

BS 5324 : 1976

Guide to
**Application of statistics
to rubber testing**

Guide pour l'application des statistiques aux essais des élastomères

Richtlinie für die Anwendung von Statistiken auf die Gummiprüfung

Contents

	Page	Tables	Page
Foreword	Inside front cover	1. Factors for estimating standard deviation from a range of values in a Normal population	4
Cooperating organizations	Back cover	2. Criteria for rejecting outlying values by Dixon's test	5
Section one. General		3. Factor interaction on compression set	12
1. Scope	2	4. Response values for combinations of the three independent variables	21
2. References	2	5. Correlation coefficient versus numbers of observations	21
Section two. General principles		6. Friedman's test critical values K_{cr} for $\alpha = 0.05$ level of significance	23
3. A practical problem from rubber testing	2	7. Ranking of 10 rubber vulcanizates by five observers	24
4. Introduction to statistical methods	3	8. Plot positions (double exponential frequency distribution function)	26
5. Application of statistical tests to the example in clause 3	10	9. Weighting factors w for calculating mode	27
6. Design and analysis of experiments	11	10. Weighting factors d for calculating standard deviation	27
7. Analysis of variance	15	11. The Normal (Gaussian) distribution function	35
8. Application of regression analysis	17	12. Percentage points of Student's t -distribution	36
9. Reporting of results	22	13. 5 % points of F distribution	
10. Ranking methods	23	Figures	
Section three. Statistical techniques applicable to specific tests		1. Normal distribution curve	3
11. Introduction	24	2. Confidence limits versus true coefficient of variation	6
12. Tensile testing	24	3. Real difference versus true coefficient of variation	7
13. Tear strength (crescent)	27	4. Confidence limits versus estimated coefficient of variation	8
14. Abrasion resistance	28	5. Real difference versus estimated coefficient of variation	9
15. Crack growth and fatigue testing	28	6. Compression set versus post-cure for different durations of test	12
16. Resistance to low temperatures	31	7. Dependency of flexing life on age of material	18
17. Ozone resistance	31	8. Illustration of simple linear regression	19
Appendices		9. Histogram of results	22
A. Formulae for easy reference	32	10. Double exponential distribution	25
B. References (useful books and tables)	32	11. Analysis of tensile strength results	25
C. Glossary of statistical terms and symbols	33	12. Normal (Gaussian) distribution function	35
D. Statistical reference tables	35		

Foreword

In the preparation of this British Standard, under the authority of the Rubber Industry Standards Committee, use has been made of the guidance given in BS 2987 'Notes on the application of statistics to paper testing'.

The problems of sampling and the reliability of results obtained in rubber testing are clearly cases for the application of statistical methods, and in recent years a number of papers have been published indicating the growing interest in statistical methods in the rubber industry.

The results of rubber tests are usually subject to appreciable variation, due not only to experimental error, but also to variation in the raw materials and to the operation of many unknown or random factors in the manufacturing process. The interpretation of test results therefore requires some care and the chief value of statistical methods in this connection lies in their ability to replace purely subjective judgements of the data by objective criteria such as the tests described in the text. Statistical analysis of the data cannot add to the accuracy of an experiment by turning an uncertain result into a certainty, but it can enable the conclusions to be expressed more precisely in terms of a definite probability. Neither can statistical analysis be used to obtain meaningful results from experiments designed without reference to underlying statistical principles.

Although statistical methods have their foundation in mathematical theories of probability, a knowledge of such theories is not required for the successful application of these methods. Only straightforward arithmetical processes are involved and, in many cases, the calculation does not require the use of a machine. One or two shortcut methods of simplifying calculations are given.

The purpose of the present guide is to give only a brief outline of the subject, but with sufficient detail to enable

the reader to carry out the simpler statistical calculations on his test data, such as those for ascertaining the precision of test means and the significance of differences in test means (a list of formulae for easy reference is given in appendix A). For the more advanced techniques, such as detailed application of analysis of variance, which can be of great utility for the comparison of instruments, for compounding investigations and for the analysis of causes of variation in test readings, reference should be made to appropriate textbooks (see appendix B).

A secondary purpose of this guide is to secure greater uniformity in the presentation of test results and in the use of statistical terms and symbols (see clause 4 and glossary, appendix C).

The problems of sampling depend too much on particular conditions (such as manufacture and storage) to be capable of detailed treatment in this standard. Consideration has been given to these matters in the preparation of BS 1673 'Method of testing raw rubber and unvulcanized compounded rubber': Part 1 'Sampling' and the subject is given an extensive general treatment in BS 5309 'Sampling chemical products': Part 1 'Introduction and general principles' and Part 4 'Sampling of solids'.

In preparing this guide account has been taken of the work of Technical Committee TC 69, Applications of statistical methods, of the International Organization for Standardization (ISO). Attention has been particularly directed to ensure the correspondence of the definitions for statistical terms given in this guide with those approved by TC 69 and published in ISO/R 645. 'Statistical vocabulary and symbols. First series of terms and symbols': Part 1 'Statistical vocabulary' and ISO/R 1786 'Statistical vocabulary and symbols. Second series of terms and symbols'.

British Standard Guide to

Application of statistics to rubber testing

Section one. General

1. Scope

This standard has two purposes. One is to arouse the interest of persons engaged in rubber testing in the use of statistical methods, the other is to provide an easy source of reference to the formulae required for the treatment of data, so as to satisfy the general reporting clauses included in most of the rubber testing methods. It is complementary to the British Standards on the testing of rubber.

This standard is not intended to be used for quality control purposes. That subject is dealt with in BS 600 and BS 2564.

2. References

The titles of the British Standards referred to in this standard are listed on the inside back cover.

Section two. General principles

3. A practical problem from rubber testing

Imagine that a rubber manufacturer is accustomed to supplying a certain type of rubber A for a particular job. He knows from experience over a long period of time that this material is of satisfactory modulus, but recently he has developed rubbers B, C and D at an attractive cost per kilogram. As a safeguard against complaints, he decides to test the rubbers before attempting to pass them on to his customers. He carries out tensile tests, ten for each rubber (because he knows how misleading three results can be), under comparable conditions and gets the following results.

Modulus at 300 % elongation (MPa)

A	B	C	D
19.1	19.1	18.4	21.9
18.3	19.2	23.8	20.2
17.6	20.6	21.0	17.5
21.7	19.2	17.0	16.6
17.7	18.7	25.0	11.5
20.7	23.3	18.4	21.8
19.7	22.0	27.5	20.5
19.3	21.8	27.5	18.4
16.3	18.0	20.0	18.8
20.0	19.0	23.0	18.2

The purpose of these tests is to show up differences of modulus of the four rubbers, and probably the first thing the manufacturer will do with the figures is to run his eye up and down the columns. This may be sufficient to answer the questions, but, if the figures are difficult to compare, he will calculate the means or average values of the sets of ten results.

	A	B	C	D
Mean	19.04	20.09	22.16	18.64

In doing this, although he perhaps neither knows nor cares, he has applied statistical methods of analysis to the sets of figures. He has deduced a derived quantity or statistic which gives information in a succinct form about the readings.

What do these mean values tell him? One person may conclude from the four means that rubbers B and C are of higher modulus than A and rubber D is of lower modulus. Another may be more cautious. He knows that any test on rubber is subject to considerable variability (that is why he has done ten tests instead of one) and there is therefore some uncertainty about the mean values. He may conclude that, since the test average of C is 16 % higher than that of A, C is of appreciably higher modulus, and B possibly of higher modulus than A. The result for D is about 3 % lower and, although he may have doubts about the significance of this result, he will perhaps reject rubber D, just to be on the safe side.

The mean value is not enough. The variability of the readings has to be considered. In the unrealistic case where all ten readings on rubber A happened to be 19.04 and those on B 20.09 it could be confidently asserted that B is of higher modulus than A. In practice, some variability is to be expected, and the greater this is the more doubtful does the conclusion become. As the variability increases further, it becomes obvious that there is no reliable evidence of a difference of modulus between the two rubbers.

Ways of assessing variability. A rough impression can be obtained by running over the figures by eye, but conclusions drawn from this procedure are subjective. If the variability has to be expressed in a report the highest and lowest values can be used, or the range, but again the interpretation of these statements is open to doubt. The mean deviation (see glossary, appendix C) is a better expression of variability, but it is of limited usefulness.

The most satisfactory and efficient way of expressing variability is by means of the standard deviation, because it remains more constant than any other measure in repeated sets of figures. The range, for example, may vary relatively more than the standard deviation from one set of readings to another because it depends only on the two extreme values in each set. The standard deviation is more difficult to calculate than the range or mean deviation, but this is more than compensated by its versatility. It is fundamental to statistical analysis.

The coefficient of variation is closely related. It is the standard deviation expressed as a percentage of the mean value. It has the useful property of typifying a material where the standard deviation increases in proportion to the mean value. Under these circumstances the coefficient of variation is constant. This is often the case in rubber testing.

The test portion itself is another variable. Only a small part of the consignment of rubber A was tested and this test portion has been assumed to be representative of the rubber as a whole. If another test portion were to be tested, it is likely that a different mean value of tensile modulus would be obtained and this might lead to different conclusions. When a large series of similar test portions is tested, however, a definite pattern begins to arise. As more and more test portion means are taken, the pattern of the distribution of these means will more closely approach a smooth curve.

On the basis of this phenomenon, statements may be made about a single test portion which take account of their variability.

Why use statistics in rubber testing? In certain simple situations, conclusions may be reached without any reasonable doubt. It should be recognized, however, that personal judgement often plays a large part in arriving at a decision. This may be tempered by experience but, unless statistical methods are used, at least one guess is involved.

Statistics eliminates the guesses. Instead of a forthright statement that the tensile modulus of rubber A is 19.04 MPa (which may be proved wrong by the next sample tested), a conclusion firmly based on statistics might be 'that it is 95 % certain that the true tensile modulus of rubber A lies between 18.05 and 20.03 MPa'.

The question whether it is worthwhile to employ statistics in a particular problem depends entirely upon the importance of the answer. If the most objective conclusion is sought, for scientific or commercial reasons, then statistical methods should be considered.

4. Introduction to statistical methods

4.1 Distribution of results. A collection of values (for example, individual test results) relating to a specific property of the material being tested tends to arrange itself about the arithmetic mean in a manner which may be represented by a distribution curve (such as that illustrated in figure 1). When, as is often the case, the distribution of a large number of values approximates to a particular mathematical law, known as the Normal (Gaussian) distribution, a number of useful calculations can be made, e.g. of the proportion of test readings likely to differ by more than a given amount from the arithmetic mean (see appendix D, table 11 and figure 12).

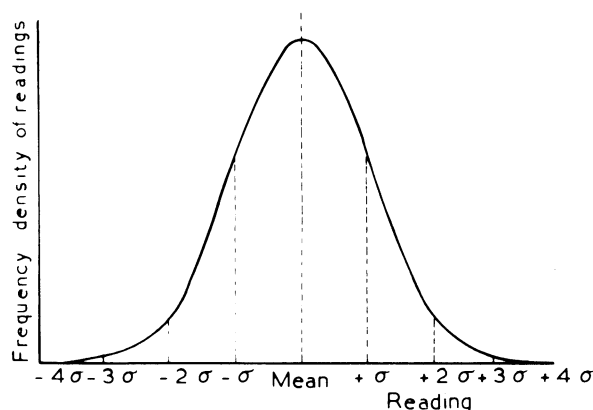


Figure 1. Normal distribution curve

Statistical tests are available to test for the departure from normality for a distribution, and details of some of these, together with the appropriate tables, can be found in *Biometrika tables for statisticians* (see appendix B, ref. 18) pages 61 to 63 and 183 to 184.

In some cases (such as in the Akron abrasion test) the distribution may be found to be markedly asymmetrical and to depart appreciably from the Normal form but, even in such cases, the distribution can often be converted into an approximately Normal form by a simple transformation of the readings, for example, by conversion

of each reading to its logarithm (which has been found effective in the Akron abrasion test). Other simple transformations which have sometimes proved effective with experimental data are the square root or the reciprocal of the readings.

Even where the Normal distribution is not found for the individual readings, the distribution of the means of small groups of readings (≥ 3), such as compose the usual tests on rubber, nearly always approximates to the Normal form.

4.2 Sampling from a population. A set of data obtained in any single investigation can be regarded as a sample from a much larger collection, or population, of results, to which the given data belong. For example, the population might consist of the results of testing every piece of material in a given batch. The usual purpose of carrying out the tests is to estimate the properties of the population distribution from the data provided by the sample. Two properties that it is usually desired to know are the arithmetic mean and the standard deviation, which is a measure of variability.

4.3 Standard deviation (σ). The standard deviation is defined by taking the average, over the whole population, of the squared deviation of each value from the mean, and then taking the square root of this average. This definition of variability serves as a basis for statistical calculations, and various properties of the distribution may be deduced from it.

The frequency of individual readings decreases as they become further from the mean. Thus, for Normal distribution, approximately 1 in 20 readings may be expected to be beyond $\pm 2\sigma$ from the mean, 1 in 80 beyond $\pm 2.5\sigma$ and 1 in 370 beyond $\pm 3\sigma$ (see figure 1).

The distance from the mean that is exceeded by any given proportion of readings can be calculated from the formula $\pm u\sigma$, where u is a tabulated quantity called the standardized Normal deviate.

4.4 Coefficient of variation (v). It is sometimes convenient to express the standard deviation as a percentage of the arithmetic mean, in which case the value so calculated, known as the coefficient of variation, can be used as a measure of the relative variability of sets of readings with different mean values.

4.5 Estimation of the mean and the standard deviation from a sample. An estimate of the population mean is provided by the sample mean, $\bar{x}$, which is simply the arithmetic mean of the results composing the sample.

The best estimate of the population standard deviation is obtained by calculating the sample standard deviation, i.e. by summing the squares of the deviations of individual results from their mean, dividing this sum by one less than the total number of readings, and taking the square root of the quotient.

Thus,

$$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}}$$

where

x is the individual reading;

$\bar{x}$ is the arithmetic mean of the group of readings;

n is the number of readings in the group;

s is the estimate of standard deviation having the same units as x .

The purpose of dividing the sum of squares by $(n - 1)$ (this $(n - 1)$ is called the 'degrees of freedom') rather than by n is to correct a tendency to underestimate the population standard deviation when n is used as the divisor, especially in the case of small-sized samples. Reference should be made to the glossary in appendix C for a description of this.

Note that the symbol Σ denotes 'the sum of . . .'. In this case the sum is that of all individual values of $(x - \bar{x})^2$.

For methods of computation of s see the worked example in clause 5.

The quantity s above is referred to as simply the standard deviation, the prefix 'estimate' being omitted, as understood. It is in several respects the best way of expressing the variability of results in a sample.

The population coefficient of variation, as estimated from the small sample, may be calculated as follows:

$$v = 100 s / |\bar{x}|$$

where $|\bar{x}|$ is the numerical value of $\bar{x}$, regardless of the algebraic sign; that is, v is always positive.

Other measures of 'central tendency' and 'variability' exist. If further statistical processing of the test data is contemplated these are not usually used. They may, however, have application in straight reporting of results and for this reason they are discussed in 4.6.

4.6 Other measures of central tendency. Other mean values, such as the harmonic and geometric means, can be regarded as the arithmetic mean of transformed data and need not be considered further.

4.6.1 Median (middle value). As a statistic this is based primarily on the order of observations and not on their values. It has a larger variance than the arithmetic mean. On the other hand, its value is not affected by an atypical value; moreover the median value may be considered to be more 'typical' than the mean value for skew distributions. In the case of a Normal distribution the mean and median are coincident.

The median may be calculated by arranging the results in ascending or descending order and crossing off the highest and lowest results together until only one (or two values) are left. If two values are left the median may be estimated as the mean of these values e.g.

Set 1. ~~8.2~~ ~~10.1~~ ~~13.7~~ 13.8 ~~14.0~~ ~~14.9~~ ~~14.9~~
median = 13.8

Set 2. 14.7 ~~17.8~~ 20.9 22.3 ~~23.8~~ ~~24.2~~
median = 21.6

4.6.2 Mode (most frequently occurring value). This value has the largest frequency density and is therefore intuitively the most typical of the central values. Its value is, however, difficult to estimate, even if the shape of the underlying distribution is mathematically known.

4.6.3 Mid-range. This value is defined as the average of the highest and lowest values. It has some drawbacks when used as a measure of central tendency since it is based only on the two extreme observations, which are usually the least reliable; furthermore, its efficiency decreases rapidly for larger series of observations.

From a theoretical point of view this is not a very meaningful statistic, especially when taking into account the fact that a large number of theoretical distributions have an infinite range.

4.7 Other measures of variability. The scatter of data can be described according to three principles:

- (a) measures of distance (e.g. the range);
- (b) measures compiled from the deviations of every observation from some central value (e.g. mean deviation);
- (c) measures compiled from deviations amongst all observations (e.g. mean difference).

4.7.1 Range. The range is the numerical difference between the highest and lowest of a set of readings. In a Normal population the average range of a number of measurements is related to the population standard deviation by a factor which depends on the number of readings. An estimate of the standard deviation is given by multiplying the range of a random sample of n observations from a Normal population, by the factor a_n . This is illustrated in table 1, which is an extract from Lindley, D.V. and Miller, J.C.P. *Cambridge elementary statistical tables* (C.U.P.) (tables 6 and 7).

Table 1. Factors for estimating standard deviation from a range of values in a Normal population

n	a_n
2	0.8862
3	0.5908
4	0.4857
5	0.4299
6	0.3946
7	0.3698
8	0.3512
9	0.3367
10	0.3249

For estimating the standard deviation, the range is almost as good as any other method when the number of readings is not larger than about ten and the readings follow a distribution which is not far from the Normal. However, since the exact calculation of the sample standard deviation is relatively quick on ten results or fewer, this would normally be carried out.

With larger numbers of readings, the use of the range is not recommended, except as a check against gross errors in calculating the standard deviation. It is better to use the standard deviation, which utilizes all the information in the data, instead of relying on the extreme readings.

4.7.2 Interquartile range. This is the distance between two values between which the central half of the observations fall; e.g. 50 % of the observations fall within the so-called 'interquartile range'.

Interquartile ranges are only used when very large sample sizes are available; they reveal very little, however, about the way the bulk of the observations is condensed inside this range.

4.7.3 Mean deviation. This is the mean of the absolute value of the deviations around the central value. As it is excluded from more advanced statistical techniques, its usefulness is restricted; nor does it give a more comprehensible measure of the scatter than standard deviation. Obviously the value obtained will depend on the type of central value chosen (mean, median, or mode).

4.7.4 Mean difference. This is the average of the absolute values of the differences of all possible pairs of observations. Again because of the absolute nature of mean difference, this measure of dispersion is excluded from statistical techniques while it virtually offers little more than the standard deviation.

4.8 Unexpectedly high or low results. It is often required to know whether an apparently high or low result in a set of data should be rejected.

As a general principle, a reading should never be rejected unless there are grounds outside the data for so doing. An occasional high or low reading is to be expected, to a varying degree, in the distribution of any rubber property, due to the nature of the rubber itself, and rejection of such readings may lead to a distorted picture of the distribution; the standard deviation, in particular, will be underestimated.

On the other hand, high or low readings are sometimes caused by errors of operation or damaged test pieces, in which case the readings can be legitimately rejected; or again, they may be due to flaws in the test piece, such as thin spots or pin holes. The rejection may then depend on the purpose of the test. For example, in testing the air permeability of rubber sheeting the presence of occasional high readings might be taken to indicate the presence of pin holes in the rubber. If the rubber is being tested as a functional material, then such readings should be retained; but if the air permeability is only required as an index of the rubber's technical properties, possibly in relation to the compounding, then the reading might fairly be rejected.

When a rejectable reading is suspected, and there are no confirmatory indications, statistical criteria outlined by W.J. Dixon (*Processing data for outliers: biometrics*, 1953, 9 (1) 74) can sometimes be used to decide whether or not the reading may be legitimately rejected. An alternative method is the use of Grubb's test described in *Annal. math. stat.* 21 (1) (March 1950).

The problem is approached statistically by considering the sample values to be drawn from a mixed population, consisting of a main population together with a small proportion of values from another, interfering population, which differs from the first in mean, or variance, or in both, so as to yield occasional high (or low) sample values. The object of the test criteria is to remove the more extreme members of the interfering population in order to be able to calculate relatively unbiased estimates of the mean and standard deviation of the main population. It is assumed that both populations follow the Normal distribution law. The definition of the test criteria, and their critical values, are given in table 2. In this table, x_n is the extreme value to be tested, x_{n-1} is the nearest neighbour to x_n , and so on, the readings being arranged in order of magnitude from x_1 to x_n .

If the value of r calculated from the sample data exceeds the tabulated critical value, the extreme reading may be rejected with a fair degree of confidence (see clause 5 for a worked example).

The critical values correspond to a significance level of 5 %. This means that, if the data follow a pure Normal distribution, only one extreme reading in 20 will, on the average, be rejected unnecessarily.

The method should not be used for test methods where there is reason to think that the distribution departs from the Normal form. Data of this type, however, may often be transformed to nearly Normal form (see 4.1).

To summarize, no reading should ever be rejected unless either there is evidence of a definite source of error affecting the reading, or the reading is found rejectable according to the above statistical test; the latter may be applied only when the data are known to follow a Normal distribution law, or have been transformed to this form. Indiscriminate use of the rejection test will result in underestimation of the standard deviation.

Table 2. Criteria for rejecting outlying values by Dixon's test

Number of readings, n	Criterion	Critical values
5 6 7	$r_n = \frac{x_n - x_{n-1}}{x_n - x_1}$	0.642 0.560 0.507
8 9 10	$r_n = \frac{x_n - x_{n-1}}{x_n - x_2}$	0.554 0.512 0.477
11 12 13	$r_n = \frac{x_n - x_{n-2}}{x_n - x_2}$	0.576 0.546 0.521

4.9 Confidence interval for the mean, when the population standard deviation (σ) is known. The means of groups of n readings (for example 5, 10 or 20) may be expected to vary about the true mean in a Normal manner, but they will naturally be less variable than individual readings. The distance from the true mean within which a single observed mean may be expected to lie, with a given probability, is given by:

$$\pm u\sigma/\sqrt{n}$$

where

u is the standardized Normal deviate and has the value 1.96 for a probability of 95 % and

$\sigma/\sqrt{n}$ is the standard error of the mean.

Values of the standardized Normal deviate corresponding to other levels of probability may be obtained from tables of the Normal distribution.

The limits given above refer to the distribution of observed means about a known true mean. By using the same limits in an inverse sense, it is possible to state, with a good degree of confidence, the range, about the observed mean, within which the true but unknown mean lies. When used in this sense, the limits are called 'confidence limits', the degree of confidence being expressed by a probability, which is usually taken as 95 %. Thus the 95 % confidence limits of a population mean are given by:

$$\bar{x} \pm \frac{1.96\sigma}{\sqrt{n}}$$

Confidence limits are the usual methods of expressing the precision of an estimate. The use of 95 % limits ensures that the statement that the true mean lies within these limits will be correct, on the average, in 19 cases out of 20. It will be noted that the precision, or closeness of the limits, depends on the number of readings and their standard deviation and also on the probability level (usually 95 %) selected.

The lines in figure 2 show the distance of the 95 % confidence limits from the mean, for typical small numbers (n) of tests. These lines are based on the above formula, but, to make them of more general application, the distance has been expressed as a percentage of the mean, and plotted against coefficient of variation, in the place of standard deviation.

The limits are expressed as the percentage distance from the sample mean, which is based on n readings; these limits are only applicable when the population standard deviation, σ , is known. Limits in the case where the standard deviation, σ , is estimated by s from the sample are given in figure 4.

4.10 Number of tests required for a given precision of the mean, when the population standard deviation (σ) is known.

The formula $u\sigma/\sqrt{n}$ given in 4.9 may be used to calculate the number of readings required to estimate the mean with a given degree of precision. For example, in a rubber test having a coefficient of variation of 7 %, the number of readings required to estimate the mean within ± 5 % (95 % confidence limits) is given by the equation:

$$(1.96) (7)/\sqrt{n} = 5$$

whence $n = 7.5$.

Therefore eight readings should be sufficient to give the required precision.

4.11 Significance of difference between means of two tests, when the population standard deviations (σ) are known.

In comparing the means of two tests it is necessary to have some criterion to decide whether the difference is likely to have arisen by chance, or indicates a real difference between the means of the populations from which the samples were drawn. This criterion is called the least significant difference.

It is so chosen that, as long as no real difference exists between the means of the two populations, there is only a small probability of it being attained or exceeded. This probability, or 'significance level', is usually taken as 1 in 20, but circumstances may require other level to be used.

When the observed difference between means exceeds the least significant difference it is unlikely that the respective true means are the same. The conclusion that a real difference between them exists will be incorrect in fewer than 1 case out of 20 (if the 5 % level of significance is chosen).

A difference that is less than the least significant value, on the other hand, should be taken to indicate not that the true means are necessarily the same, but merely that the difference is 'insignificant', that is, small enough to escape detection by the given data.

For the general case, where the number of readings in the two tests differ, the least significant difference is given by the following formula:

$$\begin{aligned} &\text{least significant difference between means} \\ &= 1.96\sigma\sqrt{(1/n_1 + 1/n_2)} \quad (5 \% \text{ level of significance}) \end{aligned}$$

where

n_1 and n_2 are the numbers of readings in the two tests, and

σ is the known standard deviation of the population.

For the usual case, where each test contains the same number (n) of readings, the formula becomes:

$$\begin{aligned} &\text{least significant difference between means} \\ &= 1.96\sigma\sqrt{(2/n)} \quad (5 \% \text{ level of significance}) \end{aligned}$$

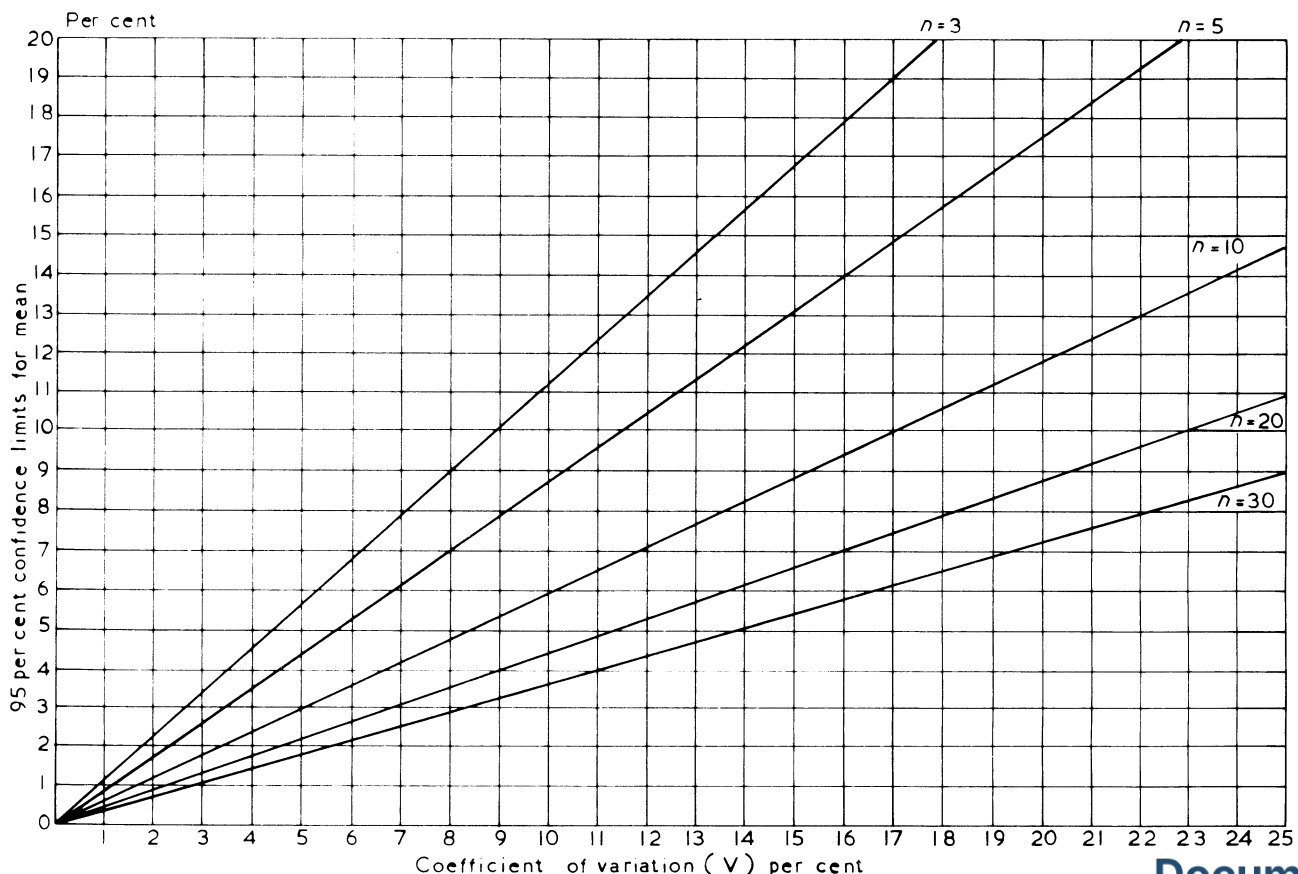


Figure 2. Confidence limits versus true coefficient of variation

The least significant difference, according to the last formula, is plotted in figure 3 for $n = 5, 10, 20$ and 30 readings in each test. As in figure 2 the coefficient of variation is used in place of the standard deviation; the least significant difference is accordingly expressed as percentage of the mean.

When the standard deviation, s , is estimated from the sample, figure 5 should be used.

4.12 Confidence interval for the mean, when the population standard deviation (σ) is unknown. In 4.10 and 4.11 the precision of the mean and the least significant difference between means have been calculated on the assumption that the variability is accurately known; this is not usually the case, however. It is often not sufficient to rely on 'typical' values of variability based on previous data. The precision of a test mean or the least significant difference between two test means should preferably be calculated from the estimate of standard deviation provided by the test readings themselves. This estimate of standard deviation is subject to some error, since the number of readings available is generally small (10 or 20) and, to allow for this error, the foregoing formulae require some modification, as follows.

Confidence limits for the mean $= \bar{x} \pm ts/\sqrt{n}$
where

t is given, for any chosen probability level, by Student's t -distribution (double-sided) (tables of which are included in nearly all the references; an abbreviated version is also given in table 12, appendix D);

s is the estimate of the standard deviation, calculated from the test readings;

n is the number of readings in the group.

Values of t depend on the number of readings, but, in the tables, the different values of t are usually arranged according to a number γ which is called the number of degrees of freedom. In the present case, the number of degrees of freedom is one less than the number of readings, so that the tables should be entered at $\gamma = (n - 1)$.

For calculating 95 % limits, the appropriate values of t , corresponding to typical values of n , are as follows:

$t = 2.78$ for $n = 5$ readings;

$t = 2.26$ for $n = 10$ readings;

$t = 2.09$ for $n = 20$ readings.

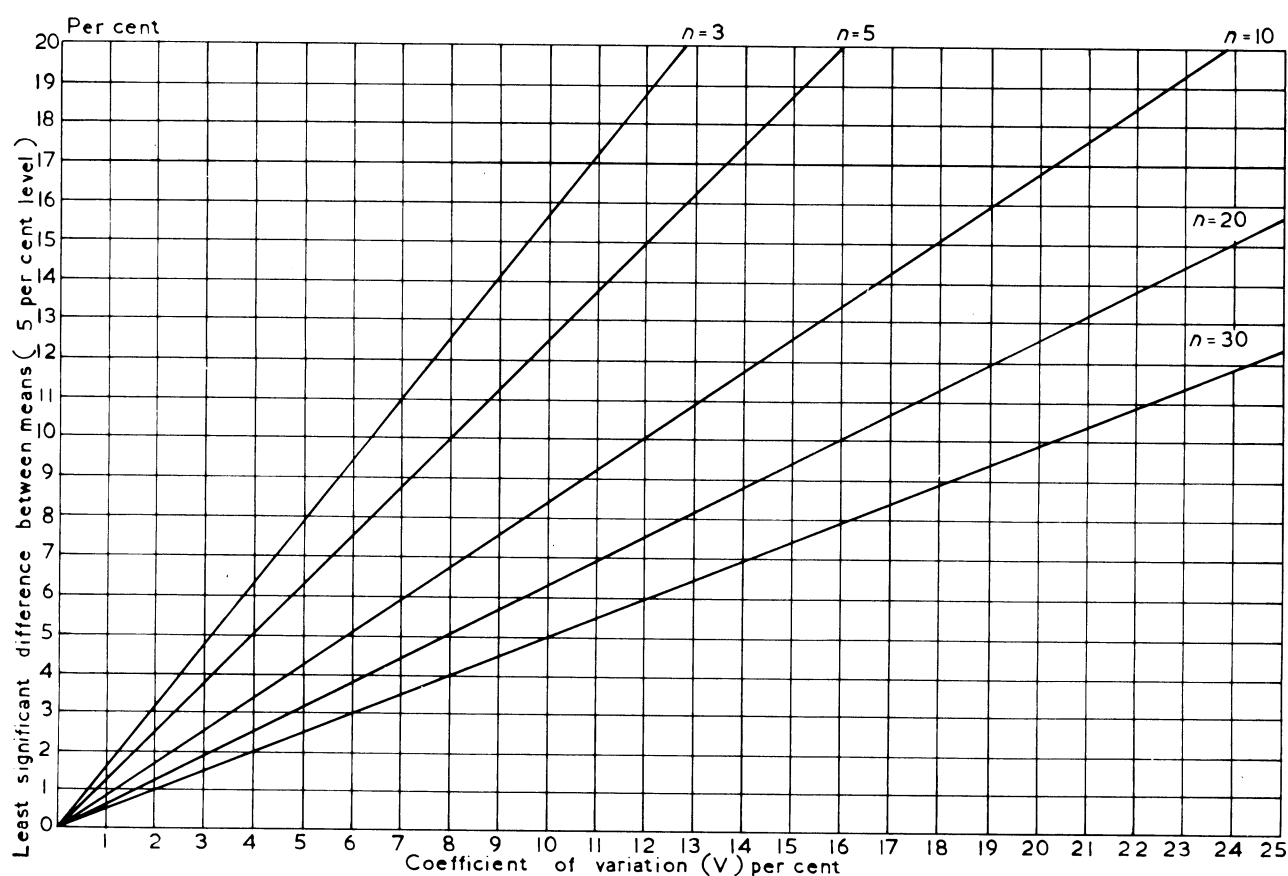


Figure 3. Real difference versus true coefficient of variation

It will be noticed that, owing to the uncertainty in the value of the standard deviation estimated from a small number of readings, the values of t are somewhat larger than the corresponding values of u given by the Normal distribution ($u = 1.96$, see 4.9) and the 95 % limits for the mean are correspondingly wider.

Figure 2 refers to the precision of the mean when the standard deviation is known from independent data, as may be the case in quality control. Figure 4 gives the corresponding information when the standard deviation has to be calculated from the test data, which will be the more common occurrence.

4.13 Significance of difference between means of two tests when the population standard deviations are known (Student's t -test). For the general case, where the number of readings in the two tests differ:

Least significant difference between means

$$= ts \sqrt{(1/n_1 + 1/n_2)}$$

where

t is given by Student's t -distribution (double-sided) (the tables should be entered at $\gamma = n_1 + n_2 - 2$);

s is a pooled estimate of standard deviation, calculated from the two sets of readings (see 4.15 and worked example in clause 5);

n_1, n_2 are the numbers of readings in the two test.

For the usual case, where each test contains the same number (n) of readings:

Least significant difference between means

$$= ts \sqrt{(2/n)}$$

(The t tables should be entered at $\gamma = 2n - 2$).

For a 5 % level of significance, typical values of t are:

$t = 2.31$ for $n = 5$ readings in each test;

$t = 2.10$ for $n = 10$ readings in each test;

$t = 2.02$ for $n = 20$ readings in each test.

Figure 5 gives the corresponding information to that in figure 3.

Before this particular significance test can be applied, the validity of obtaining a pooled estimate of the standard deviation should be checked. This is done by applying a test for significance of difference between variances, described in 4.14. If the variances are not significantly different, then the pooled estimate of the standard deviation is derived as shown in 4.15, and the calculated value of the pooled standard deviation (s) is used in the above formulae.

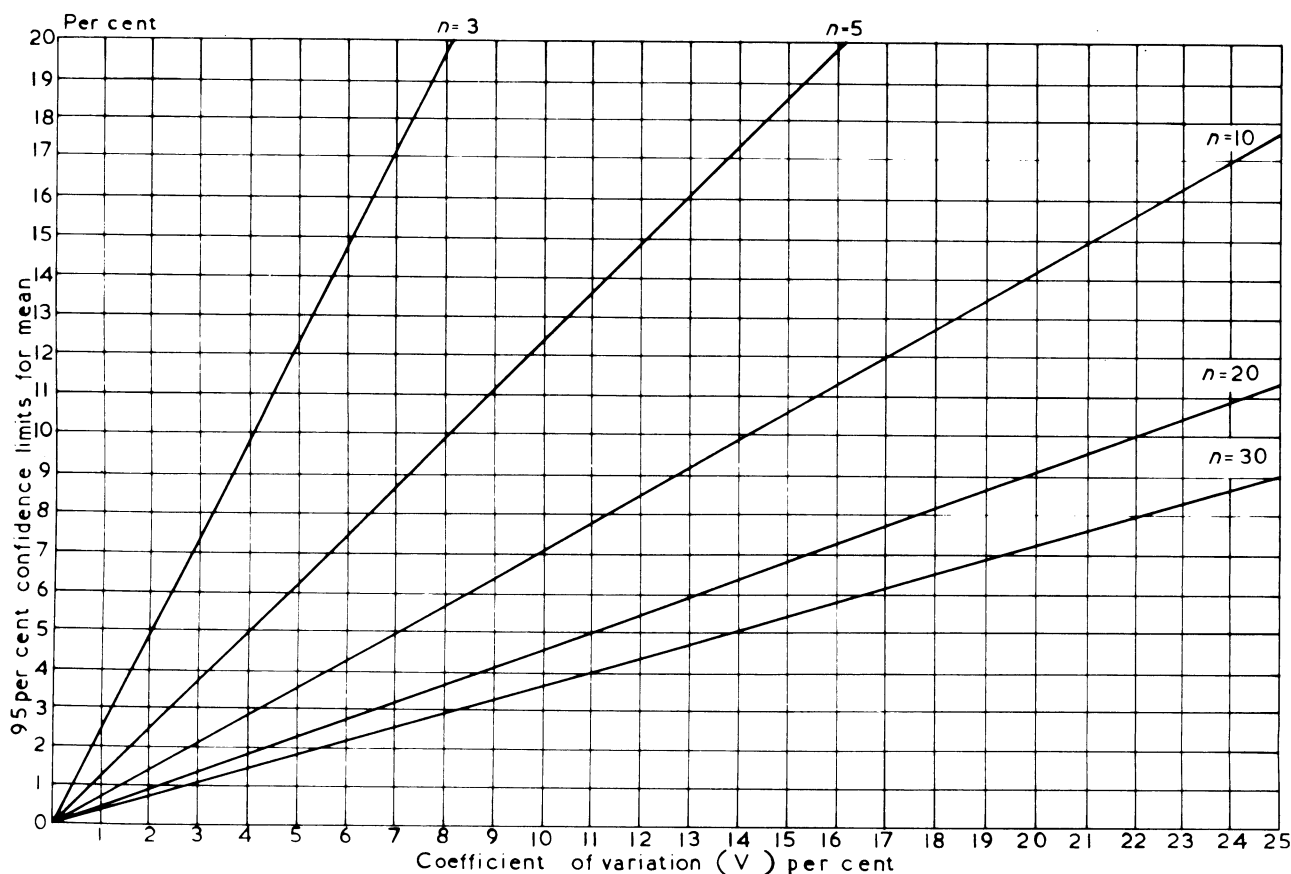


Figure 4. Confidence limits versus estimated coefficient of variation

If the F-test of 4.14 is significant, indicating that the population variances are really different, but it is still essential to compare the means, then a very approximate test would be still to combine the variances, calculate s , and test for a significant difference in the means in the same way as above. *However it is stressed that the test in this case would only be a rough guide.*

4.14 Significance of difference between variances. On some occasions it is not a comparison of mean value which is of major importance, but a comparison of variances. For example, if two operators are each carrying out a number of repeat determinations on a particular sample it will often be useful to assess whether there is any real difference in the repeatability of the two operators; i.e. whether the results of one operator are much more closely grouped around his mean value than the results of the other operator are around his.

(It may also be of interest to test whether there is any bias between the two operators, i.e. whether the means are significantly different; but this is not considered in this subclause.)

Let the consideration be confined to the following specific situation:

Operator 1 carries out n_1 determinations and the variance of the n_1 results is s_1^2 .

Operator 2 carries out n_2 determinations and the variance of the n_2 results is s_2^2 .

The ratio $\frac{s^2 \text{ (larger)}}{s^2 \text{ (smaller)}}$ is calculated, where s^2 (larger) is the larger of s_1^2 and s_2^2 , and s^2 (smaller) is the smaller of s_1^2 and s_2^2 .

Appendix D, table 13 is then entered at the appropriate point, where

γ_1 is the number of degrees of freedom on which s^2 (larger) has been based;

γ_2 is the number of degrees of freedom on which s^2 (smaller) has been based.

If the ratio $\frac{s^2 \text{ (larger)}}{s^2 \text{ (smaller)}}$ is greater than the appropriate value of the tabulated F-distribution (appendix D, table 13) then it can be stated with 95 % confidence that the population variance σ_L^2 (of which s^2 (largest) is an estimate) is greater than the population variance σ_S^2 (of which s^2 (smallest) is an estimate).

If the ratio $\frac{s^2 \text{ (larger)}}{s^2 \text{ (smaller)}}$ is less than the appropriate F-distribution value, then we have not sufficient evidence to refute the hypothesis that the two population variances are equal.

This test (known as the F-test or 'variance ratio' test) could also be applied in the comparison of the variances of various methods of analysis.

For strict correctness this type of test should be applied to the variances s_1^2 and s_2^2 mentioned in 4.15.

The ideas which have been discussed so far are sufficient for the solution of many common types of problem, and in clause 5 they are applied to an example. Clause 5 also contains hints and precautions on computing which will lighten the arithmetical work.

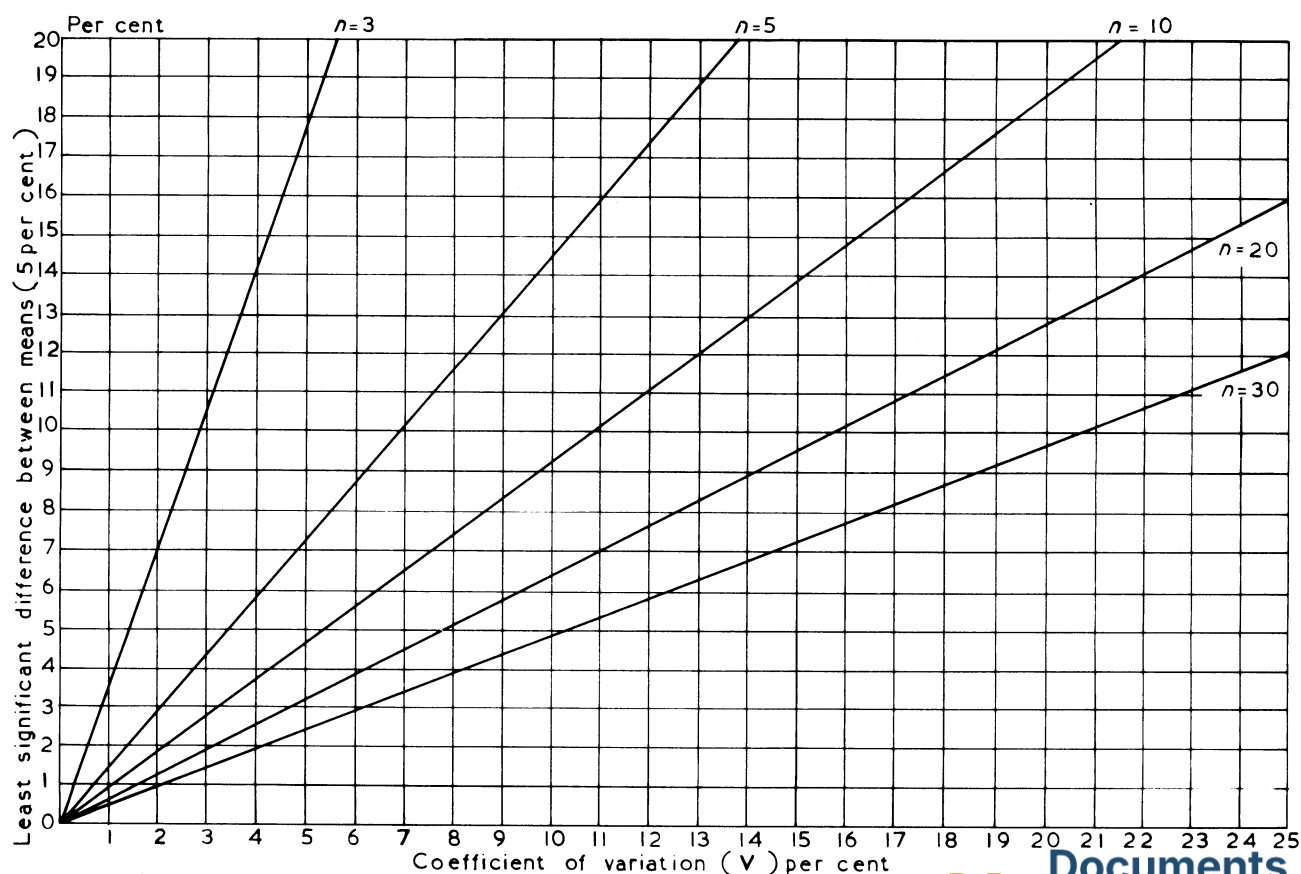


Figure 5. Real differences versus estimated coefficient of variation

4.15 Pooled estimate of standard deviation from two sets of readings. The 'pooled' estimate should be formed only if the difference in variances as described in 4.14 is not significant. When two sets of readings are available, the best 'pooled' estimate of standard deviation is given by the following formula, rather than by the mean of the two separate estimates:

$$s = \sqrt{\frac{\Sigma(x_1 - \bar{x}_1)^2 + \Sigma(x_2 - \bar{x}_2)^2}{n_1 + n_2 - 2}}$$

When the numbers of readings (n_1 and n_2) in the two tests differ, more weight is given, by this formula, to the test containing the greater number of readings.

When the standard deviations have already been calculated separately for each sample, giving s_1 and s_2 for example, it is seen that the last formula gives the 'pooled' estimate of standard deviation as:

$$s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

Sometimes the variability will be very different in the two sources of material supplied and pooling of the two estimates may not be justified. It would be beyond the scope of this guide to deal here with the procedure to be followed in this but it may be remarked that, provided that the two samples are of nearly the same size, the error arising from pooling will never be serious.

4.16 Single sided significance tests and confidence intervals. A 'double sided significance test' is one in which significant differences in either direction are of importance. A situation in which only a difference in one specific direction is of importance requires what is called a 'single-sided significance test'. The type of test to be used depends completely on the practical application, and is always decided before the experimentation.

In the situations described in 4.11 and 4.13 the general question being asked is: 'Is there a significant difference between . . . ?' In that case a double-sided significance test was appropriate. However, in practical applications a typical question might be: 'Is the mean of this batch significantly less than the specification?' In such a situation it is of no importance if the batch mean is significantly greater than the specification; only if it is significantly below the specification value need action be taken, and so a single sided test should be used.

(a) If the population standard deviation (σ) is known the following applies.

When a single sided test is appropriate and the test is being carried out at the 5 % significance level, the 1.96 which occurs in the formulae in 4.11 for 'least significant difference between means', is replaced by the value of 1.64. Also a single-sided 95 % confidence interval for the population mean can be calculated. This can be expressed as 'it is 95 % certain that the population mean is greater than $\bar{x} - \frac{1.64\sigma}{\sqrt{n}}$,

or as 'it is 95 % certain that the population mean is less than $\bar{x} + \frac{1.64\sigma}{\sqrt{n}}$,

depending on which is the statement relevant to the practical application.

(b) If the population standard deviation (σ) is unknown the following applies.

When a single-sided test is appropriate, the t -distribution values referred to in 4.13 would be replaced by the corresponding figure from the single-sided t -distribution tables. For a 5 % level of significance and a single sided test, the appropriate t -value would be

$t = 1.86$ for $n = 5$ readings in each test

$t = 1.73$ for $n = 10$ readings in each test

$t = 1.68$ for $n = 20$ readings in each test.

Also a single-sided 95 % confidence interval for the population mean can be calculated. This can be expressed as 'it is 95 % certain that the population mean is greater

than $\bar{x} - \frac{ts}{\sqrt{n}}$,

or as 'it is 95 % certain that the population mean is less

than $\bar{x} + \frac{ts}{\sqrt{n}}$,

again depending on which statement is relevant to the practical situation. The value of t in these confidence intervals would be that selected from the 5 % column of the single-sided distribution, with the appropriate degrees of freedom.

5. Application of statistical tests to the example in clause 3

The figures in the example in clause 3 can now be treated by the statistical methods developed in clause 4.

The standard deviation is calculated by the formula in 4.5.

It is not necessary to calculate and list the individual deviations ($x - \bar{x}$) and their squares in order to estimate the standard deviation. Computation is simplified by the use of the identity:

$$\Sigma(x - \bar{x})^2 = \Sigma x^2 - (\Sigma x)^2/n,$$

where

Σx^2 is the sum of squares, and

$(\Sigma x)^2/n$ is the correction factor,

so that the estimate of standard deviation becomes:

$$s = \sqrt{\frac{\Sigma x^2 - (\Sigma x)^2/n}{n - 1}}$$

The application of this formula will be clear from the worked example. The computing proceeds by the following steps.

(a) Arithmetical work may be reduced by the use of an arbitrary origin, chosen at or near the general level of the readings. In the worked example below, this is done by subtracting 20 from each reading before proceeding with the calculations. The standard deviation is unaffected by this procedure.

(b) Compute the sums Σx and Σx^2 , remembering that x may now be the distance of the reading from the arbitrary origin. For the squares of the readings, the use of tables is recommended (for example, Barlow's tables of squares) if a calculating machine is not available. In some cases, as in parts of the example, the numbers are small enough to be squared in the head.

(c) Before proceeding further Σx and Σx^2 should be checked. Rather than merely repeating what has already been done, with the risk of repeating the same error,