INTERNATIONAL STANDARD

ISO 19003

First edition
2006-10-01

Rubber and rubber products — Guidance on the application of statistics to physical testing

Caoutchouc et produits à base de caoutchouc — Lignes directrices pour l'application des statistiques aux essais physiques

Reference number
ISO 19003:2006(E)

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

Published in Switzerland

Contents

Page

Foreword

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

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

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

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights. ISO shall not be held responsible for identifying any or all such patent rights.

ISO 19003 was prepared by Technical Committee ISO/TC 45, *Rubber and rubber products*, Subcommittee SC 2, *Testing and analysis*.

Introduction

Statistical methods have an important role at all stages of the testing process, from the design of the experiment to the interpretation of results. Hence, those involved in testing require a basic understanding of statistical principles and knowledge of the statistical techniques which need to be applied.

There are many text books and International Standards which describe statistical methods, but it is convenient to have a guide which is a single, easy source of reference to the most commonly used methods and formulae, and which also considers their particular application to the various rubber test methods. This International Standard is therefore complementary both to the general standards on statistics and to the standards on methods of test for rubber.

The approach taken in this International Standard is that, for each subject, the text is structured into principles, methodology and applications to rubber testing. Under principles, the basic concepts of the subject are briefly outlined. Methodology considers the statistical techniques which can be applied; basic procedures and formulae are given but, as appropriate, more detailed matter is placed in annexes and, for less commonly used methods or more advanced treatment, reference is made to other publications. “Applications to rubber testing” indicates how and where the methods may be applied, and gives examples which are particular to rubber properties and tests.

Rubber and rubber products — Guidance on the application of statistics to physical testing

1 Scope

This International Standard provides guidance on the application of statistics to rubber testing. It is intended not to conflict with or replace existing International Standards covering basic statistical techniques, but rather to complement them and provide examples of those techniques applied to particular rubber testing situations.

2 References

This International Standard refers to other publications that provide information or guidance. These standards are listed in the Bibliography.

3 Terms and definitions

For the purposes of this document, the following terms and definitions apply.

NOTE These definitions, which are expressed as far as possible in non-mathematical terms, apply to the main statistical terminology used. More comprehensive and rigorous lists can be found in the various parts of ISO 3534 and in the standards dealing with specific statistical techniques indicated in the Bibliography.

3.1
population
totality of data that could (theoretically) be obtained to characterize the property of the rubber, compounding ingredient or process being measured

3.2
sample
data actually available from the population as a result of an experimental test programme having been undertaken

3.3
variability
tendency for tests performed on nominally identical test pieces to produce different test results

3.4
arithmetic mean
sum of the (population or sample) data divided by the number of values used

NOTE The “average” is the statistic most frequently used to describe a group of data. There are several kinds of average and they are often used in common parlance without specifying the type, which can be a source of confusion. Averages fall into two categories: computational and positional. The arithmetic mean is the most frequently used computational average. Others are considered in Annex B. Positional averages are the median and mode. The calculation of the arithmetic mean is given in Equations (1) and (2) in 6.2.2.2.

3.5
median
middle value (or average of the two middle values) when the data in a sample are arranged in numerically increasing value

3.6
mode
value of the property being measured which occurs with the greatest frequency

3.7
residual
difference (+ or −) between each value and the mean

NOTE The sum of the residuals must be 0.

3.8
variance
arithmetic mean of the squared residuals

3.9
standard deviation
square root of the variance

NOTE The calculation of standard deviation is given in Equations (5) and (6) in 6.2.3.2.1.

3.10
coefficient of variation
ratio of the standard deviation to the mean, generally expressed as a percentage

NOTE The calculation of coefficient of variation is given in Equation (8) in 6.2.3.4.

3.11
range
maximum value minus the minimum value

3.12
standard error
standard deviation of the estimate of the population mean

NOTE The calculation of standard error is given in Equation (7) in 6.2.3.2.3.

3.13
bias
difference between the average statistic and the true value of the parameter it is estimating, arising out of one or more systematic errors

3.14
accuracy
closeness of agreement between a test result and the accepted reference value

3.15
trueness
closeness of agreement between the average value of a large number of test results and the true or accepted reference value

NOTE It is usually expressed in terms of bias.

3.16
precision
closeness of agreement between test results

3.17
repeatability
precision obtained under conditions where 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 a short interval of time

3.18
reproducibility
precision obtained under conditions where independent test results are produced with the same method on identical test material in different laboratories with different operators using different equipment

3.19
level of significance
probability of error associated with a significance test

3.20
distribution function
function describing the probability that a random variable will take a value less than or equal to a number x

3.21
density distribution
slope of the distribution function at every value, i.e. the first derivative of the distribution function

3.22
normal distribution
symmetrical “bell-shaped” density distribution which is fully defined by its mean and standard deviation

NOTE It is also known as the Laplace Gauss or Gaussian distribution.

3.23
double exponential distribution
asymmetrical distribution, fully defined by a single “shape” parameter, which has been used to characterize the distribution of tensile strengths in rubber compounds

3.24
Weibull distribution
symmetrical distribution fully defined by three parameters and found to be useful in characterizing lifetime tests such as fatigue

3.25
degrees of freedom
number of independent differences between the readings available for an estimate of standard deviation

3.26
confidence interval
range within which a value or parameter can be expected to lie with a given probability

3.27
confidence limits
extreme values of the confidence interval

4 Symbols

$a,b,c\ldots$	Constant coefficients in a regression line
C	The coefficient of concordance in Friedman's test, or Cochran's quotient when testing variances for the presence of outliers
C_i	The ith cusum value
C_{pq}	Factors used in the derivation of regression coefficients
C_v	The coefficient of variation
$f(x)$	A property or parameter which is a function of x or a density distribution function
F	The observed value of Snedecor's F-ratio in a given case
F_{cr}	The statistically critical value for F at a given confidence level and for the given degrees of freedom for the lesser and greater mean squares
F_r	The F-value for a regression line
H_0/H_a	The null/alternative hypothesis parameter
K	Friedman's statistic for a rank correlation test
M_z	The mean square for factor z
n	The number of values in a series
$p(x)$	A probability distribution function
P_m	The plot positions for the graphical presentation of a series of values
Q	Dixon's quotient when testing values or means for outliers
r	The repeatability of a test method for a particular test or series of tests
(r)	The repeatability expressed as a percentage of the mean from a test or series of tests
R	The reproducibility of a test method for a particular test or series of tests
(R)	The reproducibility expressed as a percentage of the mean from a test or series of tests
s	The estimate of the population standard deviation from the available sample
s'	the standard deviation of a series of numbers
S	The weighted standard error for the combination of two series of values, or the rank sum for a sample in Friedman's test
S_t	The total sum of the squares of the differences between individual values and their mean
S_z	The sums of squares for factor z
t_α	Student's t-value for a given probability (or confidence level) α
U_r	The random uncertainty in a measurement
U_s	The systematic uncertainty in a measurement
v_z	The number of degrees of freedom for factor z
x	An individual numerical value, such as the tensile strength of a single test piece
x_i	A single value in a series of values, such as a tensile strength in a set of five replicate values
x_{ij}	A single value in a series of values in which two factors are present, such as the tensile strength in sets of replicates obtained at different temperatures

$\bar{x}$ The arithmetic mean of a series of numbers, x_i

Z The Z-score in hypothesis testing

α, β The probability of an event occurring

μ The population mean of a distribution

$\hat{\mu}$ The estimate of the population mean from the available sample

σ The population standard deviation of a distribution

5 Limitations of test results

5.1 Variability

5.1.1 All measurements are subject to variability. It is necessary to know the sources of variability and make a reliable estimate of its magnitude. From this information, it should then be possible to judge the reliability of the results and hence their uncertainty and significance.

5.1.2 The term population is, expressed simply, the total number of objects in a large group (see 3.1). In testing terms, a population may be, for example, the total number of possible tensile strength results which could be obtained on a particular rubber compound if every piece of the material made was tested.

5.1.3 A sample is a selected number of, for example, parts or tensile results taken from the population.

NOTE 1 To avoid confusion, sample should not be used to mean test piece.

NOTE 2 Sample can have two meanings:

a) in the physical sense, as in taking five parts from a boxful;

b) in the statistical sense, as in taking five test results.

5.1.4 If five tensile strength measurements are made from a sheet taken from a batch of rubber, an example of the results which might be obtained is shown in Table 1.

Table 1 — Tensile strength measurements from one batch of rubber

Measurement number	Tensile strength MPa
1	16,8
2	15,4
3	16,3
4	17,7
5	17,6

The sources of variability are:

a) the intrinsic variability of the sheet rubber, arising from the fact that it is not perfectly homogeneous;

b) the variability due to the testing procedure, including test piece preparation, machine accuracy and operation error.

If several sheets are tested, there is an additional source of variability due to variations in moulding.

If several batches are mixed, two more sources of variation are added:

1) that from the mixing procedure;

2) any variation in compounding ingredients.

If sheets which are nominally the same are given to a number of operators, there is variability due to the operators.

Similarly, if a number of different test apparatuses are used, variability due to the machines is introduced. Taking things further, sheets may be tested in different laboratories and between-laboratory variability introduced.

5.1.5 In practice, the magnitude of variability is minimized by carefully controlling the processing operations and the test apparatus and procedures. It is never eliminated altogether and inter-laboratory comparisons have demonstrated that for many rubber tests it can be far greater than was previously thought.

Whatever test is carried out, there is genuine variation due to the material and also variation due to uncontrolled testing errors. It is often very difficult to separate the two. For example, testing errors can arise from

a) random variations in test piece geometry due to limitations in cutting precision;

b) variations in the response of the test apparatus;

c) fluctuations in the operator's performance.

These errors may be large or small and of indeterminate direction so that eventually they tend to cancel out. More serious is systematic error or bias which is unidirectional, for example the error due to a machine being wrongly calibrated or an operator consistently misreading a scale.

5.1.6 Testing error apart, the sample of results will not be representative of the whole population if the physical sample is not representative. Differences between repeat mixes and between repeat mouldings should be expected because of some variation in the quantities and quality of ingredients used, the efficiency of mixing and the time of curing, etc. If gross errors are made, some very atypical results are recorded and it is dangerous to rely heavily on one small sample unless certain that it is representative.

The evaluation of an alternative ingredient by comparison with the standard formula may be considered. The mixes are uniform, the tester follows the procedures correctly and it is concluded, using statistical methods, that the new ingredient is an improvement. It is easily forgotten that this conclusion assumes that the samples of each compound were truly representative of the population. If the variability which would arise from repeat mixings is rather larger than the testing error, as is often the case, then tests on a series of repeat mixes may show no difference between the ingredients or even that the new ingredient was worse.

5.2 Accuracy, trueness and precision

Accuracy is the closeness of agreement between a test result and the accepted reference value (see 3.14), while trueness is the closeness of agreement between the average value of a large number of test results and the true or accepted reference value (see 3.15). Precision, on the other hand, is the closeness of agreement between the test results (see 3.16), independent of any reference value that may exist. To keep variability to a minimum, the test method should be as reproducible as possible, i.e. it should have good precision. However, having high precision may be of little value if the test has a large bias and hence poor accuracy. Both are required and indeed they are related in that poor precision (poor reproducibility) will contribute to lowering the accuracy.

Reproducibility (see 3.18) is the term generally reserved to describe the variation found between different laboratories, and perhaps also at different times. Repeatability (see 3.17) is used to describe the variation between repeats in the same laboratory at essentially the same time. It follows that laboratories may exhibit very good repeatability but, because of bias, the reproducibility between the laboratories is poor.

5.3 Relevance and significance

5.3.1 If accuracy or repeatability were the only interest, testing would be limited to the most accurate or precise methods. However, the test should be relevant in the sense that the results have a useful meaning in terms of material or product performance. All tests are not equal: some have more relevance than others in terms of product performance, material consistency or value as design data. The word significance is sometimes used to mean relevance and applied to the actual test or property measured, but significance is used in this International Standard in the statistical sense as in one material being significantly stronger, for example, than another.

Significance in this sense is concerned with whether observed differences in results are likely to be real or can reasonably be attributed to chance alone. If the probability of obtaining the observed difference through pure chance is small, for example less than 1 in 20, then the difference is said to be significant.

5.3.2 The set of tensile strength results quoted in 5.1 could be compared to other sets obtained on different materials on the same occasion giving, for example, three sets as in Table 2.

Table 2 — Tensile strength measurements from three materials

Measurement number	Tensile strength MPa		
	Material A	Material B	Material C
1	16,8	15,6	16,4
2	15,4	16,4	15,4
3	16,3	14,5	14,3
4	17,7	15,8	14,7
5	17,6	16,0	14,4

The averages of the results for materials A and B are higher than that for C but an assessment should be made as to whether or not they are significantly higher. Without the use of statistical tools it is rather difficult to make this assessment. In fact, using a test for significance as discussed in 7.2.2 it can be proved that A is significantly greater than C with 95 % confidence but that A is not significantly different from B, again with 95 % confidence. This is a useful conclusion but its limitations should be appreciated. The statistical tests prove (with a 1 in 20 chance of being wrong) that results A are significantly greater than C. They do not prove that material A is stronger than material C. It is known that results from one sheet of one mix may not be representative of a formulation and these results from a very small test programme should be treated with caution.

5.3.3 In the above example the differences between the average results were relatively small but tensile strength can be measured accurately with reasonably small variability so that it is not surprising that 10 % difference could be proved significant. For other, less reproducible tests a much greater percentage difference may be needed before the difference can be proved significant. For example, in an electrical resistivity test the mean value for one material was several times higher than that for a second material but the difference could not be proved significant. The deduction can be made that significance is not only dependent on the difference between mean values but also on the amount of variability which is inherent in the test.

6 Distribution of results and measures of central tendency

6.1 Principles

A collection of values, for example individual test results relating to a specific property, are arranged about a mean value. Usually the distribution of results may be represented by a particular mathematical law such as the curve shown in Figure 1.

From the shape of the curve it is possible to obtain useful measures of the tendency towards a central value, the degree of the spread of results and the proportion of results likely to differ by more than a certain amount from the centre.

6.2 Methodology

6.2.1 Types of distribution

6.2.1.1 The normal distribution

The most widely used distribution function is called the normal, or Gaussian, distribution function (see 3.22) which can be completely characterized by two parameters, the mean μ (see 3.4), and the standard deviation σ (see 3.9). These parameters are considered further in 6.2.2 and 6.2.3 respectively. The density distribution is a symmetrical bell-shaped function, the mathematical description of which can be found in Annex A.

Values of the ordinate, i.e. the density of the function, $f(z)$, at given values of z have been tabulated and can be found in any statistics text book. In order to make the tables suitable for general application, the abscissa z, is presented in reduced form, such that z is the number of standard deviations x (the value measured in the experiment) is away from the mean. Since the curve is symmetrical it is usual to find only the positive values of z that are tabulated since $f(z) = f(-z)$.

The proportion of the whole distribution which lies between two values, x_1 and x_2 (i.e. the probability distribution function) can be determined from the integral of the density distribution (see Annex A), but since this integral cannot be expressed analytically, it is more convenient to use tabulated values, which again are available in the form of the reduced variable Z, in any standard text on statistics.

In these tables the value of x_1 is usually set to μ with only the positive reduced variable Z quoted.

For these tables:

when $z = 0$, $p(Z) = 0{,}0$;

when $z \to \infty$, $p(Z) \to 0{,}5$.

If z is negative, $p(Z) = -p(+Z)$.

To find the proportion of the curve (i.e. the probability of the observation) lying between x_1 and x_2, $(x_2 > x_1)$ the procedure is as follows:

a) Determine Z_1 and Z_2 where

$$Z_1 = (x_1 - \mu)/\sigma;$$

$$Z_2 = (x_2 - \mu)/\sigma.$$

b) Find $p(Z_1)$ and $p(Z_2)$ from the tables.

c) Determine the required probability $p(Z_2 - Z_1)$ where

$$p(Z_2 - Z_1) = p(Z_2) - p(Z_1).$$

NOTE The signs should be taken into account.

6.2.1.2 The double exponential distribution

In the case of the distribution of tensile strengths (and elongations at break) for vulcanizates, there is considerable evidence that the density distribution function is not symmetrical but is skewed towards lower strength values (see References [1] to [6]), although this has been questioned by some investigators (see Reference [7]). The density distribution function which has been found to give a good representation of these skewed data is given by the double exponential distribution (see 3.23), shown in Figure 2 and described mathematically in Annex A.

Although of theoretical interest, the double exponential function has not found widespread application, both because of its complexity and the fact that, with the small number of test pieces normally considered in a tensile test, there is no significant difference in the mean and standard deviation derived from the double exponential function and from the normal Gaussian function.

6.2.1.3 The Weibull distribution

A distribution function that frequently arises out of fatigue data and similar lifetime testing is the Weibull distribution function (see 3.24), the form of which is illustrated in Figure 3 and described mathematically in Annex A.

The distribution is characterized by the following three parameters:

a) The parameter a represents the minimum life parameter for x at which the probability of failure just reaches zero (that is, giving an infinite lifetime). In most practical applications, a is taken to be zero but, where there is a genuine fatigue limit, a can take a finite, non-zero, value.

b) The parameter b affects both the spread of results and the peak position of the density function.

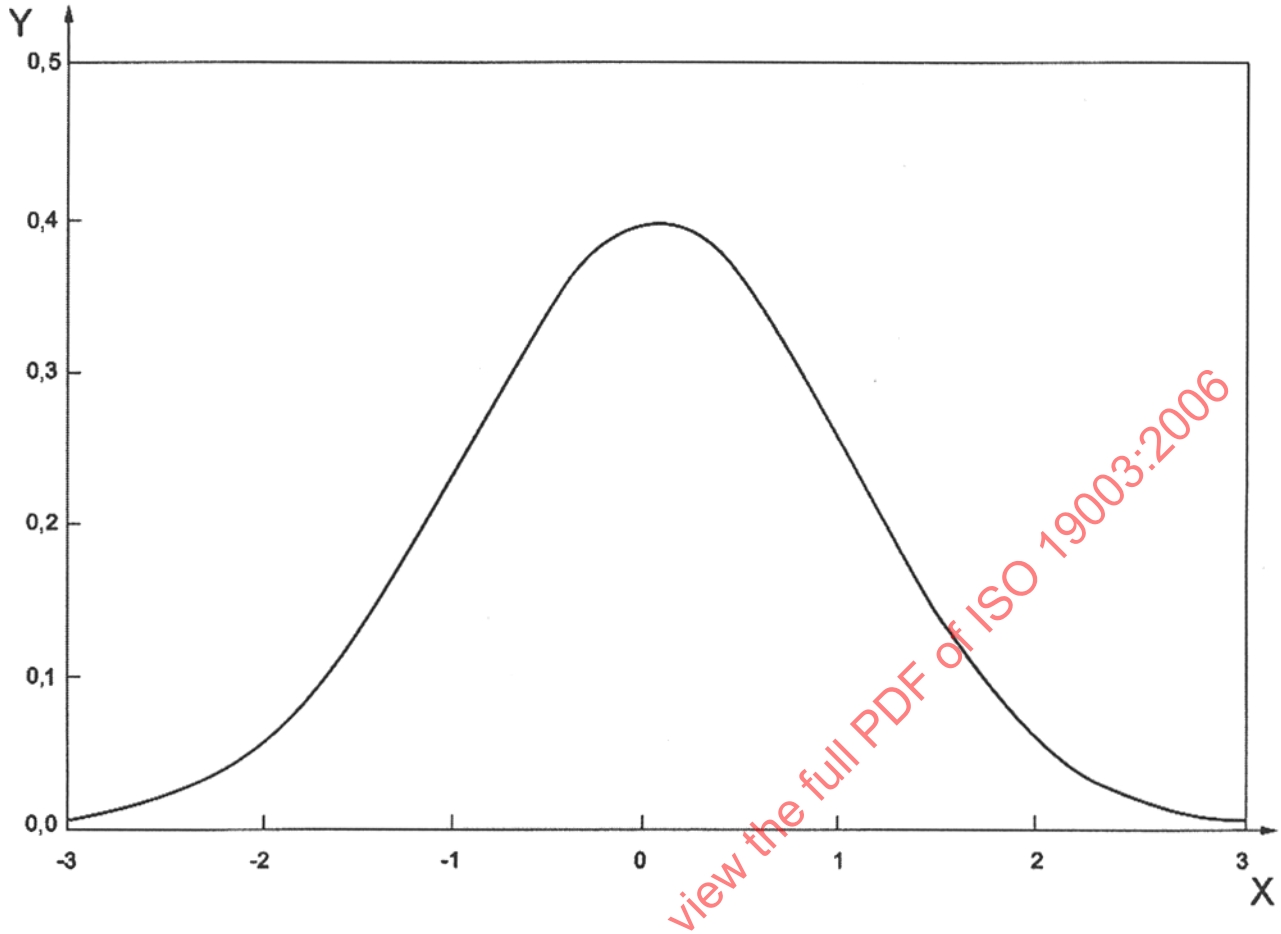
c) The parameter k alters the shape of the density distribution.

   When k = 1, the function is a simple exponential.

   When $k > 1$, the distribution increases from zero (at $x = a$), reaches a maximum and then decreases monotonically, reaching zero again at $x =$ infinity.

   When k is approximately 3,44, the distribution is approximately Gaussian, with the mean and median equal to each other.

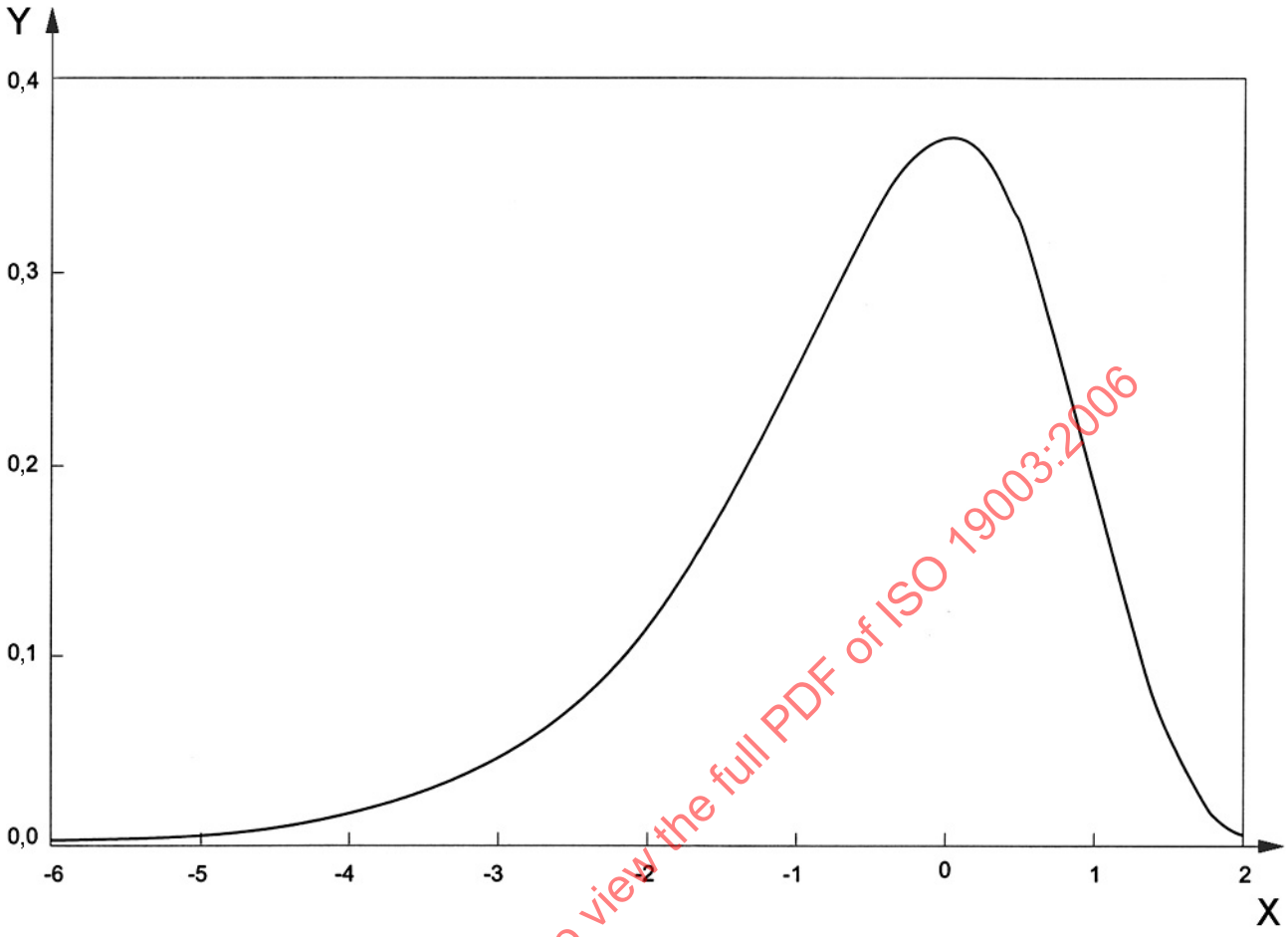
   Generally, the Weibull distribution is positively skewed (i.e. skewed towards longer lifetimes).



Key

X parameter y (see Annex A)
Y probability

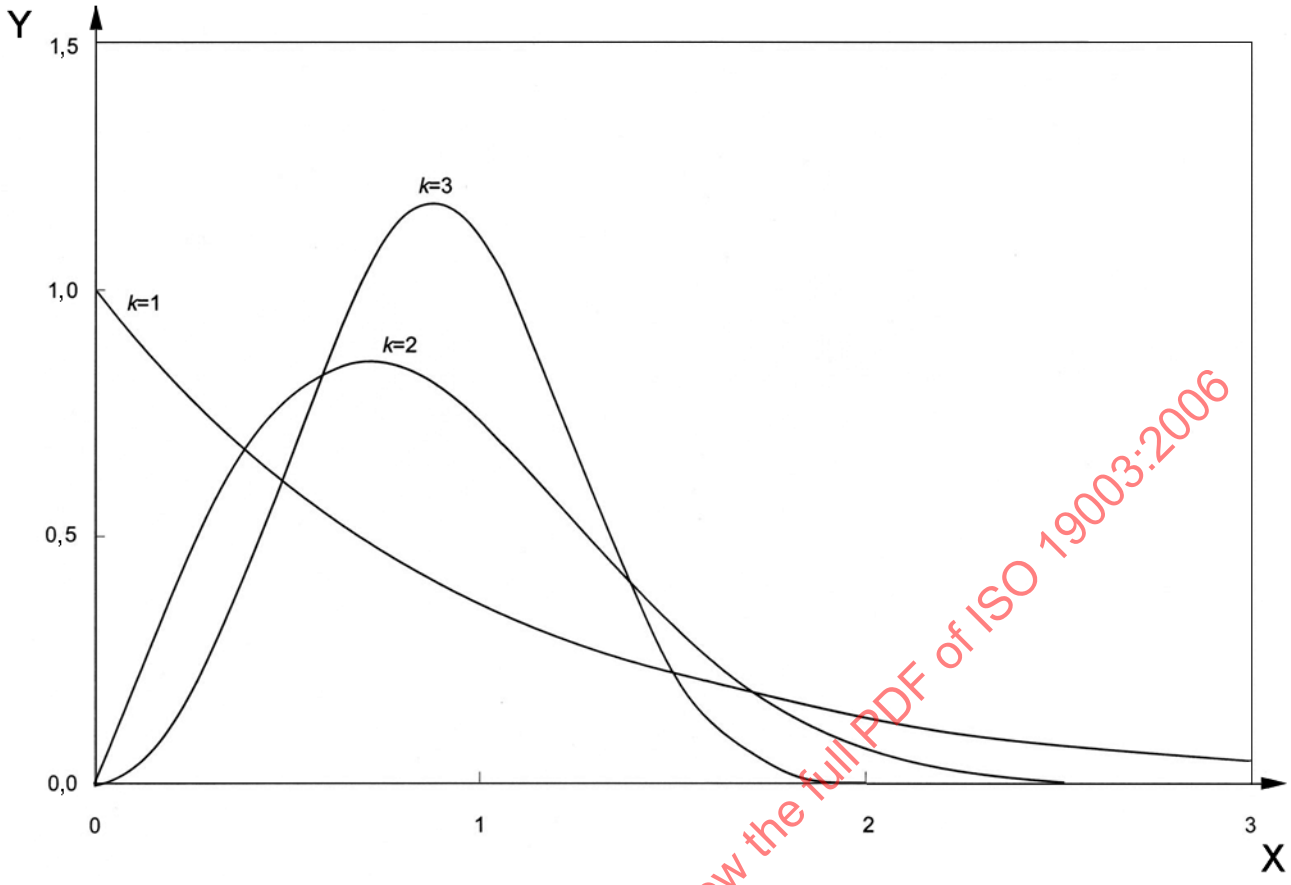
Figure 1 — Gaussian (normal) density function



Key

X parameter ε (see Annex A)
Y probability

Figure 2 — Double exponential density function



Key

X parameter $(x - a)/b$ (see Annex A)
Y probability

Figure 3 — Weibull density function

6.2.2 Measures of central tendency

6.2.2.1 General

Even under the most careful experimental conditions, carrying out repeat measurements on identical material produces a scatter of results. It is useful, therefore, to have some idea of the average or typical value that can be expected of those results. Since typical values tend to lie towards the middle of the data when these are arranged in numerical order, such numbers are also called measures of central tendency.

The small number of experimental results obtained is a sample of the infinite number of results, i.e. the population, that would fully encompass the measurement being made. The more results that are available, the more accurately sample statistics match the population statistics. Clearly, the average of the sample can be determined precisely but, generally, it is the average of the population from which the sample is drawn that is of greater interest and this can only be estimated from the sample, or samples, available. It is appropriate to distinguish sample and population statistics and hence different symbols are used.

6.2.2.2 The mean

The mean (see 3.4) is the most commonly encountered measure of central tendency and it is often also called the average. The precise meaning should be clear when these words are used as there are several ways of averaging a set of numerical values. The arithmetic average or mean is the one that is generally meant and so the word arithmetic is often omitted but, where there is the possibility of confusion, it ought to be included. In

this International Standard, where the word mean is used the arithmetic mean is intended unless otherwise stated.

The mean of a sample, $\bar{x}$, is defined as the sum of the individual numerical values in the sample divided by the number of values in the sample and is given mathematically by the equation

$$\bar{x} = \frac{(x_1 + x_2 + x_3 + ... + x_n)}{n} \quad (1)$$

which in the shorthand sigma notation becomes

$$\bar{x} = (\Sigma x) / n \quad (2)$$

As noted previously, the mean of the population from which the sample was taken is given the symbol μ. This value is almost never known in practice, but has to be estimated from the sample. The estimated mean of the population $\hat{\mu}$, based on the available sample, is taken to be equal to the sample mean. In other words

$$\hat{\mu} = \bar{x} \quad (3)$$

where $\hat{\mu}$ is an estimate of μ.

Where there are a large number of results having discrete values, it may be more convenient to record the number of occurrences of each value. If each value x occurs f times, then

$$\bar{x} = \left[\Sigma(fx)\right] / n \quad (4)$$

NOTE In this case, $n = \Sigma f$.

The same technique can be applied where an infinite variation in value x can occur, but it is more convenient to group the data into bands, counting the number of results in each band.

Other types of mean sometimes encountered are briefly described in Annex B.

6.2.2.3 The median

If the data in the set of results are arranged in numerical order, then the middle value (or the mean of the two middle values where there is an even number of values) is the median (see 3.5).

Geometrically, the median of a density distribution function is the value of the abscissa corresponding to that vertical line which divides the distribution into two equal areas.

For tensile strength or elongation at break, where the double exponential distribution function is expected to apply, it is strongly recommended that the median value be quoted for the measure of central tendency. The reason for this is three-fold:

a) for the small number of replicate results normally involved in a tensile test (typically five or even three), the median is not influenced by any extreme values that may have arisen;

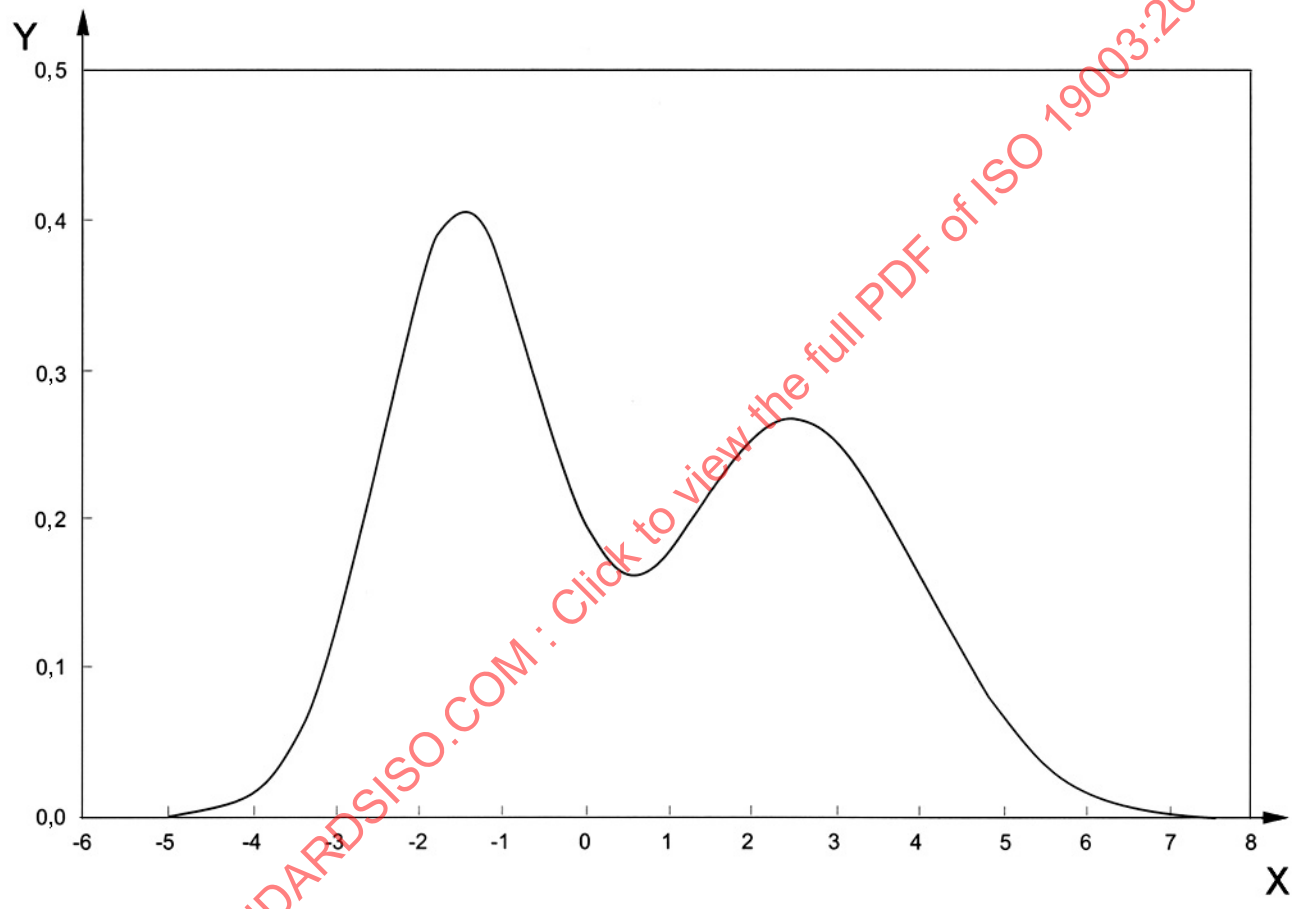
b) for a small number of replicates, the median can be deduced by inspection without recourse to any calculation;

c) there is an equal probability of an individual observation being greater than or being less than the median.

6.2.2.4 The mode

The mode (see 3.6) of a density distribution function is the value of the abscissa at which the maximum of the function occurs. Although little used in practice, it is referenced here for completeness as it has been used to characterize the double exponential distribution enunciated by Kase from earlier work by Fisher and Tippet [1].

Almost all distributions encountered in the rubber industry are unimodal, i.e. having a single maximum value. However, bimodal (having two maxima) or multi-modal distributions can result where two, or more, different mechanisms are occurring simultaneously in a process. Thus, tensile testing at a temperature close to the glass transition of the rubber can lead to a mixture of brittle and rubbery failure. This mixed mode of failure is best analysed by separating the results from the two types of failure and analysing each independently. An example of a bimodal distribution is given in Figure 4.



Key

X parameter y (see Annex A)
Y probability

Figure 4 — Bimodal density function

6.2.2.5 Inter-relationships

It is clear from the definitions of these various measures of central tendency that for symmetrical distributions

Mean = Median = Mode

The inter-relationships for non-symmetrical distributions are necessarily more complex and those for the double exponential and the Weibull distribution are given in Annex C.

6.2.3 Measures of dispersion

6.2.3.1 General

Just as it is useful to know the average value of a set of numbers, so also is their spread or dispersion important. The more loosely packed the numbers are about the mean, the less discrimination there is between two sets of numbers or between the experimental values and, for example, the specification requirement.

The variance and standard deviation are used with the mean, whilst the range is normally used with the median and mode. Also, the quartiles indicate 25 % of the data is higher or lower and the percentiles a specified percent is higher or lower.

6.2.3.2 Standard deviation

6.2.3.2.1 By definition, the standard deviation s' of a sample of results having values of x_i is given as the square root of the average squared deviation of each of the values from their mean ($\bar{x}$). It is given by the equation:

$$s' = \left[\frac{\sum (x_i - \bar{x})^2}{n} \right]^{1/2} \tag{5}$$

An alternative form that is sometimes more convenient to use is given in Annex D. Some of the risks associated with its use are also given.

The symbol s' is used here to represent the standard deviation of the sample of n results. If this sample is representative of the population from which it has been drawn, then the true standard deviation σ can be estimated by equating it with the statistic s defined by applying Bessel's correction to Equation (5). s is given by the equation:

$$s = \left[\frac{\sum (x_i - \bar{x})^2}{n-1} \right]^{1/2} \tag{6}$$

This is the form of the standard deviation that should normally be used when performing statistical tests since it is an estimate for the population as a whole and not simply the particular sample chosen.

6.2.3.2.2 It can be shown that

a) 68,26 % of values for a normal distribution lie within ± 1 standard deviation of the mean;

b) 95,44 % lie within ± 2 standard deviations;

c) 99,73 % lie within ± 3 standard devations.

For all practical purposes, in a normal distribution (or the distribution of the means of sets of values), the whole population is covered by six standard deviations. Therefore this interval is used in the setting of control charts (see Clause 18).

6.2.3.2.3 Related to the standard deviation is the standard error of the mean, which is determined from the standard deviation by dividing the standard deviation by the square root of the number of observations in the sample. It is, therefore, the standard deviation of the estimate of the population mean (see 3.12). Thus

$$S = s/\sqrt{n} \quad (7)$$

where S is the standard error of the mean.

The standard error is a measure of the expected spread of a series of mean values in the same way that the standard deviation is a measure of the expected spread of the individual values, and it is the standard error which should be used when making statistical comparisons between groups of numbers which are themselves the means of a group of numbers (see 6.3.2.4 and 7.2.2.2).

6.2.3.3 The range

While the standard deviation has valuable mathematical properties, historically it was somewhat cumbersome to calculate without computers and on occasion this might outweigh its value. Where a less precise estimate will suffice, the range, i.e. the maximum value minus the minimum value in the sample, may be used. It is possible to estimate the standard deviation from the range by multiplying the range by a factor which depends on the number of results in the set. For values of n between 2 and 11, the factors A_n are given in Table 3.

Table 3 — Table of factors for converting range to standard deviation

n	A_n
2	0,886
3	0,591
4	0,486
5	0,430
6	0,395
7	0,370
8	0,351
9	0,337
10	0,325
11	0,315

Thus, $s \approx$ range $\times A_n$.

The range and the quartiles can also be used as measures of spread around the median in an analogous manner to the standard deviation around the mean.

6.2.3.4 Coefficient of variation

Where the relative dispersion of results about their mean is of interest as, for example, in the comparison of the variabilities of the volume swell test with the density test, the ratio of the standard deviation to the mean can be used to normalize the effects of having very different numerical values for the means. It is usual to express this ratio as a percentage and to call it the coefficient of variation (see 3.10). It is given by the equation:

$$C_V = (s/\bar{x} \times 100) \quad (8)$$

where C_V is the coefficient of variation.

6.2.4 Transformation to normal distribution

The individual results obtained from a rubber test may not immediately conform to the normal distribution function. Other possible distribution functions such as the double exponential or Weibull may be theoretically

(or empirically) found to give a better representation of those data (see 6.2.5). Where certain statistical interferences need to be made concerning a set of data, for example the determination of confidence intervals or limits (see Clause 7), knowledge of the distribution function which describes the data is required. Because of the extensive range of tests and techniques that have been developed for the normal distribution, it is worth investigating whether a simple transformation of the data will result in a normally distributed data set to an accuracy sufficient for the situation being analysed.

It is also important to bear in mind that, even where the normal distribution is not found for the individual readings, the distribution of the means of groups of readings (as low as three per group) such as those obtained in the usual tests on rubber nearly always approximate to the Gaussian form (central limit theorem).

The transformation most commonly found to be effective in this regard is to take the logarithms of the values and treat these as the variable to be analysed. The use of log-probability graph paper or computer transformation makes this a very quick and simple test.

Other transformations that have been found to work on occasion include

a) taking the square root;

b) taking the reciprocal of the value.

Sometimes the addition of a constant to (or subtraction of a constant from) the value prior to taking logarithms, roots or reciprocals is required. It may be possible to deduce a suitable value for this constant from knowledge of the process being examined, but often it should be established empirically.

6.2.5 Test of departure from normality

6.2.5.1 Normal distribution function

6.2.5.1.1 The simplest way of testing a series of observations for the normality of its distribution is by plotting the results on probability paper. A normally distributed set of observations results in a straight line from which the mean and standard deviation can be derived.

6.2.5.1.2 The procedure is as follows:

a) Sort the data into ascending numerical order.

b) As probability paper is printed in the form of a percentage function, calculate the plotting position P_m for point m out of a total of n results using the equation:

$$P_m = 100m/(n + 1) \quad (9)$$

c) Plot the value of the mth point as ordinate against P_m as abscissa.

6.2.5.1.3 A more or less straight line indicates that the distribution is normal, but a marked deviation from linearity indicates that the distribution is non-Gaussian. Under these circumstances, the nature of the deviation may indicate the kind of distribution function that is more appropriate. In particular, if the larger values are systematically higher than the straight line defined by the lower values, the use of a logarithm or root transformation (see 6.2.4) will often result in a linear plot.

6.2.5.1.4 The above check does not provide a true test for normality in the statistical sense, but does give a rapid indication of the suitability, or not, of the normal distribution function as the model for the data observed. If the plot is not reasonably linear even after transformations have been applied, then more detailed symmetry and kurtosis tests may be required (taking into account the purpose for which the data are being used). These are outside the scope of this International Standard and the interested user is referred to ISO 5479.

Log-normal and Weibull graph papers are also available.

6.2.5.2 Double exponential distribution function

6.2.5.2.1 Where the distribution is expected to follow the double exponential function

a) order the data into descending numerical value;

b) plot the value as ordinates against an abscissa of plot positions which are given by Table 4.

6.2.5.2.2 If the double exponential function is valid, then a straight line will result and the following values may be obtained as described:

a) the mode, which corresponds to the ordinate at which the abscissa is zero;

b) the standard deviation, which is the difference in ordinates corresponding to a unit difference in the abscissa (i.e. it is the slope of the line);

c) the median strength, which is the value corresponding to an abscissa of − 0,366 5;

d) the mean, which is the value corresponding to an abscissa of − 0,577 2.

Table 4 — Plot positions for the double exponential distribution

n [a]	Total number of test results, N																				
	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
1	0,89	0,97	1,92	1,06	1,10	1,13	1,17	1,19	1,21	1,23	1,25	1,27	1,28	1,29	1,31	1,32	1,33	1,35	1,36	1,37	1,37
2	0,21	0,38	0,48	0,57	0,63	0,69	0,74	0,79	0,81	0,84	0,88	0,91	0,93	0,95	0,97	0,99	1,01	1,02	1,04	1,06	1,07
3	– 0,40	– 0,14	0,04	0,18	0,28	0,36	0,44	0,50	0,55	0,59	0,62	0,67	0,70	0,71	0,74	0,77	0,79	0,81	0,83	0,85	0,87
4	– 1,15	– 0,68	– 0,39	– 0,19	– 0,05	0,07	0,16	0,24	0,30	0,36	0,41	0,45	0,50	0,53	0,56	0,59	0,62	0,64	0,67	0,69	0,71
5	– 2,54	– 1,37	– 0,89	– 0,59	– 0,38	– 0,23	– 0,11	0,00	0,09	0,15	0,21	0,26	0,31	0,34	0,38	0,43	0,46	0,49	0,52	0,55	0,57
6		– 2,61	– 1,55	– 1,06	– 0,75	– 0,54	– 0,38	– 0,25	– 0,15	– 0,06	0,02	0,09	0,15	0,18	0,23	0,28	0,31	0,35	0,38	0,42	0,44
7			– 2,76	– 1,70	– 1,20	– 0,90	– 0,67	– 0,50	– 0,38	– 0,27	– 0,18	– 0,10	– 0,19	0,02	0,08	0,13	0,18	0,22	0,26	0,29	0,32
8				– 2,88	– 1,83	– 1,33	– 1,02	– 0,79	– 0,63	– 0,49	– 0,37	– 0,27	– 0,29	– 0,12	– 0,07	– 0,01	0,04	0,09	0,13	0,17	0,20
9					– 2,99	– 1,94	– 1,44	– 1,13	– 0,91	– 0,73	– 0,58	– 0,47	– 0,37	– 0,29	– 0,21	– 0,15	– 0,10	– 0,04	0,01	0,05	0,09
10						– 3,09	– 2,05	– 1,53	– 1,23	– 1,00	– 0,82	– 0,67	– 0,55	– 0,46	– 0,38	– 0,30	– 0,23	– 0,18	– 0,12	– 0,07	– 0,02
11							– 3,18	– 2,14	– 1,63	– 1,31	– 1,08	– 0,90	– 0,76	– 0,64	– 0,54	– 0,45	– 0,37	– 0,31	– 0,24	– 0,19	– 0,14
12								– 3,26	– 2,22	– 1,71	– 1,39	– 1,16	– 0,98	– 0,84	– 0,71	– 0,61	– 0,52	– 0,44	– 0,37	– 0,31	– 0,25
13									– 3,34	– 2,29	– 1,79	– 1,47	– 1,23	– 1,05	– 0,90	– 0,78	– 0,67	– 0,58	– 0,59	– 0,43	– 0,37
14										– 3,40	– 2,36	– 1,86	– 1,53	– 1,30	– 1,12	– 0,97	– 0,85	– 0,74	– 0,65	– 0,56	– 0,49
15											– 3,47	– 2,43	– 1,92	– 1,60	– 1,36	– 1,18	– 1,03	– 0,90	– 0,79	– 0,70	– 0,62
16												– 3,53	– 2,49	– 1,99	– 1,66	– 1,42	– 1,24	– 1,08	– 0,96	– 0,84	– 0,76
17													– 3,58	– 2,55	– 2,05	– 1,72	– 1,47	– 1,29	– 1,13	– 1,01	– 0,90
18														– 3,64	– 2,61	– 2,10	– 1,77	– 1,52	– 1,34	– 1,18	– 1,06
19															– 3,69	– 2,65	– 2,15	– 1,82	– 157	– 1,39	– 1,24
20																– 3,74	– 2,70	– 2,20	– 1,87	– 1,62	– 1,44
21																	– 3,78	– 2,75	– 2,23	– 1,91	– 1,67
22																		– 3,82	– 2,79	– 2,28	– 1,96
23																			– 3,86	– 2,83	– 2,33
24																				– 3,90	– 2,88
25																					– 3,94

[a] n = serial number of test result when arranged in decreasing magnitude.

6.2.5.3 Weibull distribution function

If the Weibull distribution function is assumed to be valid, the procedure is as follows:

a) sort the data into ascending order;

b) calculate the plotting positions P_m as for the normal distribution above;

c) plot the values for P_m against the observed lifetime on special Weibull probability paper.

NOTE Although this can be purchased directly, it is readily constructed from normal log-log graph paper and Annex E shows how this can be done.

6.3 Applications to rubber testing

6.3.1 General

For many tests, results will approximate to a normal distribution and it is appropriate to express results as the arithmetic mean and the standard deviation as routine practice. Typical examples are given in 6.3.2 to 6.3.5.

6.3.2 Tensile testing

6.3.2.1 The following three sets of 12 replicate tensile strengths shown in Table 5 were observed after testing in accordance with ISO 37 and arranging the results in descending order.

As there are 12 results, the median in each case is the average of the middle two values.

Table 5 — Tensile strength measurements

Measurements in MPa

Compound A		Compound B		Compound C	
Tensile strength	Median	Tensile strength	Median	Tensile strength	Median
26,7		28,4		19,7	
26,2		27,9		19,6	
26,1		27,4		19,2	
26,1		27,1		19,0	
25,9		26,8		18,7	
25,8		26,5		18,4	
	25,8		26,4		18,3
25,8		26,3		18,1	
25,8		26,2		17,3	
25,7		26,0		16,4	
25,6		25,9		15,6	
25,1		24,6		15,1	
25,0		24,1		13,5	

6.3.2.2 The information contained in Table 6 has been calculated by the following methods:

a) the mean has been calculated as defined in 6.2.2.2;

b) the standard deviation and the standard error of the mean have been calculated as defined in 6.2.3;

c) the calculated median has been calculated as defined in C.1 (i.e. for a double exponential distribution) using the values obtained in a) and b).

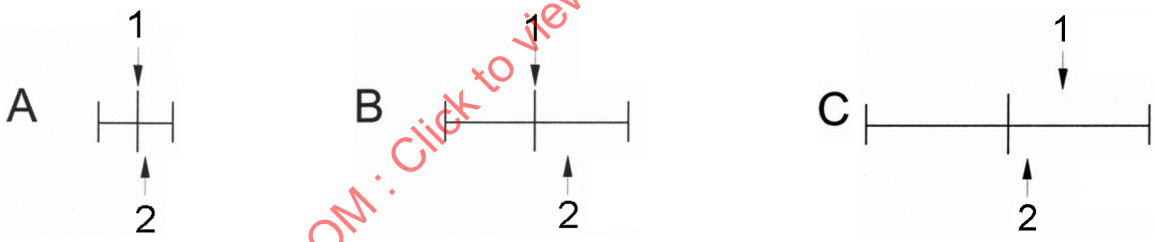
Table 6 — Calculated values for tensile strength measurements

Compound	Mean	Standard deviation	Standard error of the mean	Observed median	Calculated median
A	25,8	0,46	0,13	25,8	26,6
B	26,4	1,24	0,36	26,4	26,6
C	17,6	1,99	0,57	18,3	17,9

6.3.2.3 On comparing the two sets of median values, it is clear that the median obtained from inspection of the data is essentially the same as that calculated from the mean, the standard deviation and Annex C.

6.3.2.4 Examining the mean values of the three compounds (in relation to their standard errors) shows A and B to be within experimental error of each other, but compound C to be significantly different.

6.3.2.5 All the results have been summarized graphically in Figure 5.



Key

1 median 1 (observed value)

2 median 2 (calculated value)

Figure 5 — Graphical representation of mean and median data for three compounds
(using the tensile strength data of Table 6)

6.3.3 Fatigue

6.3.3.1 A tension fatigue test carried out in accordance with ISO 6943 on 10 replicate test pieces gave the observations listed in Table 7.

Table 7 — Tension fatigue test measurements

Test piece	Cycles to failure
1	219
2	347
3	494
4	593
5	700
6	858
7	1 037
8	1 146
9	1 461
10	1 795

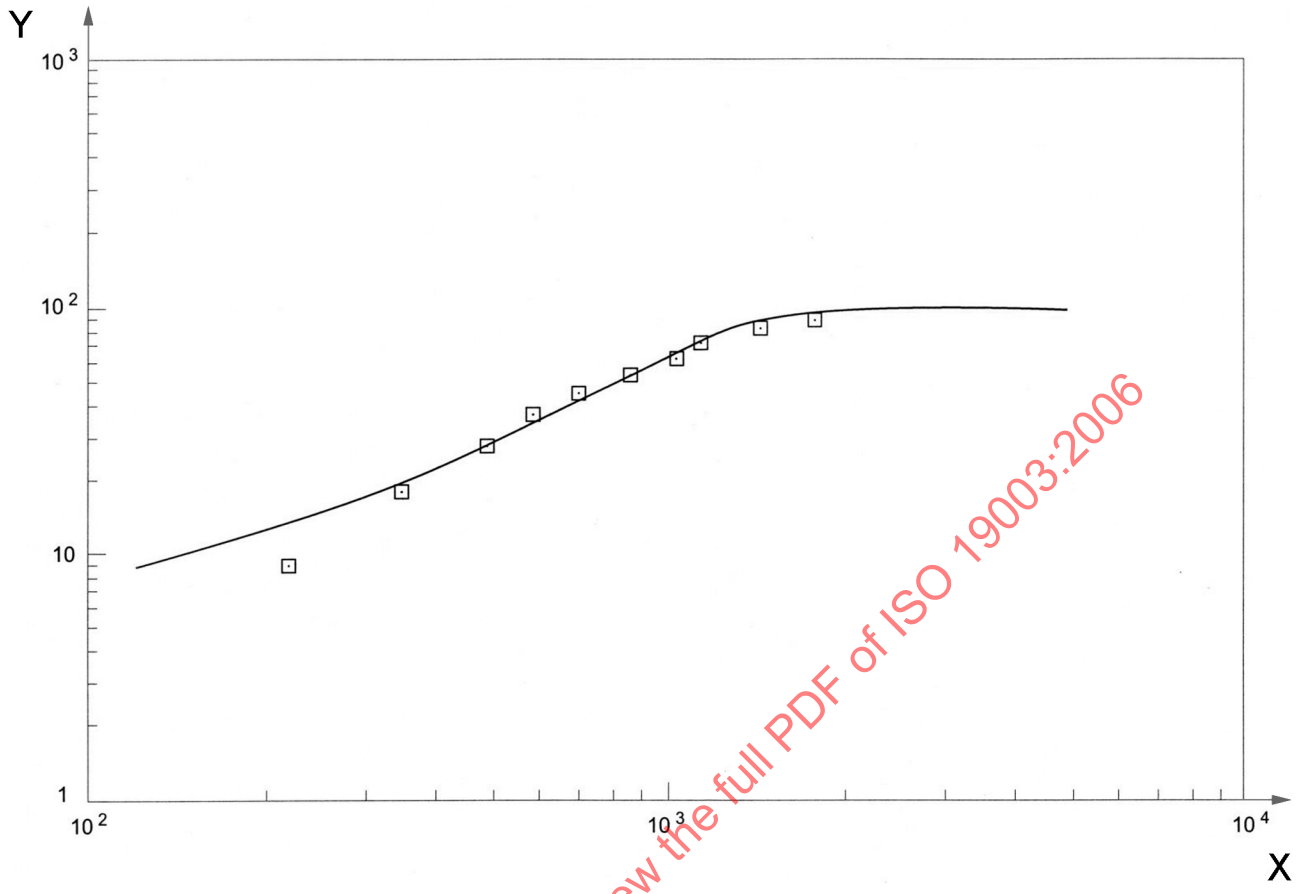
6.3.3.2 When the data in Table 7 are plotted as a normal distribution function, a systematic departure from the expected curve is observed, as shown in Figure 6 where a log-log scale has been used for convenience. The expected normal distribution of lifetimes is shown as the full curve.

6.3.3.3 From the type of test being performed, it would be expected that a Weibull distribution would give a good representation of the data. Hence a Weibull plot, constructed in accordance with Annex E, results in the plot shown in Figure 7. A linear regression analysis (see Clause 11) using the Weibull ordinate for the y-values and the logarithm of the fatigue life as the x-values produces the following values:

$k = 1{,}53$;

$b = 1\ 006$.

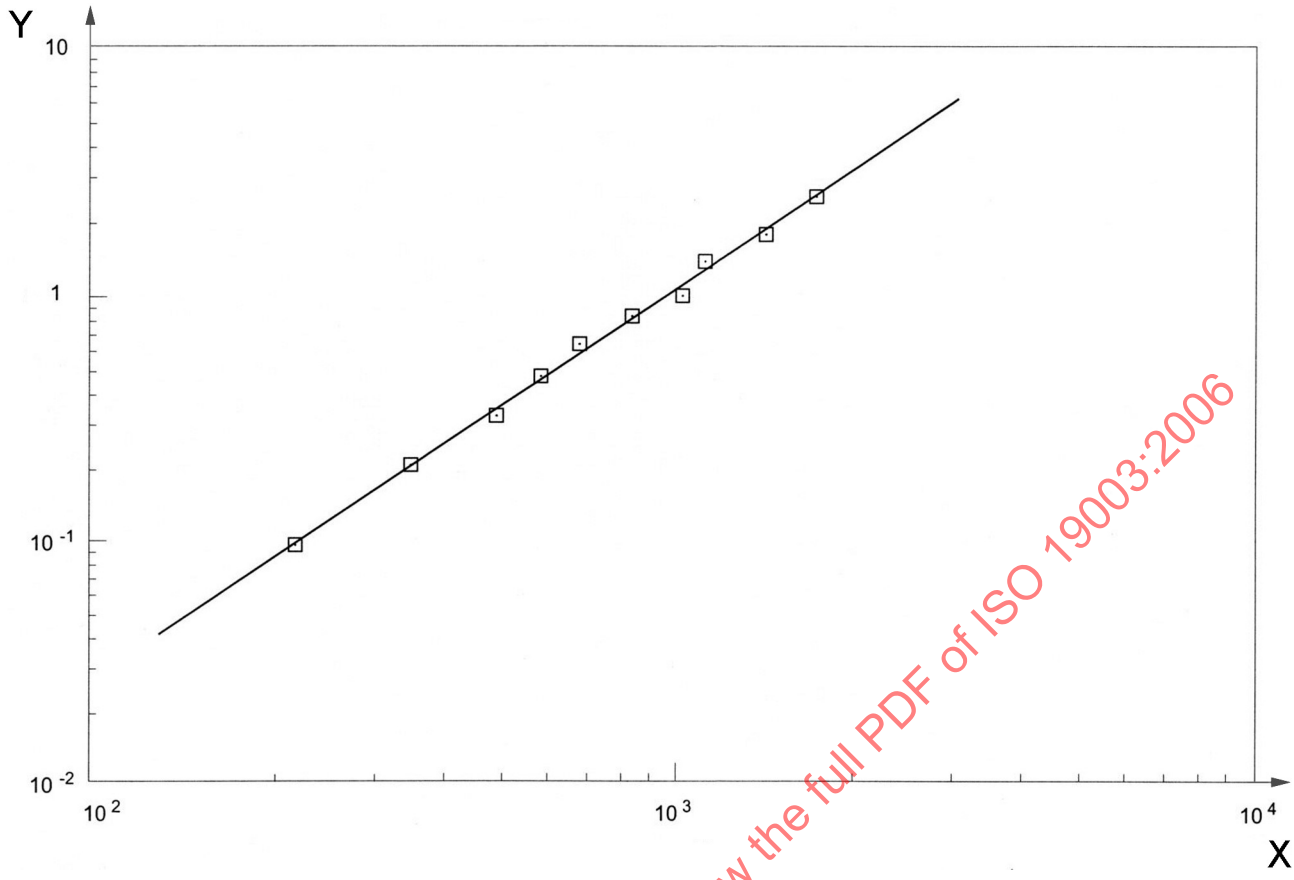
The strong linearity of the plot (the value of the variance ratio is over 5 000 and so the regression is highly significant) indicates that the coefficient a in the Weibull equation is zero.



Key

X fatigue life
Y normal ordinate

Figure 6 — Fatigue data: normal plot (using the data of Table 7)



Key

X fatigue life
Y Weibull ordinate

Figure 7 — Fatigue data: Weibull plot (using the data of Table 7)

6.3.4 Conversion to normal distribution

6.3.4.1 Data for electrical resistivity testing gave the results shown in Table 8.

Table 8 — Electrical resistivity measurements

Test piece	Resistivity Ω·cm
1	$2{,}81 \times 10^{11}$
2	$3{,}54 \times 10^{8}$
3	$2{,}68 \times 10^{10}$
4	$2{,}75 \times 10^{9}$
5	$1{,}20 \times 10^{10}$

If the assumption were made that the data are normally distributed, the analysis would give the values:

a) mean = $6{,}46 \times 10^{10}$;

b) standard deviation, $s = 1{,}21 \times 10^{11}$.

6.3.4.2 Clearly, from the observed mean and standard deviation the data are not normally distributed, a fact which is readily confirmed by plotting the resistivities sorted into ascending value against the plot positions 17, 33, 50, 67, 83 on standard probability paper. A rapidly increasing slope with increasing probability value is observed instead of a straight line.

6.3.4.3 When the same data are plotted as the logarithm of resistivity (on log-probability paper for convenience) a very good straight line is achieved having a mean at approximately 10^{10} Ω·cm. This is very close to the value of the geometric mean, $9{,}75 \times 10^9$, as would be expected from Annex B.

6.3.5 Other uses of the median

6.3.5.1 Measurements of the hardness of rubber compounds are probably the most frequently made of all the tests. The technique is not a highly precise one and results are usually quoted in whole numbers. Specifications of hardness are almost always given as the required nominal value ± 5 hardness degrees.

6.3.5.2 In this context, there is no mathematical advantage in calculating the mean from the set of five results (almost invariably) taken on a single test piece. Instead, the median, which can be deduced in a few seconds without a calculator, is the preferred measure of central tendency to use. For example, the results in Table 9 were derived on three compounds of nominal hardness 50 IRHD when tested in accordance with ISO 48.

Table 9 — Hardness measurements

Measurements in IRHD

Result	Compound 1	Compound 2	Compound 3
1	50	52	49
2	51	53	48
3	49	51	47
4	51	50	47
5	50	55	46

6.3.5.3 The values given in Table 10 were calculated from those in Table 9.

Table 10 — Values calculated from hardness measurements

Measurements in IRHD

Compound	Mean	Standard deviation, s	Median
1	50	1	50
2	52	2	52
3	47	1	47

In all cases, the median is within the 95 % confidence interval (see 7.2.1) of the mean and gives, for all practical purposes, the same result. In some instances, it may be necessary to carry out formal statistical testing in which case the standard deviation as well as the mean are required and there is no advantage in abstracting the median from the list. Also, if large numbers of replicate hardness values are to be processed, the median can become more tedious to determine than the mean.

NOTE The use of the median is purely pragmatic. There is no suggestion that the hardness is anything other than normally distributed about its mean.

7 Confidence limits and significant difference

7.1 Principles

As was stated in Clause 6, the mean and standard deviation derived for a given set of observations can only be estimates of the true mean and standard deviation of the whole population from which these observations are a random selection. To the extent that there is no systematic bias in the observations, the greater the number of results available, the less uncertainty there is over the accuracy of these estimates.

Unless the true mean (or standard deviation) can be deduced via some *a priori* reasoning, it is impossible to state how close a given calculated mean (or standard deviation) is to the true value. But it is possible to indicate with a known degree of uncertainty (the confidence level) that the true value will lie within a particular interval about that calculated value. The greater the degree of certainty required, the larger does this confidence interval become and the further apart are the limits (the confidence limits) of this interval.

Since any calculated measures of central tendency or dispersion are subject to uncertainty, when two such measures are compared, they cannot be expected to agree precisely. The difference between them becomes significant in the statistical sense only when it exceeds a limiting value which could have occurred, with a given probability, purely by chance.

It should be noted that a statistically significant difference between two measures of a property does not imply that the difference has any practical significance. The latter can only be judged in the context of the application being studied and the sensitivity of the application to the measured property. Thus, two compounds having tear strengths of 10 N/mm and 15 N/mm with standard deviations of 1 N/mm based on five results are significantly different at the 99 % level of confidence but, if the specification calls for a minimum tear strength of 25 N/mm, neither of them meets the specification and hence the difference between them is insignificant (and irrelevant) in practical terms.

Significance tests such as Student's t-test make several assumptions about the data being analysed, and an awareness of these assumptions is important in order to use the technique successfully. Firstly, the data must represent actual values, such as tensile strength or length; subjective values and grades cannot be treated in this way. Secondly, they must be normally distributed. Finally, they must be independent so that an error in one observation does not influence the error in another observation. If the data are not normally distributed, they can often be made to approximate to normality by a transformation (for example taking logarithms). The tests also assume that the two means being compared come from similar populations, i.e. their variances are not significantly different (see 7.2.2.2 for more information).

7.2 Methodology

7.2.1 Confidence limits and confidence intervals

7.2.1.1 For the mean

7.2.1.1.1 If the population mean, μ, and standard deviation, σ, are known, then it is a simple matter (see 6.2.1.1) to work out what the probability is that the mean of n results lies within, for example, three standard deviations of μ. Using the normal notation of $\bar{x}$ to represent the mean of the n results, then on 99,7 % of occasions

$$\mu - 3\sigma/\sqrt{n} < \bar{x} < \mu + 3\sigma/\sqrt{n} \tag{10}$$

By similar reasoning, if μ is unknown and σ is known, it is logical to assert that

$$\bar{x} - 3\sigma/\sqrt{n} < \hat{\mu} < \bar{x} + 3\sigma/\sqrt{n} \tag{11}$$

Thus, at a confidence level of 99,7 %, it is expected that the population mean would lie within the confidence interval of $\pm 3\sigma/\sqrt{n}$ about the sample mean. If a smaller confidence level, say 95 %, were chosen, the confidence interval would be $\pm 1{,}96\sigma/\sqrt{n}$.

7.2.1.1.2 It is unfortunate that the exact values of μ and of σ are unknown. Only the estimate s is known from the sample tested. For the same confidence level 99,7 %, the value $3\sigma/\sqrt{n}$ has to be increased to take account of this uncertainty in σ, the amount of the increase being dependent on the value of n. The factors used are given by the Student's t-distribution rather than the normal distribution (which is the limiting value of the t-distribution when n reaches infinity). Further consideration of the t-distribution is outside the scope of this International Standard and reference should be made to one of the many textbooks on statistics that are available. A selection of useful reference works is included in the Bibliography.

7.2.1.1.3 It is, therefore, assumed in the following clauses that the true mean and standard deviation are unknown and that only the estimates of the mean $\bar{x}$ and of the standard deviation s are known. For an exposition of situations in which one or other of the true parameters is known, reference should be made to ISO 2602 and ISO 2854.

The confidence limits for the mean are normally required for the 95 % and the 99 % confidence levels. In either case, the limits are given by the equations:

$$c_L = \bar{x} - (t_\alpha s)/\sqrt{n} \quad (12)$$

$$c_U = \bar{x} + (t_\alpha s)/\sqrt{n} \quad (13)$$

$$c_I = 2(t_\alpha s)/\sqrt{n} \quad (14)$$

where

c_L = lower confidence limit;

c_U = upper confidence limit;

c_I = confidence level.

It is therefore possible to be 95 % (or 99 %, etc.) confident that the true mean μ of the population does lie within this interval about the calculated mean (i.e. the estimated population mean).

In these equations, n is the number of observations and t is the appropriate Student's t-value, obtained from Table 11.

Table 11 — A selected table of Student's t-values

n	Confidence level Two-sided case		Confidence level One-sided level	
	95 %	99 %	95 %	99 %
	$t_{0,975}$	$t_{0,995}$	$t_{0,95}$	$t_{0,99}$
2	12,71	63,66	6,314	31,82
3	4,303	9,925	2,920	6,965
4	3,182	5,841	2,353	4,541
5	2,776	4,604	2,132	3,747
6	2,571	4,032	2,015	3,365
7	2,447	3,707	1,943	3,143
8	2,365	3,499	1,895	2,998
9	2,306	3,355	1,860	2,896
10	2,262	3,250	1,833	2,821
11	2,228	3,169	1,812	2,764
12	2,201	3,106	1,796	2,718
13	2,179	3,055	1,782	2,681
14	2,160	3,012	1,771	2,650
15	2,145	2,977	1,761	2,624
16	2,131	2,947	1,753	2,602
17	2,120	2,921	1,746	2,583
18	2,110	2,898	1,740	2,567
19	2,101	2,878	1,734	2,552
20	2,093	2,861	1,729	2,539
21	2,086	2,845	1,725	2,528
22	2,080	2,831	1,721	2,518
23	2,074	2,819	1,717	2,508
24	2,069	2,807	1,714	2,500
25	2,064	2,797	1,711	2,492
26	2,060	2,787	1,708	2,485
27	2,056	2,779	1,706	2,479
28	2,052	2,771	1,703	2,473
29	2,048	2,763	1,701	2,467
30	2,045	2,756	1,699	2,462
40	2,024	2,707	1,682	2,430
50	2,008	2,680	1,676	2,404
60	2,000	2,664	1,673	2,393
120	1,980	2,617	1,658	2,358
inf.	1,960	2,576	1,645	2,326

7.2.1.1.4 As noted in 6.2.3.2.3, the quantity $s/\sqrt{n}$ is called the standard error of the estimate (of the mean). It can thus be seen that to halve the confidence interval approximately four times as many observations have to be taken.

NOTE t is approximately constant except at very small values of n.

7.2.1.1.5 The value of t_α depends on the confidence level required, the number of observations (or, more precisely, the number of degrees of freedom) and whether a single-sided or a two-sided confidence interval is being sought.

In the simple cases being considered here, the number of degrees of freedom is $(n - 1)$.

7.2.1.1.6 A single-sided confidence interval is used when, for example, a comparison is being made between an observed mean value for a test, such as compression set, and the specification maximum (or minimum) to which it is being tested. This is because the only concern is with those values that might exceed (or not reach) the requirement and there is no interest in the values at the other side of the distribution function, these being those that conform to the specification limit. In this case, t_α is given in the tables under the columns for $t_{0,95}$ for the 95 % and $t_{0,99}$ for the 99 % confidence limits.

7.2.1.1.7 A two-sided confidence interval is used when, for example, it is necessary to know the interval within which the true mean could be expected to lie with the given degree of confidence. In this case, both sides of the distribution function are equally important and will contribute equally to the probability. Thus, for 95 % confidence the $t_{0,975}$ column is required and for 99 % confidence the $t_{0,995}$ column is required.

7.2.1.1.8 There are situations where it might be more convenient to calculate the value of Student's t-factor, for a given probability and number of degrees of freedom, rather than using reference tables. Provided that an error not exceeding 0,5 % of the true t-value is acceptable, then the following equation may be used:

$$t_\alpha = A + BC^{[1/(n-1)]} \tag{15}$$

where the constants A, B and C are as given in Table 12.

Table 12 — t-value constants

t	A	B	C
$t_{0,95}$	0,875 7	0,770 03	7,062 3
$t_{0,975}$	1,053 1	0,909 30	12,819 2
$t_{0,99}$	1,264 0	1,069 9	28,559 0
$t_{0,995}$	1,418 7	1,171 7	53,120 9

7.2.1.2 For the standard deviation

7.2.1.2.1 As in the case of the mean, the standard deviation calculated from a set of data can only be an estimate of the true standard deviation for the population as a whole and as such will have a measure of uncertainty associated with it. Confidence limits can therefore be set which, to a stated degree of confidence, contain the population standard deviation.

7.2.1.2.2 Unlike the confidence limits for the mean, the limits for the standard deviation are not symmetrical about the estimate s. This arises out of the fact that the standard deviation, unlike the mean, cannot be negative.

7.2.1.2.3 As discussed in 7.2.1.1.6 and 7.2.1.1.7 when considering the confidence limits for the mean, there are two cases to be considered. These are the single-sided case and the two-sided case depending on whether just an upper (or lower) limit, or both, is being considered.

$$c_{Us} = \left[\frac{ns^2}{\chi_{\alpha}^2} \right]^{1/2} \tag{16}$$

where c_{Us} is the upper confidence level for s.

$$c_{Ls} = \left[\frac{ns^2}{\chi_{1-\alpha}^2} \right]^{1/2} \tag{17}$$

where c_{Ls} is the lower confidence level for s.

The denominator in these equations comes from the chi-squared distribution function which is defined as the distribution of the sums of the squares of independent standardized normal variants. Some values are given in Table 13 and more comprehensive tables are available in the list of references.

For the single-sided case, α = 0,95 or 0,99 for the 95 % or 99 % confidence limits, respectively.

For the two-sided case, α = 0,975 or 0,995 for the 95 % or 99 % confidence limits, respectively.

As stated in 7.2.1.1.5, the number of degrees of freedom to be entered to find the chi-squared factor is $(n - 1)$.

Table 13 — A selected table of chi-squared values

n	Chi-squared for two-sided case				Chi-squared for one-sided case			
	95 % $\chi_{0,025}^2$	95 % $\chi_{0,975}^2$	99 % $\chi_{0,005}^2$	99 % $\chi_{0,995}^2$	95 % $\chi_{0,05}^2$	95 % $\chi_{0,95}^2$	99 % $\chi_{0,01}^2$	99 % $\chi_{0,99}^2$
1	0,001	5,023	0,000 039	7,879	0,004	3,841	0,000 2	6,635
2	0,051	7,378	0,010	10,597	0,103	5,991	0,020	9,210
3	0,216	9,348	0,072	12,838	0,352	7,815	0,115	11,345
4	0,484	11,143	0,207	14,860	0,711	9,488	0,297	13,277
5	0,831	12,833	0,412	16,750	1,145	11,071	0,554	15,086
6	1,237	14,449	0,676	18,548	1,635	12,592	0,872	16,812
7	1,690	16,013	0,989	20,278	2,167	14,067	1,239	18,475
8	2,180	17,535	1,344	21,955	2,733	15,507	1,646	20,090
9	2,700	19,023	1,735	23,589	3,325	16,919	2,088	21,666
10	3,247	20,483	2,156	25,188	3,940	18,307	2,558	23,209
11	3,816	21,920	2,603	26,757	4,575	19,675	3,053	24,725
12	4,404	23,337	3,074	28,300	5,226	21,026	3,571	26,217
13	5,009	24,736	3,565	29,819	5,892	22,362	4,107	27,688
14	5,629	26,119	4,075	31,319	6,571	23,685	4,660	29,141
15	6,262	27,488	4,601	32,801	7,261	24,996	5,229	30,578
16	6,908	28,845	5,142	34,267	7,962	26,296	5,812	32,000
17	7,564	30,191	5,697	35,719	8,672	27,587	6,408	33,409
18	8,231	31,526	6,265	37,156	9,390	28,869	7,015	34,805
19	8,907	32,852	6,844	38,582	10,117	30,144	7,633	36,191
20	9,591	34,170	7,434	39,997	10,851	31,410	8,260	37,566
21	10,283	35,479	8,034	41,401	11,591	32,671	8,897	38,932
22	10,982	36,781	8,643	42,796	12,338	33,924	9,542	40,289
23	11,689	38,076	9,260	44,181	13,091	35,173	10,196	41,638
24	12,401	39,364	9,886	45,559	13,848	36,415	10,856	42,980
25	13,120	40,647	10,520	46,928	14,611	37,653	11,524	44,314
26	13,844	41,923	11,160	48,290	15,379	38,885	12,198	45,642
27	14,573	43,194	11,808	49,645	16,151	40,113	12,879	46,963
28	15,308	44,461	12,461	50,993	16,928	41,337	13,565	48,278
29	16,047	45,722	13,121	52,336	17,708	42,557	14,257	49,588
30	16,791	46,979	13,787	53,672	18,493	43,773	14,954	50,892

7.2.2 Significant difference

7.2.2.1 General

Closely related to the concept of confidence limits is that of significant difference, where a comparison needs to be made either between two means or two standard deviations. In the following subclauses, it is assumed that there are two sets of observations with the statistics shown in Table 14.

Table 14 — Statistics for observation sets

Observation set	Mean value	Estimated standard deviation	Number of observations
1	$\bar{x}_1$	s_1	n_1
2	$\bar{x}_2$	s_2	n_2

7.2.2.2 For the mean

The means cannot be regarded as likely to be different at the given confidence level α if

$$|\bar{x}_1 - \bar{x}_2| > t_\alpha S \tag{18}$$

where

$|\bar{x}_1 - \bar{x}_2|$ signifies the absolute value of the difference;

t_α is the value of Student's t (Table 11) for the two-sided case, entered for $(n_1 + n_2 - 2)$ degrees of freedom, being 97,5 for the 95 % confidence level or 99,5 for the 99 % confidence level;

S is the weighted standard error for the combined sets of observations, which is calculated by the equation:

$$S = \left[\frac{n_1 + n_2}{n_1 n_2} \times \frac{(n_1 - 1)s_1^{\,2} + (n_2 - 1)s_2^{\,2}}{n_1 + n_2 - 2} \right]^{1/2} \tag{19}$$

In most cases, $n_1 = n_2$ and the equation can be considerably simplified to the form:

$$S = \frac{\sqrt{(s_1^{\,2} + s_2^{\,2})}}{\sqrt{n}} \tag{20}$$

It is assumed in this analysis that there is no significant difference (at a stated level) in the standard deviations s_1 and s_2. If such a difference is significant, then the two sets of observations cannot be considered to have come from the same population and hence their means would not usually be compared. The test for the significance of the difference between two standard deviations is given in 7.2.2.3.

7.2.2.3 For the standard deviation

7.2.2.3.1 The procedure is as follows:

a) Determine the ratio of the variances using the equation:

$$F = (s_1/s_2)^2 \tag{21}$$

it being taken that $s_1 > s_2$.

b) Consult Table 15 where the critical values for Snedecor's F-quotient at the 95 % and the 99 % confidence levels are given.

c) Use one of the following:

1) α = 0,05 for a single-sided 95 % confidence level;

2) α = 0,025 for a two-sided 95 % confidence level;

3) α = 0,01 for a single-sided 99 % confidence level;

4) α = 0,005 for a two-sided 99 % confidence level.

d) Establish the critical F-value for the $(n_1 - 1)$ degrees of freedom for s_1 and the $(n_2 - 1)$ degrees of freedom for s_2 by finding the intersection of the column with $(n_1 - 1)$ degrees of freedom and the row with $(n_2 - 1)$ degrees of freedom. If the calculated F-value is greater than this tabulated critical F-value, then the two standard deviations are different at the chosen confidence level.

Table 15 — Snedecor's F-values for selected degrees of freedom
(DF_l = lesser degrees of freedom, DF_g = greater degrees of freedom)

a) F_{95} for one-sided test

DF_l	DF_g														
	3	4	5	6	7	8	10	12	15	20	24	30	40	60	120
3	9,28	9,12	9,01	8,94	8,89	8,85	8,79	8,74	8,80	8,66	8,64	8,62	8,59	8,57	8,55
4	6,59	6,39	6,26	6,16	6,09	6,04	5,96	5,91	5,86	5,80	5,77	5,75	5,72	5,69	5,66
5	5,41	5,19	5,05	4,95	4,88	4,82	4,74	4,68	4,62	4,56	4,53	4,50	4,46	4,43	4,40
6	4,76	4,53	4,39	4,28	4,21	4,15	4,06	4,00	3,94	3,87	3,84	3,81	3,77	3,74	3,70
7	4,35	4,12	3,97	3,87	3,79	3,73	3,64	3,57	3,51	3,44	3,41	3,38	3,34	3,30	3,27
8	4,07	3,84	3,69	3,58	3,50	3,44	3,35	3,28	3,22	3,15	3,12	3,08	3,04	3,01	2,97
10	3,71	3,48	3,33	3,22	3,14	3,07	2,98	2,91	2,85	2,77	2,74	2,70	2,66	2,62	2,58
12	3,49	3,26	3,11	3,00	2,91	2,85	2,75	2,69	2,62	2,54	2,51	2,47	2,43	2,38	2,34
15	3,29	3,06	2,90	2,79	2,71	2,64	2,54	2,48	2,40	2,33	2,29	2,25	2,20	2,46	2,11
20	3,10	2,87	2,71	2,60	2,51	2,45	2,35	2,28	2,20	2,12	2,08	2,04	1,99	1,95	1,90
24	3,01	2,78	2,62	2,51	2,42	2,36	2,25	2,18	2,11	2,03	1,98	1,94	1,89	1,84	1,79
30	2,92	2,69	2,53	2,42	2,33	2,27	2,16	2,09	2,01	1,93	1,89	1,84	1,79	1,74	1,68
40	2,84	2,61	2,45	2,34	2,25	2,18	2,08	2,00	1,92	1,84	1,79	1,74	1,69	1,64	1,58
60	2,76	2,53	2,37	2,25	2,17	2,10	1,99	1,92	1,84	1,75	1,70	1,65	1,59	1,53	1,47
120	2,68	2,45	2,29	2,17	2,09	2,02	1,91	1,83	1,75	1,66	1,61	1,55	1,50	1,43	1,35

Table 15 (*continued*)

b) F_{95} **for two-sided test**

DF_l	DF_g														
	3	4	5	6	7	8	10	12	15	20	24	30	40	60	120
3	15,44	15,10	14,88	14,43	14,62	14,54	14,42	14,34	14,25	14,17	14,12	14,08	14,04	13,99	13,95
4	9,98	9,60	9,36	9,20	9,07	8,98	8,84	8,75	8,66	8,56	8,51	8,46	8,41	8,36	8,31
5	7,76	7,39	7,15	6,98	6,85	6,76	6,62	6,52	6,43	6,33	6,28	6,23	6,18	6,12	6,07
6	6,60	6,23	5,99	5,82	5,70	5,60	5,46	5,37	5,27	5,17	5,12	5,07	5,01	4,96	4,90
7	5,89	5,52	5,29	5,12	4,99	4,90	4,76	4,67	4,57	4,47	4,42	4,36	4,31	4,25	4,20
8	5,42	5,05	4,82	4,65	4,53	4,43	4,30	4,20	4,10	4,00	3,95	3,89	3,84	3,78	3,73
10	4,83	4,47	4,24	4,07	3,95	3,85	3,72	3,62	3,52	3,42	3,37	3,31	3,26	3,20	3,14
12	4,47	4,12	3,89	3,73	3,61	3,51	3,37	3,28	3,18	3,07	3,02	2,96	2,91	2,85	2,79
15	4,15	3,80	3,58	3,41	3,29	3,20	3,06	2,96	2,86	2,76	2,70	2,64	2,59	2,52	2,46
20	3,86	3,51	3,29	3,13	3,01	2,91	2,77	2,68	2,57	2,46	2,41	2,35	2,29	2,22	2,16
24	3,72	3,38	3,15	2,99	2,87	2,78	2,64	2,54	2,44	2,33	2,27	2,21	2,15	2,08	2,01
30	3,59	3,25	3,03	2,87	2,75	2,65	2,51	2,41	2,31	2,20	2,14	2,07	2,01	1,94	1,87
40	3,46	3,13	2,90	2,74	2,62	2,53	2,39	2,29	2,18	2,07	2,01	1,94	1,88	1,80	1,72
60	3,34	3,01	2,79	2,63	2,51	2,41	2,27	2,17	2,06	1,94	1,88	1,82	1,74	1,67	1,58
120	3,23	2,89	2,67	2,52	2,39	2,30	2,16	2,05	1,94	1,82	1,75	1,69	1,61	1,53	1,43

c) F_{99} **for one-sided test**

DF_l	DF_g														
	3	4	5	6	7	8	10	12	15	20	24	30	40	60	120
3	29,46	28,7	28,24	27,91	27,67	27,49	27,23	27,05	26,87	26,69	26,60	26,50	26,41	26,32	26,22
4	16,69	15,98	15,52	15,21	14,98	15,80	14,55	14,37	14,20	14,02	13,93	13,84	13,75	13,65	13,56
5	12,06	11,39	10,97	10,67	10,46	10,29	10,05	9,89	9,72	9,55	9,47	9,38	9,29	9,20	9,11
6	9,78	9,15	8,75	8,47	8,26	8,10	7,87	7,72	7,56	7,40	7,31	7,23	7,14	7,06	6,97
7	8,45	7,85	7,46	7,19	6,99	6,84	6,62	6,47	6,31	6,16	6,07	5,99	5,91	5,82	5,74
8	7,59	7,01	6,63	6,37	6,18	6,03	5,81	5,67	5,52	5,36	5,28	5,20	5,12	5,03	4,95
10	6,55	5,99	5,64	5,39	5,20	5,06	4,85	4,71	4,56	4,41	4,33	4,25	4,17	4,08	4,00
12	5,95	5,41	5,06	4,82	4,64	4,50	4,30	4,16	4,01	3,86	3,78	3,70	3,62	3,54	3,45
15	5,42	4,89	4,56	4,32	4,14	4,00	3,80	3,67	3,52	3,37	3,29	3,21	3,13	3,05	2,96
20	4,94	4,43	4,10	3,87	3,70	3,56	3,37	3,23	3,09	2,94	2,86	2,78	2,69	2,61	2,52
24	4,72	4,22	3,90	3,67	3,50	3,36	3,17	3,03	2,89	2,74	2,66	2,58	2,49	2,40	2,31
30	4,51	4,02	3,70	3,47	3,30	3,17	2,98	2,84	2,70	2,55	2,47	2,39	2,30	2,21	2,11
40	4,31	3,83	3,51	3,29	3,12	2,99	2,80	2,66	2,52	2,37	2,29	2,20	2,11	2,02	1,92
60	4,13	3,65	3,34	3,12	2,95	2,82	2,63	2,50	2,35	2,20	2,12	2,03	1,94	1,84	1,73
120	3,95	3,48	3,17	2,96	2,79	2,66	2,47	2,34	2,19	2,03	1,95	1,86	1,76	1,66	1,53

Table 15 (*continued*)

d) F_{99} for two-sided test															
DF_l	DF_g														
	3	4	5	6	7	8	10	12	15	20	24	30	40	60	120
3	47,47	46,19	45,39	44,84	44,43	44,13	43,69	43,39	43,08	42,78	42,62	42,47	42,31	42,15	41,99
4	24,26	23,15	22,46	21,97	21,62	21,35	20,97	20,70	20,44	20,17	20,03	19,89	19,75	19,61	19,47
5	16,53	15,56	14,94	14,51	14,20	13,96	13,62	13,38	13,15	12,90	12,78	12,66	12,53	12,40	12,27
6	12,92	12,03	11,46	11,07	10,76	10,57	10,25	10,03	9,81	9,59	9,47	9,36	9,24	9,12	9,00
7	10,88	10,05	9,52	9,16	8,89	8,68	8,38	8,18	7,97	7,75	7,65	7,53	7,42	7,31	7,19
8	9,60	8,81	8,30	7,95	7,69	7,50	7,21	7,01	6,81	6,61	6,50	6,40	6,29	6,18	6,06
10	8,08	7,34	6,87	6,54	6,30	6,12	5,85	5,66	5,47	5,27	5,17	5,07	4,97	4,86	4,75
12	7,23	6,52	6,07	5,76	5,52	5,35	5,09	4,91	4,72	4,53	4,43	4,33	4,23	4,12	4,01
15	6,48	5,80	5,37	5,07	4,85	4,67	4,42	4,25	4,07	3,88	3,79	3,69	3,58	3,48	3,37
20	5,82	5,17	4,76	4,47	4,26	4,09	3,85	3,68	3,50	3,32	3,22	3,12	3,02	2,92	2,81
24	5,52	4,89	4,49	4,20	3,99	3,83	3,59	3,42	3,25	3,06	2,97	2,87	2,77	2,66	2,55
30	5,24	4,62	4,23	3,95	3,74	3,58	3,34	3,18	3,01	2,82	2,73	2,63	2,52	2,42	2,30
40	4,98	4,37	3,99	3,71	3,51	3,35	3,12	2,95	2,78	2,60	2,50	2,40	2,30	2,18	2,06
60	4,73	4,14	3,76	3,49	3,29	3,13	2,90	2,74	2,57	2,39	2,29	2,19	2,08	1,96	1,83
120	4,50	3,92	3,55	3,28	3,09	2,93	2,71	2,54	2,37	2,19	2,09	1,98	1,87	1,75	1,61

7.3 Applications to rubber testing

7.3.1 General

Knowledge of the confidence limits for data enables objective assessments to be made of differences in that data. Examples of this are given in 7.3.2 and 7.3.3.

7.3.2 Confidence limits and specification limits

In a stress relaxation test carried out in accordance with ISO 3384, three compounds were tested against a specification requiring a maximum stress relaxation of 20 % over the 7 day test duration. The results obtained were as shown in Table 16.

Table 16 — Percentage stress relaxation measurements

Compound	Test piece			Mean	Standard deviation, s
	1	2	3		
1	22,1	22,6	22,8	22,5	0,36
2	17,5	19,7	18,5	18,6	1,10
3	13,7	14,3	15,9	14,6	1,14

Visual inspection of the results suggests that compound 1 fails, and compounds 2 and 3 pass. For compound 2, however, the mean is close to the limit (i.e. one standard deviation) which makes its true status less clear.

The lower confidence limit of each compound at the 95 % significance level is obtained from the equation:

$$c_L = \bar{x} - (t_{0,95}s)/\sqrt{n} \quad (22)$$

Therefore, for compound 1 with a value for s of 0,36,

$$c_L = 22,5 - (2,92 \times 0,36/\sqrt{3})$$

$$c_L = 21,9$$

In a similar way, Table 17 can be derived for all the compounds.

Table 17 — Lower confidence limits for percentage stress relaxation

Limit	Compound		
%	1	2	3
90	22,1	19,8	15,8
95	21,9	20,5	16,5
99	21,1	23,0	19,2

From these limits, it is over 99 % certain that compound 1 fails and compound 3 conforms to the specification. However, compound 2 conforms to the specification with only a 90 % certainty.

If a further three test pieces of compound 2 were tested, a more definite conclusion could probably be reached. (For example, if the same mean and standard deviation were obtained on the additional tests, the corresponding confidence limits would be 19,3, 19,5 and 20,1. This gives between 95 % and 99 % confidence that compound 2 does pass.)

7.3.3 Comparison of results

The supplier and purchaser of a grade of rubber compound (compound 1) each carry out tear tests in accordance with ISO 34-1:2004, method B, on the same batches to assess their degree of agreement. The results obtained by the two laboratories are presented in Table 18. In considering the results for compound 1, it can be seen that the difference in the means exceeds the $t_{0,95}S$ product so it is more than 95 % certain that the two laboratories are not producing statistically equivalent data.

The same tests were also carried out on a different compound (compound 2). These results are also presented in Table 18.

In the case of compound 2, the difference in the means is bordering on the 99 % significance level. Since laboratory 1 on both occasions has produced the lower strengths, it is probable, though by no means certain on the basis of only two sets of results, that there is a systematic difference between the laboratories. This may be as a result of a different depth of the nick or errors in the force transducer. In such a case, an effort should be made to trace and rectify any deficiency.

Table 18 — Tear tests

Tear measurements in N/mm

Compound	Test piece	Tear measurement		Mean $\bar{x}$		Standard deviation s		Difference between means (absolute value) $\lvert\bar{x}_1 - \bar{x}_2\rvert$	Standard error S	Degrees of freedom	$t_{0,95}$	$t_{0,95}S$
		Lab.1	Lab.2	Lab.1	Lab.2	Lab.1	Lab.2					
1	1	19,0	21,0									
	2	20,6	19,9									
	3	20,2	21,4	19,8	20,9	0,63	0,64	1,1	0,40	8	2,31	0,92
	4	19,4	21,5									
	5	19,9	20,8									
2	1	22,1	25,3									
	2	18,4	23,3									
	3	22,3	24,0	21,4	24,7	1,69	1,64	3,3	1,05	8	2,31	2,43
	4	22,1	27,3									
	5	22,2	23,6									

8 Ranking methods

8.1 Principles

Sometimes, observations cannot be quantified precisely and subjective judgements of merit have to be made. In these cases, the usual quantitative techniques cannot be applied and it is necessary to resort to ranking methods.

8.2 Methodology

8.2.1 Friedman's test

8.2.1.1 The following steps should be taken in this test:

The members of a group of observers independently rank the same samples into order of increasing merit according to previously defined criteria.

The sum of the squares of the differences between each rank sum and the mean rank sum is determined and compared with the critical value corresponding to a given level of significance.

If the observed factor is greater than the critical factor, then there is a significant difference (at the given confidence level).

8.2.1.2 Thus, if there are m observers and n samples, each observer independently assigns the number 1 to the best sample, 2 to the next best and so on down to n to the poorest. If two or more samples are judged to be equally good, then they are assigned the same rank number, this being simply the average rank of the group. For each sample, the rank sum S_R is determined from the individual rank values R by the equation:

$$S_i = \sum_{j=i}^{m} R_{ij} \tag{23}$$

The mean rank sum, $\overline{S}_R$, is the average of the n rank sums as given by the equation:

$$\overline{S} = \frac{\sum_{i=1}^{n} S_i}{n} \tag{24}$$

Friedman's statistic, K, is then given by the equation:

$$K = \sum_{i-1}^{n} (S_i - \overline{S})^2 \tag{25}$$

If $K > K_{cr}$, the samples are significantly different. Values of K_{cr} for the 95 % confidence level are tabulated in Table 19.

Table 19 — Friedman's test: critical values K for a level of significance of 0,05

m[a]	Number of observations in the sample N												
	3	4	5	6	7	8	9	10	11	12	13	14	15
2	—	20	38	64	96	138	192	258	336	429	538	664	808
3	18	37	64	104	158	225	311	416	542	691	865	1 063	1 292
4	26	52	89	144	217	311	429	574	747	950	1 189	1 460	1 770
5	32	65	113	183	277	396	547	731	950	1 210	1 512	1 859	2 254
6	42	76	137	222	336	482	664	887	1 155	1 469	1 831	2 253	2 738
7	50	92	167	272	412	591	815	1 086	1 410	1 791	2 233	2 740	3 316
8	50	105	190	310	471	676	931	1 241	1 612	2 047	2 552	3 131	3 790
9	56	118	214	349	529	760	1 047	1 396	1 813	2 302	2 871	3 523	4 264
10	62	131	238	388	588	845	1 164	1 551	2 014	2 558	3 189	3 914	4 737
11	66	144	261	427	647	929	1 280	1 706	2 216	2 814	3 508	4 305	5 211
12	72	157	285	465	706	1 013	1 396	1 862	2 417	3 070	3 827	4 697	5 685
13	78	170	309	504	764	1 098	1 512	2 017	2 618	3 326	4 146	5 088	6 159
14	84	183	333	543	823	1 182	1 629	2 172	2 820	3 581	4 465	5 479	6 632
15	90	196	356	582	882	1 267	1 745	2 327	3 021	3 837	4 784	5 871	7 106

[a] m = number of observers ranking the observations in the sample.

8.2.1.3 Where a significant difference is shown, the mean rank for each sample can be determined from the equation:

$$\overline{R}_i = S_i / m \tag{26}$$

although it should not be assumed that there is necessarily a significant difference between any pair of mean rank values even though there is a significant difference taken across the n samples as a whole.

8.2.1.4 Whether or not significance is obtained depends on the differences between the samples as well as on the degree of agreement between the observers. The coefficient of concordance between the observers is given by the equation:

$$C = \frac{12K}{nm^2(n^2-1)} \tag{27}$$

8.2.1.5 This parameter may take any value between 0 (no agreement) and 1 (complete agreement). In order for high degrees of concordance to be achieved, the rankings should be based on a single criterion which has been clearly described.

8.2.2 The outside count test

8.2.2.1 This is a rough and ready method for the comparison of two specific samples out of a total of n. It can be particularly useful where one of the two samples is a reference material being used for comparison purposes.

8.2.2.2 The procedure is as follows:

a) in the sample containing the highest value, count the number of values which are higher than the highest value in the other sample;

b) count the number of values in the other sample which are lower than the lowest value in the first sample.

If the sum of these two counts is greater than six, it can be concluded that the two samples are different at the 95 % confidence level.

8.3 Applications to rubber testing

Ten vulcanizates containing different antiozonants were simultaneously tested for ozone resistance in accordance with ISO 1431-1, after which five observers independently ranked the 10 compounds for degree of cracking using crack length as the criterion. Table 20 and Table 21, respectively, give the results and statistical calculations from this test.

Table 20 — Ozone resistance test results

Vulcanizate	Observer				
	A	B	C	D	E
1	4	3½	3	3	4
2	1	2	2	3	2
3	5½	5	4	6	4
4	5½	6	6	5	6
5	2	1	1	1	1
6	3	3½	5	3	4
7	8	7	7½	9	10
8	10	9	9	8	8
9	7	8	7½	7	7
10	9	10	10	10	9

Table 21 — Statistical calculations for ozone resistance tests

Vulcanizate	Sum	Mean sum	Difference	Mean rank
1	171/2	271/2	− 10	3,5
2	10	271/2	− 171/2	2,0
3	241/2	271/2	− 3	4,9
4	281/2	271/2	+ 1	5,7
5	6	271/2	− 211/2	1,2
6	181/2	271/2	− 9	3,7
7	411/2	271/2	+ 14	8,3
8	44	271/2	+ 161/2	8,8
9	361/2	271/2	+ 9	7,3
10	48	271/2	+ 201/2	9,6

From these data

$$K = (-10)^2 + (-171/2)^2 + (-3)^2 + \ldots + (201/2)^2 = 1\,929$$

where K = Friedman's K-value

For $n = 10$ and $m = 5$, $K_{cr} = 731$, hence the different antiozonants are producing statistically significant differences in ozone resistance between the compounds.

The coefficient of concordance C can be calculated as

$$C = 12 \times 192\,9 / [5^2 \times (10^3 - 10)]$$

$$= 0{,}94$$

This indicates that there is a high degree of agreement in the judgements of the five observers.

9 Criteria for rejecting outliers

9.1 Principles

9.1.1 There are occasions when a single result in a test sequence can appear to be out of line with the rest of the data. The rejection of such a result as an outlier, which would otherwise distort what is considered to be the true data represented by the other results, is sometimes considered. This is a course of action which should be avoided. Rejection of results without good cause can lead to serious distortion of the true distribution and will lead particularly to a significant under-estimate of the standard deviation.

9.1.2 A result should not be rejected unless one of the following cases applies:

a) There is clear physical evidence that the result has been caused by some recognizable fault in the sample.

b) An objective statistical test gives a strong indication that the result is unlikely to have arisen purely by chance. As with any statistical test, a given confidence level should be arbitrarily assigned as the criterion for rejection.

9.1.3 A result that is shown to be unusual at between the 95 % to 99 % confidence level should be marked as a straggler (see the ISO 5725 series).

A result that exceeds the 99 % confidence level should be marked as an outlier and should then be eliminated from the analysis (some workers use four standard deviations as the level).

In both cases, the test piece which gave rise to the suspect result should be examined for evidence of its abnormality. In the case of a straggler, lack of any such evidence should cause the data to be retained in the analysis but, if there is clear physical evidence of abnormality, then its result can be discarded.

9.1.4 In addition to the testing of individual observations in a set, it is possibly appropriate, as for example in inter-laboratory trials, to test for outliers in terms of the means of the series of tests performed. However, prior to this a test for standard deviation should be made. If, for example, one laboratory's standard deviation is significantly different to that which could be expected on the basis of the other laboratories' standard deviations, this laboratory's results cannot be taken as coming from the same population as the other laboratories and should, therefore, be discarded. As before, rejected or straggling data should be critically examined to try to ascertain the cause with the view to correcting any faults.

9.1.5 Examination of the outlying data can show that a simple calculation error or similar quantifiable fault had occurred before the results were reported and that this can be corrected at source. The corrected data can then be entered in place of the originals and the statistical tests re-applied.

9.2 Methodology

9.2.1 General

The assumption is made in the following tests that the data being examined are normally distributed (see 6.2.5) or have been transformed into a form that is normally distributed (see 6.2.4).

9.2.2 Dixon's test

9.2.2.1 Dixon's test is applied to individual observations or to the means of sets of observations and it is assumed that both abnormally large or small observations are to be equally tested for rejection against Dixon's criterion.

9.2.2.2 The procedure is as follows:

a) Arrange the n observations in ascending numerical order; i.e. x_1 the smallest through to x_n the largest.

b) Derive Dixon's quotient, Q, from step c), d) or e) as appropriate.

c) If $3 \leqslant n \leqslant 7$, then record the larger of

$$\frac{x_2 - x_1}{x_n - x_1} \text{ and } \frac{x_n - x_{n-1}}{x_n - x_1} \tag{28}$$

d) If $8 \leqslant n \leqslant 12$, then record the larger of

$$\frac{x_2 - x_1}{x_{n-1} - x_1} \text{ and } \frac{x_n - x_{n-1}}{x_n - x_2} \tag{29}$$

e) If $n > 12$, then record the larger of

$$\frac{x_3 - x_1}{x_{n-2} - x_1} \text{ and } \frac{x_n - x_{n-2}}{x_n - x_3} \tag{30}$$

f) Compare Dixon's quotient, Q, so derived, to the data given in Table 22. The following conclusions can be made:

1) if Q exceeds the 5 % value but is less than the 1 % value, the first or last (according to which of the two ratio calculations gave the higher Q) is marked as a straggler;

2) if Q exceeds the 1 % value, the result is marked as an outlier and rejected. In this case, the test can be repeated with the $(n - 1)$ remaining results.

NOTE In Table 22, n equals the number of observations. This version of Dixon's test is as published in *Statistical Manual*, Edited by E.L. Crow, F.A. Davis and M.W. Maxfield, Dover Publications, 1960.

Table 22 — Critical values for Dixon's test

Criterion	n	Critical values	
		5 %	1 %
$Q10$	3	0,970	0,994
	4	0,829	0,926
	5	0,710	0,821
	6	0,628	0,740
	7	0,569	0,680
$Q11$	8	0,608	0,717
	9	0,564	0,672
	10	0,530	0,635
	11	0,502	0,605
	12	0,479	0,579
$Q22$	13	0,611	0,697
	14	0,586	0,670
	15	0,565	0,647
	16	0,546	0,627
	17	0,529	0,610
	18	0,514	0,594
	19	0,501	0,580
	20	0,489	0,567
	21	0,478	0,555
	22	0,468	0,544
	23	0,459	0,535
	24	0,451	0,526
	25	0,443	0,517
	26	0,436	0,510
	27	0,429	0,502
	28	0,423	0,495
	29	0,417	0,489
	30	0,412	0,483
	31	0,407	0,477
	32	0,402	0,472
	33	0,397	0,467
	34	0,393	0,462
	35	0,388	0,458
	36	0,384	0,454
	37	0,381	0,450
	38	0,377	0,446
	39	0,374	0,442
	40	0,371	0,438

9.2.3 Cochran's test for variance

9.2.3.1 Cochran's test is applied to the variances of sets of observations and should be applied before Dixon's test for means. The following assumptions are made:

a) It is assumed that there are the same number of observations (replicates) in each set. Some relaxation of this condition is possible without seriously compromising the test, but every effort should be made to satisfy it and the number of exceptions should be kept small.

b) It is assumed that only abnormally large variances are to be examined for rejection (i.e. it is a one-sided test, unlike Dixon's test).

9.2.3.2 The procedure is as follows.

a) Given a set of n standard deviations, s_i, calculate Cochran's quotient, C, as follows:

$$C = s_{\max}{}^2 / \sum_{i=1}^{n} s_i{}^2 \tag{31}$$

where $s_{\max}$ is the largest standard deviation in the group of n.

NOTE An exception is in the case where the number of replicates is 2 when the range, w, is substituted for the standard deviation, s.

b) Compare C with the critical values given in Table 23, which can be used for replicates between 2 and 6 inclusive and for up to 40 sets of results. The following deductions can be made:

1) if the observed C-value is greater than the 5 % critical value and less than the 1 % critical value, the standard deviation is marked as a straggler;

2) if the observed C-value exceeds the 1 % value the set of data producing that standard deviation is rejected as an outlier.

NOTE For results outside the scope of Table 23, reference should be made to more extensive statistical tables.

Table 23 — Critical values for Cochran's test

p	$n = 2$		$n = 3$		$n = 4$		$n = 5$		$n = 6$	
	1 %	5 %	1 %	5 %	1 %	5 %	1 %	5 %	1 %	5 %
2	—	—	0,995	0,975	0,797	0,939	0,959	0,906	0,937	0,877
3	0,993	0,967	0,942	0,871	0,883	0,798	0,834	0,746	0,793	0,707
4	0,968	0,906	0,864	0,768	0,781	0,684	0,721	0,629	0,676	0,590
5	0,928	0,841	0,788	0,684	0,696	0,598	0,633	0,544	0,588	0,506
6	0,883	0,781	0,722	0,616	0,626	0,532	0,564	0,480	0,520	0,445
7	0,838	0,727	0,664	0,561	0,568	0,480	0,508	0,431	0,466	0,397
8	0,794	0,680	0,615	0,516	0,521	0,438	0,463	0,391	0,423	0,360
9	0,754	0,638	0,573	0,478	0,481	0,403	0,425	0,358	0,387	0,329
10	0,718	0,602	0,536	0,445	0,447	0,373	0,393	0,331	0,357	0,303
11	0,684	0,570	0,504	0,417	0,418	0,348	0,366	0,308	0,332	0,281
12	0,653	0,541	0,475	0,392	0,392	0,326	0,343	0,288	0,310	0,262
13	0,624	0,515	0,450	0,371	0,369	0,307	0,322	0,271	0,291	0,246
14	0,599	0,492	0,427	0,352	0,349	0,291	0,304	0,255	0,274	0,232
15	0,574	0,471	0,407	0,335	0,332	0,276	0,288	0,242	0,259	0,220
16	0,553	0,452	0,388	0,319	0,316	0,262	0,274	0,230	0,246	0,208
17	0,532	0,434	0,372	0,305	0,301	0,250	0,261	0,219	0,234	0,198
18	0,514	0,418	0,356	0,293	0,288	0,240	0,249	0,209	0,223	0,189
19	0,496	0,403	0,343	0,281	0,276	0,230	0,238	0,200	0,214	0,181
20	0,480	0,389	0,330	0,270	0,265	0,220	0,229	0,192	0,205	0,174
21	0,465	0,377	0,318	0,261	0,255	0,212	0,220	0,185	0,197	0,167
22	0,450	0,365	0,307	0,252	0,246	0,204	0,212	0,178	0,189	0,160
23	0,437	0,354	0,297	0,243	0,238	0,197	0,204	0,172	0,182	0,155
24	0,425	0,343	0,287	0,235	0,230	0,191	0,197	0,166	0,176	0,149
25	0,413	0,334	0,278	0,228	0,222	0,185	0,190	0,160	0,170	0,144
26	0,402	0,325	0,270	0,221	0,215	0,179	0,184	0,155	0,164	0,140
27	0,391	0,316	0,262	0,215	0,209	0,173	0,179	0,150	0,159	0,135
28	0,382	0,308	0,255	0,209	0,202	0,168	0,173	0,146	0,154	0,131
29	0,372	0,300	0,248	0,203	0,196	0,164	0,168	0,142	0,150	0,127
30	0,363	0,293	0,241	0,198	0,191	0,159	0,164	0,138	0,145	0,124
31	0,355	0,286	0,235	0,193	0,186	0,155	0,159	0,134	0,141	0,120
32	0,347	0,280	0,229	0,188	0,181	0,151	0,155	0,131	0,138	0,117
33	0,339	0,273	0,224	0,184	0,177	0,147	0,151	0,127	0,134	0,114
34	0,332	0,267	0,218	0,179	0,172	0,144	0,147	0,124	0,131	0,111
35	0,325	0,262	0,213	0,175	0,168	0,140	0,144	0,121	0,127	0,108
36	0,318	0,256	0,208	0,172	0,165	0,137	0,140	0,119	0,124	0,106
37	0,312	0,251	0,204	0,168	0,161	0,134	0,137	0,116	0,121	0,103
38	0,306	0,246	0,200	0,164	0,157	0,131	0,134	0,113	0,119	0,101
39	0,300	0,242	0,196	0,161	0,154	0,129	0,131	0,111	0,116	0,099
40	0,294	0,237	0,192	0,158	0,151	0,126	0,128	0,108	0,114	0,097

NOTE n is the number of results per cell;

p is the number of laboratories at the given level.

9.3 Applications to rubber testing

9.3.1 General

The application of tests for outliers can, in principle, be applied to any set of data, but it is most often applied in the case of inter-laboratory testing trials.

9.3.2 Dixon's test applied to individual results

A series of eight replicate compression set results were obtained on type 1 test pieces in accordance with ISO 815 as shown in Table 24.

Table 24 — Compression set results

Result number	Result %
1	24,1
2	25,9
3	24,2
4	25,1
5	10,1
6	28,1
7	18,3
8	26,9

An initial brief examination of these data suggests that the 10,1 result is so far out of line that it ought to be ignored in calculating the mean and standard deviation. However, results such as these should be checked against Dixon's criterion. In this case, the following conclusion is reached.

Table 25 shows the results sorted into ascending order.

Table 25 — Sorted compression set results

Ascending order	Result %
1	10,1
2	18,3
3	24,1
4	24,2
5	25,1
6	25,9
7	26,9
8	28,1

From Table 25 it is seen that:

$x_1 = 10{,}1;$

$x_2 = 18{,}3;$

$x_7 = 26{,}9;$

$x_8 = 28{,}1.$

Dixon's quotients for eight replicates are

$$\frac{18{,}3 - 10{,}1}{26{,}9 - 10{,}1} \text{ and } \frac{28{,}1 - 26{,}9}{28{,}1 - 18{,}3} \quad (= 0{,}488 \text{ and } 0{,}122)$$

The larger of these is taken and compared to the critical value for the 95 % confidence level, which is 0,608 according to Table 22. Since the calculated statistic is less than the critical value, there is no justification for rejecting the low result.

9.3.3 Cochran's variance test

An inter-laboratory trial involving seven laboratories produced the results shown in Table 26 for a volume swell test carried out in accordance with ISO 1817.

Table 26 — Volume swell test 1

Laboratory	Result			Mean	Standard deviation
	1 %	2 %	3 %		
1	17,8	18,1	18,1	18,0	0,173
2	19,6	19,5	19,6	19,6	0,058
3	22,9	22,9	22,4	22,7	0,289
4	19,9	19,7	19,7	19,8	0,115
5	13,4	14,2	15,1	14,2	0,850
6	22,5	22,1	22,0	22,2	0,265
7	20,8	20,5	20,7	20,7	0,153

The result for laboratory 5 appears to have a suspiciously low mean and a high standard deviation. Therefore when Cochran's test is applied first of all to the standard deviations the following results are obtained:

$s_{max}^2 = 0{,}852^2$

$= 0{,}722\ 5$

$\Sigma s^2 = 0{,}173^2 + 0{,}058^2 + \ldots + 0{,}153^2$

$= 0{,}946\ 2$

Cochran's ratio $= 0{,}722\ 5 / 0{,}946\ 2$

$= 0{,}764$

For three replicates and seven laboratories, Cochran's critical value for the 99 % confidence level is 0,664 and so, as the test statistic is greater than this, the rejection of the data from laboratory 5 on statistical grounds is justified and it is not necessary to test the low mean value. The results should then be checked back to their source to see if an explanation can be found and possible corrective action taken.

9.3.4 Dixon's test applied to a group of mean values

In an inter-laboratory trial involving six laboratories, the results for a volume swell test were as shown in Table 27.

Table 27 — Volume swell test 2

Laboratory	Result			Mean	Standard deviation
	1 %	2 %	3 %		
1	13,5	13,8	13,8	13,7	0,173
2	10,8	13,0	12,6	12,1	1,173
3	12,9	13,0	12,7	12,9	0,153
4	10,9	11,2	14,2	12,1	1,825
5	14,2	14,2	14,4	14,3	0,115
6	19,7	20,8	18,9	19,8	0,954

The result for laboratory 6 appears to have a high mean value and the standard deviations appear to be quite variable, but no one result stands out as being abnormally large.

Testing by Cochran's test, first of all, confirms that no standard deviation is so large as to justify rejection of the data. Therefore Dixon's test is applied and the following is calculated:

$$\frac{x_2 - x_1}{x_6 - x_1} = 0 \quad \text{and} \quad \frac{x_6 - x_5}{x_6 - x_1} = 0{,}714 \tag{32}$$

The critical value for

— the 95 % confidence level is 0,628;

— the 99 % confidence level is 0,740.

Hence laboratory 6 is seen to be a straggler, but its data should not be rejected unless an investigation shows some fault in the procedure or equipment used.

10 Analysis of variance (ANOVA)

10.1 Principles

The variability in the results observed from a test arises from a number of sources (in practice, a very large number of sources), ranging from variations in the quality of the raw materials from which the sample was made, through the compounding and moulding processes, into the sampling and testing procedure itself. Analysis of variance is a technique which can be used to isolate and estimate the effect of those sources of variation which are having a significant effect on the measurements.

In practice, it is neither possible, nor necessary, to quantify the effect of every conceivable source of variation. Instead, it is sufficient, for any particular case, to examine the effect of the variables that are regarded as

being the most likely to have an influence or which it is desired to test for their influence (for example, the effect of different sources of carbon black, mixing time and moulding temperature on the abrasion resistance of the compound). All other factors are kept as constant as possible and the factors of interest are varied in some known way. Replicate tests carried out at each level of each factor then determine the within-sample variation, often referred to as the experimental error, against which the effects due to the factors of interest taken individually and in combination can be assessed.

10.2 Methodology

10.2.1 General

A full development of the methodology is outside the scope of this International Standard, and reference should be made to any of the excellent texts on the subject for details. The Bibliography at the end of this International Standard lists a selection of useful reference works. Many statistical software packages exist, and many spreadsheet packages have built-in statistical functions which enable the mathematics to be evaluated quickly without the need for detailed understanding of the underlying equations. It is recommended that appropriate computer programmes be used wherever possible. However, some specific examples are enumerated for the benefit of users without access to such software.

10.2.2 One factor with an equal number of replicates

10.2.2.1 The simplest case to consider is that of one factor (e.g. carbon black) at n levels (parts per hundred of rubber) each replicated r times, giving a total number of observations N. The total number of observations is calculated from the equation:

$$N = r \times n \tag{33}$$

The following sequence of calculations should be followed.

NOTE See the comments in Annex D on truncation errors.

a) Calculate the total sums of squares, S_t, given by the equation:

$$S_t = \Sigma(x_{ij} - \bar{x})^2 \tag{34}$$

where x_{ij} is the value of the jth replicate ($1 \leqslant j \leqslant r$) of the ith factor ($1 \leqslant i \leqslant n$).

b) Calculate the total degrees of freedom, v_t, given by the equation:

$$v_t = N - 1 \tag{35}$$

c) Calculate the total mean square, M_t, from a) and b) using the equation:

$$M_t = S_t / v_t \tag{36}$$

d) Calculate the between-factor sums of squares, S_b, given by the equation:

$$S_b = \frac{1}{r}\Sigma t_i^{\,2} - \frac{1}{N}(\Sigma x_{ij})^2 \tag{37}$$

where t_i is the sum of the r replicates of the ith level of the factor calculated from the equation:

$$t_i = \sum_{j=1}^{r} x_{ij} \tag{38}$$

e) Calculate the degrees of freedom, v_b, associated with the between-factor sums of squares given by the equation:

$$v_b = n - 1 \tag{39}$$

f) Calculate the between-factor mean square, M_b, from d) and e) using the equation:

$$M_b = S_b/v_b \tag{40}$$

g) Calculate the within-factor sums of squares, S_w, given by the equation:

$$S_w = S_t - S_b \tag{41}$$

h) Calculate the within-factor degrees of freedom, v_w, given by the equation:

$$v_w = v_t - v_b \tag{42}$$

i) Calculate the within-factor mean square, M_w, given by the equation:

$$M_w = S_w/v_w \tag{43}$$

10.2.2.2 These statistics can be usefully summarized in Table 28.

Table 28 — One-factor statistics summary

Source of variation	Between-factor	Within-factor	Total
Sum of squares	S_b	S_w	S_t
Degrees of freedom	v_b	v_w	v_t
Mean square	M_b	M_w	M_t

10.2.2.3 Snedecor's F-test is then applied to the ratio of M_b to M_w, with v_b as the degrees of freedom for the greater mean square and v_w as the degrees of freedom for the lesser mean square. (Clearly, if $M_b < M_w$, the between-factor variation is insignificant compared to the experimental error and the effect of the factor is to have no measurable influence on the property being determined.)

If $M_b/M_w > F(5, v_b, v_w)$, then there is a greater than 95 % probability that the different levels of the factor are having a significant effect on the property.

10.2.3 One factor with a variable number of replicates

Where the number of replicates, r, is not a constant for each level of the factor, as it is in 10.2.2, the analysis proceeds as in 10.2.2, but with the following modifications.

The total number of observations, N, is given by the equation:

$$N = \sum_{i=1}^{n} r_i \tag{44}$$

The between-factor sum of squares, S_b, is given by the equation:

$$S_b \sum \frac{t_i^2}{r_i} - \frac{1}{N} \sum x_{ij}^2 \tag{45}$$

All the other factors and the F-test are as before.

10.2.4 Two (and over) factor analysis of variance

With the addition of extra factors in the analysis, not only is there the potential for each factor to influence the measured property, but also the factors can be influenced by each other, giving an interaction (synergistic) effect.

An everyday example which illustrates this is the sweetness of a cup of tea. This depends on

a) how much sugar is added to the tea; and

b) how much the tea is stirred.

Although the analysis proceeds in a similar way to that described in 10.2.2 or 10.2.3, the detailed process is more complex and the method for two factors is given in Annex G. The same process can be extended to three and more factors, and Annex G also illustrates how the sums of squares for a three-factor analysis can be processed for possible interaction effects.

10.3 Applications to rubber testing

10.3.1 Analysis of variance is a powerful tool in assessing the separate importance particular components of a rubber compound, its processing, etc., have on the resulting properties.

10.3.2 A series of compounds having differing levels of carbon black and processing oil were tested for abrasion resistance in accordance with ISO 4649. It was expected that increasing the black level would improve the abrasion resistance but increasing the oil level would make for better processing. The results obtained are shown in Table 29. It was necessary to determine how significant the two factors were from these results.

Table 29 — Abrasion volume loss results

All results in mm^3

a) Original results					
Result	**Oil level**	**Black level**			
		60	**80**	**100**	**120**
1	0	273	256	202	188
2		233	262	215	195
3		273	242	261	177
1	5	288	257	244	242
2		260	271	229	203
3		313	311	245	201
1	10	269	247	249	217
2		317	253	220	215
3		245	262	232	203
1	20	231	270	222	230
2		298	307	227	214
3		287	278	203	242
b) Modified results					
Result	**Oil level**	**Black level**			
		60	**80**	**100**	**120**
1	0	2,73	2,56	2,02	1,88
2		2,33	2,62	2,15	1,95
3		2,73	2,42	2,61	1,77
1	5	2,88	2,57	2,44	2,42
2		2,60	2,71	2,29	2,03
3		3,13	3,11	2,45	2,01
1	10	2,69	2,47	2,49	2,17
2		3,17	2,53	2,20	2,15
3		2,45	2,62	2,32	2,03
1	20	2,31	2,70	2,22	2,30
2		2,98	3,07	2,27	2,14
3		2,87	2,78	2,03	2,42

NOTE The modified results in b) are the original results divided by 100. This is for the convenience of tabulating small numbers.

10.3.3 The calculation procedure described in G.1 was applied to the modified results and Table 30 constructed.

Table 30 — Table of sums

Oil level (factor B)	**Black level** (factor A) 60	80	100	120	**Sums of A** $({}_{B}X_{j})$
0	7,79	7,60	6,78	5,60	27,77
5	8,61	8,39	7,18	6,46	30,64
10	8,31	7,62	7,01	6,35	29,29
20	8,16	8,55	6,52	6,86	30,09
Sums of B $({}_{A}X_{i})$	32,87	32,16	27,49	25,27	

The following factors are derived from Table 29 and Table 30:

— factor T = 17,79;

— number of values of A = 4;

— number of values of B = 4;

— number of replicates = 3;

— correction factor CF = 289,07;

— ${}_{AB}X^{2}$ = 879,43.

Therefore the following values are calculated:

— S_a = 3,352;

— S_b = 0,388;

— S_{ab} = 0,337;

— S_r = 1,483;

— S_t = 5,559;

— DF_a = 3;

— DF_b = 3;

— DF_{ab} = 9;

— DF_r = 33;

— DF_t = 48;

— M_a = 1,117;

— M_b = 0,129;

— M_{ab} = 0,037;

— M_r = 0,045;

— M_t = 0,116.

10.3.4 It is conventional to summarize the data in the form of an analysis of variance table, as shown in Table 31.

Table 31 — Analysis of variance

Source	Sums of squares	Degrees of freedom	Variance estimate
Factor A	3,352	3	1,117
Factor B	0,388	3	0,129
Interaction	0,337	9	0,037
Residual	1,483	32	0,045
Total	5,560	47	0,116

10.3.5 When the interaction term is considered first and the ratio of M_{ab} to M_r taken:

$M_{ab}/M_r = 0{,}83$

As this is less than 1, it is not significant and S_{ab} can be pooled with S_r to give:

$S'_r = 1{,}820$;

$DF'_r = 42$;

$M'_r = 0{,}043$.

where

S'_r is the new value of S_r;

DF'_r is the new value of DF_r;

M'_r is the new value of M_r.

10.3.6 When the two main factors, A and B, are considered:

$M_a / M'_r = 25{,}98$;

$M_b / M'_r = 3{,}00$.

The critical F-values for 3 by 41 degrees of freedom are:

$F_{cr} = 2{,}84$ for a 95 % confidence level;

$F_{cr} = 4{,}31$ for a 99 % confidence level.

10.3.7 The conclusion is that both factors are significant, although the oil factor is only just significant at the 95 % level while the carbon black factor is significant at well over the 99 % level. There is no significant interaction between the factors. Thus, the oil level may be increased in order to improve the processability of the compound without having too damaging an effect on the abrasion resistance of that compound.

NOTE The above example could also be analysed using the least-squares regression method which would enable a quantitative estimate of the relationship between the variables and their interaction to be determined.

11 Regression analysis

11.1 Principles

11.1.1 When a series of tests is undertaken in which a test parameter, for example compression set, is measured at different values of an independent variable such as time or temperature, it is to be expected that some form of functional relationship will exist between them. However, as shown in Clause 10, there are many sources of variation in the process, with the net result that the observed data do not fit perfectly to a single curvilinear function but are scattered more or less around the function of choice. The functional relationship between the dependent (measured) and the independent (controlled) variables is known as the regression line.

11.1.2 The model equation to be chosen can be deduced from scientific laws, but generally this is not the case and an empirical relationship should be resorted to. Under these circumstances, the simplest functional relationship which adequately describes the observations should be used. Thus, for compression set as a function of compression time (temperature being kept constant) a linear relationship between set and the logarithm of time can be expected to give an excellent representation of the data within the experimental error observed. Clearly, however, the true functional relationship cannot be linear as compression sets below 0 % or above 100 % are not possible. Thus, a transition function would be better able to describe the relationship over a wider time span than has been encountered in the experiments. A simple alternative is to replace the compression set (cs) value as the ordinate by the function:

$$\log[cs/(100 - cs)]$$

11.1.3 In considering the form of function to use, account should be taken of three points:

a) The coefficients of the function can possibly be derived analytically (e.g. by the method of least squares) or, if not, an iterative method used as an alternative.

b) The benefit to be gained from the more complicated function should be sufficient to justify the extra effort in deriving its coefficients.

c) An assessment should be made as to whether or not the observed data will have to be extrapolated to reach the conditions of particular interest.

   Examples of tests where this applies are:

   1) ageing for short periods at high temperature to predict behaviour over long periods at lower temperatures;

   2) estimating the stress relaxation at long times from short time tests.

If interpolation or very short extrapolations are required, as for example in the estimation of the temperature at which 70 % retraction occurs in a temperature of retraction test, then the smallest-order polynomial that gives the correct trends in the data should be chosen.

For extrapolation, however, polynomials are notoriously dangerous and the higher the order of the polynomial, the worse this tendency becomes. In these circumstances, it may be necessary to resort to the use of more complex functions in order to avoid predictions which would be wildly inaccurate. Another reason for not using a more complex function, however, apart from the difficulty of deriving its coefficients, is that it does not necessarily represent the observed data quite as well as the simpler function does, even though it is safer to use it for extrapolation.

11.1.4 There are now available several powerful computer programmes for personal computers which can make curve fitting little more cumbersome than entering the data and selecting a function or functions from the library of built-in functions. It is recommended that such programmes be used wherever possible to reduce the time and effort required in performing the analysis.

It is also worth noting that many relationships can be reduced to linear form, which is especially easy to solve, by means of transformations such as logarithm, reciprocal or roots.

11.2 Methodology

11.2.1 General

The method of least squares is presented here for polynomials (or any functions that can be reduced to polynomials) up to the third order as an illustration of the technique. These will cover most applications in the rubber industry. For the derivation of the equations and also for the development of higher-order polynomials, reference should be made to standard mathematical textbooks. (See the Bibliography at the end of this International Standard which lists a selection of useful reference works.)

There are many iterative techniques available for fitting curves to functions that cannot be processed by the least-squares method, but again these are outside the scope of this International Standard and reference should be made to mathematical textbooks.

11.2.2 Linear least squares

11.2.2.1 Consider a set of results, y, obtained at a set of conditions, x, there being a total of n data pairs. The summation of terms is carried out over all n data pairs.

The simplest, linear, form of regression line can be written as:

$$y = a + bx \tag{46}$$

where

a is the intercept on the y-axis when $x = 0$;

b is the slope of the regression.

11.2.2.2 To calculate the best estimates for a and b, first calculate the following factors:

C_{11} as given by the equation:

$$C_{11} = \Sigma(x^2) - \frac{(\Sigma x)^2}{n} \tag{47}$$

C_{yy} as given by the equation:

$$C_{yy} = \Sigma(y^2) - \frac{(\Sigma y)^2}{n} \tag{48}$$

C_{y1} as given by the equation:

$$C_{y1} = \Sigma(xy) - \frac{(\Sigma x \Sigma y)}{n} \tag{49}$$

11.2.2.3 The coefficients are then calculated using the equations:

$$a = \frac{(\Sigma y - b\Sigma x)}{n} \tag{50}$$

$$b = \frac{C_{y1}}{C_{11}} \tag{51}$$

Whether this regression is statistically significant can be tested by calculating a further factor, D, which is given by the equation:

$$D = b\Sigma x \tag{52}$$

The variance ratio for the regression is then given by the equation:

$$F_r = D\frac{n-2}{C_{yy} - D} \tag{53}$$

This F_r value should then be compared with tables of Snedecor's F-values with 1 degree of freedom for the greater mean square and $(n - 2)$ degrees of freedom for the lesser mean square. The regression is significant at the given confidence level if F_r is greater than the tabulated value of F.

11.2.3 Quadratic least squares

The regression line is here assumed to be of the form:

$$y = a + bx + cx^2 \tag{54}$$

Calculation of the factors required for this analysis are given in Annex H. The value of F_r in this case is compared to the tabulated F-values for 2 degrees of freedom for the greater mean square and $(n - 3)$ degrees of freedom for the lesser mean square.

11.2.4 Cubic least squares

The regression line is here assumed to be of the form:

$$y = a + bx + cx^2 + dx^3 \tag{55}$$

Calculation of the factors required for this analysis are given in Annex H and the value of F_r is compared with the tabulated F-values for 3 degrees of freedom for the greater mean square and $(n - 4)$ degrees of freedom for the lesser mean square.

11.3 Applications to rubber testing

11.3.1 General

Regression analysis allows the quantitative relationships derived between compounding or experimental features and the physical properties to be derived.

11.3.2 The effect of temperature on compression set

11.3.2.1 In a series of tests in accordance with ISO 815 examining the value of compression set after 7 days ageing at various temperatures, the data given in Table 32 were recorded.

Table 32 — Compression set measurements after 7 days' ageing

Temperature	Result			Mean
°C	1 %	2 %	3 %	
70	21,3	27,4	25,5	24,7
85	29,6	29,2	33,3	30,7
100	36,8	34,7	38,5	36,7
125	47,2	44,8	48,0	46,6
150	57,7	58,5	56,7	57,6

11.3.2.2 From the laws of chemical kinetics, it is reasonable to postulate that a functional relationship of the Arrhenius kind can be applicable to the data. Thus the compression set, cs, can take the form shown in the equation:

$$\text{cs} = \alpha\exp(\beta/T) \tag{56}$$

where

α and β are constants;

T is the temperature, in degrees Kelvin (absolute temperature).

This function is not directly accessible to a least-squares method of determining α and β, but it is readily transformed into one by taking natural logarithms as shown in the equation:

$$\ln(\text{cs}) = \ln(\alpha) + \frac{\beta}{\theta + 273} \tag{57}$$

where θ is the temperature, in degrees Celsius.

This function has the same form as the equation:

$$y = a + bx \tag{58}$$

where

$y = \ln(\text{cs})$;

$x = 1/(\theta + 273)$.

Thus, using the mean values for compression set as the source of the dependent variable, y, the transformed table is as shown in Table 33.

Table 33 — Transformed compression set variables

x 10^{-3}	y
2,92	3,21
2,79	3,42
2,68	3,60
2,51	3,84
2,36	4,05

11.3.2.3 From the various summation terms given in 11.2.2, the following factors are derived:

$C_{11} = 1{,}93 \times 10^{-7}$;

$C_{yy} = 0{,}446$;

$C_{y1} = -\,2{,}93 \times 10^{-4}$;

$D = 0{,}445$;

$F_r = 1\,265$.

The regression coefficients are:

$a = 7{,}66$;

$b = -1\ 520$.

11.3.2.4 From the transformation applied to linearize the function as shown in 11.3.2.2:

$a = \ln(\alpha)$;

$b = \beta$.

Hence

$\alpha = 2\ 120$;

$\beta = -1\ 520$.

The regression equation is:

$$cs = 2\ 120 \exp\left(\frac{-1\ 520}{\theta + 273}\right) \quad (59)$$

11.3.2.5 The variance ratio is significant at well over the 95 % confidence level and, to illustrate the goodness of fit, the regression value of the compression set can be calculated and compared to the experimental value as shown in Table 34.

Table 34 — Comparison of compression set values

Temperature °C	Observed set %	Calculated set %
70	24,7	25,2
85	30,7	30,4
100	36,7	36,0
125	46,6	46,5
150	57,6	58,3

11.3.3 Effect of ageing on tensile strength

11.3.3.1 A rubber compound was heat-aged at 70 °C for a period of 1 month. At weekly intervals, a sample of five dumb-bells was removed from the oven, cooled overnight and tested at 23 °C with the results shown in Table 35.

Table 35 — Tensile strengths after ageing

Tensile strengths in MPa

Ageing time days	Tensile strength for test piece No.					
	1	2	3	4	5	Median
0	13,0	11,1	11,0	11,6	13,4	11,6
7	18,1	17,3	16,6	16,8	18,7	17,3
14	17,3	17,6	17,5	18,9	16,7	17,5
21	13,3	13,7	12,3	14,3	13,7	13,7
28	4,24	4,09	3,83	3,87	3,86	3,87

11.3.3.2 There is clearly an initial increase in strength, probably as a result of increasing cross-link density. This is followed by a rapid decrease in strength as degradation takes hold.

The simplest function to fit such data is the quadratic. Proceeding to calculate the various factors given in Annex H results in a regression line of the form defined by the equation:

$$TS = 11{,}6 + 1{,}16t - 0{,}0511t^2 \quad (60)$$

where

TS is the tensile strength, in megapascals;

t is the time, in days.

11.3.3.3 The variance ratio is found to be 602 which is above the 95 % confidence level for 2 by 2 degrees of freedom. Note that this regression should not be used to extrapolate to longer ageing times. For instance, in this case the regression predicts a tensile strength of – 10 MPa at the next weekly interval of 35 days.

It can, however, be used to estimate the time it takes for the tensile strength to fall to 50 % of its original value. Thus, for a tensile strength of 5,8 MPa, the quadratic equation can be solved to give a value of t of

$$t = \frac{-1{,}16 - \sqrt{1{,}16^2 - 4\left[-0{,}051\ 1 \times (11{,}6 - 5{,}8)\right]}}{2 \times -0{,}051\ 1}$$

The impossibility of negative time makes the other root inadmissible, which gives

$t = 26{,}9$ days

11.3.4 Temperature of retraction test

11.3.4.1 In a temperature of retraction test carried out in accordance with ISO 2921, the percentage retraction of three test pieces was measured every 2 min as the temperature in the heat exchange bath rose from –70 °C to ambient. An aim of the test was to estimate the temperature at which 10 % (TR10), 50 % (TR50) and 70 % (TR70) recovery had occurred. A total of 44 data pairs for each of the three test pieces was produced and an abbreviated table for the mean value only is given in Table 36.

Table 36 — Measurements of temperature of retraction

Temperature °C	Retraction %	Temperature °C	Retraction %	Temperature °C	Retraction %
– 68,3	0,0	– 32,9	17,0	3,8	66,0
– 62,2	0,0	– 26,7	23,3	9,2	75,7
– 56,6	0,7	– 20,6	29,0	15,3	85,3
– 50,6	2,7	– 14,9	36,0	22,2	90,7
– 44,8	7,0	– 8,4	41,3		
– 38,7	11,0	– 0,8	55,7		

11.3.4.2 Consideration of the mathematics of this test show that the retraction should be contained within the boundaries of 0 % and 100 % as the temperature varies from low to high values and hence a sigmoidal-shaped function would be expected to produce an accurate regression line. However, as only interpolation of the data needs to be made, it is safe to use the much simpler cubic regression given in 11.2.4.

Determination of the factors given in Annex H therefore produces a regression of

$$R = 57{,}8 + 1{,}57\theta + 0{,}007\ 177\theta^2 - 0{,}000\ 051\ 3\theta^3 \quad (61)$$

where

the retraction value, R, is expressed as a percentage;

the temperature, θ, is in degrees Celsius.

11.3.4.3 The value F_r was found to be 16,24 which is well in excess of the 95 % confidence level for 3 by 43 degrees of freedom.

11.3.4.4 For the given values of the retraction value (10 %, 50 % and 70 %), it is a relatively easy trial and error calculation to find the corresponding temperature since this is not required to be known to a high precision (the nearest degree being quite adequate). The outcome of the test, along with the TR-values estimated using a sigmoidal function (cumulative normal distribution function), is as given in Table 37.

Table 37 — Retraction value results

Retraction value	Temperature °C	
	Cubic	Sigmoidal
10	− 39	− 37
50	− 5	− 6
70	+ 8	+ 7

Thus the very much simpler cubic regression gives results for the test which, when compared with those obtained using the more complex sigmoidal function, are within the accuracy that can be expected from this particular test.

12 Uncertainty of measurement

12.1 Principles

12.1.1 It is recognized that any statement of the result of a measurement is incomplete without the inclusion of a statement of the uncertainty associated with that measurement. This uncertainty is a statement giving the limits within which the true value of the measurement is considered to lie. To be complete, there should also be a confidence level concerning the probability of the true value being inside the limits of the stated uncertainty.

12.1.2 In practice, it is neither necessary nor practical to consider confidence levels of 100 %, although at first sight this might be considered desirable, as this would result in infinitely large uncertainty. The usually accepted confidence level is 95 %, and this should be adopted whenever possible. For example, calibration results are now quoted to 95 % confidence level when associated with accredited calibrations and a clause stating this is included in their accompanying certificates.

12.1.3 A distinction should be made between uncertainty and error, the latter being the difference between the indicated value or result and the true value. A systematic error can be corrected if additional information concerning its magnitude and direction is available via sources external to the experiment.

12.1.4 A variety of factors can influence the uncertainty of the stated results, and these should all be taken into account to produce a single value of uncertainty. The factors range from the fact that a single measurement can have any value in the observed measurement distribution, to the uncertainties associated

with the measurement of the temperature at which the measurement result was made, and the uncertainty of the calibration of the measuring equipment used in achieving the result.

12.2 Methodology

12.2.1 Compilation of a single value for uncertainty

In order to calculate the single value of uncertainty for a measurement, it should be appreciated that there are two important contributors to this uncertainty referred to as random uncertainty and systematic uncertainty.

12.2.2 Random uncertainty (U_r)

12.2.2.1 If a number of measurements is made under the same conditions, a range of actual values is obtained in practice.

The variations are the result of independent random influences ranging from electrical noise producing variations in meter readings to operator reading errors due to difficulties in reading the printed or engraved scales frequently associated with rubber and plastics testing equipment.

12.2.2.2 An analysis of a sample of experimental measurements will usually be found to produce a Gaussian or normal distribution curve. Examination of such a curve would show that 68,3 % of all possible measured values in the population for this distribution would fall between limits $\pm\,\sigma$, where σ is defined as the standard deviation. Further, it can be shown that 95 % lie between $\pm\,1{,}96\sigma$, and that $\pm\,3\sigma$ value would include 99,7 % of all measured values.

Thus the uncertainty, $\pm U$, can be referred to as equal to $\pm 1{,}96\sigma$ for a confidence level of 95 % when a normal distribution for the whole population is being considered. In practice, this can be treated as $\pm\,2\sigma$.

12.2.2.3 When experimental measurements are made, only a limited number of results is actually taken, and it can be shown that an estimated standard deviation can be calculated as shown in 6.2.3.2.

It is then possible to calculate a random uncertainty for such a finite measurement sample to any given confidence level by using the Student's t distribution method.

Table 11 gives a t-value for any number of measurements, n, at the selected confidence level (95 % in most practical cases). The t-value is that value by which the standard deviation of a finite set of values, n, should be multiplied when producing an uncertainty at the selected confidence level.

The values found at $n = \infty$ are referred to as k-values; these values differ with the probability, P, but show clearly that the selection of $\pm 2{,}00\sigma$ is quite justifiable in practice as this would only change the confidence level from 0,950 to 0,955.

12.2.2.4 The random uncertainty, U_r, of the mean value is then obtained using the equation:

$$U_r = \frac{\sigma t}{\sqrt{n}} \tag{62}$$

The above formula is used when n is small, e.g. four.

If a large number of results is available, e.g. 10 or more, it is possible to regard t as equivalent to 1,96 for most purposes at 95 % confidence level, thus

$$U_r = \frac{k\sigma}{\sqrt{n}} \tag{63}$$

12.2.3 Systematic uncertainty (U_s)

12.2.3.1 After the calculation of random uncertainty of the measurement and the application of any known corrections, consideration should be given to other uncertainties that can influence the results.

12.2.3.2 Systematic uncertainty can result from factors as different as the use of the wrong corrections, temperature effects, and calibration uncertainties.

Calibration uncertainties are readily recognized and can be obtained quantitatively from the calibration certificates accompanying the test or measuring apparatus.

Careful examination of all sources of systematic error can sometimes lead to the elimination of problems such as the use of incorrect corrections, the possibility of transcription errors and sometimes software errors.

12.2.3.3 Individual systematic uncertainties can often only be assessed by knowledge of the realistic limits, in other words the uncertainty is presumed to have a rectangular distribution (as opposed to the Gaussian or normal distribution considered earlier), or the measurement value has an equal chance of occurring anywhere between the limits.

In a single rectangular distribution, the standard deviation can be shown to be:

$$\sigma = \frac{a}{\sqrt{3}} \tag{64}$$

where a is equal to half the maximum range (i.e. a is the semi-range) of the observed values.

If there are a number of independent contributions, all having rectangular distributions, with semi-ranges a_1, a_2 to a_m, the resultant standard deviation is given by the equation:

$$\sigma_s = \left(\frac{a_1^2 + a_2^2 + ...a_m^2}{3} \right)^{1/2} \tag{65}$$

If the concept of confidence levels is now introduced, a systematic uncertainty, U_s, can be obtained from the equation:

$$U_s = k\sigma_s \tag{66}$$

Where more information on the distribution of results is available, the semi-range a can be replaced by the standard deviation σ and therefore, more generally, if a number of uncorrelated contributions to the systematic uncertainty, with standard deviations σ_{s1}, σ_{s2}, etc. are present, the resulting systematic uncertainty is given by the equation:

$$U_s = k(\sigma_{s1}^2 + \sigma_{s2}^2 + ... + \sigma_{sm}^2)^{1/2} \tag{67}$$

In some cases, it is found that individual uncertainties are already available hence the above can be replaced by the equation:

$$U_s = (U_{s1}^2 + U_{s2}^2 + ... + U_{sm}^2)^{1/2} \tag{68}$$

12.2.3.4 In practice, the calibration uncertainty can be the only one available in the form of an actual uncertainty (e.g. from the calibration certificates) and the following version of the equation applies:

$$U_s = [U^2_{\text{calibration}} + k^2(\sigma_{s1}^2 + \sigma_{s2}^2 + ... + \sigma_{sm}^2)]^{1/2} \tag{69}$$

where $U_{\text{calibration}}$ is equivalent to $k\sigma_{\text{calibration}}$, but the standard deviation $\sigma_{\text{calibration}}$ associated with the calibration does not need to be calculated.

12.2.3.5 In some cases when realistic limits of uncertainty are estimated, a dominant contribution to the systematic uncertainty can be present such that the uncertainty, as calculated from Equation (67), gives a value that is greater than the arithmetic sum of the semi-ranges of the contribution.

If this is the case, then the dominant contribution should be separated from the calculation and the total systematic uncertainty given as:

$$U_s = a_d + U_s' \tag{70}$$

where U_s' is calculated from the remaining terms after exclusion of a_d.

Reference should be made to UKAS document NIS 3003 [9] for a more detailed description of this effect.

12.2.4 Deviation of a single value of total uncertainty

Once the overall random and systematic uncertainties have been obtained and all contributions to total uncertainty U have been accounted for it is possible to calculate the total uncertainty by:

$$U = U_r^2 + (U_s^2)^{1/2} \tag{71}$$

This should be modified if a dominant contribution to systematic uncertainty is present that meets the criterion mentioned at the end of 12.2.3. In this case

$$U = a_d + [U_r^2 + (U_s')^2]^{1/2} \tag{72}$$

where U_s' is obtained from:

$$U_s' = k\sigma_s' \tag{73}$$

with σ_s' being the standard deviation after omitting a_d.

It is of course imperative that all contributions have been calculated to the same confidence level, which in most cases will be 95 %.

12.2.5 Reporting of results

12.2.5.1 Once the overall uncertainty has been calculated, the final corrected value for the measurement result under consideration (expressed as the mean value $\bar{x}$) can be reported along with the overall uncertainty in the form:

$$\bar{x} \pm U \tag{74}$$

Such a statement is of limited value unless the confidence level is stated in an accompanying clause, e.g. “This uncertainty is for an estimated confidence probability of not less than 95 %”.

12.2.5.2 In practice, the uncertainty should have a resolution that is meaningful in the context of the test being carried out. It is normally justifiable to quote an uncertainty to more than two significant figures.

The number of significant figures in the stated value of the uncertainty should reflect the smallest resolution that can be observed for the particular test measurement.

12.3 Applications to rubber testing

The following step-by-step procedure for obtaining measurement uncertainty should prove helpful in practice:

a) Prior to any statement of the result, all corrections to the result should be applied.

b) Random uncertainties U_r should be obtained through the use of the standard deviation of the results. This can be available from previous work, but it is better to make at least four measurements and then use Equation (62).

NOTE If 10 measurement results are available, it is more appropriate to calculate U_r from Equation (63).

c) All systematic uncertainties should be obtained and considered.

NOTE Some, such as calibration uncertainty, are available from certificates accompanying the test apparatus.

d) Investigate whether previous experimental work can provide the information, for example standard deviations, associated with certain of the systematic contributions.

e) If only realistic limits are available for certain systematic uncertainties (i.e. a rectangular distribution in contrast to a Gaussian or normal distribution as for random uncertainty), the standard deviation is calculated by using Equation (65).

f) The overall systematic uncertainty U_s is then calculated from Equation (68) or (69).

NOTE If a dominant contribution is present when realistic limits are estimated, U_s is obtained from Equation (70) (this is when the requirements at the end of 12.2.3 are taken into account).

g) The total uncertainty is then calculated from Equation (71) or (72) and the results reported in the form given in Expression (74).

13 Sampling

13.1 Principles

13.1.1 When quantities of a product are transferred from a producer to a customer, it is unrealistic to expect every component of every item to be 100 % error-free every time. Hence some form of inspection of out-going (or in-coming) quality is needed. However, it is rarely possible, or even desirable, for every item to be fully inspected for conformity to the specification against which the product has been made. In many cases, inspection would result in the destruction of the product and, even where this were not so, the cost of inspection has to be carried by someone. Ultimately, this would be the customer.

13.1.2 It is therefore necessary to take a representative sample of a consignment (a lot) of the product being supplied and to test this sample. The lot would be accepted if the sample conforms to the inspection programme or rejected if the sample does not conform.

Even if the sample is truly representative of the lot from which it was taken, its properties will still only be an estimate of those of the lot and very rarely identical. Therefore, there will inevitably be a risk that many might be accepted which ought to have been rejected and *vice versa*. The number of nonconforming items that can be tolerated by the customer and the degree of risk associated with a wrong outcome of the sampling test should be agreed between producer and customer. Such factors cannot be objectively determined by the application of statistical tests.

13.1.3 However, given the criteria described in 13.1.2, the size of sample to take in relation to the lot size and the number of nonconforming items found in the sample that will cause the lot to be accepted or rejected can be objectively determined and is the subject of sampling theory.

13.2 Methodology

13.2.1 General

13.2.1.1 The subject of sampling is a large one and is well covered in other International Standards. No more than an outline is given here, and for details the interested user should refer to the various parts of ISO 2859 for sampling by attributes and to ISO 3951 for sampling by variables.

13.2.1.2 In sampling by attributes, it is the number of nonconformities (defined in a test or series of tests) that determines the acceptability or otherwise of the lot.

13.2.1.3 In sampling by variables, it is the estimates of the location and variability of the distributed measurements of a lot in relation to the specification limits that determine the lot's acceptability.

13.2.2 Acceptable quality level and limiting quality

13.2.2.1 The most significant statistic for the producer and customer to agree is the acceptable quality level, or AQL (see ISO 2859-1). This is an indexing device to set the limits of nonconforming items in the sample at which the lot is either accepted or rejected. It should not be inferred from this that any percentage of nonconforming items is wanted. Clearly, it is always desired that the number of nonconforming items in a lot is zero, while in practice a certain percentage of defective items can be tolerated. The AQL should be set realistically to reflect both the requirements of the customer's process needs and the quality that the producer's process is capable of achieving. (Guidance on setting an AQL can be found in ISO/TR 8550.)

13.2.2.2 The AQL is appropriate for use where a sequence of lots is being supplied. When a lot is to be considered in isolation, then the limiting quality, or LQ, is the statistic to be agreed upon (see ISO 2859-2). LQ is a quality level in either per cent nonconforming or nonconformities per 100 items. The value of LQ is really the limiting value of what is unacceptable, and in practice the actual number of nonconformities in a sample should be much less than LQ (generally less than a quarter) if the lot is not to be regularly rejected.

13.2.3 Assessment of nonconformity

13.2.3.1 The percent nonconforming and the number of nonconformities per 100 items are only numerically the same when a single test is applied in the inspection process. Where multiple tests are involved, a decision has to be made which of the two criteria is applied.

13.2.3.2 For example, consider a sample of 50 pipe sealing rings manufactured in accordance with ISO 4633 which are to be inspected for

a) outside diameter (d_o);

b) cord diameter (d_c);

c) hardness (H);

d) tensile strength (TS);

e) elongation at break (E_b).

13.2.3.3 The inspection shows that

a) 45 pipe sealing rings conform in all these respects;

b) three fail on d_o;

c) one fails on d_o and H;

d) one fails on d_o, H, TS and E_b.

Therefore the number of nonconforming rings is five out of 50, giving 10 % nonconforming.

13.2.3.4 From the results given in 13.2.3.3

$$N_c = (3 \times 1) + (1 \times 2) + (1 \times 4) = 9$$

where N_c = the total number of nonconformities in a sample of 50 items.

Therefore there are 18 nonconformities per 100 items.

13.2.3.5 In some processes, any single nonconformity can render the item unsuitable. Other nonconformities in the same item then become irrelevant. In other cases, it can be more appropriate to count the total number of times a failure to meet the specification is encountered irrespective of the item in which the failure is found.

13.2.3.6 It is implicit in the discussion of sampling so far that each nonconformity is equally important whereas in practice some can have more serious consequences than others. A discussion of this case is outside the scope of this International Standard and reference should be made to ISO 2859. Again, therefore, it is essential for the producer and customer to agree their criteria before the inspection takes place.

13.2.4 Inspection levels

Ideally, once an AQL has been agreed, it could be guaranteed that a lot with a quality greater than this would always be accepted and one with a quality less than this would always be rejected. However, this ideal is not attainable and a compromise should be set by means of the level of inspection to be applied. Three such levels are standardized:

a) Normal inspection is designed to give the producer a high degree of protection from having his lots rejected when in fact they have a quality better than the AQL.

b) Tightened inspection is designed to give the customer a high degree of protection from having lots accepted which in fact have a quality lower than the AQL.

c) Reduced inspection is designed to enable cost savings to be made in the inspection process, to the benefit of both producer and customer, where the product quality is consistently shown to be better than the required AQL.

Rules for switching from one level of inspection to another are to be found in the standards dealing with specific sampling procedures (see ISO 2859 and ISO 3951).

13.2.5 Plans for sampling by attributes

13.2.5.1 In this International Standard, only sampling by attributes is considered as this is generally the one most often employed within the rubber industry. Inspection may also be carried out by means of variables, but this technique is often more expensive and elaborate, and can lead to the rejection of a lot which, itself, contains no defective items. Such rejection can be difficult to explain to other employees, customers, suppliers, etc., who are not statistically literate. For further details on this technique, refer to ISO 3951.

13.2.5.2 Within a particular level of inspection at a given AQL, there are several methods by which the sample can be chosen from the lot.

a) The single sampling plan, in which the appropriate number of items is chosen at random from the lot and inspected, is the simplest. On the basis of the number of nonconformities recorded, the lot will either be rejected or accepted because the rejection number is always one greater than the acceptance number.

NOTE 1 The acceptance number is the maximum number of nonconformities allowed in a sample which results in a lot being accepted.

NOTE 2 The rejection number is the minimum number of nonconformities present in a sample which results in a lot being rejected.

b) The double sampling plan, in which a smaller number of items (for the same lot size) is chosen as the sample and inspected, is the next simplest.

1) If the number of nonconformities is less than or equal to the acceptance number, then the lot is accepted.

2) If it is greater than or equal to the rejection number (which is more than one greater than the acceptance number) the lot is rejected.

3) If the number is intermediate between the acceptance and rejection numbers, a second sample (of the same size as the first) is chosen at random and similarly inspected. Then the total number of nonconformities from both samples is compared with the acceptance and rejection numbers for the total sample to assess the status of the lot.

NOTE 3 This process can be extended to multiple sampling plans with up to seven sets of samples taken from a single lot.

13.2.5.3 The purpose of these more administratively involved sampling plans is to reduce the total amount of inspection that will ultimately be needed. It can, however, lead to increased inspection in addition to the extra administrative complexity, and so such plans should be used only when there is a high probability of savings being made, i.e. when there is evidence to suggest that either the lot is particularly good or particularly bad in relation to the AQL chosen. Some factors, amongst others, which will influence the choice of sampling plan are:

a) the ease with which items can be selected from the lot;

b) the resources available for undertaking the inspections;

c) the time it takes to complete the testing of a sample;

d) the number of potential nonconformities being monitored.

13.2.5.4 In addition to sampling by means of a pre-selected number of items, it is also possible to choose a sequential sampling plan in which items are chosen and inspected one at a time. A cumulative count is kept of the number of items inspected and the number of nonconformities recorded.

Decision rules are provided for establishing the status of the lot as the evidence accumulates. In principle, a lot with a quality similar to the chosen AQL could have to continue being tested until the whole lot had been tested. In practice, therefore, an upper limit is provided at which point the rejection number is set to one more than the acceptance number so that an unambiguous outcome is assured.

13.2.5.5 Where a continuing series of lots is being received and the quality of previous lots has been of a consistently high standard relative to the AQL, then skip-lot sampling may be introduced if the appropriate criteria are met. For details of this, reference should be made to ISO 2859-3.

13.2.6 Random sampling

13.2.6.1 It has been implicit in the previous sub-clauses that the sample drawn from the lot is fully representative of that lot. This can only be achieved if the items making up the sample are drawn at random. Unfortunately, the ability of people to choose randomly is very poor. There is a strong bias at the subconscious level to look for and use patterns. For this reason, it is strongly recommended that tables of random numbers be used when selecting random items. A table of random numbers can be found in ISO 2859-10, in numerous statistical text books and often as functions in computer programmes such as spreadsheets and programming languages.

13.2.6.2 Where possible, the items in a lot should be ordered or numbered and then individual items chosen according to a table of random numbers covering the range of interest (any numbers in the table which are outside this range are simply ignored).

If a published table is used, the starting point and direction within the table should also be chosen at random so that the same sequence of numbers is not used for each successive lot inspected.

In the case of small items, it can be impossible to number each one, and it is then necessary to resort to intuitive methods of selection or bulk sampling techniques.

13.2.6.3 If a given lot can be logically divided into sub-lots, then it is desirable to select items at random from the sub-lots in proportion to the size of the sub-lot relative to the lot.

For example, if a lot of 4 500 O-rings has been received in four packages of 1 000 and one package of 500 and a sample of 200 rings is to be taken for inspection, the most appropriate action would be to select

a) 44 rings at random from each of two of the four larger packages containing 1 000 rings;

b) 45 rings from each of the other two larger packages;

c) 22 from the smaller package of 500 rings.

13.3 Applications to rubber testing

Sampling plans allow a small number of representative items to indicate the level of quality in the whole consignment.

For example, a consignment of 5 000 babies' soothers is to be supplied to a customer on a regular basis, the soothers having successfully completed full type testing to the appropriate specification. It has been agreed with the customer that each lot will be tested for hydrochloric acid extractables and bite-through resistance. Since a failure in either of these would be unacceptable, the percent nonconforming criterion is to be applied. An AQL of 0,10 at normal inspection level is to be used.

From Table 1 of ISO 2859-1:1999, a lot size of 5 000 at a normal inspection level (level II) gives a code-letter L. Table 2-A for a single sampling plan at normal inspection level shows the sample size to be 200. At the intersection of the sample code-letter = L line and the AQL = 0,10 column, an upward pointing arrow is found. This means that, to achieve the level of risks inherent in the given AQL and inspection level, it is only necessary to test 125 (code K) items. However, the lot will be rejected if a single nonconformity is observed.

Tables 3 and 4 of ISO 2859-1:1999 show that, for this lot size and AQL value, there are no double or multiple sampling plans available which have characteristics equivalent to those of the single sampling plan.

14 Number of test pieces

14.1 Principles

The number of test pieces normally to be used in a test method is defined in the relevant standard, and generally these are three or five, but this number should always be regarded as the minimum that ought to be taken in the context of a routine quality control environment. At times, there will be a need to improve the statistical precision by conducting the test on a larger number of test pieces. Improving the precision of the estimates of the mean (or median) and the standard deviation should always be balanced against the extra effort and cost of achieving that precision. The number of test pieces should never be more than is sufficient for the purpose being investigated. The following simple procedures may be adopted as a guide, but more exact procedures are described in Clause 17.

14.2 Methodology

14.2.1 The confidence limits about the mean were shown in 7.2.1 to be

$$L = \pm ts/\sqrt{n} \tag{75}$$

Hence, if the confidence limits are to be no more than a certain percentage, c, away from the mean, the number of test pieces required to achieve this can be estimated by noting that

$$n \approx \left(\frac{2C_V}{c} \right)^2 \tag{76}$$

where C_V is the coefficient of variation [see Equation (8)] for the observations obtained (probably from the standard number of test pieces) so far.

The factor 2 is the approximate value of t for the 95 % confidence limits. Unfortunately, t is a function of n and so the expression cannot be solved except by iteration. However, since t varies only slowly with n once this is greater than about 10, it is usually good enough to make t constant. For instance, the factor of 2 is accurate to better than 10 % if the value of n is greater than 10 and is usually sufficiently accurate to indicate the approximate size of n when it is considered that C_V itself is only an estimate which has been based on a smaller sample. Carrying out extra tests on the extra number of test pieces required to bring the total number tested up to this value of n should then give confidence limits very close to those needed. For 99 % confidence limits, the factor to use is 3. A step-by-step procedure is given in 17.2.2.

14.2.2 Where a standard test has been performed and there is some doubt over the pass/fail status of the material because the result is close to the specification limit, then carrying out further tests can help to resolve the uncertainty. If the mean is $\bar{x}$ and the limit is M, then

$$n \approx \left(\frac{1{,}75s}{M - \bar{x}} \right)^2 \tag{77}$$

where 1,75 is the factor for the 95 % confidence level, and 2,5 for the 99 % confidence level.

The reason that the factors are different in this case is that it is the one-sided distribution which should be considered.

14.3 Applications to rubber testing

14.3.1 General

From the statistical point of view, the greater the number of test pieces the better. Time and cost considerations, however, indicate that generally three to five test pieces are sufficient for most situations.

14.3.2 Refinement of confidence limits

A 99 % certainty is required so that the true mean stress relaxation for a compound lies within 5 % of the observed mean for a set of results. If the observed mean, based on the normal triplicate test, is 6,8 % (in units of percentage per decade) with a standard deviation of 0,31 % (in the same units), the approximate number of extra replicate tests which would have to be performed can be calculated.

According to the equation given in 14.2, this is given by

$$n \approx \left[\frac{3(100 \times 0{,}31/6{,}8)}{5} \right]^2 \approx 7$$

Thus about four more replicates are needed to achieve the desired confidence in the mean. (The more precise iterative procedure described in 17.2.2 indicates the need for 10 test pieces.)

NOTE Taking into account the variation of Student's t-value with n leads to $n \approx 9$, but in practice the precise number required depends on the actual results obtained. Hence if the 99 % confidence level could be relaxed, probably only a further three replicates would be tested and the results would be considered satisfactory. If the greater degree of confidence were felt to justify the extra cost, then an additional six would probably be tested.

14.3.3 Refinement of a pass/fail status

The permeability of a rubber membrane is required to be not less than 5×10^{-15} $m^2 \cdot s^{-1} \cdot Pa^{-1}$ to a particular gas. A triplicate determination of the permeability under the given conditions gave:

a) a mean value of $6{,}1 \times 10^{-15}$ $m^2 \cdot s^{-1} \cdot Pa^{-1}$;

b) a standard deviation of $1{,}45 \times 10^{-15}$ $m^2 \cdot s^{-1} \cdot Pa^{-1}$.

The number of test pieces which is likely to be needed in order to be 95 % certain that the membrane does meet this requirement can be calculated.

From 14.2

$$n \approx \left(\frac{1{,}75 \times 1{,}45}{5 - 6{,}1} \right)^2 \approx 5$$

As tests are normally carried out in triplicate, a further three test pieces would probably be tested.

15 Expression of results

15.1 Principles

The results obtained from the application of a statistical technique need to be presented in a meaningful form and at a level of precision appropriate to the precision of the data from which they are derived.

15.2 Methodology

15.2.1 The test report

15.2.1.1 Any report presenting the results of a processing of numbers should give sufficient reference to the method of processing, the assumptions made, etc., to enable an independent check on the outcome to be made. Often it is sufficient to make reference to the standard that has been used. This standard can be the test method itself when it lays down how the data is to be handled or it can be a statistical standard written for general application. (Examples would be ISO 36 for peel adhesion and ISO 6133 for the analysis of multi-peak traces.) However, where options are included in the standard, it is essential that the options chosen are reported with the resulting data in the same way as physical parameters such as speed of testing and test piece type.

15.2.1.2 The test method used will generally indicate the statistics to be quoted in any report, and these should always be adhered to so that compliance with the standard is maintained. Where the test method is not specific, the following are recommended for inclusion in the report:

a) the mean value, $\bar{x}$ (see 6.2.2.2);

b) the estimated standard deviation of the population, s (see 6.2.3.2);

c) the coefficient of variation, C_v (see 6.2.3.4);

d) the individual test results;

e) the estimate of uncertainty, where available.

15.2.1.3 Where there is good reason to expect a non-Gaussian distribution of individual test results and where further statistical testing is not required, the median alone should be quoted in place of the mean, standard deviation, etc. If further statistical testing using individual results is required, then transformation techniques should be considered.

NOTE The central limit theorem should also be considered (see 6.2.4).

15.2.1.4 In many instances, the number of items of data in a set is small (less than a dozen) and indicating the individual values in a report is not cumbersome. It is also good practice as it easily allows further analysis to be made.

15.2.1.5 Where large data sets are encountered, some form of chart can be usefully substituted. This is especially the case where quality control is being considered (see Clause 18) and there is an on-going time element implicit in the process. If a large data set refers to a single group of results, then producing a histogram rather then quoting individual values can be more helpful.

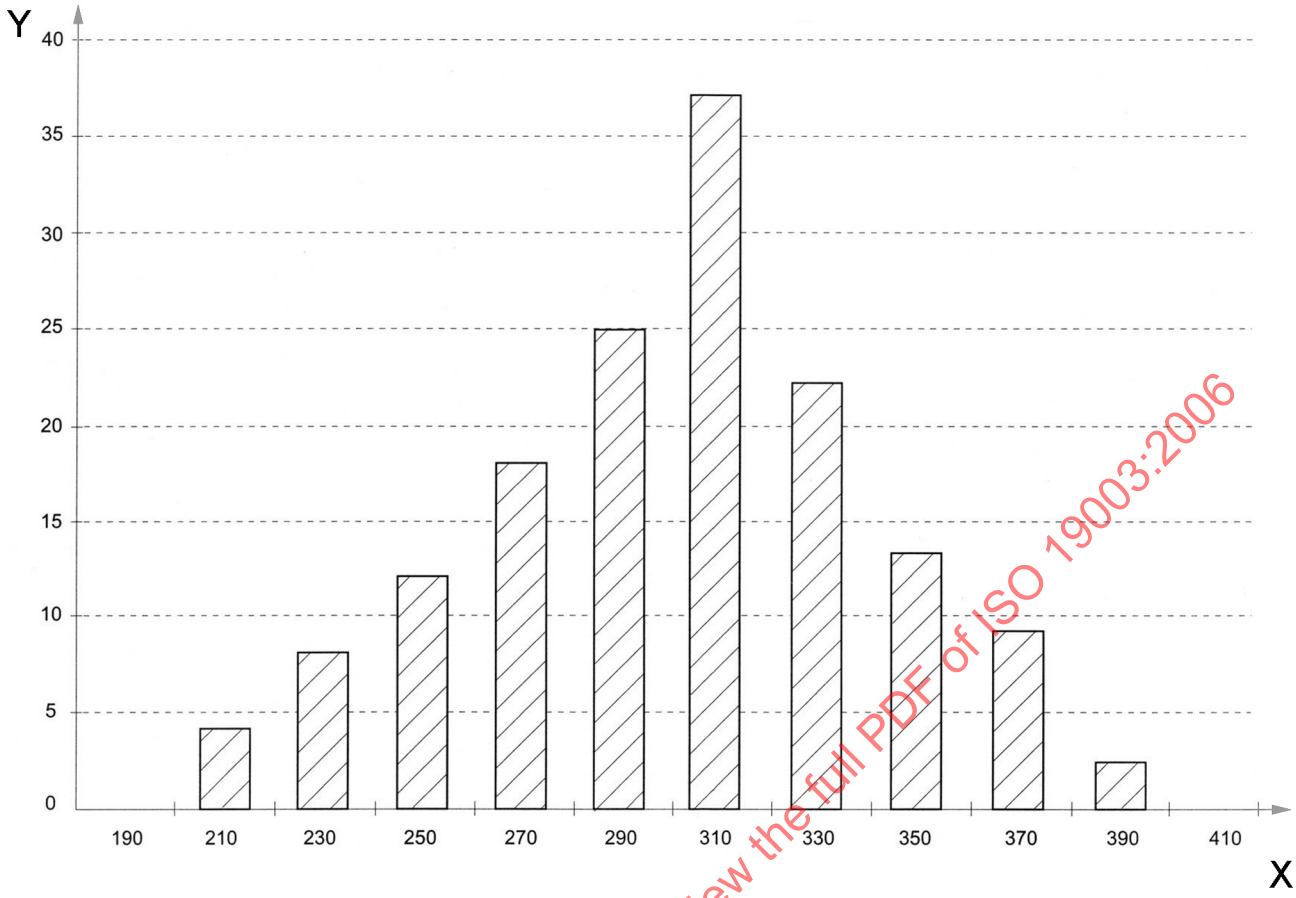
15.2.1.6 To produce a histogram from a set of data, the procedure is as follows:

a) divide the set into an appropriate number of intervals (approximately 10 is generally convenient) covering the range observed);

b) determine the number of results lying within each interval and plot this number for each interval band.

Up to approximately 5 % of results can lie outside the chosen interval range if stragglers produce end groups with very few and irregular numbers of entries. In such cases, the end intervals are left open on one side.

For example, the intervals for elongation at break might be set every 20 %, starting at 200 % and ending at 400 %. The first interval would then be defined as being for all those values having an elongation at break less than 200 % and the last interval for the values having an elongation at break greater than or equal to 400 %. Intermediate intervals would be 200 % to < 220 %, 220 % to < 240 %, etc. (see Figure 8).



Key

X elongation at break (%)
Y number of observations

Figure 8 — Histogram of elongation at break data (see Table 38)

15.2.2 Rounding

15.2.2.1 The measurement of any parameter, for example length, mass, force, concentration, can only be made to a given precision that is governed by the characteristics of the equipment being used to make that measurement. Thus, mass can generally be measured very precisely using an ordinary laboratory balance.

For example, a 100 g sample can, typically, be measured to a precision better than 1 mg. Assuming that the mass is measured to a precision of 1 mg, then it is known to 0,001 % of its value. Again, a dial gauge comparator might measure to 0,01 mm hence a thickness measurement of 2 mm cannot be known to a precision better than 0,5 %. (See Clause 12 for details of uncertainty of measurements.)

15.2.2.2 When calculations are performed on the parameters that have been directly measured, the potential for a false sense of precision almost invariably arises. It is implicit in a mathematical process that the numbers being handled are exact, when in fact they are not. Consider the thickness measurement above. Three readings are taken, giving results 2,01 mm, 2,03 mm, 2,03 mm. The calculation of the mean produces the result 2,023 333 333... mm. If the original values had been absolutely correct, then this mean would also be absolutely correct. However, re-measurement with a more precise instrument yielded thickness values of 2,013 mm, 2,025 mm, 2,027 mm. Now the mean appears to be 2,021 666 66... mm. The difference in the two means is due entirely to differences in the precision of the data being used in their derivation. Similar considerations apply to non-statistical derived functions such as stress, set and relaxation.

15.2.2.3 It follows from 15.2.2.2 that, when reporting the results of calculations, there should be no more significant figure in the derived statistic than the number of significant figures in the least precise

measurement used in its derivation. In the above example, the first mean should be reported as 2,02 and the second as 2,022.

15.2.2.4 When, however, a derived statistic is itself going to be used in further calculations, then at least one more significant figure should be retained, purely for calculation purposes, to avoid the accumulation of errors that can arise from the rounding process. In the example in 15.2.2.2, if the thickness is to be used in the calculation of stress, then for the first mean 2,023 should be used and for the second 2,0217.

15.2.2.5 The rules given in the previous sub-clauses are useful generalizations, but common-sense considerations should also be taken into account. Thus, where measured parameters with different intrinsic precisions are to be combined, there is little value in retaining the highest level of precision in the parameter which has the greatest inherent precision already. Using the previous example again, the force parameter is unlikely to be measurable to greater than three significant figures, hence using the mean thickness quoted to five significant figures will not improve the precision of the calculated stress. In this case, the mean of 2,022 could be retained without loss of accuracy.

15.3 Applications to rubber testing

15.3.1 General

As a general rule when reporting the values of parameters:

a) linear dimensions, volumes, forces and stresses should not be reported to more than three significant figures;

b) strain and energy should not be reported to more than two significant figures;

c) mass may be reported to four significant figures although three is often sufficient.

15.3.2 Construction of a histogram

The tensile testing of a large number (150) of dumb-bells in a single batch of compound resulted in a spread between 200 % and 400 % in elongation at break values. This interval was divided into 10 bands at 20 % intervals and Table 38 obtained.

Table 38 — Table of elongation at break values

Elongation range %	Number of observations
< 200	0
$\geqslant 200$ but < 220	4
$\geqslant 220$ but < 240	8
$\geqslant 240$ but < 260	12
$\geqslant 260$ but < 280	18
$\geqslant 280$ but < 300	25
$\geqslant 300$ but < 320	37
$\geqslant 320$ but < 340	22
$\geqslant 340$ but < 360	13
$\geqslant 360$ but < 380	9
$\geqslant 380$ but < 400	2
$\geqslant 400$	0

A histogram of the data is given in Figure 8, and the graphical representation of the data conveys clearly and simply the breadth and centre of the observed distribution.

15.3.3 Examples of rounding

Consider an intermittent stress relaxation test in which a test piece 10,02 mm wide by 1,03 mm thick had an initial force of 150 N applied to it when extended by 50 %. Both dimensions were measured to the nearest 0,01 mm, and the force was measured to the nearest 1 N. Thus the initial stress is calculated as

$$\text{Stress} = \frac{150}{10,02 \times 1,03} = 14,534\,038...$$

The width was measured to four significant figures, while the thickness and force only to three, hence the final stress should be reported to three figures also, giving 14,5 MPa.

After some time, the force had decreased to 67 N, so the stress had now become

$$\text{Stress} = \frac{67}{10,02 \times 1,03} = 6,491\,871...$$

The precision of the width and thickness is unchanged, but now the force had only been measured to two significant figures and so the resulting stress should only be reported to two figures, giving 6,5 MPa.

If the percentage decrease in stress with time is to be plotted, then the stress values should be further processed to obtain this percentage. Using the rounded figures as reported above results in

$$L_s = \frac{14,5 - 6,5}{14,5} \times 100 = 55,172\ 4$$

where L_s is the decrease in stress, expressed as a percentage.

Using the extra significant figure (the more correct procedure) results in

$$L_s = \frac{14,53 - 6,49}{14,53} \times 100 = 55,333\ 8$$

It would be reasonable to quote the percentage loss to the first place of decimals but no more. Under these circumstances, the difference in the two results is small and of no practical significance. This is not always so, and in all cases repeated rounding followed by further calculations builds up the errors arising from the earlier rounding processes.

16 Precision statements

16.1 General

Increasingly, test methods are including a statement of the precision which can be expected of them when they are carried out in accordance with their requirements. Test methods for rubber are included in this movement.

16.2 Principles

16.2.1 It is well known that tests performed on nominally identical material under nominally identical conditions do not, in general, produce the same result. This variability between repeated tests is given the general name precision, and the sources of variation can be attributed to many factors such as the time between measurements, the environment, the operator, the equipment used and its state of calibration.

It is observed that the variability between different operators and/or different equipment is generally greater than that for a given operator using given equipment within a short time scale. Thus, it is useful to distinguish two measures of precision, repeatability and reproducibility.

As with all statistical factors, the repeatability and reproducibility are estimates based on an accepted level of confidence. The 95 % confidence level is almost always applied and may be assumed if there is no indication to the contrary. Since there is a measure of uncertainty in the estimates of repeatability and reproducibility, it is possible to assign a confidence interval to the estimates obtained from a given experimental programme. Such an analysis is outside the scope of this International Standard.

These two, then, represent the probable extremes of precision that might be expected under practical circumstances, although other, intermediate, forms could be envisaged.

16.2.2 The term accuracy is also encountered and at one time was taken to be simply the bias (i.e. the systematic rather than the random errors) of a particular measurement (refer, for example, to ISO/TR 9272). However, current practice is to use the term accuracy to mean the combined systematic and random errors in a set of observations.

More detailed information on these and other terms relating to precision statements can be found in ISO 5725-1.

16.2.3 In the rubber industry, the term trueness is seldom appropriate, but the effect of systematic discrepancies between laboratories is nonetheless real and has been addressed by a technique called the intercal method developed in the United States. A brief description of this is provided in Annex I.

16.3 Methodology

16.3.1 In order to generate repeatability and reproducibility (often abbreviated to r and R, respectively) data, an inter-laboratory test programme should be organized. It is essential that the material being used in the test is consistent when it is despatched to the participating laboratories and that it remains consistent during the transportation and storage phases prior to it being tested in the laboratory.

16.3.2 The laboratories chosen should be selected at random, but in practice there may overriding considerations, such as the need to cover a wide range of environmental conditions or the small number of laboratories willing to be participants. Where a small number of laboratories are involved, particular care should be exercised in not having a preponderance of specially skilled laboratories or ones recognized as being reference laboratories, as the inclusion of such laboratories will tend to yield underestimates for the r and R data. In either event, the r and R values derived are specific to the group of participants at the time of the trial and do not necessarily accurately reflect the true r and R of the total population of laboratories over extended time scales.

16.3.3 In designing the test programme, it should be agreed and understood by the participants as to what will constitute a test result. This might be, for example, a single tensile-strength value or the median of a set of three or five individual values. In the former case, the replicate level will ordinarily be that which is normal for the test method being evaluated. In the latter case, it is most commonly two with, for example, one set of dumb-bells being tested one day and the next either later that same day or possibly within the next day or two.

16.3.4 In order to establish the repeatability data, tests should be made under constant conditions and, to minimize the danger of environmental or equipment drift from influencing the outcome, the time interval between repeat tests should be as small as practicable.

Since the time interval should be small (as in the example given in 16.3.3) and the same equipment should be used by the same operator, there can be the risk of unintentional bias. This might arise from the operator anticipating the result in the performance of some tests. The following precautions should be taken:

a) the test programme should be randomized to break up any patterns in the test sequence;

b) the date and possibly time of the various tests should be recorded as part of the information supplied to the coordinator of the trial;

c) equipment should not be re-calibrated during a test sequence.

16.3.5 The resulting data are arranged in tabular form similar to that for the analysis of variance (Clause 10) and the following statistics derived:

a) the estimate of the between-laboratory variance, s_L^2;

b) the estimate of the within-laboratory variance, s_w^2;

c) the average of the between-laboratory variances for all the laboratories in the test programme, s_m^2;

d) the estimate of the repeatability variance, s_r^2;

e) the estimate of the reproducibility variance, s_R^2, which is given by the equation:

$$s_R^2 = s_L^2 + s_r^2 \tag{78}$$

f) the repeatability, r, which is given by the equation:

$$r = 2\sqrt{2}s_r \tag{79}$$

g) the reproducibility, R, which is given by:

$$R = 2\sqrt{2}s_R \tag{80}$$

16.3.6 It is frequently convenient to refer to the r and R values in relative terms since they often vary with the mean value of the property being measured. The percentage value of r (or R) with respect to the mean is used which is analogous to the coefficient of variation (see 6.2.3.4). However, as certain properties are measured in percentage units, this percentage value of r or R is written in parentheses, i.e. (r) or (R), to avoid ambiguity over the units. Thus, for a test such as compression set, the mean might be 35 % and the r-value 5,2 %, making the (r)-value 14,9 %.

16.3.7 Before the test data are processed to find the within- and between-laboratory variances, they should first be examined for outliers using, for example, the Dixon and Cochran tests (see Clause 9).

16.3.8 The detailed calculations for r and R via the variances indicated above can be referred to in, for example, ISO 5725-2 and ISO/TR 9272.

16.3.9 One possible problem with the normal inter-laboratory test programme, as described in the references cited, occurs when the operator can be influenced, whether consciously or not, between the first observation and subsequent ones. In these circumstances, the replicates are not independent of each other and bias is introduced into the data. A solution to this difficulty is presented by means of the paired-sample (or split-level) method developed by W.J. Youden (see ISO 5725-5).

In essence, the method measures a single replicate result of a property for each of two materials of similar, but different, values of that property for each level of the experiment being conducted. At each level, the pair of materials should be selected carefully or the analysis produces repeatability values significantly greater than the true repeatability values for the test method. Such a pair of materials is known as a Youden pair. Obtaining suitable Youden pairs for a particular test can be difficult, and the technique is not, therefore, recommended for general use.

16.3.10 Once the values of r and R are obtained for the various levels of the property being measured, they should be examined to see if there is a correlation between them and the mean value. This can be carried out using the regression analysis techniques indicated in Clause 11. Abnormalities to the general trend of r or R with mean level should be examined to try to ascertain the cause, as this can give important information on weaknesses in the test method to particular conditions which can then be addressed.

16.4 Applications to rubber testing

16.4.1 Precision statements are becoming the norm for inclusion in standards for rubber test methods.

16.4.2 An investigation of the measurement of volume swell was undertaken by eight laboratories. Several different types of rubber and fluid were used to give a range of swelling characteristics. A summary of the r and R values associated with each mean level is given in Table 39.

Table 39 — Volume swell measurements

Level number	Mean swell %	r	(r) %	R	(R) %
1	3,66	0,76	20,8	1,86	50,8
2	10,1	0,64	6,3	2,27	22,5
3	14,0	2,65	18,9	6,96	49,7
4	20,6	0,57	2,8	4,67	22,7
5	21,5	0,44	2,0	0,80	3,7
6	42,8	0,89	2,1	4,46	10,4
7	55,3	2,36	4,3	7,34	13,3
8	96,9	5,84	6,0	12,13	12,5
9	115	8,85	7,7	26,20	22,8

16.4.3 An examination of the r and R results shows that there is a general trend towards an increasing r or R with increasing mean level, but several anomalies in the trend do occur. This effect is even more pronounced when the (r) and (R) values are compared with the mean level.

There is an indication that the (r) and (R) values go through a minimum, which suggests a quadratic relationship and, while the correlation coefficient can be shown to be statistically significant at the 95 % level or greater, the accuracy in predicting the (r) or (R) value for a given volume swell does not justify the extra calculation effort that would have to be made.

A simple average (r) and (R) for the test method is sufficiently accurate for the purpose to which the information would be put. Thus the data provides an estimate of percentage repeatability of 8 % and of percentage reproducibility of 23 % for the volume swell test.

17 Design of experiments

17.1 General information and principles

17.1.1 General information

17.1.1.1 Introduction

Every clause so far has been about the analysis of experimental data, i.e. the estimation and testing of statistics representing the system that is being measured. These are essential parts of the scientific method. No less a part of the scientific method is experimental design, i.e. the specification of the conditions under which the experimental data are observed.

Experimental design is a major part of applied statistics, about which there is an immense literature. In this chapter, only those aspects of experimental design are presented which have most to contribute to the physical sciences, specifically to the testing of rubber and rubber products.

Descriptions are necessarily brief and selective. They may be prescriptive rather than pedagogical. Readers are advised to consult the literature to reach a better understanding. Some references are given in the Bibliography.

There are many text books on the subject and there are many types of experimental design. In this subclause, a range of designs are described, selected for their usefulness to rubber technology, and leading from the simplest to the more complex.

17.1.1.2 Descriptive designs

A statistical sample of several test pieces, all randomly selected from a standard material, is tested to determine the elementary statistics of a characteristic of that material. For example, the mean and the standard deviation of the tensile strength of that standard material could be reported.

17.1.1.3 Comparative designs

17.1.1.3.1 Comparison against a standard

The characteristic of a new material can be compared against a specified industry standard. A sample of several pieces would be tested and an assessment made as to whether or not there was sufficient evidence to conclude that the measured characteristic of this material was different from the standard specification.

17.1.1.3.2 Comparison of two materials with independent samples

Two materials can be

a) of different compositions;

b) made by slightly different processes;

c) made at different places, even if they are claimed to be of the same composition and made by exactly the same process.

In order to determine if they have the same or different properties, a sample of several pieces from each material should be tested. These samples should be selected independently of each other.

17.1.1.3.3 Comparison of two materials by paired samples

As in 17.1.1.3.2, it could be necessary to determine if two materials have the same or different properties. However, in the presence of uncontrollable outside influences, a fair comparison should be ensured.

For example, samples of rubber could be exposed to the weather and their deterioration measured. One approach would be to expose test pieces in pairs, each pair comprising one item or piece of each material, thus ensuring that both members of the pair experience the same weather conditions. The data to be analysed would be the difference in deterioration measured between each pair.

17.1.1.4 Response designs

17.1.1.4.1 Factorial experiments

When new materials or manufacturing processes are being developed, there are usually several variables, or factors, that can influence a material property. Experiments to investigate the effects of several variables should be designed to allow all of those variables to be set at several levels. There is a widespread belief that the best approach is to experiment with one variable at a time and to fix all the others. That approach is inefficient, uneconomic and will not provide information about interactions between variables. Two-level factorial experiments are widely used during development studies. Since the final stage of development study requires multilevel experiments (three or more levels), if the number of controlled variables is small it may be useful to consider using three-level experimental designs in the development stage.

17.1.1.4.2 Response surface exploration with composite designs

In the final stage of a development study, when the conditions (such as the values of composition and process variables) that will yield the best value of a material property (such as the highest value of tensile strength) are being sought, additional points should be added to factorial experiments so that curvature of the response can be estimated. These designs are known as augmented or composite designs.

17.1.1.4.3 Inter-laboratory trials

Another class of experiment used in industry is the inter-laboratory trial. This has the purpose of estimating the repeatability of test results within each of a set of laboratories and the reproducibility of test results between laboratories. These are not described fully in this International Standard as they are described in ISO 5725 and ISO/TR 9272, but Clause 16 outlines some of the principles involved in making precision statements.

17.1.2 Principles

17.1.2.1 General

Statistical analysis of experimental results is necessary because of variation. All test results vary. The reasons for this variation include the following:

a) there is inherent variability of material;

b) there are imperfections in the measuring instruments and their calibrations;

c) there is sampling variation.

This variation should therefore be considered when experiments are designed.

17.1.2.2 Descriptive experiments

17.1.2.2.1 In a descriptive experiment, a characteristic of a standard material is reported from the analysis of measurements on several test results. For example, the mean tensile strength of a sample of several test pieces is calculated. This is unlikely to be the true mean value of all possible test pieces from the standard material. If the mean tensile strength of another sample of several test pieces were calculated, it would be different. The calculated sample mean is therefore only an estimate, a point estimate, of the underlying population mean. In reporting it, an interval should be reported within which the population mean can confidently be expected to lie. This interval is the confidence interval for the population mean. Clause 7 gives further details.

17.1.2.2.2 This confidence interval depends on three things:

a) the variation between measurements of the test pieces within the sample, expressed as the variance or standard deviation of the measured material property;

b) the number of test pieces tested in the sample;

c) the degree of confidence of the interval, expressed as the probability that the population mean is truly in that interval, usually as a percentage (for example, a 95 % confidence interval).

NOTE 1 The variation will usually be determined from the experiment.

NOTE 2 The number of test pieces should be specified before the experiment is carried out.

NOTE 3 The degree of confidence is the choice of the experimenter and again should be specified before the experiment is carried out.

17.1.2.2.3 Ideally, the experimenter should specify the size of the confidence interval and the confidence. For example, in the case of tensile strength, the experimenter can specify the following:

a) the size of the confidence interval is (sample mean value $\pm$ 1,0) MPa.

b) the confidence is 95 %.

The experiment would then proceed in the following four stages:

1) a preliminary experiment is carried out to enable the unknown variance of all possible test pieces for the whole of the standard material to be estimated;

2) the sample size N is calculated and this is used to estimate the specified confidence interval, using the variance estimated from the results of the preliminary experiment;

3) test measurements are made on a sample of N test pieces.

The sample mean, standard deviation and confidence interval are calculated.

These four stages are described in 17.2.

17.1.2.3 Comparative experiments

17.1.2.3.1 Statistical analysis of test results should never be regarded simply as a set of calculations leading to clear-cut statements that the effect is, or is not, significant. Such statements have no meaning without an explanation in terms of the purpose of the experiment. The conclusion depends on the circumstances of the experiment and on the intentions of the experimenter which should be declared before the tests are done. For example, the circumstances of an experiment can ordain whether or not a statistically significant effect can be detected if it exists. The intentions of the experimenter will include a statement of what he or she considers to be a technically significant effect.

17.1.2.3.2 Four major steps should be taken before starting an experiment to compare the underlying values of a characteristic for two materials:

a) Step one, in which the alternative inferences that can be made from the experiment are stated.

b) Step two, in which the acceptable risks for making the wrong inference are specified.

c) Step three, in which the difference between the two values of the characteristic is specified. This should be demonstrated statistically so as to be of technical significance.

d) Step four, in which the necessary sample size is computed.

These four steps are described more fully in 17.1.2.3.3 to 17.1.2.3.6 and then presented in greater detail in 17.2.

17.1.2.3.3 In step one, the alternative inferences that can be made from the experiment are stated. These should be stated as alternative prior hypotheses.

When two materials are to be compared according to some property, the most usual comparison is between the mean values of that property. An assessment should be made as to whether there is sufficient evidence to infer that the underlying mean values for the whole of each standard material of the underlying populations (μ_1 and μ_2) differ, even though the sample mean values, $\bar{x}_1$ and $\bar{x}_2$, differ (see 6.2.2). If there is not sufficient evidence, it should be assumed that the means of the underlying populations are the same. The assumption that they are the same is known as the null hypothesis (H_0). The assumption that they are different is known as the alternative hypothesis (H_a).

These can be stated symbolically as:

H_0 for which $\mu_1 = \mu_2$

H_a for which $\mu_1 \neq \mu_2$

In this case, the experimenter is not concerned about which of the two populations has the greater mean, only that they could be different. This will lead to a two-sided test.

If the experimenter is interested in showing that a new material has a greater mean strength than the standard material, a one-sided test can be used and the alternative hypotheses will have the form:

H_0 for which $\mu_1 = \mu_2$

H_a for which $\mu_1 > \mu_2$

The distinction should be made before the experiment is commenced. The calculation of the sample size depends on the distinction.

17.1.2.3.4 In step two, the acceptable risks for making the wrong inference are specified. The wrong inferences are called the type 1 error and the type 2 error, with probabilities α and β, respectively.

The possible inferences from a two-sided test can be understood from Table 40.

Table 40 — Inferences from a two-sided test

Truth	Inference	Correct	Error	Probability
$\mu_1 = \mu_2$	$\mu_1 = \mu_2$	Yes	—	$< (1 - \alpha)$
	$\mu_1 \neq \mu_2$	No	Type 1	$\leqslant \alpha$
$\mu_1 \neq \mu_2$	$\mu_1 = \mu_2$	No	Type 2	$\leqslant \beta$
	$\mu_1 \neq \mu_2$	Yes	—	$\leqslant (1 - \beta)$

A type 1 error occurs when the experimenter accepts the alternative hypothesis (H_a) although the null hypothesis (H_0) is true. The probability of this occurring is α. This is known as the size of the test.

Usually, α is specified as 0,05 (a 5 % chance) (see Clause 7).

A type 2 error occurs when the experimenter accepts the null hypothesis (H_0) although the alternative hypothesis is true. The probability of this occurring is β. Usually β is specified as 0,05 or 0,10. The probability of detecting a true difference is $(1 - \beta)$. Thus if β is specified as 0,05 and the experiment is designed

accordingly, there is a strong chance (a probability of 0,95) that a difference will be detected if a difference truly exists. This is known as the power of the test.

Unfortunately, it is common for experiments to be done without consideration of β or the power, and consequently true effects can remain undetected. For example, consider the tensile strengths of compounds A and B in 6.3.2. Suppose that a purpose of the experiment was to show a statistically significant difference (α = 0,05) of 1 MPa. A power calculation shows that with only 12 sample measurements for each compound there is a probability of 0,75 of detecting that difference if it exists. Sample sizes of 23 would be needed to give a probability of 0,95 of detecting that difference.

17.1.2.3.5 In step three, the difference, which should be demonstrated statistically so as to be of technical significance, is specified.

The purpose of many experiments is to discover an improvement in the material property which is being tested. In other experiments, the purpose can be to show that, under different circumstances, there is no difference in the material property.

In either case, the experimenter should be able to state the smallest difference which should be regarded as likely to have a practical or technical significance. For example, a determination can be made of how much stronger, in terms of tensile strength, one material should be over another to make its selection preferable for a particular application. Whether this should be 1 MPa or 2 MPa or 0,5 MPa can depend on the application.

The specification of this smallest difference, δ, is essential to the design of a comparative experiment.

17.1.2.3.6 In step four, the necessary sample size is computed.

There are several formulae for calculating sample size. The correct choice of formula depends on the type of comparative experiment. These can be:

a) comparison against a standard;

b) comparison of two materials with independent samples;

c) comparison of two materials by paired samples;

It will also depend on whether the proposed test of comparison is one-tailed or two-tailed (see 17.2, 17.3 and Table 42).

The information needed in any of the calculations is the choice of α, β, δ and an estimate of the variance, σ^2, of all possible test pieces for the whole of the standard material.

The choices of α, β, δ depend entirely on the opinions and purposes of the experimenter, as already explained.

An estimate of the population variance, σ^2, can be obtained:

1) from earlier experiments;

2) from literature;

3) from a preliminary experiment with at least five test pieces.

17.1.2.4 Response experiments

17.1.2.4.1 Properties and factors

Much research and development in the materials sciences is intended to establish relationships between the properties of materials and suspected influencing factors which the technologist can control in the production of those materials.

The properties are called response variables (they are also called dependent variables because they are thought to depend on the factors).

Influencing factors are called control variables (they are also called independent variables). The control variables are usually composition variables and process variables. All of these variables should be measurable.

Sometimes there are other variables which can influence the response variables but cannot be controlled although they can be identified and measured. Common examples are the temperature and humidity of a factory workshop atmosphere. These variables are called concomitant variables (they are also called covariates).

17.1.2.4.2 Experimental design

The experimental design is the specification, before the experiment is commenced, of the values of the control variables at which the response variables are measured. The experiment should be designed according to the expected relationship between the response variable and the control variables. The expected relationship is a hypothesis. The hypothesis should be formulated as an algebraic model that can be represented in terms of the measurable variables.

If the model were simple and exact, then the experimental design would be simple. Few test pieces would be needed to provide exact values of the model coefficients. The fitted model could then be used to predict response values exactly for any choice of settings of the influencing factors. However, there are several reasons why this cannot be achieved and these include the following:

a) The exact relationship can never be known because the model can be only an approximation to reality.

b) All measurements are subject to time-dependent deviations which cannot be identified but which show their presence by trends in the observed values of response variables.

c) All measurements are subject to random deviations. These represent other unidentified variables which, taken together, show no pattern or trend.

d) The effects of concomitant variables, which are those variables which are identified as possible influencing factors but can be measured only during or after an experiment.

The experiment should therefore be designed so as to reduce the influence of these unknowns.

17.1.2.4.3 Statistical objectives

The statistical objectives of designed response experiments are to specify:

a) an algebraic model representing the expected relationship between the response variables and the influencing factors;

b) the number of observations;

c) the values of the control variables at every observation;

d) the order of the observations.

The intention is to:

1) ensure that all effects in the model can be estimated from the observed data;

2) test the reality of those effects by comparison with random variation;

3) ensure that all effects can be estimated with the greatest possible precision, thereby reducing the influence of random variation;

4) ensure that all effects can be estimated with the least possible bias, or greatest accuracy, thereby reducing the effects of time-dependent errors;

5) suggest improvements to the model;

6) keep within a budget of effort and cost.

17.1.2.4.4 Experimental designs

The following experimental designs are discussed in 17.1.2.4.5 to 17.1.2.4.7:

a) two-level factorial experiments;

b) two-level fractional factorial experiments;

c) composite designs.

17.1.2.4.5 Two-level factorial experiments

The two-level factorial design is fundamental to experimental design for the physical sciences. This design is based on the assumption that a linear model will approximate to the true relationship fairly well in some restricted range of those factors. The basic idea can be illustrated with a linear model based on a single influencing factor (Figure 9):

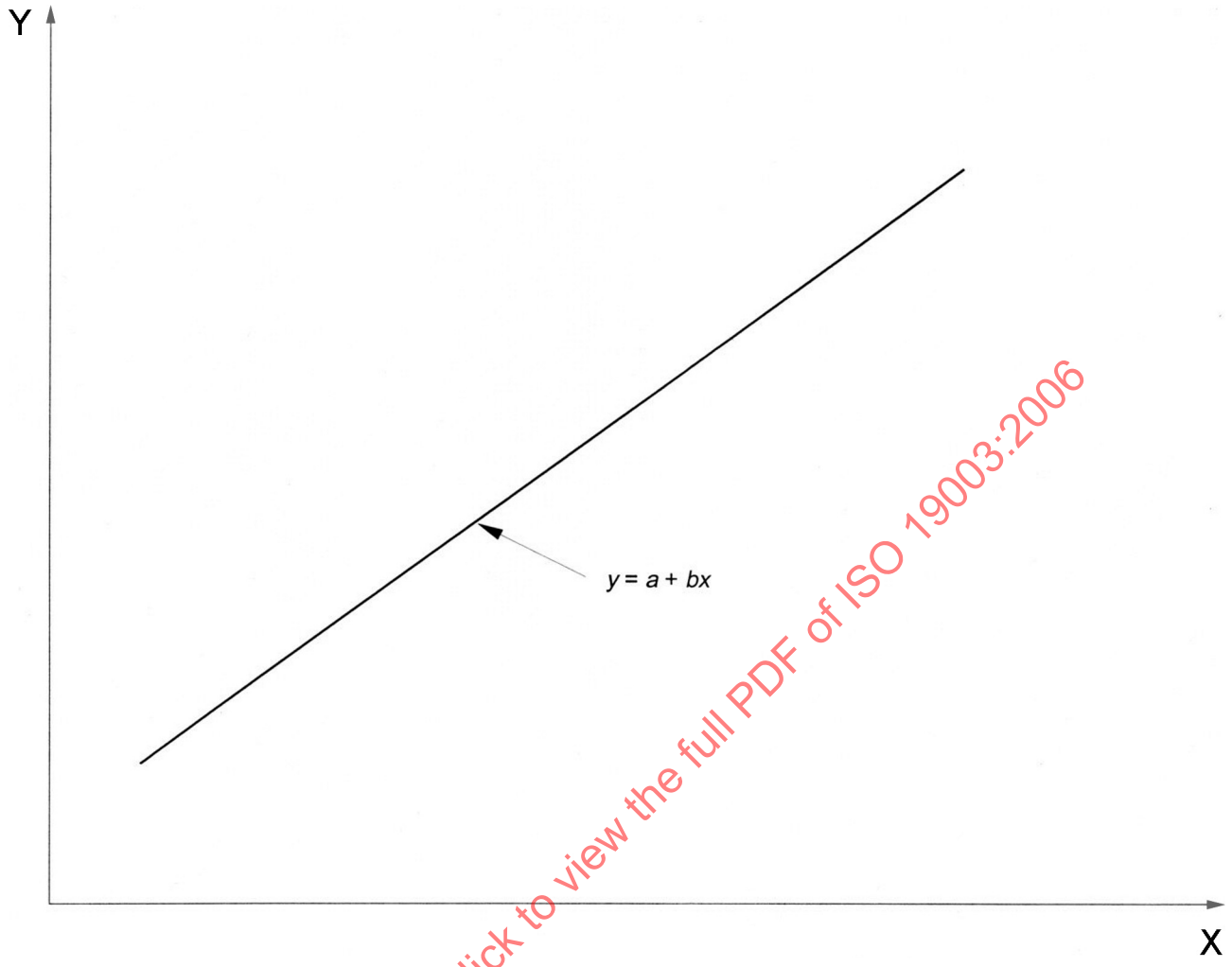
$$y = a + bx \tag{81}$$

where

y is the response variable;

x is the control variable;

a and b are the coefficients to be estimated.



Key

X control variable, x

Y response variable, y

Figure 9 — Linear relationship between the dependent variable and a control variable

This model is deemed to be roughly adequate to approximate the true relationship for x in the range x_L to x_U. Deviations between expected values and the observed values can be described by adding a term e to the right-hand side of Equation (81).

The object is to estimate a and b with the greatest precision if all observations are divided equally between the two ends of the range of x. A common fault among experimenters is to divide the range into $(N - 1)$ equal parts (where N is the number of planned observations) and to make one observation at each end and at each of the division points. This choice of factor values is a design which would not give the most precise estimates of a and b in the presence of e. It would be preferable to do half the trials at x_L and half at x_U. This gives rise to the expression “a two-level experiment”.

In Equation (81), the effect on y of a change in x of one unit is represented by the coefficient b which is the slope of the line.

The relationship between x and y may also be represented using a different notation. In this notation, the independent variables (the x's) are called factors A, B, C, ..., the effects of which are represented by the symbols A, B, C, ...

The range of a factor is specified by the two ends of the range, i.e. the high and the low values of the factor. These are represented by lower-case letters with suffixes. For example, in the single-factor experiment, the high and low values of factor A would be a_1 and a_0 respectively. This lower case notation is also used to represent the observed values of the dependent variable at the corresponding observation points [see Figure 10 a)].

Thus the effect of factor A on the dependent variable y over the complete range of factor A is equal to:

(Value of y at point a_1) – (Value of y at point a_0)

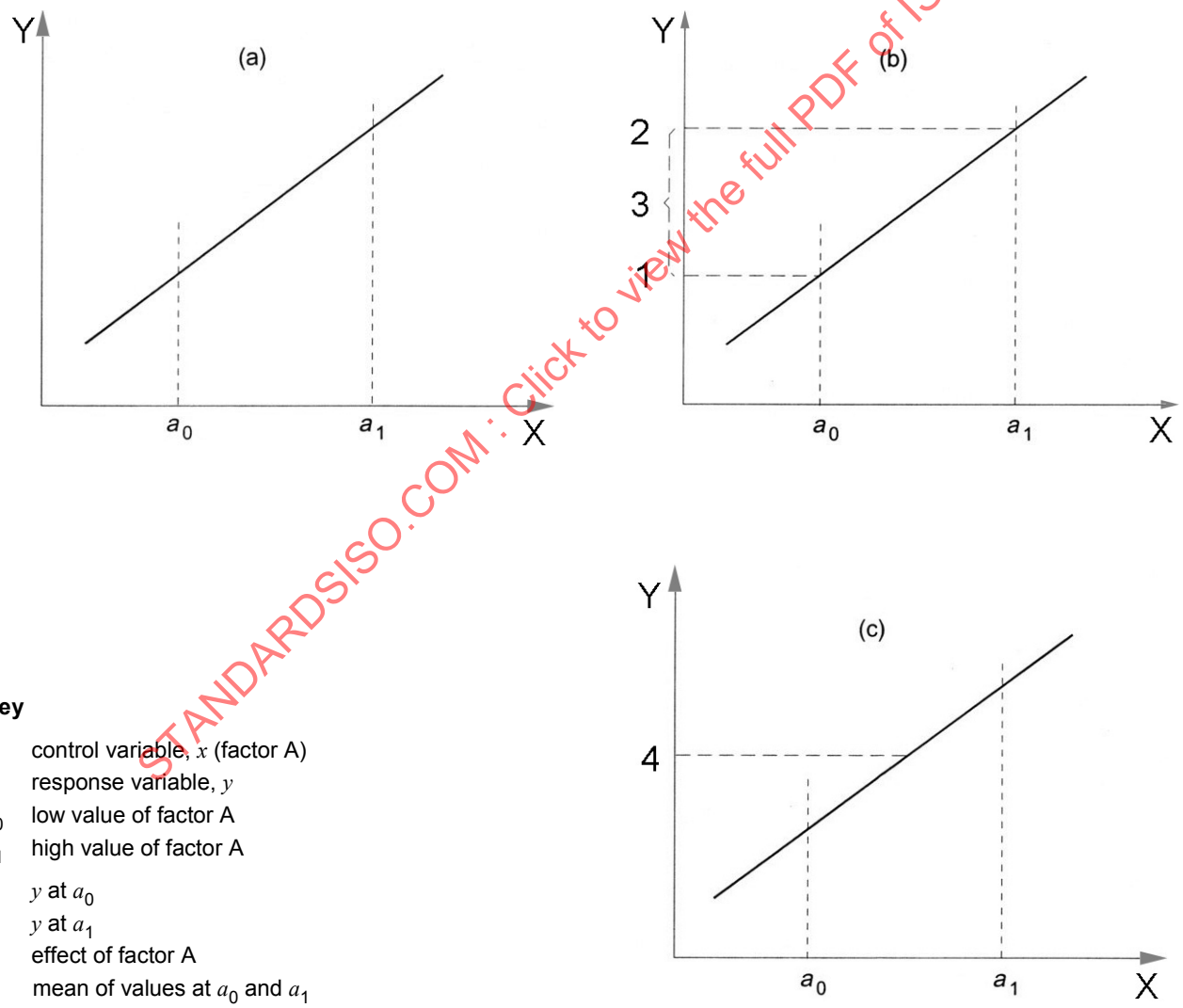
as shown in Figure 10 b), or, more briefly,

$$A = a_1 - a_0 \tag{82}$$

where the symbol A is used to denote the effect of factor A.

Similarly, the mean value of y [see Figure 10 c)] is simply

$$M = (a_1 + a_0)/2 \tag{83}$$



Key

X control variable, x (factor A)
Y response variable, y
a_0 low value of factor A
a_1 high value of factor A
1 y at a_0
2 y at a_1
3 effect of factor A
4 mean of values at a_0 and a_1

Figure 10 — A two-level single-factor experimental design

Now consider two factors, A and B, which can be represented as two variables in a plane with the dependent variable y along a third dimension perpendicular to the plane [see Figure 11 a)]. The high and low values of B are b_1 and b_0. If observations of y are made only at points defined by the extreme ranges of the two factors, there are four points which can be denoted by the combinations of letters a_0b_0, a_1b_0, a_0b_1 and a_1b_1. The notation can be abbreviated further to represent these four points as:

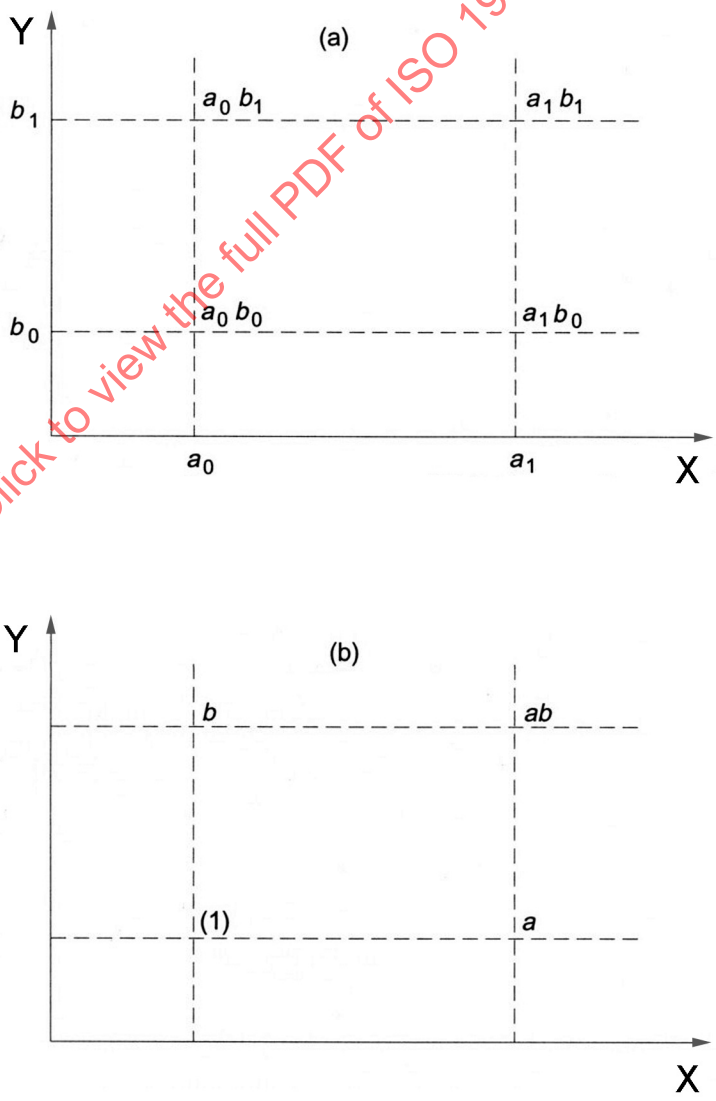
(1), a, b, ab

where

the symbol (1) denotes the observation point at which all the factors are at their low levels (see Figure 11 b);

the point a is where factor A is at its high level but factor B is at its low level;

the point ab is where both factors are at their high levels.



Key

X factor A
Y factor B
a_0 low value of factor A
a_1 high value of factor A
b_0 low value of factor B
b_1 high value of factor B

NOTE The lower drawing uses the simplified (abbreviated) notation.

Figure 11 — A two-level two-factor experimental design

The rule is that the high and low levels of factors are represented by the presence or absence, respectively, of lower-case letters.

Analysis is almost as easy as in the single-factor case and the following conclusions can be reached:

a) Using the combinations of lower-case letters to represent the values of y observed at the corresponding points, the average effect of factor A is:

$$A = \frac{a + ab}{2} - \frac{(1) + b}{2} \tag{84}$$

That is, the effect of A is the difference between the mean value of y observed at all the points where A was at its high level and the mean value of y observed at all the points where A was at its low level.

b) The effect of B is calculated in a similar manner to that of A and is given by:

$$B = \frac{b + ab}{2} - \frac{(1) + a}{2} \tag{85}$$

c) The interaction of factors A and B can be defined as the difference between the effect of A at the high level of B and the effect of A at the low level of B. It is denoted by AB. Thus:

$$AB = (ab - b) - [a - (1)] \tag{86}$$

This is exactly the same as: the difference between the effect of B at the high level of A and the effect of B at the low level of A.

The estimation of these effects is equivalent to fitting the algebraic model:

$$y = a_0 + a_1x_1 + a_2x_2 + a_{12}x_1x_2 \tag{87}$$

where y is the response variable, x_1 and x_2 are two control variables and a_0, a_1, a_2 and a_{12} are algebraic coefficients. (These should not be confused with the a used in Equations (82) to (86) to denote variables.)

Least-squares regression analysis (Clause 11) is widely used for analysis of these and other experiments to be described. Computer software is available for this analysis which includes the estimation and testing of coefficients in equations such as (87).

17.1.2.4.6 Two-level fractional factorial experiments

These principles of design and analysis of two-level factorial experiments can be extended to experiments involving any number of factors.

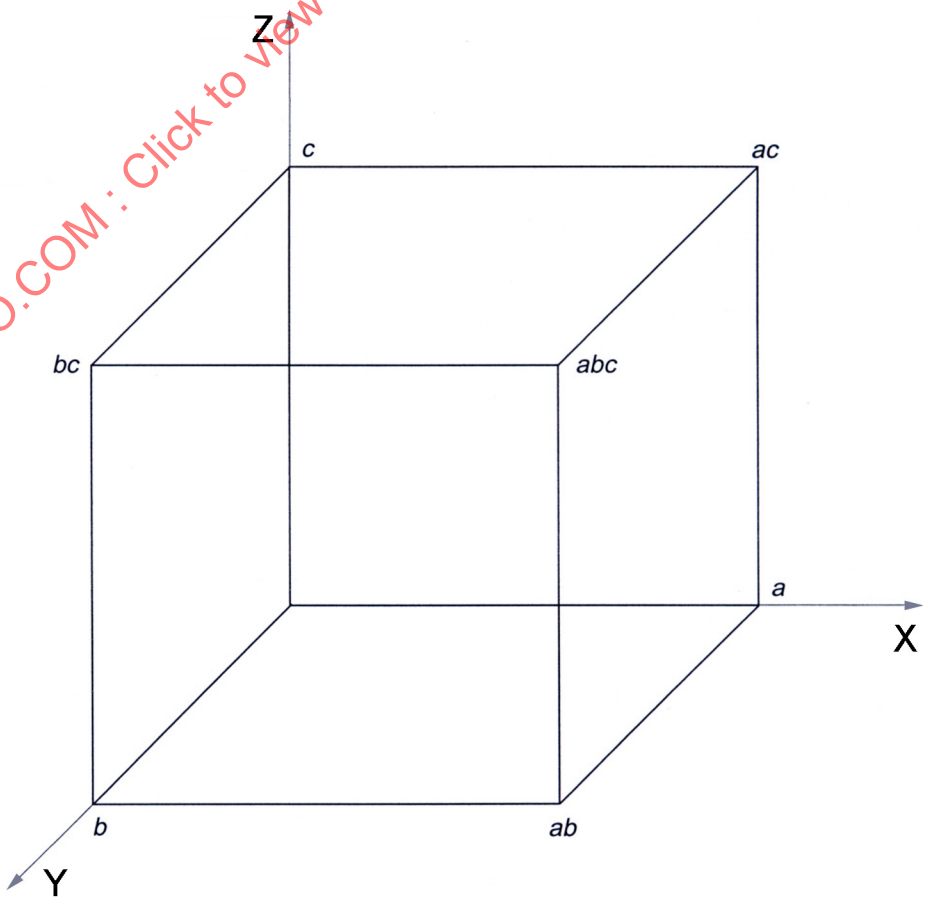
See Figure 12 for an illustration of a three-factor situation. However, the number of observations in such an experiment increases exponentially with the number of factors. If there are n factors, the number of test pieces required in an experiment is proportional to 2^n. Table 41 illustrates the exponential increase.

Table 41 — Observation points

Number of factors	Number of points at which observations are made
1	2
2	4
3	8
4	16
5	32
6	64
7	128
8	256
9	512
10	1 024

It is not unusual to have experiments with seven or more factors (control variables). Thrift demands an experiment with only a fraction of the experiments in a full design but which can still supply information on important features of the model. If a suitable fraction can be found, the resulting experiment is called a two-level fractional factorial.

The theory and method of constructing these fractional experiments is described in textbooks. Also software is available for the automatic design and analysis of these experiments (see Bibliography).



Key

X factor A
Y factor B
Z factor C

Figure 12 — A two-level three-factor experimental design

17.1.2.4.7 Composite designs

Whereas two-level factorial experiments, and their fractions, are suitable for fitting models that are linear in the main effects and including interactions, they are not suitable for estimating curvature of response if it exists. For example, if there is a single control variable, Equation (81) can be suitable either if the relationship is genuinely linear for all values of x [see Figure 13 a)] or on the rising or decreasing slope of a quadratic response [see Figure 13 b)].

However, if the experiment is to be done for a range of x which is close to the peak (or trough) of the quadratic response, as in Figure 13 c), curvature will have a major effect and should be estimated. This is particularly important if a purpose of the experiment is to estimate the value of x for which y is a maximum (or minimum).

Equation (81) should then be augmented as follows:

$$y = a + bx + cx^2 \tag{88}$$

Similarly, Equation (87) should be augmented as follows:

$$y = a_0 + a_1x_1 + a_2x_2 + a_{12}x_1x_2 + a_{11}x_1^2 + a_{22}x_2^2 \tag{89}$$

Designs for these augmented relationships are called augmented or composite designs. The theory and methodology of constructing them is described in several textbooks. Software is available for constructing and analysing them. Analysis is usually by least-squares regression.

NOTE Since most relationships are not linear, two-level experiments are best used only if

a) the range between the low and the high levels is so small that the curvature of the response can be neglected [see Figure 13 b)];

b) the number of controlled variables is large and a screening test is required to identify the important ones;

c) qualitative variables are involved (e.g. N550 vs N330 carbon black, or straining/not straining the stock).

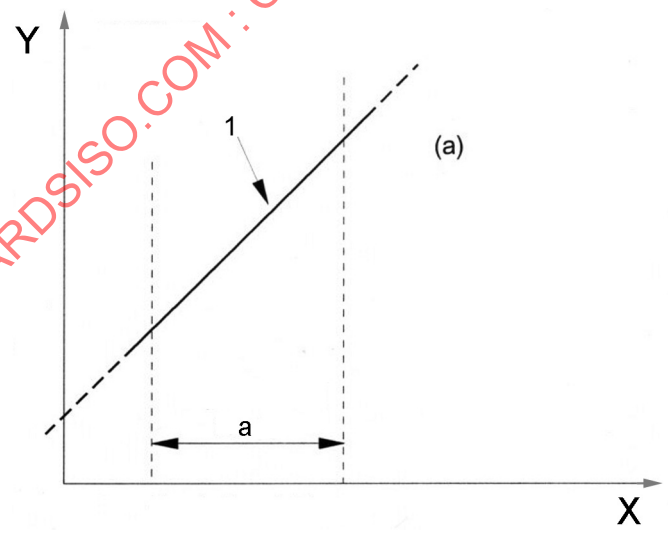
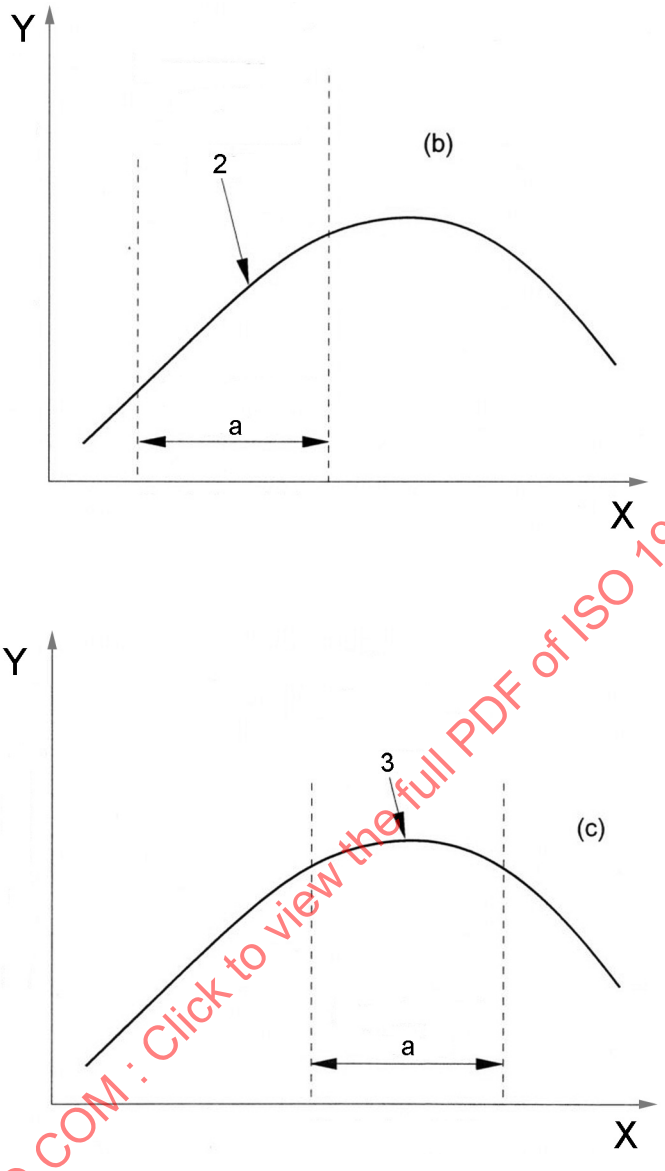


Figure 13 — Quadratic relationship between the dependent variable and a control variable



Key

X control variable, x
Y response variable, y
1 linear response
2 approximately linear response over the experimental range
3 quadratic response over the experimental range

[a] Experimental range.

Figure 13 (*continued*)

17.2 Methodology

17.2.1 General

17.2.1.1 Procedures are presented here for designing and analysing descriptive and comparative experiments, the principles of which have been explained in 17.1. Examples are shown in boxes.

17.2.1.2 Procedures for the design and analysis of two-level (fractional) factorial experiments, or for composite experiments, are not included. The necessary theory and methodology are beyond the scope of this International Standard. Software exists for these procedures to be handled automatically.

17.2.1.3 Z-scores are used in all of the following procedures (see 6.2.1.1). These are standardized normal variates and are available from published statistical tables. However, for convenience in using this International Standard, the extract given in Table 42 should be sufficient.

Table 42 — Z-scores

α (or β)	Z_α (one-tail)	Z_α (two-tail)
0,01	2,326	2,576
0,05	1,645	1,960
0,10	1,282	1,645
0,15	1,036	1,440
0,20	0,842	1,282
0,25	0,675	1,150
0,30	0,524	1,036
0,35	0,385	0,935
0,40	0,253	0,842

17.2.2 Descriptive experiments

The mean of a standard material should be reported with a specified confidence interval, $2c_l$. The confidence interval should be stated as a percentage confidence as given by the equation:

$$2c_l = 100(1 - \alpha) \quad (90)$$

The procedure is as follows:

a) Obtain an estimate of the underlying population variance, σ^2, by a preliminary experiment in which at least five randomly selected pieces are tested.

$\sigma^2 = 2$

b) State the desired half width of the confidence interval, c_l.

$c_l = 1$

c) State α as representative of the required confidence interval.

For a 95 % confidence interval:

$\alpha = 0{,}05$

d) Look up the corresponding two-tail value of Z_α.

$Z_\alpha = 1{,}96$

e) Calculate N as the integer nearest to the value of $0{,}5 + (Z_\alpha \sigma / c_l)^2$.

$N = 8$

f) Obtain a more conservative value of N by repeating the calculation using the corresponding t_α value with the first estimate of N as the entry point in the table of t-values (see Table 11). This is achieved in two steps as follows:

1) Look up the corresponding two-tail value of t_α for N degrees of freedom.

$t_\alpha = 2{,}12$

2) Calculate N as the integer nearest to the value of $0{,}5 + (t_\alpha \sigma / c_l)^2$.

$N = 9$

g) Repeat steps 1) and 2) in f) until a stable value of N is achieved.

17.2.3 Comparative experiments

17.2.3.1 Comparison against a standard

The underlying population of a property measure of the standard material has a known mean μ_0 and known variance σ^2. Assuming that the property of the new material has the same variance (which can be checked later), the size of the sample needed to detect an improvement of at least δ can be calculated.

The procedure is as follows:

a) State the object of the trial.

H_0: $\mu_1 = \mu_0$

H_a: $\mu_1 > \mu_0$

(single-sided)

b) Choose α, β, δ.

$\alpha = 0{,}05$

$\beta = 0{,}01$

$\delta = 1$

c) Look up the value of Z_α (single-sided normal).

$Z_\alpha = 1{,}645$

d) Look up the value of Z_β.

$Z_\beta = 2{,}326$

e) Compute N, where N is the integer nearest to the value of the expression:

$$0,5 + (Z_\alpha + Z_\beta)^2\ \sigma^2/\delta^2$$

if $\sigma^2 = 2$,
then $N = 32$

17.2.3.2 Comparison of two materials with independent samples

Two samples are drawn from different populations. The variances, σ^2, are equal and known. The comparison will indicate if the two populations are the same. The procedure is as follows.

a) State the object of the trial.

H_0: $\mu_1 = \mu_2$

H_a: $\mu_1 \neq \mu_2$

(double-sided)

Equivalently: H_{a1}: $\mu_2 < \mu_1$ with $\alpha/2$ risk

H_{a2}: $\mu_2 > \mu_1$ with $\alpha/2$ risk

b) Choose α, β, δ^2, σ^2.

$\alpha = 0,01$
$\beta = 0,02$
$\delta^2 = 1$
$\sigma^2 = 2$

c) Look up the value of Z_α (double-sided normal).

$Z_\alpha = 2,576$

d) Look up the value of Z_β (single-sided normal).

$Z_\beta = 2,054$

e) Compute N_1, where

$$N_1 = N_2 = N$$

and is the integer nearest to the value of the expression:

$$0,5 + 2(Z_\alpha + Z_\beta)^2\ \sigma^2/\delta^2$$

$N = 86$

17.2.3.3 Comparison of two materials with paired samples

The procedure is as follows:

a) State the object of the trial.

H_0: $\mu_{diff} = 0$

H_a: $\mu_{diff} > 0$

b) Choose α, β, δ and estimate $(\sigma_{diff})^2$.

NOTE If the variance of the underlying population is σ^2 and the variances of both populations are assumed to be the same, then the variance of the difference between two values, one from each population, is $2\sigma^2$.

$\alpha = 0{,}10$

$\beta = 0{,}05$

$\delta = 1$

$\sigma_{diff}^2 = 2\sigma^2$

$= 2$

c) Look up the value of Z_α (single-sided normal).

$Z_\alpha = 1{,}282$

d) Look up the value of Z_β (double-sided normal).

$Z_\beta = 1{,}645$

e) Compute N pairs, where N is the integer nearest to the value of the expression:

$0{,}5 + (Z_\alpha + Z_\beta)^2 \, (2\sigma^2)/\delta^2$

$N = 18$

17.2.4 Response experiments

These include two-level factorials and composite designs. They are not covered in this International Standard and the references given should be consulted. Some examples are given in 17.3.3.

17.3 Applications to rubber testing

17.3.1 Descriptive experiments

17.3.1.1 Refinement of confidence limits

Reference should be made to 14.3.2, which considers an example of stress relaxation. When the procedure given in 17.2.2 is followed, the following results are obtained:

a) $\sigma = 0{,}31$ (Standard deviation of 0,31 % per decade.)

b) $\delta = 0{,}34$ (The half width of the requested confidence interval is 5 % of 6,8 % which is 0,34 %.)

c) $\alpha = 0{,}01$ (A 99 % confidence interval is required.)

d) $Z_\alpha = 2{,}576$ (A two-tail figure is needed so that α is split between the two extremes of the distribution.)

e) $(Z_\alpha \sigma / \delta)^2 = 5{,}52$

Rounding up: $N = 6$

f) When steps 1) and 2) are repeated as necessary, the following results are obtained:

1) $t_\alpha = 4{,}032$ (Table 11: two-sided: 99 %: $n = 6$)

2) $N = 14$

First repeat:

3) $t_\alpha = 3{,}012$ (Table 11: two-sided: 99 %: $n = 14$)

4) $N = 8$

Second repeat:

5) $t_\alpha = 3{,}499$ (Table 11: two-sided: 99 %: $n = 8$)

6) $N = 11$

Third repeat:

7) $t_\alpha = 3{,}169$ (Table 11: two-sided: 99 %: $n = 11$)

8) $N = 9$

Fourth repeat:

9) $t_\alpha = 3{,}355$ (Table 11: two-sided: 99 %: $n = 9$)

10) $N = 10$

Fifth repeat:

11) $t_\alpha = 3{,}25$ (Table 11: two-sided: 99 %: $n = 10$)

12) $N = 9$

Thus repeated iteration gives a value of N that oscillates between 9 and 10. The more reliable choice would be 10.

17.3.1.2 Refinement of a pass/fail status

Reference should be made to 14.3.3, which considers an example of permeability. When the procedure given in 17.2.2 is followed, the following results are obtained:

a) $\sigma = 1{,}45$ ($1{,}45 \times 10^{-15}$ $m^2 \cdot s^{-1} \cdot Pa^{-1}$)

b) $\delta = 1{,}1$ [$(6{,}1 - 5) \times 10^{-15}$ $m^2 \cdot s^{-1} \cdot Pa^{-1}$]

c) $\alpha = 0{,}05$ (95 % confidence required.)

d) $Z_\alpha = 1{,}645$ (A one-tail figure is needed because the concern is only that the permeability should not be less than 5×10^{-15} m^2 $s^{-1} \cdot Pa^{-1}$.)

e) $(Z_\alpha \sigma/\delta)^2 = 4{,}7$

Rounding up: $N = 5$

f) When steps 1) and 2) are repeated as necessary, the following results are obtained:

1) $t_\alpha = 2{,}132$ (Table 11: one-sided: 95 %: $n = 5$)

2) $N = 8$

First repeat:

3) $t_\alpha = 1{,}895$ (Table 11: one-sided: 95 %: $n = 8$)

4) $N = 7$

Second repeat:

5) $t_\alpha = 1{,}943$ (Table 11: one-sided: 95 %: $n = 7$)

6) $N = 7$

Thus seven test pieces are needed.

17.3.2 Comparative experiments

17.3.2.1 Comparison of a new material against a standard material

Reference should be made to 14.3.3. In this application, let the standard rubber membrane have a permeability of 5×10^{-15} m^2·s^{-1}·Pa^{-1} with a known standard deviation of $\sigma = 1{,}45 \times 10^{-15}$ m^2·s^{-1}·Pa^{-1}. Assuming that the permeability of the new material has the same variance, the size of the sample needed to detect an improvement, δ, of at least 1×10^{-15} m^2·s^{-1}·Pa^{-1} can be calculated. When the procedure given in 17.2.3.1 is followed, the following results are obtained:

a) Let the true mean permeability of the standard material be μ_0 and the true mean permeability of the new material be μ_1.

Then

H_0: $\mu_1 = \mu_0$ (Null hypothesis is that there is no difference.)

H_a: $\mu_1 > \mu_0$ (Alternative hypothesis is that there is an increase in permeability: a one-sided change.)

b) α, β, δ are chosen.

$\alpha = 0{,}05$ (In order to look for a 95 % confidence interval.)

$\beta = 0{,}05$ (In order to have a 95 % chance of detecting the difference if it exists.)

$\delta = 1$ (Only a difference of at least 1×10^{-15} m^2·s^{-1}·Pa^{-1} would be deemed to have any technical merit.)

c) $Z_\alpha = 1{,}645$

d) $Z_\beta = 1{,}645$

e) $[(Z_\alpha + Z_\beta)\sigma/\delta]^2 = 22{,}5$

Rounding up: $N = 23$

17.3.2.2 Comparison of two materials using independent samples

Reference should be made to 7.3.3. A single compound material is tear-tested by two laboratories to discover if there is any bias one way or another in the tear-testing methods. It is assumed that there is no difference in the test pieces supplied to the two laboratories. The true mean values and variances are the same. However, there can be a difference between the measured results, caused by a difference in the measuring equipment or method.

Let δ_1 be the true mean measured value obtained by laboratory 1 and μ_2 be the true mean measured value obtained by laboratory 2.

From experience, or a preliminary trial, it is known that with this material a standard deviation, σ, of 0,9 can be expected.

A prior specification could be that a difference, δ, of more than 1 would trigger a technical enquiry.

When the procedure given in 17.2.3.2 is followed, the following results are obtained:

a) The alternative hypotheses are:

H_0: $\mu_1 = \mu_0$ (The laboratory test methods produce the same results.)

H_a: $\mu_1 \neq \mu_0$ (The laboratory test methods produce different results even when testing the same materials.)

b) α, β, δ, σ are chosen.

$\alpha = 0{,}05$ (In order to look for a 95 % confidence interval.)

$\beta = 0{,}1$ (In order to have a 90 % chance of detecting the difference if it exists.)

$\delta = 1$

$\sigma = 0{,}9$

c) $Z_\alpha = 1{,}960$ (two-sided)

d) $Z_\beta = 1{,}282$ (one-sided)

e) $2(Z_\alpha + Z_\beta)2\sigma^2/\delta^2 = 17{,}08$

Rounding up: $N = 18$

Thus 36 test pieces should be cut from the source material and randomly allocated, 18 to each laboratory. The original positions of the 36 pieces should be recorded in case widely divergent values suggest that there could have been segregation of material composition and properties in the source material.

17.3.2.3 Comparison of two materials with paired samples

17.3.2.3.1 Example 1

Random allocation was recommended in 17.3.2.2. An alternative would be to cut sample test pieces in pairs. Their adjacency would ensure that the effect of any segregation of properties for the source material would be minimized.

The two pieces in each pair should be labelled A and B, then a random sequence of As and Bs should be used to allocate the test pieces to the two laboratories.

Thus, if $N = 12$, a random sequence could be:

laboratory 1: B A B B A B A A B B B A

laboratory 2: A B A A B A B B A A A B

In this experiment, the variable of interest is the difference between the measured values of each pair of test pieces.

When the procedure given in 17.2.3.3 is applied, the following results are obtained:

a) H_0: $\mu_{diff} = 0$

NOTE 1 This is the null hypothesis that there is no mean difference between the measured values of the two laboratories.

H_a: $\mu_{diff} > 0$

NOTE 2 The alternative hypothesis is that there is a difference.

b) $\sigma_{diff} = \sigma\sqrt{2}$

$= 0{,}90 \times 1{,}414$

$= 1{,}273$

Using the same values as those in 17.3.2.2:

$\alpha = 0{,}05$;

$\beta = 0{,}1$;

$\delta = 1$.

c) $Z_\alpha = 1{,}645$

NOTE This is a single-sided value.

d) $Z_\beta = 1{,}282$

NOTE This is a single-sided value.

e) $(Z_\alpha + Z_\beta)^2(2\sigma^2)/\delta^2 = 13{,}88$

Rounding up: $N = 14$

Thus 14 pairs of test pieces should be cut from the source material. In this case a comparison with paired samples is cheaper than a comparison with independent samples.

17.3.2.3.2 Example 2: A comparison of resistance to wear

There is a choice of two materials, A and B, for making rubber soles for boys' shoes. One experimental design would be to provide one group of boys with shoes soled with material A and a second group with shoes soled with material B. However, if the boys vary greatly in the rates at which they wear out shoes, any difference between the two materials could be hidden. A paired design would remove much of that variability. Each boy would be given a pair of shoes with the sole of one shoe made from material A and the sole of the other from material B. The choice of which was left or right for each boy would be determined by randomization.

A preliminary trial reveals that the standard deviation, σ, of wear after a month is 2 mm. It is agreed that a difference, δ, between means of 1 mm will be sufficient to rule that one material is better than the other. It is necessary to calculate how many pairs of shoes should be made in order to demonstrate with a confidence of 95 % ($\alpha = 0{,}05$) that a difference of 1 mm is statistically significant. It is also specified that there should be a 95 % chance ($\beta = 0{,}05$) of detecting that difference if it exists.

When the procedure given in 17.2.3.3 is applied, the following results are obtained:

a) H_0: $\mu_{\text{mdiff}} = 0$

NOTE 1 The null hypothesis is that the wear rates of the two materials are the same.

H_a: $\mu_{\text{mdiff}} > 0$

NOTE 2 The alternative hypothesis is that there is a difference.

b) $\sigma = 2$

Therefore

$$\sigma_{\text{diff}} = \sigma\sqrt{2}$$

$$= 2 \times 1{,}414$$

$$= 2{,}828$$

In addition

$\alpha = 0{,}05$;

$\beta = 0{,}05$;

$\delta = 1$.

c) $Z_\alpha = 1{,}645$

NOTE This is a single-sided value.

d) $Z_\beta = 1{,}645$

NOTE This is a single-sided value.

e) $(Z_\alpha + Z_\beta)^2(2\sigma^2)/\delta^2 = 86{,}59$

Rounding up: $N = 87$

Thus 87 pairs of shoes should be made and allocated to boys for a month's wear.

It is instructive to repeat this calculation using different values of α, β, δ, σ.

17.3.3 Response experiments

17.3.3.1 Two-level factorial designs

17.3.3.1.1 A nitrile rubber compound is being developed to have:

a) Good fluid resistance.

NOTE 1 The volume swell should be as low as possible when a test piece is immersed in a standard oil in accordance with ISO 1817.

b) Good low-temperature characteristics.

NOTE 2 The brittleness temperature should be as low as possible when measured in accordance with ISO 812.

The types of factor that the technologist would want to examine would be:

1) the grade of nitrile rubber as characterized by the acrylonitrile (ACN) content, typically grade 1 (28 %), grade 2 (34 %) and grade 3 (40 %);

2) the type of plasticizer used, typically dioctyl phthalate (DOP) and butylcarbitoladipate (BCA);

3) the amount of plasticizer, typically from 10 parts per hundred to 30 parts per hundred of rubber (p.h.r.);

4) the type of carbon black, typically N550 and N330;

5) the amount of carbon black, typically from 30 p.h.r. to 70 p.h.r.

The grading of the nitrile rubber into three distinct grades presents a problem. They are ordered into three levels of ACN and can crudely be treated as a continuous variable. However, a better design and analysis would follow if the experimental levels of ACN could be chosen at any points between 28 % and 40 %. Also, since this is a wide range of ACN, a linear model appropriate to a two-level design is better served by choosing two grades whose ACN values are close together.

17.3.3.1.2 The specification for the experimental design is:

a) Response variables:

1) Y_1 is the fluid resistance as measured by fluid swell;

2) Y_2 is the brittleness temperature.

b) Control variables:

1) X_1 = GRADE: low level 1, high level 2;

2) X_2 = PLAS: two levels DOP (level 1) and BCA (level 2);

3) X_3 = PX: low level 10 p.h.r., high level 15 p.h.r.;

4) X_4 = BLACK: two levels N550 (level 1) and N330 (level 2);

5) X_5 = BX: low level 30 p.h.r., high level 40 p.h.r.

PX and BX are the amounts of plasticizer and black, respectively.

NOTE The ranges of the continuous variables, X_1, X_3, X_5, have been chosen so that they are wide enough for some effects to be observed, but not so wide as to cover all possibilities. Wider ranges can include large quadratic (curvature) effects that would obscure the main linear effects. See 17.1.2.4.7.

17.3.3.1.3 Without any prior knowledge about interactions, the experiment should be designed to permit the estimation of all 10 first-order interactions:

— GRADE.PLAS

— GRADE.PX

— GRADE.BLACK

— GRADE.BX

— PLAS.PX

— PLAS.BLACK

— PLAS.BX

— PX.BLACK

— PX.BX

— BLACK.BX

If there were more factors, some consideration could be given to selecting interactions for inclusion according to the results of earlier experiments or a deep knowledge of the controlling physics and chemistry.

17.3.3.1.4 With this specification, the experimental design given in Table 43 is produced.

Table 43 — Experimental design for nitrile rubber compound development (full factorial analysis)

Observation	GRADE	PLAS	PX	BLACK	BX
1	1,0	1,0	10,0	1,0	30,0
2	2,0	1,0	10,0	1,0	30,0
3	1,0	2,0	10,0	1,0	30,0
4	2,0	2,0	10,0	1,0	30,0
5	1,0	1,0	15,0	1,0	30,0
6	2,0	1,0	15,0	1,0	30,0
7	1,0	2,0	15,0	1,0	30,0
8	2,0	2,0	15,0	1,0	30,0
9	1,0	1,0	10,0	2,0	30,0
10	2,0	1,0	10,0	2,0	30,0
11	1,0	2,0	10,0	2,0	30,0
12	2,0	2,0	10,0	2,0	30,0
13	1,0	1,0	15,0	2,0	30,0
14	2,0	1,0	15,0	2,0	30,0
15	1,0	2,0	15,0	2,0	30,0
16	2,0	2,0	15,0	2,0	30,0
17	1,0	1,0	10,0	1,0	40,0
18	2,0	1,0	10,0	1,0	40,0
19	1,0	2,0	10,0	1,0	40,0
20	2,0	2,0	10,0	1,0	40,0
21	1,0	1,0	15,0	1,0	40,0
22	2,0	1,0	15,0	1,0	40,0
23	1,0	2,0	15,0	1,0	40,0
24	2,0	2,0	15,0	1,0	40,0
25	1,0	1,0	10,0	2,0	40,0
26	2,0	1,0	10,0	2,0	40,0
27	1,0	2,0	10,0	2,0	40,0
28	2,0	2,0	10,0	2,0	40,0
29	1,0	1,0	15,0	2,0	40,0
30	2,0	1,0	15,0	2,0	40,0
31	1,0	2,0	15,0	2,0	40,0
32	2,0	2,0	15,0	2,0	40,0

This is a full factorial analysis ($2^5 = 32$ observations). With fewer interactions selected, a half-factorial of 16 observations, or even a quarter-factorial of eight observations, could be possible.

The 32 test pieces should be prepared in a random order. They should be tested for fluid resistance, Y_1, in a second random order and tested for brittleness temperature, Y_2, in a third random order. The measured results should be analysed using least-squares regression analysis.

17.3.3.2 Two-level fractional factorial design

17.3.3.2.1 Suppose that, from previous experience, only one interaction (PX.BX) is believed to be effective and the available budget forces the use of the smallest possible experiment. A two-level fractional factorial experiment, a quarter-factorial, is produced. This is shown in Table 44.

Table 44 — Experimental design for nitrile rubber compound development (quarter-factorial)

Observation	GRADE	PLAS	PX	BLACK	BX
1	1,0	1,0	10,0	1,0	30,0
2	1,0	1,0	15,0	2,0	30,0
3	1,0	2,0	10,0	2,0	40,0
4	1,0	2,0	15,0	1,0	40,0
5	2,0	2,0	10,0	2,0	30,0
6	2,0	2,0	15,0	1,0	30,0
7	2,0	1,0	10,0	1,0	40,0
8	2,0	1,0	15,0	2,0	40,0

17.3.3.2.2 This is a small experiment, a quarter-factorial, and there are two dangers associated with it:

a) there could be some interactions other than the one specified (PX.BX);

b) there could be insufficient information for error analysis and for satisfactory testing of the fitted model.

These dangers should be traded against the economic advantage of a small experiment.

Alternatively, there are two approaches to dealing with these dangers:

1) more interactions can be introduced so that a half design (16 observations) would be produced;

2) the experiment can be replicated so that two test pieces would be made for each observation point.

17.3.3.3 Composite designs

17.3.3.3.1 Suppose that, after running the experiment specified in 17.3.3.1, or the one in 17.3.3.2, and analysing the results, the indications are that the optimum values of the two response variables are achieved using the plasticizer DOP and the carbon black N550, with high levels of plasticizer (PX) and high levels of carbon black (BX). At this stage, an experiment to fit a quadratic response function would be appropriate.

NOTE The two response variables are:

1) high fluid resistance as measured by low volume swell;

2) low brittleness temperature.

17.3.3.3.2 A specification for this further experiment to discover the best composition could be as given in Table 45.

Table 45 — Experimental specifications for nitrile rubber compound development (quadratic response function)

Variable	Low	High	Increment	Interactions	Quadratic terms
GRADE	1	3	1	GRADE.PX	$(GRADE)^2$
PX (using DOP)	18	30	2	GRADE.BX	$(PX)^2$
BX (using N550)	40	70	5	PX.BX	$(BX)^2$
NOTE PX and BX are measured in parts per hundred of rubber (p.h.r.).					

This would produce the composite design given in Table 46.

Table 46 — Experimental design for nitrile rubber compound development (quadratic response function)

Observation	GRADE	PX	BX
1	1,0	20,0	45,0
2	3,0	20,0	45,0
3	1,0	28,0	45,0
4	3,0	28,0	45,0
5	1,0	20,0	65,0
6	3,0	20,0	65,0
7	1,0	28,0	65,0
8	3,0	28,0	65,0
9	1,0	24,0	55,0
10	3,0	24,0	55,0
11	2,0	18,0	55,0
12	2,0	30,0	55,0
13	2,0	24,0	40,0
14	2,0	24,0	70,0
15	2,0	24,0	55,0
16	2,0	24,0	55,0

17.3.3.3.3 An experiment carried out according to this design would produce observed values (measurements of the response variables) which would permit the fitting of a model of the form:

$$y = a + b_1x_1 + b_2x_2 + b_3x_3 + c_1x_1^2 + c_2x_2^2 + c_3x_3^2 + d_1x_1x_2 + d_2x_1x_3 + d_3x_2x_3 \quad (91)$$

for each of the response variables.

NOTE Random orders of preparation and testing should be used.

These fitted models can then be used to make a close estimate of the optimum conditions, together with measures of confidence of those estimates based on analysis of variation of responses.

18 Statistical quality control

18.1 Principles

Variation in the quality of any product (or service) is inevitable, but the application of statistical principles to data systematically gathered on the product enables decisions and courses of action to be taken which can significantly reduce the amount of reject material produced. In processes of any complexity, there will be several stages of quality control applied at key points in the manufacturing process, but at each stage the principle is the same, i.e. to monitor the process so that deviations which are unlikely to have occurred by chance can be detected quickly and corrective action taken to bring the process back into statistical control. The question of sampling is dealt with in Clause 13.