducibility R

14.1 Method of analysis

In this International Standard, the method of analysis adopted involves carrying out the computation of m, r and R for each level separately. When there are q levels, the results of the computation are expressed as m_j, r_j and R_j ($j = 1, 2, \ldots, q$). Subsequently, it is investigated whether r and/or R depend on m and if so, the functional relationship is determined.

14.2 Basic data

The basic data needed for the computations are presented in the three tables in figure 1 (see 11.5) :

— table A containing the original results;

— table B containing the measures of within-cell spread;

— table C containing the cell averages.

14.3 Non-empty cells

As a consequence of the rule stated in 16.9, the number of non-empty cells to be used in the computations will, for a specified level, always be the same in tables B and C. An exception might occur if, owing to missing data, a cell in table A contains only a single test result, which will entail an empty cell in table B but not in table C. In that case, it is possible either

a) to discard the solitary test result, which will lead to empty cells in both tables B and C, or,

b) if this is considered an unwarranted loss of information, to insert a nominal value of zero (0) in table B.

If option b) is taken, the computations have to be carried out in accordance with 14.9. For a cell with a single test result, any value could be inserted in table B without influencing the final outcome, but a nominal value (0) seems most appropriate.

The number of non-empty cells may be different for different levels; hence the subscript j in p_j.

14.4 Number of replicates per cell

Owing to missing data or to the possible rejection of some of the original test results, the number of replicates per cell in table A (see 11.5.1) need not be the same, and this number is therefore denoted by n_{ij} for laboratory i and level j.

14.5 Rounding of results

The computations described in the remainder of this clause assume that the instructions for rounding specified in 11.5.2 and 11.5.3 have been observed. No further rounding should be carried out in the course of the computations, but an appropriate rounding should be applied to the final results m, r and R.

14.6 Variations of procedure

The computational procedure depends on the type of experiment and on the number of replicates in the cells. Four different situations are examined, each illustrated by a numerical example, which cover most situations likely to arise. In each example, only one level is considered so for convenience the subscript j has been omitted. Any outliers found have already been discarded and only the acceptable data are quoted. The data shown have been extracted from tables B and C only, as table A is irrelevant at this stage.

If, owing to random errors, a negative value for s_L^2 is obtained from the calculations, a value of zero should be substituted in the formula for s_R^2.

14.7 Uniform-level experiment with $n = 2$ replicates per cell

14.7.1 Basic data for one of the levels from tables B and C (see figure 1)

Laboratory	Original data from	
	table B	table C
i	w_i	$\bar{y}_i$
1	0,5	31,45
2	0,0	30,90
3	0,2	30,80
4	0,4	31,30
5	0,3	31,45
6	0,2	31,50
7	0,0	31,40

14.7.2 Computational formulae and numerical results

Number of laboratories : p	$p = 7$
Number of replicates : n	$n = 2$
$T_1 = \Sigma \bar{y}_i$	$T_1 = 218{,}80$
$T_2 = \Sigma \bar{y}_i^2$	$T_2 = 6\,839{,}555\,0$
$T_3 = \Sigma w_i^2$	$T_3 = 0{,}58$
$s_r^2 = \dfrac{T_3}{2p}$	$s_r^2 = \dfrac{0{,}58}{2 \times 7} = 0{,}041\,4$
$s_L^2 = \dfrac{pT_2 - T_1^2}{p(p-1)} - \dfrac{s_r^2}{2}$	$s_L^2 = \dfrac{7 \times 6\,839{,}555\,0 - 218{,}80^2}{7 \times 6} - \dfrac{0{,}041\,4}{2} = 0{,}061\,3$
$s_R^2 = s_L^2 + s_r^2$	$s_R^2 = 0{,}061\,3 + 0{,}041\,4 = 0{,}102\,7$
$m = \dfrac{T_1}{p}$	$m = \dfrac{218{,}80}{7} = 31{,}26$
$r = 2{,}8\sqrt{s_r^2}$	$r = 2{,}8\sqrt{0{,}041\,4} = 0{,}57$
$R = 2{,}8\sqrt{s_R^2}$	$R = 2{,}8\sqrt{0{,}102\,7} = 0{,}90$

14.8 Uniform-level experiment with a constant $n > 2$ replicates per cell

14.8.1 Basic data for one of the levels from tables B and C (see figure 1)

Laboratory	Original data from		Number of replicates
	table B	table C	
i	s_i	$\bar{y}_i$	n_i
1	0,82	28,03	3
2	1,50	21,25	3
3	3,00	22,47	3
4	0,58	25,50	3
5	1,49	33,08	3
6	0,50	24,23	3
7	2,38	20,53	3
8	0,93	30,17	3
9	1,07	22,40	3

14.8.2 Computational formulae and numerical results

Number of laboratories : p	$p = 9$
Number of replicates : n	$n = 3$
$T_1 = \Sigma \bar{y}_i$	$T_1 = 227,66$
$T_2 = \Sigma \bar{y}_i^2$	$T_2 = 5\,907,243\,4$
$T_3 = \Sigma s_i^2$	$T_3 = 22,403\,1$
$s_r^2 = \dfrac{T_3}{p}$	$s_r^2 = \dfrac{22,403\,1}{9} = 2,489\,2$
$s_L^2 = \dfrac{pT_2 - T_1^2}{p(p-1)} - \dfrac{s_r^2}{n}$	$s_L^2 = \dfrac{9 \times 5\,907,243\,4 - 227,66^2}{9 \times 8} - \dfrac{2,489\,2}{3} = 17,727\,4$
$s_R^2 = s_L^2 + s_r^2$	$s_R^2 = 17,727\,4 + 2,489\,2 = 20,216\,6$
$m = \dfrac{T_1}{p}$	$m = \dfrac{227,66}{9} = 25,30$
$r = 2,8\sqrt{s_r^2}$	$r = 2,8\sqrt{2,489\,2} = 4,42$
$R = 2,8\sqrt{s_R^2}$	$R = 2,8\sqrt{20,216\,6} = 12,6$

14.9 Uniform-level experiment with unequal number of replicates per cell

14.9.1 Basic data for one of the levels from tables B and C (see figure 1)

Laboratory	Original data from		Number of replicates
	table B	table C	
i	s_i	$\bar{y}_i$	n_i
1	0,14	21,30	2
2	0,14	21,50	2
3	0,07	20,75	2
4	0,21	21,75	2
5	0,10	20,90	3
6	0,21	21,05	2
7	0,28	21,50	4
8	0,21	20,85	2
9	0,28	21,10	2
10	0,35	20,85	2
11	(0)	21,30	1

14.9.2 Computational formulae and numerical results

Number of laboratories : p	$p = 11$
$T_1 = \Sigma n_i \bar{y}_i$	$T_1 = 508{,}30$
$T_2 = \Sigma n_i \bar{y}_i^2$	$T_2 = 10\,767{,}765\,0$
$T_3 = \Sigma n_i$	$T_3 = 24$
$T_4 = \Sigma n_i^2$	$T_4 = 58$
$T_5 = \Sigma (n_i - 1) \cdot s_i^2$	$T_5 = 0{,}632\,5$
$s_r^2 = \dfrac{T_5}{T_3 - p}$	$s_r^2 = \dfrac{0{,}632\,5}{24 - 11} = 0{,}048\,6$
$s_L^2 = \left[\dfrac{T_2 T_3 - T_1^2}{T_3 (p-1)} - s_r^2\right]\left[\dfrac{T_3 (p-1)}{T_3^2 - T_4}\right]$	$s_L^2 = \left[\dfrac{24 \times 10\,767{,}765\,0 - 508{,}30^2}{24(11-1)} - 0{,}048\,6\right]\left[\dfrac{24(11-1)}{24^2 - 58}\right] = 0{,}088\,4$
$s_R^2 = s_L^2 + s_r^2$	$s_R^2 = 0{,}088\,4 + 0{,}048\,6 = 0{,}137\,0$
$m = \dfrac{T_1}{T_3}$	$m = \dfrac{508{,}30}{24} = 21{,}18$
$r = 2{,}8\sqrt{s_r^2}$	$r = 2{,}8\sqrt{0{,}048\,6} = 0{,}62$
$R = 2{,}8\sqrt{s_R^2}$	$R = 2{,}8\sqrt{0{,}137\,0} = 1{,}04$

14.10 Split-level experiment

14.10.1 Basic data for one of the levels from tables B and C (see figure 1)

Laboratory i	Original data from table B d_i	Original data from table C $\bar{y}_i$
1	− 0,54	18,770
2	− 0,47	18,615
3	− 0,43	18,465
4	− 0,48	19,660
5	− 0,51	18,865
6	− 0,49	18,335
7	− 0,53	18,895
8	− 0,50	18,680
9	− 0,57	19,105

14.10.2 Computational formulae and numerical results

Number of laboratories : p	$p = 9$
$T_1 = \Sigma \bar{y}_i$	$T_1 = 169{,}390$
$T_2 = \Sigma \bar{y}_i^2$	$T_2 = 3\ 189{,}327\ 850$
$T_3 = \Sigma d_i$	$T_3 = -4{,}52$
$T_4 = \Sigma d_i^2$	$T_4 = 2{,}283\ 8$
$s_r^2 = \dfrac{pT_4 - T_3^2}{2p(p-1)}$	$s_r^2 = \dfrac{9 \times 2{,}283\ 8 - (-4{,}52)^2}{2 \times 9 \times 8} = 0{,}000\ 860$
$s_L^2 = \dfrac{pT_2 - T_1^2}{p(p-1)} - \dfrac{s_r^2}{2}$	$s_L^2 = \dfrac{9 \times 3\ 189{,}327\ 850 - 169{,}390^2}{9 \times 8} - \dfrac{0{,}000\ 860}{2} = 0{,}152\ 050$
$s_R^2 = s_L^2 + s_r^2$	$s_R^2 = 0{,}152\ 050 + 0{,}000\ 860 = 0{,}152\ 910$
$m = \dfrac{T_1}{p}$	$m = \dfrac{169{,}39}{9} = 18{,}821$
$r = 2{,}8\sqrt{s_r^2}$	$r = 2{,}8\sqrt{0{,}000\ 860} = 0{,}082$
$R = 2{,}8\sqrt{s_R^2}$	$R = 2{,}8\sqrt{0{,}152\ 910} = 1{,}09$

14.11 Coding data

14.11.1 The calculations can often be simplified and the risks of computational errors reduced by coding the data. The objective of coding is to reduce the number of digits to be handled and/or to reduce the number of decimal places involved. This is achieved by subtracting a suitable number (the coding constant) from the basic data and multiplying the remainder by a factor (the coding factor) which is usually an integral power of 10.

The data to be used in the calculations then become

$$x = v(y - u)$$

where

x is the coded data;

y is the original data;

u is the coding constant;

v is the coding factor.

From this, $\bar{x}_{ij} = v(\bar{y}_{ij} - u)$ and the other coded values are related to their uncoded equivalents by

coded $w_{ij} = v$(uncoded w_{ij}),

coded $s_{ij} = v$(uncoded s_{ij}),

coded $d_{ij} = v$(uncoded d_{ij}),

and any of these relationships lead to

coded $s_r^2 = v^2$(uncoded s_r^2),

coded $s_L^2 = v^2$(uncoded s_L^2).

At the end of the calculations, the coded results are translated back to the original units, first by dividing by v to remove the coding factor and then, in the case of the mean level, by adding back the coding constant u.

Coding may be introduced either at the stage of the basic data (table A) or at a later stage (tables B and C).

The example given in 14.11.2 shows the effect of coding by repeating the calculation of 14.10. The coding constant $u = 18{,}000$ and the coding factor $v = 100$.

Thus the data become

$$\bar{x}_i \text{ (coded)} = 100\ (\bar{y}_i - 18{,}000)$$

so, for laboratory 1,

$$\bar{x}_{c1} = 100(18{,}770 - 18{,}000) = 77{,}0$$

$$d_{c1} = 100(d_1) = 100(-0{,}54) = -54$$

NOTE — The coding constant u has no effect on the within-cell difference, nor would it affect the within-cell standard deviations or ranges in the other examples. Therefore only the coding factor v affects the contents of table B.

14.11.2 Basic data for one of the levels from tables B and C (see figure 1)

Laboratory i	Original data from table B d_i	Original data from table C $\bar{y}_i$	Coded data for table B d_{ci}	Coded data for table C $\bar{x}_{ci}$
1	− 0,54	18,770	− 54	77,0
2	− 0,47	18,615	− 47	61,5
3	− 0,43	18,465	− 43	46,5
4	− 0,48	19,660	− 48	166,0
5	− 0,51	18,865	− 51	86,5
6	− 0,49	18,335	− 49	33,5
7	− 0,53	18,895	− 53	89,5
8	− 0,50	18,680	− 50	68,0
9	− 0,57	19,105	− 57	110,5

14.11.3 Computational formulae and numerical results

Number of laboratories : p	$p = 9$
Coding constant : u	$u = 18{,}000$
Coding factor : v	$v = 100$
$t_1 = \Sigma \bar{x}_{ci}$	$t_1 = 739{,}0$
$t_2 = \Sigma \bar{x}_{ci}^2$	$t_2 = 72\ 878{,}50$
$t_3 = \Sigma d_{ci}$	$t_3 = -452$
$t_4 = \Sigma d_{ci}^2$	$t_4 = 22\ 838$

It can be seen by comparison with the uncoded calculation in 14.10 that :

coded $t_1 = v$(uncoded $T_1 - pu$)

coded $t_2 = v^2T_2 - 2v^2T_1 + pv^2u^2$

coded $t_3 = vT_3$

coded $t_4 = v^2T_4$

The coded values of s_r^2, s_L^2 and s_R^2 are then calculated in the usual way.

NOTE — The coded variances are exactly v^2 times the uncoded values.

$s_r^2 = \dfrac{pt_4 - t_3^2}{2p(p-1)}$	$s_r^2 = \dfrac{9 \times 22\,838 - (-452)^2}{2 \times 9 \times 8} = 8{,}60$
$s_L^2 = \dfrac{pt_2 - t_1^2}{p(p-1)} - \dfrac{s_r^2}{2}$	$s_L^2 = \dfrac{9 \times 72\,878{,}50 - 739{,}0^2}{9 \times 8} - \dfrac{8{,}60}{2} = 1\,520{,}5$
$s_R^2 = s_L^2 + s_r^2$	$s_R^2 = 1\,520{,}5 + 8{,}60 = 1\,529{,}1$

In the final stage, the values of m, r and R are decoded as follows :

$m = u + \dfrac{t_1}{vp}$	$m = 18{,}000 + \dfrac{739{,}0}{100 \times 9} = 18{,}821$
$r = \dfrac{1}{v} 2{,}8\sqrt{s_r^2}$	$r = \dfrac{1}{100} 2{,}8\sqrt{8{,}60} = 0{,}082$
$R = \dfrac{1}{v} 2{,}8\sqrt{s_R^2}$	$R = \dfrac{1}{100} 2{,}8\sqrt{1\,529{,}1} = 1{,}09$

It can be seen that the final results for m, r and R are identical to those found using the basic data without coding.

15 Establishing a functional relationship between r (or R) and m

15.1 It cannot always be taken for granted that there exists a regular functional relationship between r (or R) and m. In particular where material heterogeneity forms an inseparable part of the variability of the test results (see 4.2.4 and 4.2.6), there will only be a functional relationship if this heterogeneity is a regular function of m. With solid materials of different composition and coming from different production processes, a regular functional relationship is in no way certain. This point should be decided before the following procedure is applied. Alternatively, separate values of r and R would have to be established for each material investigated.

15.2 The reasoning and computational procedures presented in 15.3 to 15.9 apply to both r and R, but they are presented here for r only for the sake of brevity. Only three types of relationship will be considered :

a) equation I (a straight line through the origin) :

$$r = bm$$

b) equation II (a straight line with a positive intercept) :

$$r = a + bm$$

c) equation III (an exponential relationship) :

$$\log r = c + d \log m \text{ (or } r = Cm^d) : d < 1$$

It is to be expected that in the majority of cases at least one of these formulae will give a satisfactory fit. If not, the statistical expert carrying out the analysis should seek an alternative solution. To avoid confusion, the constants a, b, c, C and d, occurring in these equations may be distinguished by suffixes a_r, b_r, ... for r, and a_R, b_R, ... for R, but these have been omitted in this clause to simplify the notations.

15.3 In general, $d > 0$ so that equations I and III will lead to $r = 0$ for $m = 0$, which may seem unacceptable from an experimental point of view. However, when reporting the precision data, it should be made clear that they apply only within the levels covered by the inter-laboratory precision experiment.

15.4 For $a = 0$ and $d = 1$, all three equations are identical, so when a lies near zero and/or d lies near unity, two or all three of these equations will yield practically equivalent fits. In such a case equation I should be preferred because it permits the simple statement :

> "Two single test results are considered as suspect when they differ by more than $(100b)$ %."

In statistical terminology, this is a statement that the coefficient of variation, $(100s/m)$, is a constant for all levels.

15.5 If in a plot of r_j against m_j, or $\log r_j$ against $\log m_j$, the set of points are found to lie reasonably close to a straight line, a line drawn by hand may provide a satisfactory solution; but if, for some reason, a numerical method of fitting is preferred, the procedure of 15.6 is recommended for equations I and II, and that of 15.8 for equation III.

15.6 From a statistical viewpoint, the fitting of a straight line is complicated by the fact that both m and r are estimates and thus subject to error. However, as the slope b is usually small (of the order of 0,1 or less), then errors in m have little influence and the errors in estimating r predominate.

15.6.1 A good estimate of the parameters of the regression line requires a weighted regression because, statistically, the standard error of r is proportional to the predicted value of $r(\hat{r})$.

The weights have to be proportional to $1/\hat{r}_j^2$, where $\hat{r}_j$ is the predicted repeatability for level j. However, $\hat{r}_j$ depends on the parameters that have yet to be calculated.

A mathematically correct procedure for finding estimates corresponding to the weighted least squares of residuals of r is rather complicated; the following procedure, which has proved to be satisfactory in practical applications, is recommended.

15.6.2 With weights W_j equal to $1/r_{nj}^2$, where $n = 0, 1, 2, \ldots$, for successive iterations, then the calculated formulae are

$$T_1 = \sum_j W_j$$

$$T_2 = \sum_j m_j W_j$$

$$T_3 = \sum_j W_j m_j^2$$

$$T_4 = \sum_j W_j r_j$$

$$T_5 = \sum_j W_j m_j r_j$$

Then for equation I ($r = bm$), the value of b is given by T_5/T_3.

For equation II ($r = a + bm$), the values of a and b are given by

$$a = \frac{T_3T_4 - T_2T_5}{T_1T_3 - T_2^2}$$

and

$$b = \frac{T_1T_5 - T_2T_4}{T_1T_3 - T_2^2}$$

15.6.3 For equation I, algebraic substitution for the weights leads to the following simplified expression :

$$b = \frac{\sum_j (r_j/m_j)}{q}$$

15.6.4 For equation II, the initial values r_{0j} are the original values of r as obtained by one of the procedures laid down in clause 14. These are used to calculate

$$W_{0j} = \frac{1}{r_{0j}^2} \quad (j = 1, 2, \ldots, q)$$

and to calculate a_1 and b_1 as in 15.6.2. This leads to

$$r_1 = a_1 + b_1 m_1$$

or

$$r_{1j} = a_1 + b_1 m_j$$

The computations are then repeated with

$$W_{1j} = \frac{1}{r_{1j}^2}$$

to produce

$$r_2 = a_2 + b_2 m$$

The same procedure could now be repeated once again with weights $W_{2j} = 1/r_{2j}^2$, derived from these equations, but this will only lead to unimportant changes. The step from W_{0j} to W_{1j} is effective in eliminating gross errors in the weights, and the equations for r_2 should be considered as the final result.

15.7 The standard error of log$r$ is approximately proportional to $V(r)$, the coefficient of variation of r. As the standard error of r is proportional to the predicted value of $r(\hat{r})$, the standard error of log$r$ will be independent of r and an unweighted regression of log$r$ on log m is appropriate for equation III.

15.8 For equation III, the computational formulae are :

$$T_1 = \sum_j \log m_j$$

$$T_2 = \sum_j (\log m_j)^2$$

$$T_3 = \sum_j \log r_j$$

$$T_4 = \sum_j (\log m_j)(\log r_j)$$

and thus

$$c = \frac{T_2 T_3 - T_1 T_4}{q T_2 - T_1^2}$$

and

$$d = \frac{q T_4 - T_1 T_3}{q T_2 - T_1^2}$$

15.9 Examples of fitting equations I, II, and III of 15.2 to the same set of data are given in 15.9.1 to 15.9.3. The data are taken from the case study in clause 24 and have been used here only to illustrate the numerical procedure. They will be further discussed in clause 24.

15.9.1 Example of fitting equation I : $r = bm$

m_j	3,94	8,28	14,18	15,59	20,41
r_j	0,258	0,501	0,355	0,943	1,102
r_j/m_j	0,065 5	0,060 5	0,025 0	0,060 5	0,054 0
$b = \frac{\Sigma (r_j/m_j)}{q}$	$\frac{0{,}265\ 5}{5} = 0{,}053\ 1$				
r	$0{,}053\ 1\ m$				

15.9.2 Example of fitting equation II : $r = a + bm$

m_j and r_j	Values as in 15.9.1				
W_{0j}	15	4,0	7,9	1,1	0,82
	$r_1 = 0{,}161 + 0{,}025\ 1\ m$				
r_{1j}	0,260	0,369	0,517	0,552	0,673
W_{1j}	15	7,3	3,7	3,3	2,2
	$r_2 = 0{,}085 + 0{,}043\ 6\ m$				
r_{2j}	0,257	0,446	0,703	0,765	0,975
W_{2j}	15	5,0	2,0	1,7	1,0
	$r_3 = 0{,}090 + 0{,}043\ 0\ m$				
r_{3j}	0,261	0,449	0,704	0,765	0,974
	(The difference from r_2 is negligible)				

NOTE — The values of the weights are not of critical importance. Two significant figures suffice.

15.9.3 Example of fitting equation III : log$r$ = $c + d$ log$m$

logm_j	+ 0,595	+ 0,918	+ 1,152	+ 1,193	+ 1,310
logr_{0j}	− 0,588	− 0,300	− 0,450	− 0,025	+ 0,042
	log$r$ = − 1,057 9 + 0,767 9 log$m$ or $r = 0{,}088\ m^{0,77}$, which yields				
r_j	0,253	0,448	0,678	0,729	0,898

16 Statistical analysis as a step-by-step procedure

NOTE — Figure 2 illustrates, in a step-by-step sequence, the procedure described in this clause.

16.1 Collect all available test results in one table — table A of figure 1 (see 11.5 and 11.5.1). It is recommended that this table be arranged into p rows, indexed $i = 1, 2, \ldots, p$, representing the p laboratories that have contributed data, and q columns, indexed $j = 1, 2, \ldots, q$, representing the q levels in increasing order.

In a uniform-level experiment, the test results within a cell of table A need not be distinguished and may be put in any desired order. However, in a split-level experiment, it shall be clearly stated which of the two test results belongs to sub-level a and which to sub-level b, and the results shall be entered in that specific order. (See clause 6.)

16.2 Inspect table A for any obvious irregularities; investigate and, if necessary, discard any obviously erroneous data and report to the panel. It is sometimes immediately evident that the test results of a particular laboratory or in a particular cell lie at a level inconsistent with the other data. Such obviously discordant data should be discarded straight away, but the fact shall be reported to the panel for further consideration. (See 17.1.)

16.3 From table A, corrected in accordance with 16.2 when needed, compute table B containing measures of within-cell spread, and table C containing the cell averages. (See 11.5.1, 11.5.2 and 11.5.3.)

When a cell in table A for a uniform-level experiment contains only a single test result, one of the options of 14.3 should be adopted. A single test result in a cell for a split-level experiment shall be discarded.

16.4 Inspect tables B and C, level by level, for possible stragglers and/or statistical outliers [see 11.2.3 a)]. Apply the statistical tests of clauses 12 and 13 to all suspect items and mark the stragglers with a single asterisk and the statistical outliers with a double asterisk. If there are no stragglers or statistical outliers, ignore 16.5 to 16.9 and proceed directly with 16.10.

16.5 Investigate whether there is, or may be, some technical explanation for the stragglers and/or statistical outliers and, if possible, verify such explanations. Correct or discard, as required, those stragglers and/or statistical outliers that have been satisfactorily explained, and apply corresponding corrections to the tables. If there are no stragglers or statistical outliers left that have not been explained, ignore 16.6 to 16.9 and proceed directly with 16.10.

NOTE — A large number of stragglers and/or statistical outliers may indicate a pronounced variance in homogeneity or pronounced differences between laboratories, and thereby cast doubt on the suitability of the test method; this should be reported to the panel.

16.6 If the distribution of the unexplained stragglers or statistical outliers in tables B or C does not suggest any outlying laboratories [see 11.2.3 d)], ignore 16.7 and proceed directly with 16.8.

16.7 If the evidence against some suspected outlying laboratories is considered strong enough to justify the rejection of some or all data from these laboratories, discard the requisite data and report to the panel.

The decision to reject some or all data from a particular laboratory is the responsibility of the statistical expert carrying out the analysis, but shall be reported to the panel for further consideration. (See 17.1.)

16.8 If any stragglers and /or statistical outliers remain that have not been explained or attributed to an outlying laboratory, discard the statistical outliers but retain the stragglers.

16.9 If, in the previous steps, any entry in table B has been rejected, the corresponding entry in table C shall also be rejected, and *vice versa*.

16.10 From the entries that have been retained as correct in tables B and C, compute, by the procedures given in clause 14, for each level separately, the mean level m_j, the repeatability r_j and the reproducibility R_j.

16.11 If the experiment only used a single level, or if it has been decided that the repeatability and reproducibility should be given separately for each level (see 15.1) and not as functions of the level, ignore 16.12 to 16.17 and proceed directly with 16.18.

NOTE — The following steps 16.12 to 16.16 are applied to r and R separately, but, for the sake of brevity, they are written out in terms of r only.

16.12 Plot r_j against m_j and judge from this plot whether r may depend on m or not.

If r is judged to depend on m, ignore 16.13 and proceed directly with 16.14.

If r is judged to be independent of m, proceed with 16.13.

If in doubt, it is best to work out both cases and let the panel decide.

There exists no useful statistical test appropriate for this problem, but technical experts familiar with the test method should have sufficient experience to take a decision.

16.13 Use the average $\frac{1}{q}\sum_j r_j = r$ as the final value of the repeatability. Ignore 16.14 to 16.17 and proceed directly with 16.18.

16.14 Judge from the plot of 16.12 whether the relationship between r and m can be represented by a straight line, and, if so, whether equation I ($r = bm$) or equation II ($r = a + bm$) is more appropriate (see 15.2). Determine the parameter b, or the two parameters a and b, by the procedure of 15.6. If the linear relationship is considered satisfactory, ignore 16.15 and proceed directly with 16.16. If not, proceed with 16.15.

16.15 Plot $\log r_j$ against $\log m_j$ and judge from this plot whether the relationship between $\log r$ and $\log m$ can reasonably be represented by a straight line. If this is considered satisfactory, fit equation III ($\log r = c + d \log m$) (see 15.2), using the procedure of 15.8.

16.16 If a satisfactory relation has been established according to 16.14 or 16.15, then the final values of r (or R) are the smoothed values obtained from this relationship for given values of m. Ignore 16.17 and proceed with 16.18.

16.17 If no satisfactory relationship has been established according to 16.14 or 16.15, the statistical expert should decide whether some other relationship between r and m can be established or, alternatively, that whether the data are so irregular that the establishment of a functional relationship is considered as impossible.

16.18 When the final values of r and R have been established, it is possible to verify that they correspond to a 95 % probability, as required by the definitions of 3.1, by means of the data from which they have been computed. How this can be done is illustrated in the case studies in clauses 22 to 24.

17 Reporting to, and decisions to be taken by, the panel

17.1 Report by the statistical expert

Having completed the statistical analysis, the statistical expert should write a report to be submitted to the panel. In this report, the following information should be given :

a) a full account of the observations received from the operators and/or supervisors concerning the standard for the test method [see 10.6 c)];

b) a full account of the laboratories that have been rejected as outlying laboratories in steps 16.2 or 16.7, together with the reasons for their rejection;

c) a full account of the stragglers and/or statistical outliers that were discovered, and whether these were explained and corrected, or discarded;

d) a table of the final results m_j, r_j, and R_j, and an account of the conclusion reached in steps 16.12, 16.14 or 16.15, illustrated by one of the plots recommended in these steps;

e) tables A, B and C (see 11.5) used in the statistical analysis, possibly as an appendix.

17.2 Decisions taken by the panel

The panel should then discuss this report and take a decision concerning the following questions :

a) Are the discordant test results of rejected outlying laboratories, if any, due to defects in the description of the standard for the test method?

b) What action should be taken with respect to rejected outlying laboratories? (See 17.3.)

c) Do the results of the outlying laboratories and/or the comments received from the operators and supervisors indicate the need to improve the standard for the test method? If so, what are the improvements required?

d) Do the results of the precision experiment justify the establishment of final values of the repeatability and the reproducibility? If so, what are the final values for repeatability and reproducibility, in what form shall they be published, and what is the region in which the precision data apply?

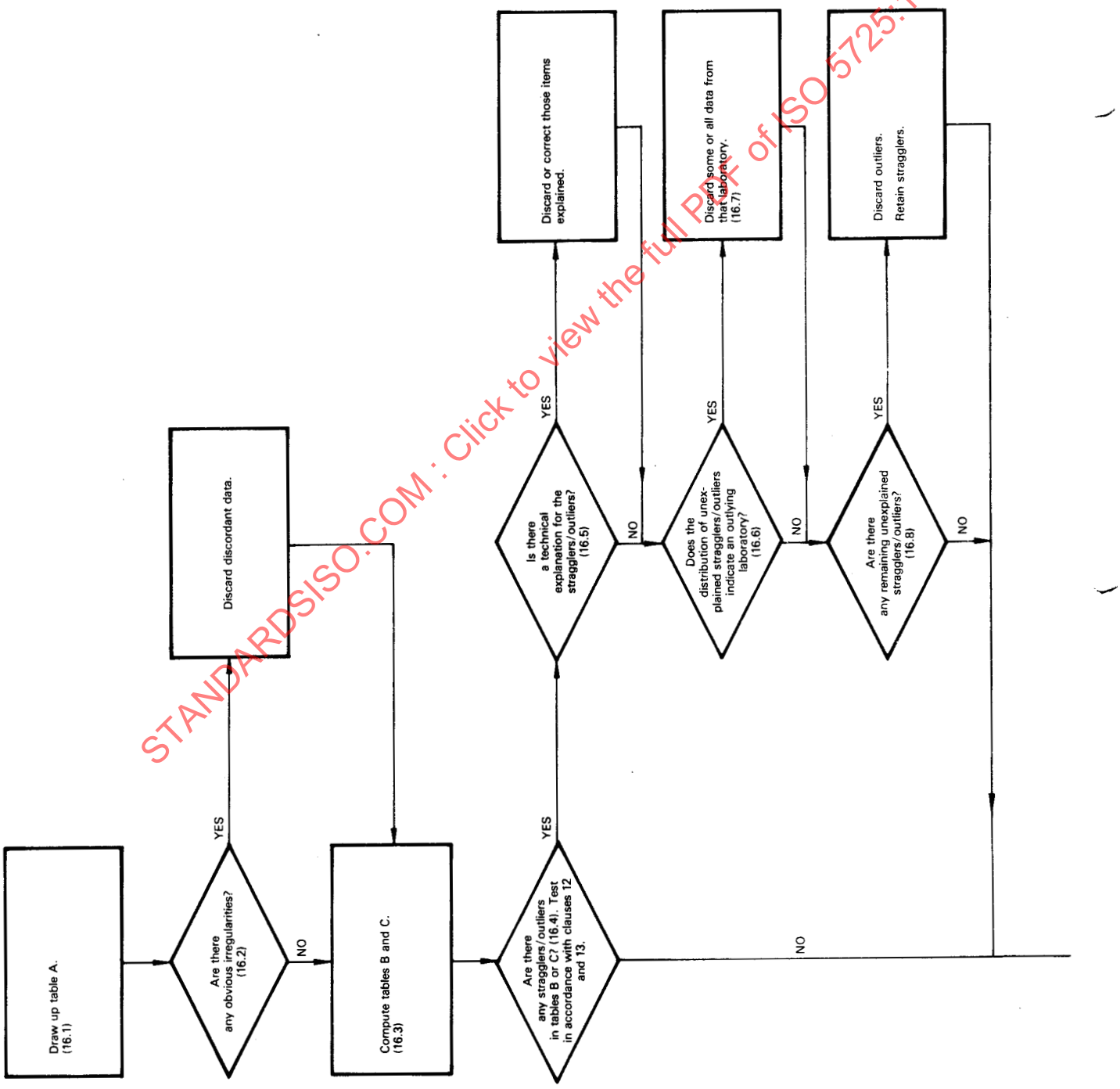
17.3 Outlying laboratories

If the method has been accepted as being satisfactory,

a) all laboratories rejected as outliers shall be informed of the fact and of the reasons for their rejection;

b) a laboratory rejected on the basis of stragglers and/or statistical outliers in table B will show too high a repeatability variance, which may be due to poor technique or lack of experience of the operator. These laboratories should be encouraged to improve their method, using the established value of the repeatability as a guide (see section four).

c) a laboratory rejected on the basis of stragglers and/or statistical outliers among the cell averages in table C may be misreading the standard, or using some instrument with a serious systematic error in its readings. This requires further investigation; the panel should discuss how this can be organized and take corresponding action.



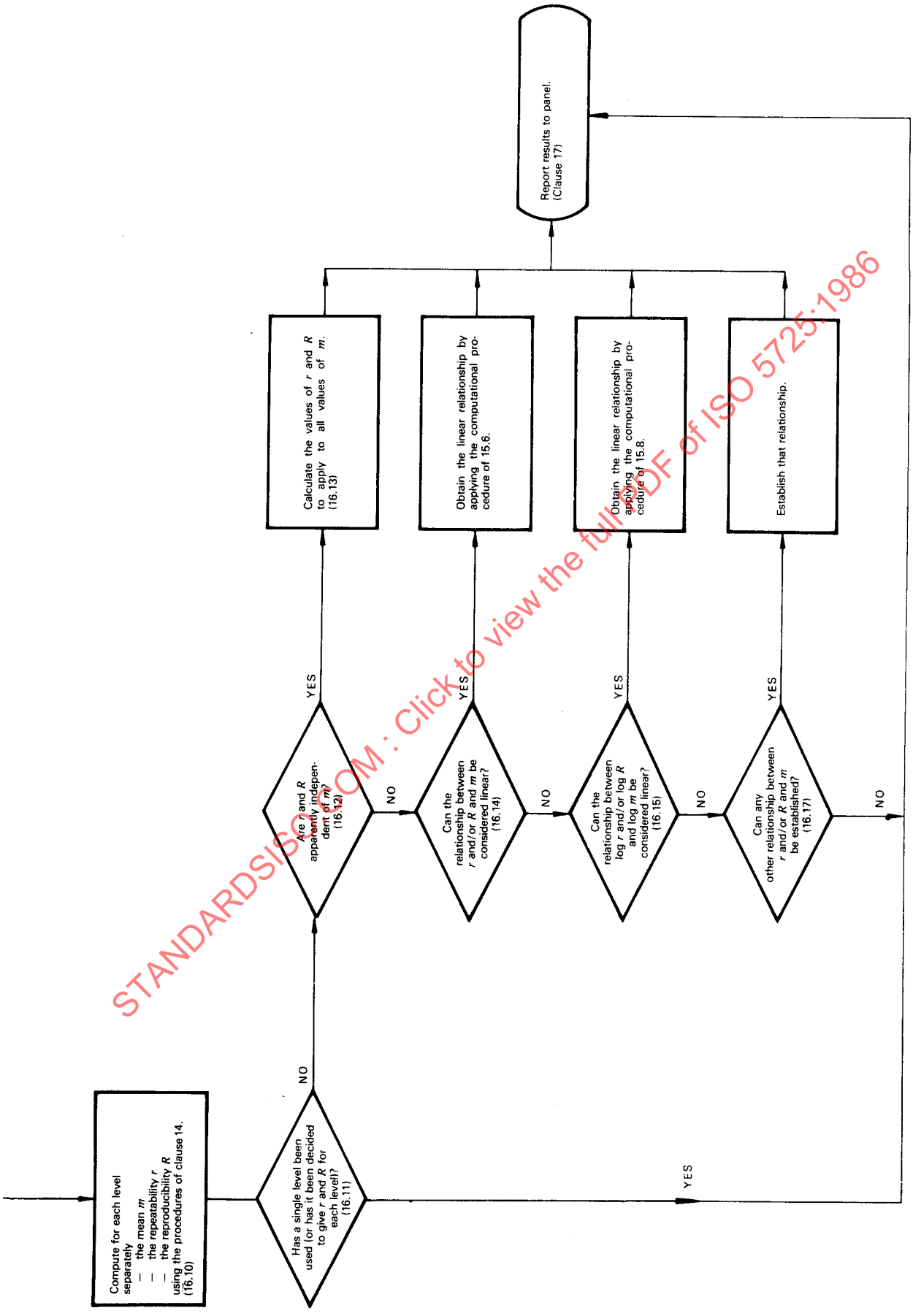


Figure 2 — Flow diagram of the principal steps in the statistical analysis (see clause 16)

Section four : Utilization of precision data

18 Publication of repeatability and reproducibility values

18.1 When a standard test method, for which precision data have been determined, is published, such data shall be included in a section of the method headed "Precision". This section is as much an integral part of the method as other sections on apparatus, reagents, etc.

18.2 The repeatability and reproducibility values should normally be published as a table of five columns giving, respectively, the range of test results (or a typical result), the repeatability for that range (or level), both as a standard deviation and as a critical difference, and the reproducibility for that range (or level) again both as a standard deviation and as a critical difference :

Range or level	Repeatability conditions		Reproducibility conditions	
	s_r	r	s_R	R
From . . . . . to				
From . . . . . to				
etc.				
From . . . . . to				

18.3 A statement should be added linking the precision to the difference between two results and to the 95 % probability level. Suggested wordings are as follows :

"The difference between two single results found on identical test material by one operator using the same apparatus within the shortest feasible time interval will exceed the repeatability value r on average not more than once in 20 cases in the normal and correct operation of the method."

"Single results on identical test material reported by two laboratories will differ by more than the reproducibility value R on average not more than once in 20 cases in the normal and correct operation of the method."

18.4 A statement can optionally be added that both results should be considered suspect if the repeatability or reproducibility value, as appropriate, is exceeded. Statements regarding subsequent actions, e.g. repetition of the test, may also be included in the section on precision.

18.5 In general, a brief mention of the precision experiment should be added to the precision section, possibly as a footnote. A suggested wording is as follows :

"The precision data were determined from an experiment conducted in (year) involving (p) laboratories and (q) samples."

19 Other critical differences derivable from r and R

19.1 The critical differences, as stated in 3.3.2, are for 95 % probability levels. It is possible, however, to derive the critical differences for other probability levels.

19.1.1 Critical difference for probability levels other than 95 %

These can be obtained by multiplying the critical differences for a level of 95 % by the multiplying factors given in table 1. These multiplying factors are only valid when the distribution of the components B and e in the model of 5.1 are normal or approximately normal.

Table 1 — Multiplying factors for finding critical differences for probability levels other than 95 %

Probability level, P %	Multiplying factor
90	0,82
95	1,00
98	1,16
99	1,29
99,5	1,40

19.2 As stated in 3.1.9, 3.1.14 and 18.3, the uses of r and R are limited to the cases of two single test results obtained under either repeatability or reproducibility conditions. It is possible, however, to derive from r and R critical differences (via their variance components) for cases other than two single test results (see 19.2.1 to 19.2.4). The examples of critical differences are given for the 95 % probability level; for other probability levels, the factors in table 1 can be used.

19.2.1 More than two single determinations carried out in one laboratory

If, in one laboratory under repeatability conditions, two groups of tests are performed, with the first group on n_1 tests giving a mean value of $\bar{y}_1$ and the second group of n_2 tests giving a mean value of $\bar{y}_2$, then with s_r still being the basic standard deviation :

$$\mathrm{CrD}_{95}\left(\left|\bar{y}_1 - \bar{y}_2\right|\right) = r\sqrt{\frac{1}{2n_1} + \frac{1}{2n_2}} \qquad \ldots (18)$$

NOTE — If n_1 and n_2 are both unity, this reduces to r, as expected.

19.2.2 Two laboratories each doing more than one determination

If the first laboratory performs n_1 determinations giving a mean value $\bar{y}_1$, while the second laboratory performs n_2 determi-

nations giving a mean value $\bar{y}_2$, then the variance of the difference $(\bar{y}_1 - \bar{y}_2)$ is given by

$$s^2_{(\bar{y}_1 - \bar{y}_2)} = s_L^2 + s_r^2\left(\frac{1}{2n_1} + \frac{1}{2n_2}\right)$$

$$= s_R^2 - s_r^2\left(1 - \frac{1}{2n_1} - \frac{1}{2n_2}\right)$$

Therefore

$$\mathrm{CrD}_{95}\left(\left|\bar{y}_1 - \bar{y}_2\right|\right) = \sqrt{R^2 - r^2\left(1 - \frac{1}{2n_1} - \frac{1}{2n_2}\right)} \quad \ldots (19)$$

In particular, if $n_1 = n_2 = 1$, this reduces to R as expected, and if $n_1 = n_2 = 2$, this produces

$$\mathrm{CrD}_{95}\left(\left|\bar{y}_1 - \bar{y}_2\right|\right) = \sqrt{R^2 - \frac{r^2}{2}}$$

19.2.3 Comparison with a reference level for one laboratory

If n determinations performed by one laboratory under repeatability conditions produce a mean value $\bar{y}$ which is to be compared with a reference value m_o, then the variance of $\bar{y} - m_o$ is given by

$$s^2_{\bar{y}} = s_L^2 + \frac{1}{n}s_r^2 = s_R^2 - s_r^2\left(\frac{n-1}{n}\right)$$

Therefore

$$\mathrm{CrD}_{95}\left(\left|\bar{y} - m_o\right|\right) = \frac{1}{\sqrt{2}}\sqrt{R^2 - r^2\left(\frac{n-1}{n}\right)} \quad \ldots (20)$$

19.2.4 Comparison with a reference level for several laboratories

If p laboratories have performed n_i determinations giving averages $\bar{y}_i$ (where $i = 1, 2, \ldots, p$), and an overall average

$$\bar{\bar{y}} = \frac{1}{p}\Sigma\,\bar{y}_i$$

to be compared with the reference value m_o, then the variance of $\bar{\bar{y}}$ is given by

$$s^2_{\bar{\bar{y}}} = \frac{1}{p}\left(s_L^2 + \frac{1}{p}\Sigma\,\frac{1}{n_i}s_r^2\right)$$

Therefore

$$\mathrm{CrD}_{95}\left(\left|\bar{\bar{y}} - m_o\right|\right) = \frac{1}{\sqrt{2p}}\sqrt{R^2 - r^2\left(1 - \frac{1}{p}\Sigma\,\frac{1}{n_i}\right)} \quad \ldots (21)$$

NOTE — When, in comparing two averages, or a single average with a reference value, the absolute difference exceeds the corresponding critical differences as given above, then the difference should be considered as suspect. There may be an assignable cause and this should be investigated. In particular, when the reference value in 19.2.3 or 19.2.4 is a "true value" or "conventional true value", a suspect difference may indicate that the test method has a bias.

20 Practical applications

There are many practical applications of repeatability and reproducibility, some of which are indicated in clause 19. Some of these applications cover such matters as conformity with specifications, the problems to be borne in mind when designing specifications, differences in results obtained by a supplier and a consumer, etc. There are also some slight modifications to be made to the procedures indicated in this International Standard to cater for such problems as the recalibration of equipment between tests, the effect on repeatability of longer time intervals between tests, the verification of the repeatability of a single laboratory, or the effect on reproducibility when the two cooperating laboratories are always the same.

Such practical applications and variations in procedure have not been discussed in detail in this International Standard. The practical applications and variations in procedures will be dealt with in future International Standards.

Section five : Examples

NOTE — Technical Committee ISO/TC 69 is grateful to the individuals and organizations that provided the practical data for these examples.

21 General information

21.1 In a report on the results of a precision experiment, full details should be given concerning the standard of the test method and the way the samples have been prepared. In most of the examples given in clauses 22 to 25 this information is missing. In the literature, data are often used to illustrate the statistical analysis, the test method by which they were obtained being considered immaterial. As the main purpose of these examples in this section is to show how the analysis by the step-by-step procedure of clause 16 works out in practice, and to illustrate the report to the panel made by the statistical expert (see 17.1), this usual procedure has been adopted here. However, some references are given so that the interested reader can find details of the standards for the test methods, if so desired.

21.2 The examples in this section show, in particular, that the application of a systematic analytical procedure by rote does not always tell the whole story. Not infrequently an attentive statistician will notice peculiarities in the data that are not covered by the tests laid down in clause 16 and this induces him to apply some further criteria or graphical presentation. As it is impractical in this International Standard to cover all possible variations, a few examples have to suffice. They demonstrate why the analysis should preferably be carried out by a statistical expert experienced in the analysis of experimental data.

21.3 It may be desirable to check the 95 % level for the repeatability and for the reproducibility on the data from which they were computed. Such checks are illustrated in the examples.

21.4 The examples have been chosen to cover most aspects of the analysis. The first three examples are of uniform-level experiments, the first giving an example with few problems as all data are complete and there are no suspect observations. The second example covers the case where some test results were missing, and the third example shows a case where the data were originally complete but some of the observations were suspect. This third example also covers the fitting of a functional relation to the results. The final example is a case of a split-level experiment.

22 Uniform-level experiment with no missing or outlying data

22.1 Background

22.1.1 Test

Determination of sulfur content in coal, with results expressed as a percentage by mass. Analysis carried out in accordance with a standardized method described in the source cited (see 22.1.2).

22.1.2 Source

TOMKINS, S.S. *Industrial and engineering chemistry*, Analytical edition, 1942, **14**, pp. 141-145.[3]

22.1.3 Description

Eight laboratories participated in the experiment. Laboratory 1 reported four test results and laboratory 5 reported five or four; the other laboratories all carried out three tests.

22.2 Original data

The original data are given, as a percentage by mass [% (*m*/*m*)], in table 2, in the format of table A in figure 1 (see 11.5) and do not invite any specific remarks.

Table 2 — Original data

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

22.3 Computation of the standard deviations, s_{ij}

The standard deviations are given, as a percentage by mass [% (m/m)], in table 3 in the format of table B in figure 1 (see 11.5).

Table 3 — Standard deviations

Laboratory i \ Level j	1		2		3		4	
	s_{ij}	n_{ij}	s_{ij}	n_{ij}	s_{ij}	n_{ij}	s_{ij}	n_{ij}
1	0,005	4	0,021	4	0,010	4	0,028	4
2	0,010	3	0,006	3	0,006	3	0,000	3
3	0,021	3	0,015	3	0,006	3	0,010	3
4	0,010	3	0,025	3	0,012	3	0,038	3
5	0,019	5	0,043	4	0,032	5	0,038	5
6	0,006	3	0,015	3	0,017	3	0,020	3
7	0,012	3	0,035	3	0,010	3	0,021	3
8	0,025	3	0,042	3	0,006	3	0,006	3

Cochran's test is applied with $n = 3$ when, for $p = 8$ laboratories, the critical values are for 5 % = 0,516 and for 1% = 0,615 :

For level 1, largest s is in laboratory 8 : $\Sigma s^2 = 0{,}001\,82$: test value = 0,347;

For level 2, largest s is in laboratory 5 : $\Sigma s^2 = 0{,}006\,36$: test value = 0,287;

For level 3, largest s is in laboratory 5 : $\Sigma s^2 = 0{,}001\,72$: test value = 0,598;

For level 4, largest s is in laboratory 4 : $\Sigma s^2 = 0{,}004\,63$: test value = 0,310.

This indicates that one cell in level 3 may be regarded as a straggler, and there are no outliers. The straggler is retained in subsequent calculations.

22.4 Computation of the cell averages, $\bar{y}_{ij}$

The cell averages are given, as a percentage by mass [% (m/m)], in table 4 in the format of table C in figure 1 (see 11.5).

Table 4 — Cell averages

Laboratory i \ Level j	1		2		3		4	
	$\bar{y}_{ij}$	n_{ij}	$\bar{y}_{ij}$	n_{ij}	$\bar{y}_{ij}$	n_{ij}	$\bar{y}_{ij}$	n_{ij}
1	0,708	4	1,205	4	1,688	4	3,240	4
2	0,680	3	1,217	3	1,643	3	3,200	3
3	0,667	3	1,297	3	1,613	3	3,370	3
4	0,660	3	1,203	3	1,667	3	3,203	3
5	0,690	5	1,248	4	1,650	5	3,216	5
6	0,733	3	1,373	3	1,720	3	3,290	3
7	0,703	3	1,240	3	1,690	3	3,247	3
8	0,677	3	1,253	3	1,673	3	3,257	3

Dixon's test is applied with $H = 8$ for which the critical values are for Q_{11} at 5 % = 0,608 and at 1 % = 0,717. The values of Q_{11} found are as follows :

At level 1, $Q_{11} = \dfrac{0,733 - 0,708}{0,733 - 0,667} = \dfrac{0,025}{0,066} = 0,379$

At level 2, $Q_{11} = \dfrac{1,373 - 1,297}{1,373 - 1,205} = \dfrac{0,076}{0,168} = 0,452$

At level 3, $Q_{11} = \dfrac{1,720 - 1,690}{1,720 - 1,643} = \dfrac{0,030}{0,077} = 0,390$

At level 4, $Q_{11} = \dfrac{3,370 - 3,290}{3,370 - 3,203} = \dfrac{0,080}{0,167} = 0,479$

and there are no stragglers or outliers.

22.5 Computation of m_j, r_j and R_j

The computed values for m_j, r_j and R_j are given, as a percentage by mass [% (m/m)], in table 5 and are calculated as described in 14.9.

Table 5 — Computed values of m_j, r_j and R_j

Level j	p_j	m_j	s_r^2	r_j	s_R^2	R_j
1	8	0,690	0,000 230	0,042	0,000 698	0,074
2	8	1,252	0,000 828	0,081	0,003 673	0,171
3	8	1,667	0,000 291	0,048	0,001 210	0,097
4	8	3,250	0,000 679	0,073	0,003 389	0,163

22.5.1 Dependence of r (or R) on m

Plots of the values of r and R, given in 22.5, against m do not indicate any dependence and the average values can be adopted.

22.5.2 Final values of r and R

Rounded to three decimal places, the final values are

$r = 0{,}061$ % (m/m)

$R = 0{,}126$ % (m/m)

22.6 Check on the values of r and R

22.6.1 Repeatability

As we have more than two tests per cell, the verification procedure is slightly more complicated than if we had exactly two, the normal criterion for repeatability. With n tests in a cell, we can derive from them $n(n - 1)/2$ differences between two single test results. Thus in eight laboratories we can get

$6 + 3 + 3 + 3 + 10 + 3 + 3 + 3 = 34$ in level 1,

and likewise

30, 34 and 34 in levels 2, 3 and 4, respectively.

Of these 132 differences, eight differences or 6,1 % lie above the repeatability and 93,9 % below, which is compatible with a 95 % probability level.

22.6.2 Reproducibility

The averages in 22.4 are based on three, four or five test results, which is again slightly more complicated than if we had just two. Therefore, the formula given in 19.2.2 shall be used, but we find that there is very little difference for different values of n, as, for example,

for $n_1 = n_2 = 3$; $CrD_{95} = 0{,}115$

for $n_1 = n_2 = 5$; $CrD_{95} = 0{,}113$

If we apply an average criterion of 0,114 to all comparisons between cell averages, at each level we can form 28 differences, that is 112 in total. Of these, nine (0 + 6 + 0 + 3) or 8,0 % lie above the critical difference, which is quite an acceptable result.

22.7 Conclusions

The precision of the test method, expressed as a percentage by mass [% (m/m)], is given as

Repeatability

Standard deviation $s_r = 0{,}022$

Value of $r = 0{,}061$

Reproducibility

Standard deviation $s_R = 0{,}045$

Value of $R = 0{,}126$

These values may be applied within a range from 0,69 to 3,25 % (m/m), being determined from a uniform-level experiment involving eight laboratories covering that range of values, in which only one straggler was detected and retained.

23 Uniform-level experiment with missing data

23.1 Background

23.1.1 Test

Determination of the softening point of pitch by ring and ball, involving temperature measurement in degrees Celsius.

23.1.2 Source

Standard methods for testing tar and its products. 7th edition, 1979. Pitch section, Method Serial No. PT3 using neutral glycerine.[4]

23.1.3 Material

This was selected from commercial batches of pitch collected and prepared as specified in the "Samples" chapter of the Pitch section of the publication referred to in 23.1.2.

23.1.4 Description

16 laboratories participated; it was intended to test four specimens at approximately 87,5 °C, 92,5 °C, 97,5 °C and 102,5 °C to cover the normal commercial range of products, but wrong material was chosen for level 2 with a mean temperature of about 96 °C which was similar to level 3. Laboratory 5 applied the method incorrectly at first on the sample for level 2 (the first one they tested) and there was then insufficient material remaining for more than one determination. Laboratory 8 found that they did not have a sample for level 1 (they had two specimens for level 4).

23.2 Original data

The original data are given, in degrees Celsius (°C), in table 6 in the format of table A in figure 1 (see 11.5).

Table 6 — Original data

Laboratory i \ Level j	1		2		3		4	
1	91,0	89,6	97,0	97,2	96,5	97,0	104,0	104,0
2	89,7	89,8	98,5	97,2	97,2	97,0	102,6	103,6
3	88,0	87,5	97,8	94,5	94,2	95,8	103,0	99,5
4	89,2	88,5	96,8	97,5	96,0	98,0	102,5	103,5
5	89,0	90,0	97,2	—	98,2	98,5	101,0	100,2
6	88,5	90,5	97,8	97,2	99,5	103,2	102,2	102,0
7	88,9	88,2	96,6	97,5	98,2	99,0	102,8	102,2
8	—	—	96,0	97,5	98,4	97,4	102,6	103,9
9	90,1	88,4	95,5	96,8	98,2	96,7	102,8	102,0
10	86,0	85,8	95,2	95,0	94,8	93,0	99,8	100,8
11	87,6	84,4	93,2	93,4	93,6	93,9	98,2	97,8
12	88,2	87,4	95,8	95,4	95,8	95,4	101,7	101,2
13	91,0	90,4	98,2	99,5	98,0	97,0	104,5	105,6
14	87,5	87,8	97,0	95,5	97,1	96,6	105,2	101,8
15	87,5	87,6	95,0	95,2	97,8	99,2	101,5	100,9
16	88,8	85,0	95,0	93,2	97,2	97,8	99,5	99,8

There are no obvious stragglers or statistical outliers, and no statistical tests are required at this stage.

23.3 Cell ranges

In this example there are two results per cell and the ranges can be used to represent the variability. The cell ranges are given, in degrees Celsius (°C), in table 7 in the format of table B in figure 1 (see 11.5).

Table 7 — Cell ranges

Laboratory i \ Level j	1	2	3	4
1	1,4	0,2	0,5	0,0
2	0,1	1,3	0,2	1,0
3	0,5	3,3	1,6	3,5
4	0,7	0,7	2,0	1,0
5	1,0	—	0,3	0,8
6	2,0	0,6	3,7	0,2
7	0,7	0,9	0,8	0,6
8	—	1,5	1,0	1,3
9	1,7	1,3	1,5	0,8
10	0,2	0,2	1,8	1,0
11	3,2	0,2	0,3	0,4
12	0,8	0,4	0,4	0,5
13	0,6	1,3	1,0	1,1
14	0,3	1,5	0,5	3,4
15	0,1	0,2	1,4	0,6
16	3,8	1,8	0,6	0,3

Application of Cochran's test leads to the values of the test statistic C given in table 8.

Table 8 — Values of Cochran's test statistic

Cochran's test statistic \ Level j	1	2	3	4
C	0,391 (15)	0,424 (15)	0,434 (16)	0,380 (16)

From annex A, the critical values at the 5 % probability level are given as 0,471 for $p = 15$ and 0,452 for $p = 16$ where $n = 2$. No stragglers are indicated.

23.4 Cell averages

The cell averages are given, in degrees Celsius (°C), in table 9 in the format of table C in figure 1 (see 11.5).

Table 9 — Cell averages

Laboratory i \ Level j	1	2	3	4
1	90,30	97,10	96,75	104,00
2	89,75	97,85	97,10	103,10
3	87,75	96,15	95,00	101,25
4	88,85	97,15	97,00	103,00
5	89,50	—	98,35	100,60
6	89,50	97,50	101,35	102,10
7	88,55	97,05	98,60	102,50
8	—	96,75	97,90	103,25
9	89,25	96,15	97,45	102,40
10	85,90	95,10	93,90	100,30
11	86,00	93,30	93,75	98,00
12	87,80	95,60	95,60	101,45
13	90,70	98,85	97,50	105,05
14	87,65	96,25	96,85	103,50
15	87,55	95,10	98,50	101,20
16	86,90	94,10	97,50	99,65

NOTE — The entry for $i = 5$, $j = 2$ has been dropped (see 14.3).

From table 9, taking $j = 3$ as an example, Dixon's test statistic is

$$Q_{22} = \text{the higher of } \frac{95,00 - 93,75}{98,50 - 93,75} = 0,263 \text{ and } \frac{101,35 - 98,50}{101,35 - 95,00} = 0,449$$

The critical value at 5 % for p (or H) = 16 is 0,546. No straggler is indicated. Similar calculations for the other three levels show values of Q_{22} of 0,260, 0,429 and 0,473, which are also not significant at the 5 % level.

23.5 Computation of m_j, r_j and R_j

The computed values for m_j, r_j and R_j are given, in degrees Celsius (°C), in table 10 and calculated as described in 14.7.

Table 10 — Computed values of m_j, r_j and R_j

Level j	p_j	m_j	s_r^2	r_j	s_R^2	R_j
1	15	88,40	1,230 3	3,11	2,787 8	4,68
2	15	96,27	0,856 0	2,59	2,550 4	4,47
3	16	97,07	0,986 9	2,78	4,041 4	5,63
4	16	101,96	1,007 8	2,81	3,667 0	5,37

23.5.1 Dependence of r (or R) on m

A plot of r or R against m shown in figure 3 does not reveal any marked dependence. The changes over the range of values of m, if any at all, are too small to be considered significant. Moreover, in view of the small range of values of m and the nature of the measurements, a dependence on m is hardly to be expected. It seems safe to conclude that r and R do not depend on m in this range, which was stated as covering normal commercial material, so that the averages may be taken as the final values for repeatability and reproducibility.

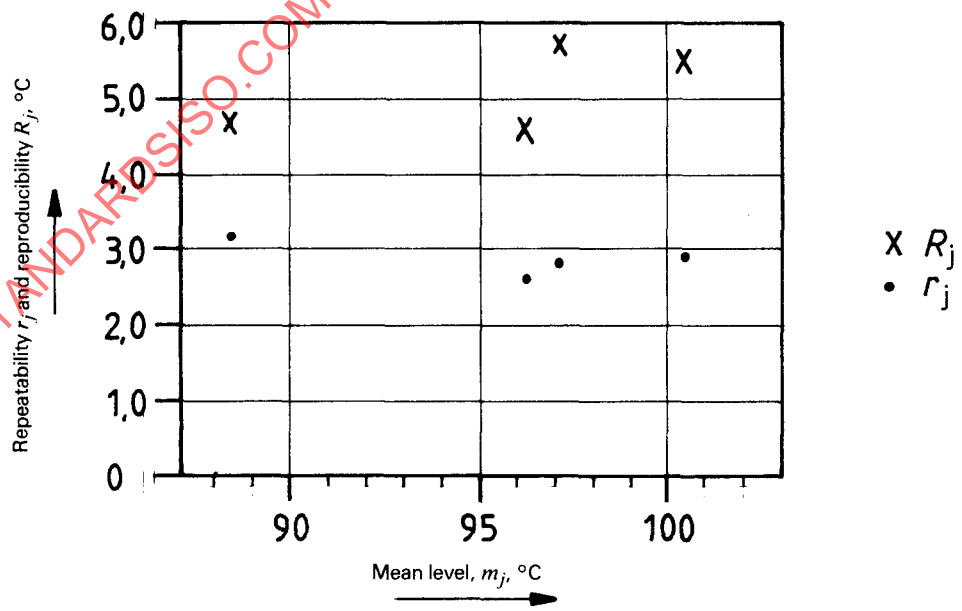


Figure 3 — Plot of r_j and R_j against m_j

23.5.2 Final values of r and R

Rounded to 0,1 °C, the final values are

$r = 2{,}8$ °C

$R = 5{,}0$ °C

23.6 Check on the values of r and R

23.6.1 Repeatability

Table 7 contains 62 absolute differences between single test results to which the repeatability should apply. Of these about 5 % should be greater than the repeatability 2,8 °C. In fact, six of the ranges or almost 10 % are greater than 2,8 °C. This could be explained as accidental and within practical limits, but a closer inspection of table 7 would be useful before drawing any final conclusions. A tally of the ranges is given in table 11.

Table 11 — Tally of ranges

Interval in °C	0,0 to 0,9	1,0 to 1,9	2,0 to 2,9	3,0 to 3,9
Tally frequency	34	20	2	6

There are only two ranges in the interval 2,0 to 2,9, both lying at the lower limit of 2,0, but there are six in the interval 3,0 to 3,9 evenly spread through the interval and scattered through table 7, not occurring particularly in one laboratory or one level.

If the errors within laboratories possessed approximately normal distributions with a common variance, then the ranges in table 7 would have approximately the distribution of the absolute values of a normal variate with mean zero. The tally given in table 11 seems to contradict this.

If all ranges in table 7 higher than 3,0 °C were eliminated, the repeatability would be reduced from 2,8 °C to 1,9 °C, which indicates the degree of improvement that might possibly be achieved. If a high precision of the test were of primary importance, the point might warrant further investigation.

23.6.2 Reproducibility

The 95 % critical value for the difference between two averages of duplicate tests calculated from the formula in 19.2.2 is 4,6 °C. From table 9 in 23.4, table 12 giving the absolute difference between the possible pairs can be drawn up.

Table 12 — Number of pairs and number of differences exceeding 4,6 °C

Level	Number of pairs	Number of differences exceeding 4,6 °C
1	105	2
2	105	2
3	120	10
4	120	7
Total	450	21

Thus 4,7 % of the differences exceed the critical difference and this is almost exactly the 5 % according to the definition of reproducibility.

23.7 Conclusions

For practical applications, the values of r and R for the test method can be considered as independent of the level of material, and, expressed in degrees Celsius (°C), are given as