Edited by John Newman • Ban Seng Choo

Advanced Concrete Technology

Testing and Quality

Advanced Concrete Technology

Advanced Concrete Technology

Constituent Materials	ISBN 0 7506 5103 2
Concrete Properties	ISBN 0 7506 5104 0
Processes	ISBN 0 7506 5105 9
Testing and Quality	ISBN 0 7506 5106 7

Advanced Concrete Technology

Testing and Quality

Edited by

John Newman
Department of Civil Engineering
Imperial College
London

Ban Seng Choo
School of the Built Environment
Napier University
Edinburgh

ELSEVIER
BUTTERWORTH
HEINEMANN

AMSTERDAM BOSTON HEIDELBERG LONDON NEW YORK OXFORD
PARIS SAN DIEGO SAN FRANCISCO SINGAPORE SYDNEY TOKYO

Butterworth-Heinemann
An imprint of Elsevier
Linacre House, Jordan Hill, Oxford OX2 8DP
200 Wheeler Road, Burlington MA 01803

First published 2003

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

Library of Congress Cataloguing in Publication Data
A catalogue record for this book is available from the Library of Congress

ISBN 0 7506 5106 7

For information on all Butterworth-Heinemann publications visit our website at www/bh/com

Typeset by Replika Press Pvt Ltd, India
Printed and bound in Great Britain

10

Statistical analysis techniques in ACT

Stephen Hibberd

10.1 Introduction

Concrete is notable as a material whose properties can vary widely depending on the choice and proportions of aggregates, cement, water, additives etc., together with the production technique. An associated feature is that even with a desired (target) mix the inherent variability in the materials and the production process will inevitably result in a final product that differs from the target requirement. Of course, provided the concrete remains within specified tolerances on key attributes then the product is acceptable. Effective management of concrete therefore must include a quantitative knowledge of the key attributes, monitoring techniques, decision methods, their limitations and an ability to interpret the measured values. Statistical techniques are consequently used extensively to understand and compare variations between concrete batches, to modify and control the production of concrete and to form the basis of Quality Control and Quality Assurance.

Within a short chapter it is impossible and not appropriate to provide an in-depth coverage of statistical theory and the underpinning ideas from probability theory. An emphasis is on statistical techniques that are exploited and applied explicitly to current practical circumstances in ACT such as trends and errors, estimation of parameters, checking test results, mix design, compliance and quality control. Initially each section will concentrate on providing fundamental understanding and competence of the background techniques that will be required and use will be made of relevant statistical tables and formulae.

10.2 Overview and objectives

The theory is divided into coherent sections that will provide a theoretical background to the applications of procedures covered in other chapters as follows:

- **Sample data and probability measures** This section aims to consolidate knowledge on the calculation of sample statistics and their relevance together with an understanding of probability typified by a normal distribution. Objectives are to:
  - consolidate terminology and calculation of sample statistics;
  - introduce concepts and measurements of probability;
  - evaluate probabilities using the normal distribution.
- **Sampling and estimation** Variations associated with the process of sampling are addressed and quantitative measures introduced. Two statistical distributions, the *t*-distribution and the *F*-distribution, are introduced and their use to provide estimates of key population parameters explained. Objectives are to:
  - understand the concept of sampling to provide an estimate of a key (population) values;
  - calculate of the precision of estimates for large sample sizes using the normal distribution;
  - evaluate estimates and precision for small sample sizes using the *t*- and *F*-distributions;
  - calculate confidence intervals and understand their application in constructing control charts.
- **Significance tests** The concept of decision making based on sample data is covered and applied to the comparison of mean and variances of key parameters. Objectives are to:
  - understand hypothesis testing in using sample data to test the validity of a statement;
  - understand the relationship between the significance level and critical values in tests;
  - evaluate a sample mean with a target (population) value;
  - compare target means or variances from two sets of sample data.
- **Regression models** An examination of possible relationships between linked parameters and the derivation of useable functional relationships will be covered in this section. Objectives are to:
  - understand the concept of correlation as a measure of association between sets of data;
  - calculate a 'least-squares' linear regression line;
  - understand and calculate correlation coefficients and residuals.
- **Statistical formulae and tables**
  A collection of some relevant formulae and tables used in ACT are provided.

10.3 Sample data and probability measures

10.3.1 Random variation

Within the physical world many natural quantities are subject to an amount of random variation in their formation and consequently provide variation in any materials for which they are a constituent part. In concrete technology random variations also occur due to changes in processing, for example due to minor chemical inconsistencies, mixing time variation and small water quantity changes. Thus, even with the most careful of measuring constitutive quantities, natural variations will occur in the properties of the resultant material. Random variations will also affect measurements of all quantities *X,* say, as these are subject to errors; if careful measurements are repeated or different instruments used then the values of *X* will lie close to some precise value but some discrepancy must be expected. Information and subsequent analysis on such variations can be obtained from a study of data collected from laboratory or on-site tests. Random variation is then often plotted in histogram form to identify principal characteristics such as central tendency, variability and shape. For quantitative analysis then an associated probability distribution for characterizing the variation in *X* is sought.

Random variations must be clearly distinguished from systematic variation that may arise from some planned change to the process or some time-varying process. For example, in the case of comparing the increased 7-day strength of concrete samples, as a result of increasing additive, then a plot of strength against quantity of additive will follow an anticipated (systematic) curve, while variation about this plot would be random error.

10.3.2 Sample data

Statistics involves dealing with information from collected data. Clearly it is important that a sufficient quantity and the correct type of information is gathered to make predictions reliable. These more advanced topics are covered later in this chapter; initially we concentrate on the key ideas associated with sampling and the representation and interpretation of data.

The most common type of 'experiment' involves taking a sample from a population, i.e. a selection of items from a whole. Ideally, the whole population would be studied, but this may be impractical for two main reasons:

- Expense – the population may be too large or testing each item may be expensive.
- Destructiveness – testing may require dismantling or running to destruction.

Generally, some form of estimation, decision or prediction is made affecting the whole population by the analysis of data from just a sample. Care must be therefore be exercised to distinguish between a

- population statistic – some value associated with the whole population (i.e. usually the quantity we need to estimate);
- sample statistic – some value obtained from a sample (i.e. a value obtained from only a part of the population).

10.3.3 Representation of data

Statistical data, obtained from surveys, experiments or any series of measurements are often so numerous that they are virtually useless unless condensed or reduced to a more suitable form. A necessary first step in any engineering situation is an investigation of available data to assess the nature and degree of variability of the physical values. Sometimes it may be satisfactory to present data 'as they are', but usually it is preferable or necessary to group the data and present the results in tabular or graphical form. For subsequent calculations or direct comparisons then some quantitative measures of data may be required. An unorganized list of data values is not easily assimilated. However, there are numerous methods of organization, presentation and reduction that can help with data interpretation and evaluation.

Histograms

Given a set of recorded data values it is useful to group the frequency of occurrence within suitable intervals. Useful histograms can be based on

- absolute frequencies – the number of data values within each interval;
- relative frequencies – the proportion of total number of data values within each interval;
- cumulative absolute frequency – the running total of absolute frequencies;
- cumulative relative frequency – the running total of relative frequencies.

Typical histograms are shown in Figures 10.1(a) and 10.1(b) corresponding to representations of data obtained from laboratory test for determining tensile strength of a concrete mix as given in example *Case 1*. The distribution of data values are made relative to intervals of width 20 and centred at mid-point values (class marks) as displayed. Figure 10.1(a), shows a typically characteristic random variation of many quantities in ACT; sample values are found to vary around some 'central location' and with some element of 'spread'.

Case 1 A sample of test results of the splitting tensile strength on 50 concrete cylinders

320	380	340	410	380	340	360	350	320	370
350	340	350	360	370	350	380	370	300	420
370	390	390	440	330	390	330	360	400	370
320	350	360	340	340	350	350	390	380	340
400	360	350	390	400	350	360	340	370	420

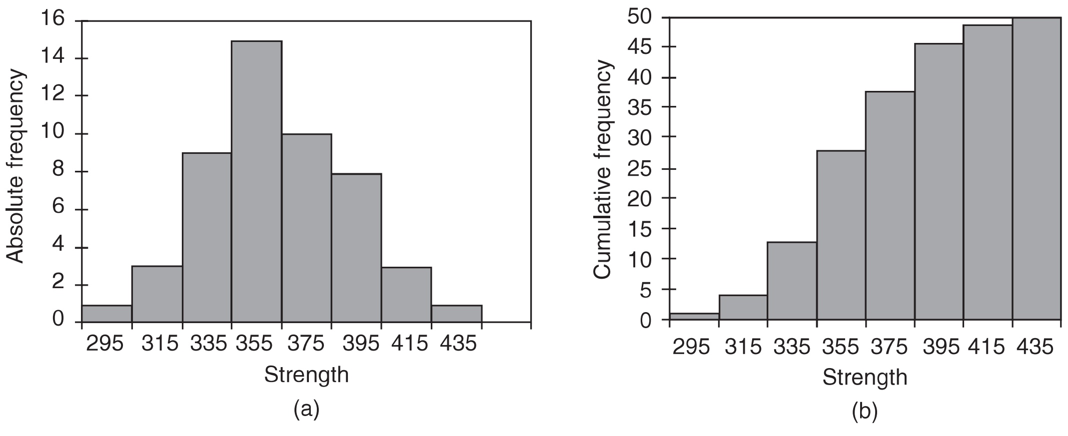


Figure 10.1 (a) Frequency histogram, (b) Cumulative frequency histogram.

Scatter diagram (Scattergram)

In measuring systematic variations of quantities X and Y, say, the recorded data values will be in the form of data pairs (x_i, y_i). An example is given in *Case* 2 that provides a set of data pairs related to measured deflections from a loaded concrete beam. When plotted on a Cartesian plot (scattergram), such as Figure 10.2, any functional relationship between the two variates may become more evident. The plot in Figure 10.2 perhaps suggests a linear relationship exists between deflection and load, with variations from an exact straight-line relationship reflecting the error associated with the data.

Case 2 The measured values of beam deflections y_i against applied loads x_i

i	1	2	3	4	5	6	7	8	9	10
x_i	100	110	120	130	140	150	160	170	180	190
y	45	52	54	54	62	68	75	75	92	88

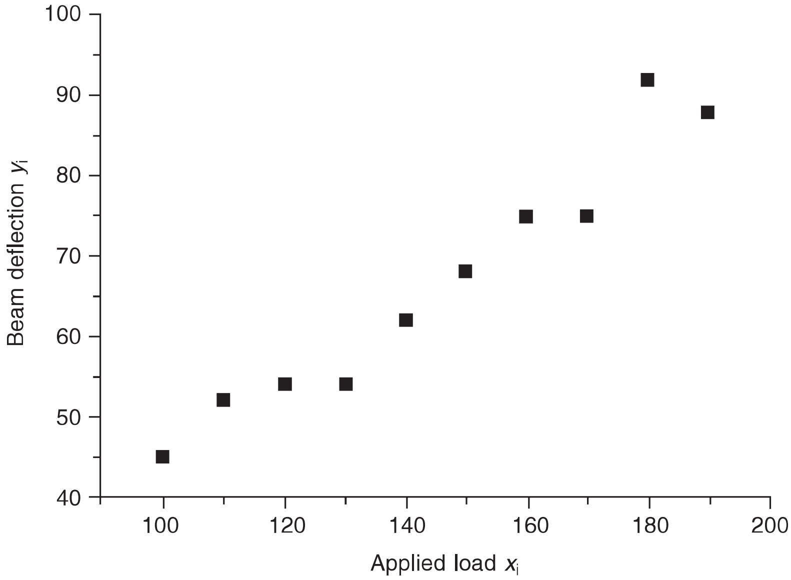


Figure 10.2 Scattergram of beam deflection data.

10.3.4 Quantitative measures

For calculation, decision making and comparison purposes it is useful to obtain standard analytic measures of the data characteristics. The two principal measures are first, that of location, given by a single representative value locating the 'centre' of the data, and second, a measure of spread or variation of data values, usually relative to the 'centre'. Given n data values, labelled x_i, say, $i = 1, n$, a measure of data values location based on the arithmetic mean of all n data values x_i is the sample mean $\bar{x}$ given by

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

A measure of the spread in data values from their mean value is given by the sample variance s^2 given by

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

and the derived value $s = \sqrt{s^2}$ termed the sample standard deviation; the latter has the same dimensions as the data values and is more frequently used in operational formulae.

The above measures are widely used but simpler measures may sometimes be more appropriate in some cases. An alternative measure for the 'centre' of the data is the median, determined from data arranged in ascending order as the middle value (odd number of data points) or the average of the middle two data values (even number of data values). The use of ordered values can be extended to give quartiles – i.e. data values divided in quarters or even finer divisions of tenths termed deciles. A simple measure of spread is given by the difference between the largest and smallest data values and called the sample range, some care is needed with this measure as it can be severely affected by a rogue data value. *Case 1* gives example data values used to obtain values of the sample mean = 363.8, sample variance = 832.2 and sample standard deviation = 28.8. By comparison the sample median = 360 and the sample range =140.

10.3.5 Population values

The above measures for mean and variance can also be applied to population values but with a significant minor alteration, for small samples in particular. For populations of a random variate X with finite number of discrete data values N, the (population) mean μ_x and the (population) variance σ_x^2 are given by

$$\mu_x = \frac{1}{N}\sum_{i=1}^{N} x_i \quad \text{and} \quad \sigma_x^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

If no ambiguity exists with the variate referenced then these may be simply written as μ and σ^2. The difference in the dividing factor in the formula between the sample variance s^2 and population variance σ^2 arises because the mean μ is exact whereas in a sample of data values then the sample mean $\bar{x}$ provides only an estimate for the actual mean μ. The population standard deviation is defined by $\sigma = \sqrt{\sigma^2}$ as might be expected.

10.3.6 Probability

A probability is a proportional chance of a particular occurrence. Perhaps the most quoted mathematical source of probability is the throwing of a single common dice. The outcome of one throw is one of six numbers 1 to 6, each with an equal chance of occurrence, i.e. the probability of obtaining a particular score is one in six – expressed as 1/6. Mathematically an event A, say, will have

$$\text{Probability of } A = P(A) = \frac{\text{Number of outcomes which result in } A}{\text{Total number of outcomes}}$$

If events are not equally likely then an appropriate modification with appropriate weightings needs to be used; the associated link between probability values and the possible outcomes is called a probability distribution. It is evident that a probability is a numerical measure that lies between 0 and 1; if the outcome is impossible then it will have a probability of 0 while a certainty will take a value of 1. In applying probability concepts to concrete technology it is usually not so straightforward as the example above to enumerate the probability values, but these often exist as a result of past experience and expertise or on the basis of laboratory testing. Probability thus provides a theory in which the uncertainty is known (or assumed known) to follow a specific probability distribution. In dealing with probability theory we will be looking at assigning proportional chances to random events through studying their associated probability distributions.

10.3.7 Probability functions

The quantities of most interest in ACT, for example compressive strength, tend to take continuous values x, of the variate X say, and lie within a range $0 < x < \infty$. However, the chance of realizing values will be associated with some underlying probability measure. In Figure 10.1(a), a distribution of tensile strengths as obtained from testing is displayed within a histogram and indicates that values falling within different intervals had differing frequencies, i.e. different probabilities of occurrence. The corresponding histogram Figure 10.1(b) identifies how the occurrences are accumulated for increasing values of the variate X.

To generate quantities in *Case 1* consistent with probability measures then frequencies need to be scaled relative to the total sample size, i.e. to graph relative frequencies rather than absolute frequencies as shown in Figures 10.3(a) and 10.3(b). In this case the area under the histogram in Figure 10.3(a) will sum to unity and the corresponding cumulative sum in Figure 10.3(b) will approach the total value of unity.

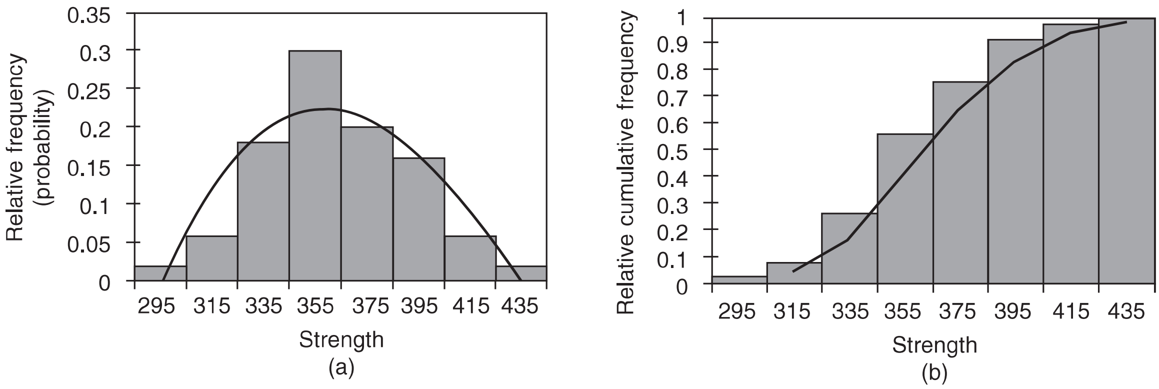


Figure 10.3 (a) Probability density, (b) probability distribution.

Although *Case 1* corresponds to a finite sample, it is straightforward to recognize that with increasing sample size, then further but smaller class intervals can be readily defined, and the histogram columns will mark out a more continuously varying area beneath characterizing continuous curves, such as those shown illustrated in Figure 10.3. This process correspondingly defines two functions in dealing with continuous variates: first, the probability density function (pdf) $f(x)$ showing the variation of probabilities for values x of the variate X and second, the cumulative distribution function (cdf) $F(x)$ that provides a summation (integration) of the probability measures for increasing values of x. These curves characterize the underlying random processes and play a crucial role in evaluating probabilities and quantifying the statistical analysis. Fortunately, a number of common distributions exist that match the characteristics of processes found in ACT situations, but even then, values are not readily obtainable from simple analytic functions but need to be evaluated from tables of values or reference to a computer-based package such as Excel or a specialized statistical package.

The cdf is given in terms of the integral of the pdf and, conversely, the derivative of the cdf is the pdf i.e.

$$\frac{dF(x)}{dx} = f(x) \quad \text{and} \quad F(x) = \int_{\text{least } x}^{x} f(x)\,dx$$

Following from our earlier work, the total probability is 1, which corresponds to the total area under the pdf curve that provides an essential constraint on the possible forms for $f(x)$. The cdf is adding probability values in the direction of increasing possible values x, so the curve will increase continually to a final value of unity. The probability value for a random variate X to lie between two values a and b, say, is the area under $f(x)$ between $x = a$ and $x = b$; evaluation is obtainable from integration of the pdf using the cdf as follows:

$$\mathrm{P}(a < X < b) = \int_{x=a}^{x=b} f(x)\,dx = \int_{x=0}^{x=b} f(x)\,dx - \int_{x=0}^{x=a} f(x)\,dx = F(b) - F(a)$$

These ideas are crucial to evaluate relevant probability values from values of the cdf as given in tables.

10.3.8 Expected values

The pdf gives detailed information on the range of values a variate might take and their appropriate chance of occurrence. It remains useful to calculate some key quantities associated with these distributions such as the most likely (mean) value or the expected measure of spread. In this instance, all possible values are available and we are dealing with population quantities. Formally, these are called expected values and can be formulated for any function $\phi(X)$, to give a weighted average value and defined by

$$E\{\phi(X)\} = \int_{\text{all } x} \phi(x) f(x)\,dx$$

Expected values of the variate X or any powers of X are termed moments; the most important are:

(i) $\phi(X) = X$, to give the population mean μ;
(ii) $\phi(X) = (X - \mu)^2$, to give the population variance σ^2.

For any pdf function then these can be evaluated and are used within the subsequent statistical analysis in comparing the predicted population values with the observed sample values.

10.3.9 Normal distribution

The normal distribution (or Gaussian distribution) is the single most important and widely known distribution in engineering and plays a central role in the theory associated with concrete technology. The pdf is given by

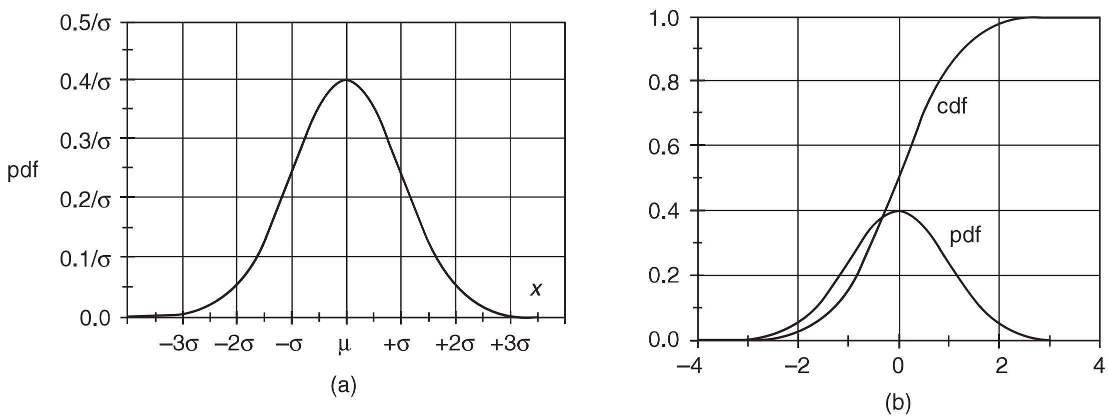


Figure 10.4 (a) pdf of $N(\mu, \sigma^2)$, (b) pdf and cdf for $N(0, 1)$.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \quad \text{with} \quad -\infty < x < \infty$$

This involves two characterizing parameters corresponding to the mean μ and standard deviation σ of the distribution and accordingly the distribution is denoted by $N(\mu, \sigma^2)$ for convenience.

The pdf has a bell-shaped curve as shown in Figure 10.4(a) which is symmetric about the mean μ, a maximum value of $1/\sigma\sqrt{2\pi}$ and with a shape that rapidly decays to zero for values away from the mean. While $f(x)$ has non-zero values for all positive and negative values of the variate x, it is negligible for most practical purposes when x is more than a distance 3σ from μ. The corresponding cdf is shown in Figure 10.4(b) and although it does not have a simple functional formulation it can be evaluated numerically. Figure 10.4(b) shows both the pdf and cdf of $N(0, 1)$, i.e a distribution with mean zero and variance unity.

The distribution for $N(0, 1)$ is particularly important as it provides a base calculation for any normal distribution and is consequently well tabulated. Statistical Table 10.1 gives a table of values shown in Figure 10.4(b).

The normal distribution is often used for its relative simplicity combined with a proven ability to provide accurate quantitative information when used in appropriate circumstances. It also has the useful 'addition' property that if $X_1 \sim N(\mu_1, \sigma_1^2)$ and $X_2 \sim N(\mu_2, \sigma_2^2)$ then the random variables

$$X_1 + X_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2) \quad \text{and} \quad X_1 - X_2 \sim N(\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2)$$

i.e. that new quantities constructed from addition of normal variables will be normal and the resulting parameters can be readily calculated.

10.3.10 Calculation of probability values (from standard tables)

Probabilities associated with the standardized normal variate $N(0,1)$ can be obtained directly from a table of values for the cdf $F(z)$, and some algebraic manipulation. The basic approaches are displayed in the following examples illustrated in Figure 10.5, where the required probability measure corresponds to evaluating the area of the marked portions of the pdf; values are determined from Statistical Table 10.1. As is typical with tables for symmetric distributions, the values corresponding to only positive values of the variate X are explicitly displayed. However, corresponding negative values are readily obtained from the symmetry of the pdf as:

$$f(-z) = f(z) \quad \text{and} \quad F(-z) = 1 - F(z)$$

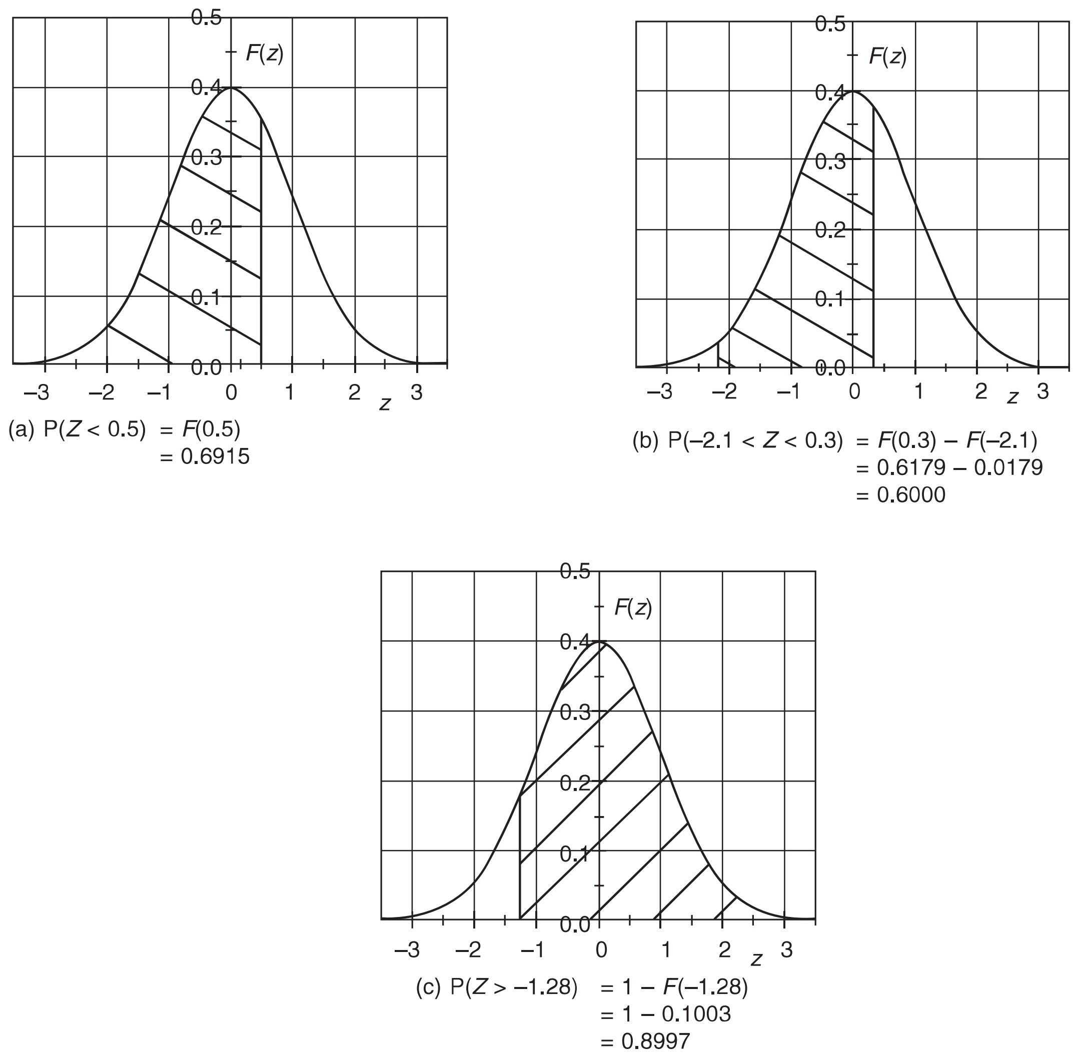


Figure 10.5 Evaluations of probability values from the standardized normal distribution $N(0, 1)$.

10.3.11 Standardized normal variate

The normal distribution has a number of special properties, one of which is that it has a simple scaling rule. This allows all calculations for probabilities from any normal distribution to be calculated from the single set of numerical values of $N(0, 1)$, called the standardized normal distribution, graphed as Figure 10.4(b) and values given in Statistical Table 10.1.

For any normal variate X, then an associated standardized variate Z is defined by

$$Z = \frac{X - \mu}{\sigma}$$

and can be shown to have the property that $Z \sim N(0, 1)$. Thus in practical calculation, probability values for a distribution $N(\mu,\sigma^2)$ can be rewritten and then evaluated in terms of $N(0, 1)$. As an example consider $P(X < a)$, where $X \sim N(\mu, \sigma^2)$. Then by simple algebra

$$P(X < a) = P\left(\frac{X - \mu}{\sigma} < \frac{a - \mu}{\sigma}\right)$$

i.e,
$$P(X < a) = P\left(Z < \frac{a - \mu}{\sigma}\right)$$

where
$$Z = \frac{X - \mu}{\sigma}$$

is the standardized variate and the value of this probability is readily obtained from the distribution $N(0, 1)$. In a similar way, any calculation of probabilities from a general normal distribution involving a random variate X can be recast into an equivalent calculation in terms of a standardized variate Z and calculation found in terms of a single set of normal values – the standardized normal values. Such values are either tabulated or held in any computer statistics package.

10.3.12 Example

A specification for the cement content of pavers is specified as 16.9 per cent from contractors. The mean and standard deviation of the cement contents of 50 pavers were tested as 17.2 per cent and 1.8 per cent, respectively. Contractors would be concerned if many pavers had cement contents below 15 per cent. Assuming the cement content follows a normal distribution, estimate the number of pavers that would be below standard.

Let X = cement content of pavers, then $X \sim N(17.2, 1.8^2)$. The probability of a single paver with cement content less than 15 per cent is calculated as

$$P(X < 15) = P\left(\frac{X - 17.2}{1.8} < \frac{15 - 17.2}{1.8}\right) = P(Z < -1.22)$$

Calculation of the probability is reduced to finding an associated probability of the standardized variate Z. Using Statistical Table 10.1, the probability is given by $P(Z < -1.22) = F(-1.22) = 1 - 0.8888 = 0.1112$. Thus the approximate total number of pavers = $50 \times 0.1112 = 5.56$.

The calculated number is approximately 6 pavers but there are several sources of error. Variations will exist in the measurements particularly as cement content analyses are not exact measurements of the actual contents of selected pavers. In statistical terms also, the sample mean and standard deviation used in the calculation are sample values and not population values. Assessment of accuracy of such estimates is covered in later sections.

10.3.13 Critical values

Many probability distributions identify that the variate can take a wide range of possible values but most of the probability is assigned within a relatively small range, e.g. for a standardized normal distribution approximately 95 per cent of the probability lies within a range $-2 < z < 2$. Correspondingly, untypical values of the variate are identified from the wider extremes of the pdf often called the 'tails' of the distribution. The extent of these regions depends on the specified probability to be designated in the tails. For computational purposes it is useful to identify these critical regions and enumerate these limits, called critical values, for given probabilities, called significance levels, at the extremes of the pdf. Illustrations of possible critical regions and associated critical values are shown for $N(0, 1)$ in Figure 10.6 for a significance level of $\alpha = 0.05$.

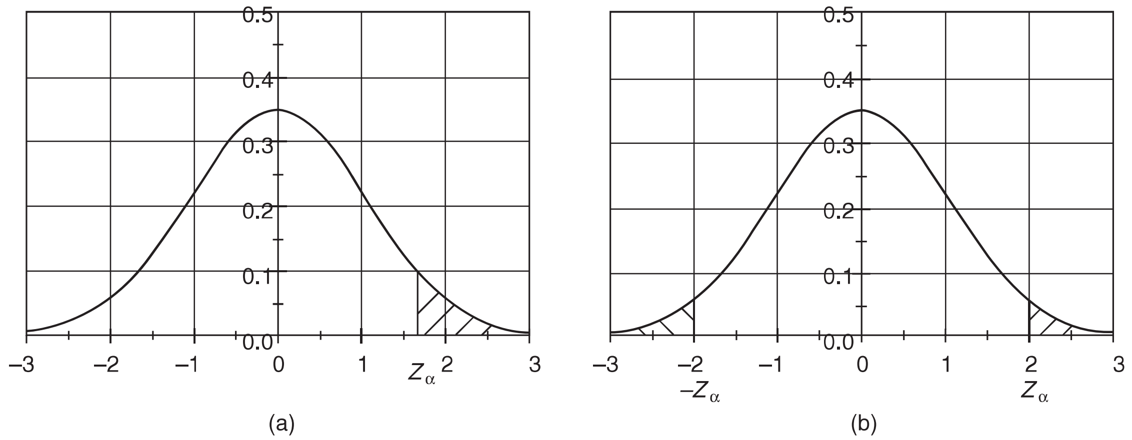


Figure 10.6 (a) Right-hand tail (α = 0.05), (b) right- and left-hand tails (α = 0.025 in each tail).

(a) *One tail* – Figure 10.6(a) shows the right-hand tail region for $N(0, 1)$ which contains a probability (significance level) of 0.05 (5 per cent). The critical value z_α is readily determined from Statistical Table 10.1 as $z_\alpha = 1.645$. For this symmetric distribution the critical values for an analogous left-hand tail are readily obtained as the negative of right-hand critical values.

(b) *Two tails* – Figure 10.6(b) shows the critical regions corresponding to a total probability (significance level) of 0.05 (5 per cent) divided equally into two tail regions, an amount $\alpha = 0.025$ is in *each* tail; the corresponding critical values are $z_\alpha = 1.960$ and $-z_\alpha = -1.960$.

Generally for two-tailed critical regions the probability is divided equally into both tails and significance level taken as 2α. Critical values z_α can be readily found from the cdf values of the distribution but for convenience it is useful to compile a separate table of critical values such as shown in Table 10.1 for the standardized normal distribution corresponding to typical significance levels $\alpha = 0.05$ (5 per cent), 0.025 (2.5 per cent) or 0.01 (1 per cent) in the right-hand (upper) tail.

Table 10.1 Critical value z_α for $N(0,1)$ with significance levels α (right-hand tail)

α	0.25	0.1	0.05	0.025	0.01	0.005	0.001
z_α	0.675	1.282	1.645	1.960	2.326	2.576	3.090

10.4 Sampling and estimation

10.4.1 Sample statistics

Statistical Inference is concerned with using probability concepts to quantitatively deal with uncertainty in obtaining representative values and making decisions. The basis is to obtain samples (from a population) to analyse and infer properties of the whole population. For example, to obtain the 'true' (i.e. population) average potential strength of concrete in a structure, one would need to put all the concrete for this structure into cubes and test them! Clearly this is not desirable or feasible, so an alternative is to take a number of random samples and obtain an estimate of the potential strength from the sample. For a given mix, the sample mean from testing the compressive strength of five concrete cubes might be calculated and taken as the strength of the whole mix. However, for more detailed analysis we might wish to know quantitatively the possible error with such a small sample size and to what level this might be improved by taking a larger sample. In this section details are provided of the underlying theory, probability distributions and how these are applied to use sample values to infer a value or parameter associated with the population. This may involve giving a range of values, called a confidence interval, consistent with a specified probability.

The underlying concept is that typically a collection of sample values X_i will be used to form a statistic Z, formed by some appropriate combination of sample values to provide an estimate of a population value. Each of the quantities X_i will take random values determined by its own distribution, but in estimation it is distribution of Z, the sample statistic, which is required. The most important statistics for use in ACT will be the sample mean $\overline{X}$ as an estimator for the population mean μ and the sample variance s^2 as an estimator for the population variance σ^2.

10.4.2 Large-sample statistics (normal distribution)

The most widely assumed distribution for a sample mean $\overline{X}$ is that it is normal or can be well approximated by a normal distribution. This is because many natural phenomena tend to vary symmetrically around some mean value and with variations that fall off rapidly from some mean value. Furthermore, use of the normal distribution for the sample mean when large samples are involved is justified by the following mathematical result (*Central Limit Theorem – CLT*). The result is exact for the case where sampling occurs for variables X_i that can each be assumed to vary as normal a normal distribution.

If n random samples are taken from a population with mean μ and standard deviation σ then the sampling distribution of $\overline{X}$ the sample mean will be approximately normal with mean μ and standard deviation $\sigma/\sqrt{n}$, the approximation improving as n becomes larger,

i.e. $\overline{X} \sim N\left(\mu, \dfrac{\sigma^2}{n}\right)$ or, in terms of the standardized variate,

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

This indicates explicitly that 'on average' the sample value $\overline{X}$ will predict the population value μ and that the accuracy will be characterized by an associated standard error, more explicitly written as $\sigma_{\overline{X}} = \sigma_X/\sqrt{n}$. Thus, the standard deviation of the sampling distribution $\sigma_{\overline{X}}$ is affected by both the standard deviation of each sample value σ_X but also with the number of sample values n. The effect of sample size can be assimilated graphically as shown in Figure 10.7 which shows the sampling distribution Z from a distribution with $\mu = 5$ and $\sigma = 1$ for increasing sample sizes $n = 5$, 25 and 100. In each case, the mean value of Z is centred around $\mu = 5$ but the probability associated with any individual test shows that the possibility of recording a sample value as inaccurate as 0.5 or greater from the population value is feasible for $n = 5$, unlikely for $n = 25$ but negligible for $n = 100$. Evaluating these probabilities using the normal tables gives respective probability values 0.2628, 0.0124, 0.0000. It also illustrates that the standard error $\sigma/\sqrt{n}$ decreases only relatively slowly, as $1/\sqrt{n}$, with increasing sample size n.

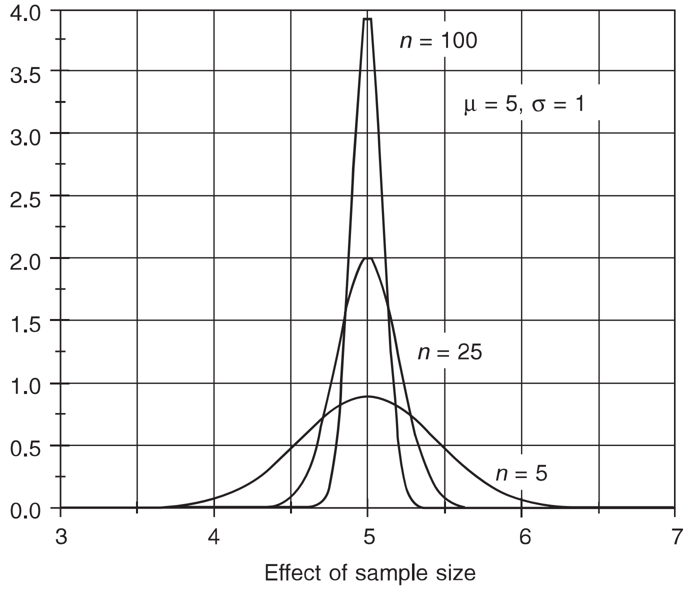


Figure 10.7 pdf of sampling distributions with $n = 5$, 25, 100.

The CLT result is useful provided a large enough sample is taken, often $n > 30$ is usually good enough, but in practice the population standard deviation σ is also unknown and needs to be approximated by the sample standard deviation s; this result remains a good approximation but for sufficient accuracy we may need an increased sample size to compensate of $n > 80$.

10.4.3 Small-sample statistics (*t*-distribution)

The statistical analysis provided in the previous section gives underpinning theory that can be applied in ACT but for practical purposes it is not feasible to always have large sample sizes. When a sample size n is not large then the distribution for the sample mean $\overline{X}$ is no longer accurately approximately by a normal distribution and a more appropriate distribution is the t-distribution. For much of the practical testing required in ACT then the appropriate statistical distributions is the *t*-distribution. The form of the probability distributions (pdf) of $t_{[2]}$ and $t_{[5]}$ as shown in Figure 10.8 compared to the pdf of the standardized normal distribution $N(0, 1)$.

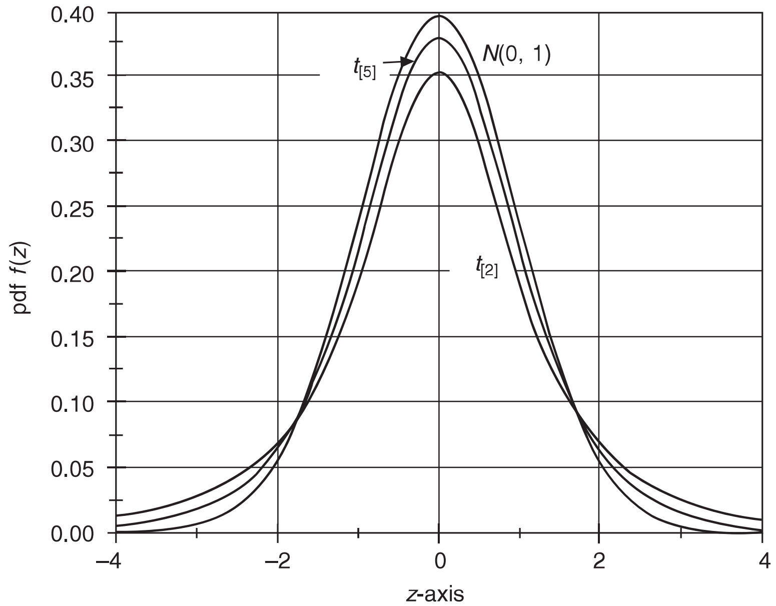


Figure 10.8 Comparison of *t*-distribution with the normal distribution.

The sampling variate for the population mean and corresponding sample distribution is defined by

$$t = \frac{\overline{X} - \mu}{s/\sqrt{n}} \sim t_{[n-1]}$$

This variate is similar to the expression from using the CLT except that it incorporates directly the sample standard deviation s to replace the usually unknown population value σ. The characterizing sample distribution is one of a family of *t*-distributions selected by a parameter $v = n - 1$, called the number of degrees of freedom. The formula for the curve of the distribution is complicated but values are tabulated in the same way as the normal distribution. Important characteristic properties of the distribution are:

- the *t*-distribution is symmetric (only positive values are usually tabulated);
- less peaked at the centre and higher probability in the tails than the normal distribution;
- a marginally different distribution exist for each value of v;
- as v becomes large, the *t*-distribution tends to the standardized normal distribution $N(0, 1)$, $(v = \infty)$;
- values are obtained from *t*-tables although a restricted set of critical values often suffice.

A selection of critical values $t_{\alpha;v}$ for the *t*-distribution, corresponding to those for the normal distribution ($V = \infty$), is given in Table 10.2. A more complete listing of cdf values is given in Statistical Table 10.2.

Table 10.2 Upper critical values $t_{\alpha;v}$ of t-distribution $t_{[v]}$ for one-tailed significance level α

v	1	2	3	4	5	6	7	8	9	10
α = 0.05	6.31	2.92	2.35	2.13	2.02	1.94	1.89	1.86	1.83	1.81
α = 0.025	12.71	4.30	3.18	2.78	2.57	2.45	2.36	2.31	2.25	2.23
α = 0.01	31.82	6.96	4.54	3.75	3.36	3.14	3.00	2.90	2.82	2.76

v	12	14	16	18	20	25	30	60	120	∞
α = 0.05	1.78	1.76	1.75	1.73	1.72	1.71	1.70	1.67	1.66	1.64
α = 0.025	2.18	2.14	2.12	2.10	2.09	2.06	2.04	2.00	1.98	1.96
α = 0.01	2.68	2.62	2.58	2.55	2.53	2.49	2.46	2.39	2.36	2.33

10.4.4 Confidence intervals

Given sample values $X_1, X_2, \ldots X_n$ then using a suitable statistic Z an estimate for a population value can be obtained from a sample value, but importantly an indication of the accuracy of this estimate may also be required explicitly. Knowledge of the distribution of the sample statistic can be used to determine an interval within which the population value might lie with a specified probability. Such a prescribed probability is called the confidence level, c say, and the resulting interval the confidence interval. The confidence level may be expressed directly as a probability, e.g. 0.95, but is often expressed as a percentage, i.e. confidence level of 95 per cent.

As an example of the technique, consider obtaining the confidence interval for the population mean μ from sample values. In this instance the sample statistic Z is usually denoted by t and is given from earlier by

$$t = \frac{\overline{X} - \mu}{s/\sqrt{n}} \sim t_{[v]}, \quad \text{where } v = n - 1$$

Sample data will provide values for $\overline{X}$, s and v. We first seek to determine an interval $-t_{\alpha;v} < t < t_{\alpha;v}$, say, that has an associated probability of 0.95 for the variate t. As the total probability is 1, then determination of values $t_{\alpha;v}$ is identical to determining the two-tailed critical values as shown in Figure 10.6(b) with a probability $(1 - c)$ distributed between each tail. With $c = 0.95$ then the corresponding critical values correspond to obtaining critical values associated with a significance level of $\alpha = 0.025$ in each tail. Hence the critical values are given by $t_\alpha = t_{\alpha;v}$ and determined from Table 10.2. Example values are:

$$n = 10\ (v = 9); \qquad t_{\alpha;9} = 2.25$$

$$n = 6\ (v = 5); \qquad t_{\alpha;5} = 2.57$$

$$n = 100\ (v = 99); \quad t_{\alpha;\infty} = 1.98$$

With n =10, the confidence interval is

$$-2.25 < \frac{|\overline{X} - \mu|}{s/\sqrt{10}} < 2.25$$

and can be rearranged as

$$\overline{X} - 2.25\,\frac{s}{\sqrt{10}} < \mu < \overline{X} + 2.25\,\frac{s}{\sqrt{10}}$$

to provide a confidence interval for the population mean μ with a confidence level of 95 per cent, once sample values for $\overline{X}$ and s are substituted.

Example

A sample of 32 concrete cubes from a certain mix were crushed and the average strength was 30 Nmm^{-2} with a standard deviation of 5 Nmm^{-2}. What is the 95 per cent confidence interval for the strength of the population mean for this type of mix?

The mix strength μ is determined from sample values $n = 32$, $\overline{X} = 30$, $s = 5$. The significance level in each tail is $\alpha = 0.025$, $v = 31$ and critical values are determined from Table 10.2 as $t_{0.025;31} = 2.04$ (nearest table values used) and hence a 95 per cent confidence interval is evaluated as

$$30 - 2.04 \times \frac{5}{\sqrt{32}} < \mu < 30 + 2.04 \times \frac{5}{\sqrt{32}}$$

to provide a 95 per cent confidence interval for mix strength as $28.20 < \mu < 31.80$.

10.4.5 Control charts

Confidence intervals can be exploited to provide the basis of control charts to monitor production values through ongoing sampling. For large sample sizes, critical values of the normal distribution can be used to determine corresponding confidence intervals for a population mean μ as

$\dfrac{|(\overline{X} - \mu)|}{\sigma/\sqrt{n}} < 1$ with probability 0.68, i.e. $\overline{X} - \sigma/\sqrt{n} < \mu < \overline{X} + \sigma/\sqrt{n}$ with 68 per cent confidence;

$\dfrac{|(\overline{X} - \mu)|}{\sigma/\sqrt{n}} < 2$ with probability 0.95, i.e. $\overline{X} - 2\sigma/\sqrt{n} < \mu < \overline{X} + 2\sigma/\sqrt{n}$ with 95 per cent confidence;

$\dfrac{|(\overline{X} - \mu)|}{\sigma/\sqrt{n}} < 3$ with probability 0.998, i.e. $\overline{X} - 3\sigma/\sqrt{n} < \mu < \overline{X} + 3\sigma/\sqrt{n}$ with 99.8 per cent confidence.

Figure 10.9 illustrates that confidence intervals can be used to obtain predetermined levels, within which sample mean values should lie with a given confidence. For example, values should lie within the lines marked ‘action’ with 95 per cent confidence. Thus, successive sample values from an ongoing process can be monitored and compared to the expected population value. Sample values will naturally vary from the population value but if recorded values lie outside set levels of confidence then increasingly strong indication is provided that the production system requires attention. Application of this concept can be further developed to provide Shewart and CUSUM charts as discussed in Chapter 9.

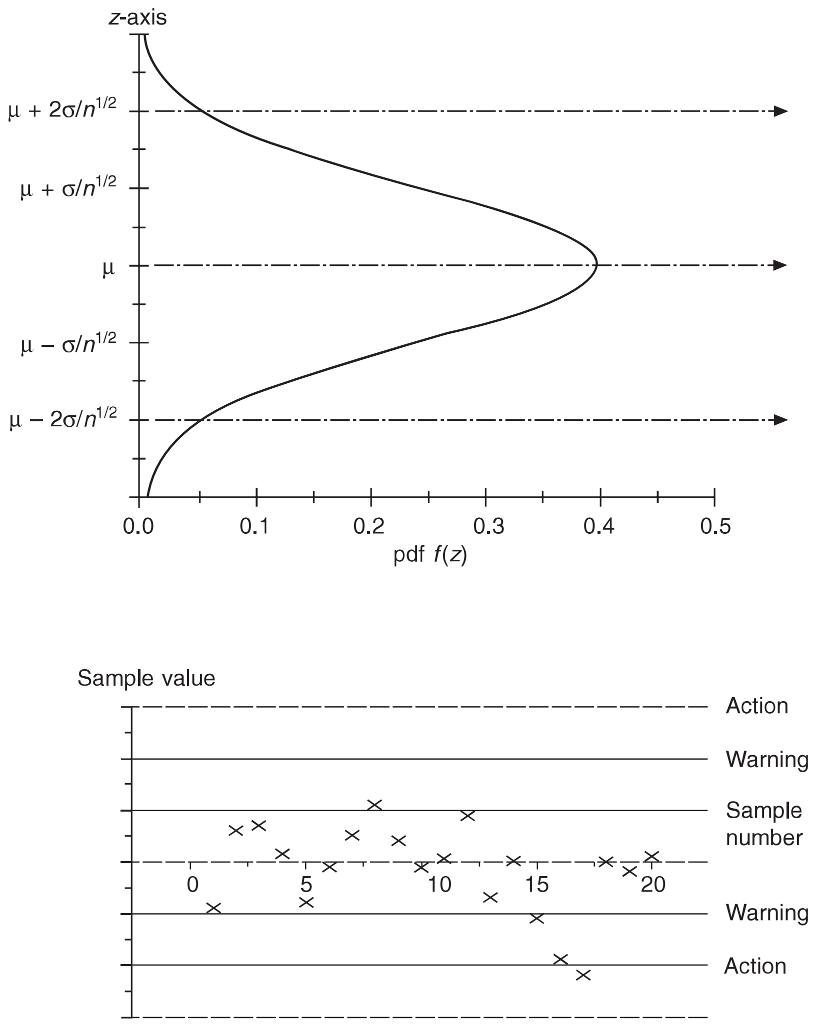


Figure 10.9 Use of confidence intervals as basis of Control Chart.

10.4.6 Comparison of means

An important technique is to be able to compare the population means between two, possibly competing, processes. This might be to ascertain, for instance, whether changing a mix formula will provide an actual increase in compressive strength. The practical difficulty is that a mean value, calculated for each process, will be obtained from sampling and so will be affected by random variation. Consequently, any difference may be accounted for entirely by natural variation and not indicating any changes within the underlying (population) values. Similarly, any measured differences in the sample values may be an over-estimate or an under-estimate of the the effect on the population values. To account for sampling then it is useful to identify a confidence level with any calculation of differences between population values.

In the case of comparing the population means μ_1 and μ_2 from separate processes X_1 and X_2 with large sample sizes n_1 and n_2 then from the CLT applied to each process the separate sample statistics for the sample means $\overline{X}_1$ and $\overline{X}_2$ are

$$\overline{X}_1 \sim N(\mu_1, \sigma_1^2/n_1) \quad \text{and} \quad \overline{X}_2 \sim N(\mu_2, \sigma_2^2/n_2)$$

involving each individual process standard deviations σ_1 and σ_2. The 'addition' property of the normal distribution can be used to obtain a sample distribution of the difference as

$$\overline{X}_1 - \overline{X}_2 \sim N(\mu_1 - \mu_2,\ \ \sigma_1^2/n_1 + \sigma_2^2/n_2)$$

i.e. that the variation measured from the difference in sample means is described by a normal distribution centred around the difference in population means. This result also highlights the importance of the standard error (standard deviation) of the sampling distribution for the comparison of means from two large populations, namely

$$\sigma_{\overline{X}_1 - \overline{X}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

For calculation then the appropriate standardized normal variate is

$$Z = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} \text{ and distributed as } Z \sim N(0,\ 1).$$

As previously in dealing directly with the CLT, the population standard deviation σ is generally unknown and must be approximated by the sample standard deviation s. In a similar way as earlier, this process can be adopted with little error for large sample sizes. For the practical case of comparing the population means from two general processes X_1 and X_2, irrespective of sample size, with sample sizes n_1 and n_2, then a modification to the sample statistic is possible that leads to a sample statistic involving a t-distribution. The sample statistic is

$$t = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{s\sqrt{1/n_1 + 1/n_2}} \quad \text{where} \quad s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

and $t \sim t_{[v]}$ with $v = n_1 + n_2 - 2$.

This result also highlights the general result for the standard error of the sampling distribution for the comparison of means from two populations, namely

$$\hat{\sigma}_e = s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \quad \text{where} \quad s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

Example

Data from two mixes were tested. From the first mix, 30 cubes were tested and found to have a mean strength of 38 Nmm^{-2} and a standard deviation of 3 Nmm^{-2}. The second mix provided 40 cubes with a mean strength of 36 Nmm^{-2} and a standard deviation of 2 Nmm^{-2}. Obtain a 95 per cent confidence interval for the difference in mix strengths.

Using the notation above, for the first mix (X_A) then $n_A = 30$, $\overline{X}_A = 38$, $s_A = 3$

and for the second mix (X_B) then $n_B = 40$, $\overline{X}_B = 36$, $s_B = 2$.

The parameter values for the t-distribution are $\overline{X}_A - \overline{X}_B = 2$, $\nu = 68$ and $s = 2.48$.

With a confidence level of 95 per cent then a corresponding two-tailed critical region is defined by

$| t | < 2.00$ following reference to Statistical Table 10.2 for the critical values for $t_{0.025;68}$

i.e. $$\left| \frac{2 - (\mu_A - \mu_B)}{2.48\sqrt{1/30 + 1/40}} \right| < 2.00 \text{ with a probability of 0.95.}$$

This gives a confidence interval for the difference in mix strengths as $0.80 < (\mu_A - \mu_B) < 3.2$, i.e. that the strength of mix A is stronger than the strength of mix B by a value of between 0.8 Nmm^{-2} and 3.2 Nmm^{-2}.

10.4.7 Comparison of variances

In comparing two processes then it may be useful to compare the amount of the underlying variation induced by the random elements of each, i.e. to compare the population variances. To confirm if given samples from a population X_1 and from a population X_2 are consistent in having equal population variances $\sigma_1^2 = \sigma_2^2$ then a statistic is available as

$$F = \frac{s_1^2}{s_2^2} \text{ which is distributed as an } F\text{-distribution, } F_{[v_1, v_2]}$$

where $v_1 = n_1 - 1$, $v_2 = n_2 - 1$ are parameters to define each distribution.

The distribution for the F-distribution depends on two degrees of freedom v_1, v_2 and has a general shape as shown in Figure 10.10

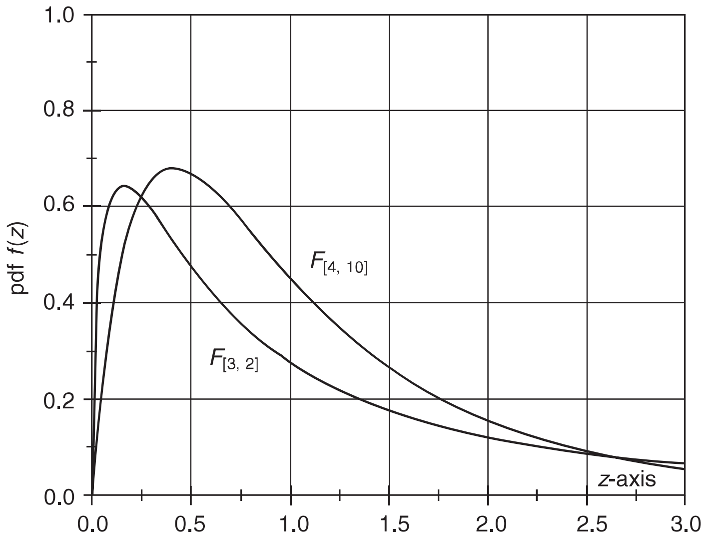


Figure 10.10 Typical shape of an *F*-distribution.

Important properties of the F-distribution are:

- the F-distribution is not symmetric;
- marginally different distributions exist for each pair of value v_1, v_2
- critical values are obtained from F-tables (see Statistical Table 10.3).

Within typical use, the requirement is to use a two-tailed test, i.e. to evaluate lower and upper critical values c_1 and c_2, say, corresponding to a significance level 2α, from

$$P(F < c_1) = \alpha \text{ and } P(F > c_2) = \alpha$$

In practice, it is usual to use only the upper critical value $c_2 = F_{\alpha;\nu_1,\nu_2}$, corresponding to a probability value α in the upper tail, by ensuring the sample statistic is taken as

$$F = \text{largest sample variance/smallest sample variance } (>1 \text{ automatically})$$

and readily accomplished by designating the sample data values corresponding to the larger sample variance as the numerator and the smaller sample variance data for the denominator. If required, values for the lower critical value c_1 are obtainable from Statistical Table 10.3 as $c_1 = 1/F_{1-\alpha;\nu_2,\nu_1}$; in this calculation the upper part of the distribution has probability $(1-\alpha)$. As an illustration, with $\alpha = 0.05$, $\nu_1 = 6$, $\nu_2 = 4$ then the upper critical value $c_2 = F_{0.05;6,4} = 6.16$ and the lower critical value $c_1 = 1/F_{0.95;4,6} = 1/4.53 = 0.22$. An example of the use of the F-variate is given in the next section.

10.5 Significance tests

10.5.1 Hypothesis testing

As part of monitoring manufacture or supply of components we may need to test the validity of a statement (or hypothesis) relevant to a population value by analysing a sample. Consider the illustrative example of a manufacturer of poker vibrators that claims that their product has an average life of 500 hours. Results from monitoring a sample of 36 such vibrators showed that the average life was 450 hours with a standard deviation of 150 hours. Does this disprove the manufacturer's claim?

In this case we are looking to evaluate the validity of the manufacturer's claim to within what might be regarded as a reasonable probability. Clearly, due to the variability in poker use it would not be anticipated that all pokers would last exactly 500 hours or even that a sample of 36 would have a mean value of 500 hours. If the mean value were 495 hours, say, then it might be suspected that the manufacturers claim was upheld while a value of 300 hours would raise significant concerns. It is perhaps 'reasonable' that a sample value of 450 hours is consistent with a target (i.e. population) lifespan of 500 hours, the difference being accountable to sampling variations. To provide quantitative measures to help then a statistical procedure termed 'hypothesis testing' is available that is linked to a stated probability value considered as 'unreasonable' – the significance level. In this example the random variables are

$$X_i \equiv \text{lifetime of each poker } i,\ i = 1, 2, \ldots, 36$$

and the stated (population) mean lifetime $\mu = 500$.

Given the sample of pokers, we test the hypothesis that $\mu = 500$. This is called the null hypothesis and denoted as H_0 (i.e. H_0: $\mu = 500$) and is characterized as giving a specific value in order to determine an appropriate *test statistic*. Statements such as $\mu < 500$ or $\mu > 500$ for the null hypothesis are not specific enough to formulate a subsequent analysis but do provide possible alternatives to accepting the null hypothesis. If the data is conclusive enough to reject H_0 then we must be accepting an alternative hypothesis H_1, say, which generally will affect the decision and so must always be clearly identified.

In the example there are two obvious choices for H_1

(i) $H_1 : \mu \neq 500$ – the poker manufacturer may find this favourable since
$\mu < 500$ means goods are under specification
$\mu > 500$ means specification could be upgraded.
(ii) $H_1 : \mu < 500$ – the consumer is interested in under-specification.

The sample data will be used to devise a test statistic in order to decide whether to accept or reject H_0. In this case it is an estimator for the population mean that is required as given in terms of the sample mean $\overline{X}$. The most appropriate sample statistic will be

$$t = \frac{\overline{X} - \mu}{s/\sqrt{n}}$$

with population mean μ and have a sample distribution $t \sim t_{[n-1]}$.

In Case 1 the hypothesis H_1 would be selected over the null hypothesis H_0 if $\overline{X}$ is sufficiently greater than 500 or sufficiently smaller than 500; these discriminating values are associated with two-tailed critical values. With Case 2 then H_0 would only be rejected in favour of H_1 if the sample value was sufficiently small with the discriminating critical value associated with a one-tail (left-hand) critical value.

The rationale for hypothesis testing is to assume that H_0 is true (and so the statistical analysis based on H_0 is valid) and to identify the related critical value(s) for a specified significance level α, associated with the choice of H_1. The value of the significance level is determined by circumstance or regulation but a typical value used for illustration is taken as $\alpha = 0.05$ (sometimes quoted as 5 per cent). The sample data is used to identify the validity of H_0 by checking if the sample value is consistent with the statistical analysis. This is determined by checking if the data value falls within the anticipated range (acceptance region) of probability of size $(1 - \alpha)$ as determined by the critical values. Critical values are illustrated in Figure 10.11; Case 1 involves two tail regions defined by a lower value c_1 and an upper value c_2 whilst Case 2 has an upper tail region identified by an upper critical value c_3. Values of the critical values are found directly from the appropriate table of critical values for the sample distribution. If the sample value falls outside a critical value (i.e. sample value falls within an appropriate tail region) then it is deemed not acceptable and the null hypothesis is considered untrue as a consequence.

In this example,

Case 1: $H_0{:}\mu = 500$, $H_1{:}\mu > 500$ or $\mu < 500$
and on taking values for μ, s and n from the null hypothesis the test statistic is

$$t = \frac{\overline{X} - 500}{150/\sqrt{36}}$$

The sample distribution is $t_{[35]}$. Taking a significance level of $\alpha = 0.05$ split between two-tails (i.e. a probability 0.025 in each tail) the critical values are readily determined from Table 10.2 as $c_1 = -2.04$ and $c_2 = 2.04$ (using symmetry of the t-distribution).

The data value for the test statistic is obtained on substituting the observed value for the sample mean as

$$\hat{t} = \frac{450 - 500}{150/\sqrt{36}} = -2.00$$

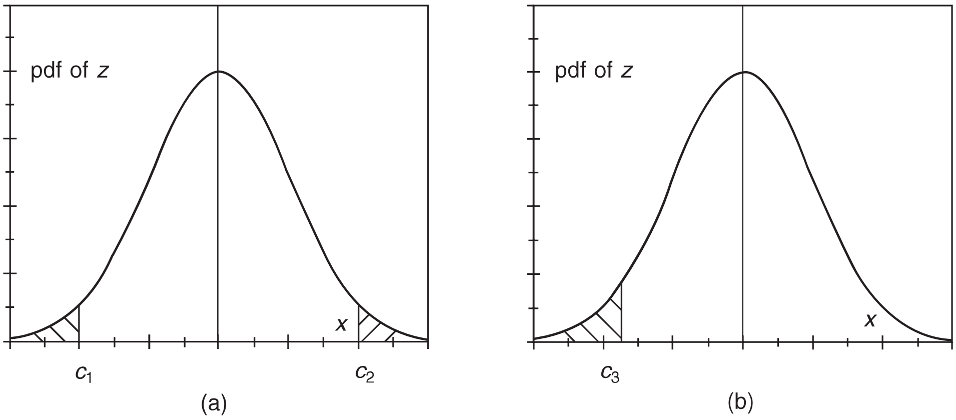


Figure 10.11 (a) Rejection regions Case 1, (b) rejection region Case 2.

This value falls within the acceptance region (marginally), i.e. within the bulk region outside of the rejection region(s). A conclusion is that, relative to the alternative hypothesis, the observed value falls within a region that is consistent with the random variation anticipated and the hypothesis $\mu = 500$ is accepted.

Case 2: $H_0{:}\mu = 500$, $H_1{:}\mu < 500$.
The test statistic and the data value is the same as Case 1, i.e

$$t = \frac{\overline{X} - 500}{150/\sqrt{36}} \quad \text{and} \quad \hat{t} = \frac{450 - 500}{150/\sqrt{36}} = -2.00$$

In this case the only rejection region for the null hypothesis is defined by a one-tailed (lower) region with significance level $\alpha = 0.05$ and determined with reference to Table 10.2 as $c_3 = -1.70$ (from symmetry upper and lower values differ only by a sign).

The observed value of the test statistic $\hat{t} = -2.00$ therefore falls within the rejection region and a conclusion is that the manufacturer's claim has not been achieved, i.e. the difference between the claimed and observed values cannot be attributed to natural variations with the stated significance level and the hypothesis $\mu = 500$ is rejected and $\mu < 500$ is accepted.

The discrepancy in conclusions between Cases 1 and 2 highlights the importance of identifying an appropriate alternative hypothesis and specification of the significance level.

Summary – general theory

The stages identified in the above example can be applied more generally to other practical situations and follow a similar approach although the choice of test statistic and its associated distribution will change accordingly. A summary of the stages is:

1 Make a null hypothesis H_0 and an alternative hypothesis H_1 (H_0 is *always* chosen to be specific to fully specify a sample statistic).
2 Assume H_0 is true and identify an appropriate test statistic t and its associated distribution.
3 Obtain a numerical value for t using the given sample data $\hat{t}$, say.
4 Specify a significance level α and determine a critical value (or values).
5 Accept H_0 or reject (i.e. accept H_1) depending on whether the sample value $\hat{t}$ falls within an acceptance or rejection region.

10.5.2 Comparison of means

Hypothesis testing can be readily used to help determine if differences on sample mean values measured between two processes are significant as shown in the following example.

Example

Concrete from two separate mixes averaged 25 Nmm^{-2} and 30 Nmm^{-2} respectively. In both cases 6 cubes were taken. Calculate if there is a significant difference between the two mixes when the standard deviation of the first was 3 Nmm^{-2} and the second was 5 Nmm^{-2}.

The variates in this case are X_1 and X_2 the strength of each mix. Associated data values for each mix are sample sizes $n_1 = 6$ and $n_2 = 6$, sample mean values $\overline{X}_1 = 25$ and $\overline{X}_2 = 30$, sample standard deviations $s_1 = 3$ and $s_2 = 5$.

The relevant test statistic is given from earlier as a comparison of sample means

$$t = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{s\sqrt{1/n_1 + 1/n_2}} \quad \text{where} \quad s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

and $\quad t \sim t_{[v]} \quad$ with $\quad v = n_1 + n_2 - 2.$

In the above μ_1 and μ_2 are the relevant population mean values which are unknown. However, an astute choice of hypothesis makes this unnecessary; the two hypotheses chosen are:

H_0: null hypothesis $\mu_1 = \mu_2$ (this will prove specific enough)
H_1: alternative hypothesis $\mu_1 \neq \mu_2$ (this defines a two-tailed test).

Assuming the null hypothesis applies then the test statistic becomes after evaluation:

$$t = \frac{(\overline{X}_1 - \overline{X}_2)}{s\sqrt{1/3}}$$

where $s = 4.123$ and $t \sim t_{[10]}$.

The data value for the test statistic is

$$\hat{t} = \frac{25 - 30}{4.123 \times 0.577} = -2.102$$

Taking a significance level of 0.05, then critical values are associated with two tails, each with a probability of 0.025 and upper value can be determined from Table 10.2 for $t_{0.025;10}$ as $c_2 = 2.23$. Thus it follows that the acceptance region for the null hypothesis is $-2.23 < t < 2.23$. The calculated data value lies within this region, from which it is possible to conclude that the null hypothesis is consistent with the data and therefore there is no significant difference between the strengths of the two mixes.

The test statistic specifically used in this section is often termed a *t*-statistic and can be usefully identified as

$$t = \frac{\text{observed difference in means}}{\hat{\sigma}_e}$$

10.5.3 Comparison of variances

Use can be made of the F-distribution to determine if any changes to a process has resulted in a smaller variation (measured by sample variances) rather than just to random choice as identified in the following example.

Example

The sample variances for the diameters of 23 nominally identical cast cylinders was 1.93 mm^2. For a random sample of size 13 taken from a second population the corresponding figure was 4.06 mm^2. Would one be justified in assuming that the two populations have diameters with the same variability?

The variates in this case are X_1 and X_2 the cylinder diameters from the two sources and the test is based on a comparison of population variances σ_1^2 and σ_2^2. A test statistic is given by

$$F = \frac{\text{largest sample value of } s_1^2}{\text{smallest sample value of } s_2^2}$$

and distributed as $F_{[v_1, v_2]}$, where $v_1 = n_1 - 1$, $v_2 = n_2 - 1$.

Comparing with the data values then $s_1^2 = 4.06$ (largest value), $n_1 = 13$ and $s_2^2 = 1.93$, $n_2 = 23$.

The relevant two hypotheses are:

H_0: null hypothesis $\sigma_1^2 = \sigma_2^2$ (i.e. $\sigma_1^2/\sigma_2^2 = 1$)
H_1: alternative hypothesis $\sigma_1^2 \neq \sigma_2^2$ (this defines a two-tailed test).

The data value for the test statistic is evaluated as $\hat{F} = 4.06/1.93 = 2.10$.

Taking a significance level of 0.05, then critical values are associated with two tails, each with a probability of 0.025 but preliminary selection of taking the largest sample variance as the numerator means only the upper value is relevant. A critical upper value associated with a probability $\alpha = 0.025$ is given from a table value for $F_{0.025;12,22}$ and determined as $c_2 = 2.60$. Thus the upper rejection region is $\hat{F} > c_2$, and comparison of values gives that $\hat{F}$ lies below the rejection region and a conclusion is no significant difference between variation in the two populations.

The test statistic specifically used in this section is termed an F-statistic and can be expressed as

$$F = \frac{\sigma_1^2}{\sigma_2^2} \quad [\text{where } \sigma_1 > \sigma_2]$$

10.5.4 Significance and errors

An obvious factor with hypothesis testing is the choice and meaning of the significance level. By nature of dealing with processes that involve variation then some error is always present but it is important to try to quantify any error and evaluate any subsequent consequences. The significance level, α say, is in fact a probability measure:

$$\text{P(rejecting } H_0 \text{ when it is true)} = \alpha$$

and is the chance that the test statistic value $\hat{Z}$ falling in the rejection region within a hypothesis test could have occurred as a rare event rather than an incorrect hypothesis H_0. Such an error is called a **Type 1** error and obviously should be made as small as possible.

However, decreasing α increases the chance of making a **Type 2** corresponding to failing to rejecting the hypothesis H_0 when it is false. The value for this error β is given by the probability

$$\text{P(accepting } H_0 \text{ when it is false)} = \beta.$$

In sampling, a decision has to be made for the value of α and/or β by taking into account the costs and penalties attached to both types of errors. However, as might be expected these values are linked; generally decreasing Type 1 errors will increase Type 2 errors and vice versa. Calculation of the links between the parameters α, β and n depend upon the specific test chosen and can be very involved. In practice a decision rule for a significance test is determined by taking a significance level α and arrive at an appropriate rule; the associated values of β can be computed for various values of n. Alternatives that may be used are:

- given a decision rule – compute the errors α and β
- decide on α and β and then arrive at a decision rule.

A related curve is that of β known as the operating characteristic curve, or O-C curve, and is described in Chapter 9.

10.6 Regression models

10.6.1 Correlation

Correlation is concerned with the amount of association between two or more sets of variables. An illustration is given earlier in *Case* 2 for the deflection of a concrete beam where the scattergram of the data displayed in Figure 10.2 shows a strong linear relationship, with increasing values of beam deflection (y) associated with proportionate increasing values of applied load (x). The amount of association for different situations may not be so discernible as illustrated in the following sets of data:

Case 3 An experimental determination of the relation between the normal stress (x) and the shear resistance (y) of a cement-stabilized soil yielded the following results:

Normal stress x kN/mm^2	10	12	14	16	18	20
Shear resistance y kN/mm^2	10.9	18.7	15.4	25.1	19.3	17.6

Case 4 Percentages of sand (y) recorded at different depths (x) from samples were:

x (mm)	0	400	800	1200	1600	2000	2400	2800	3200
y (%)	70.2	52.9	54.2	52.4	47.4	49.1	30.7	36.8	37.4

Case 5 Failure loads concrete beams with load for first crack (x) and failure load (y) were:

load x	10350	8450	7200	5100	6500	10600	6000	6000
load y	10350	9300	9600	10300	9400	10600	10100	9900
load x	9500	6500	9300	6000	6000	5800	6500	
load y	9500	10200	9300	9550	9550	10500	10200	

The relationships between values y and x can be individually plotted on a series of scattergrams as shown in Figure 10.12 from which the qualitative association between the variables x and their linked values y can be assessed.

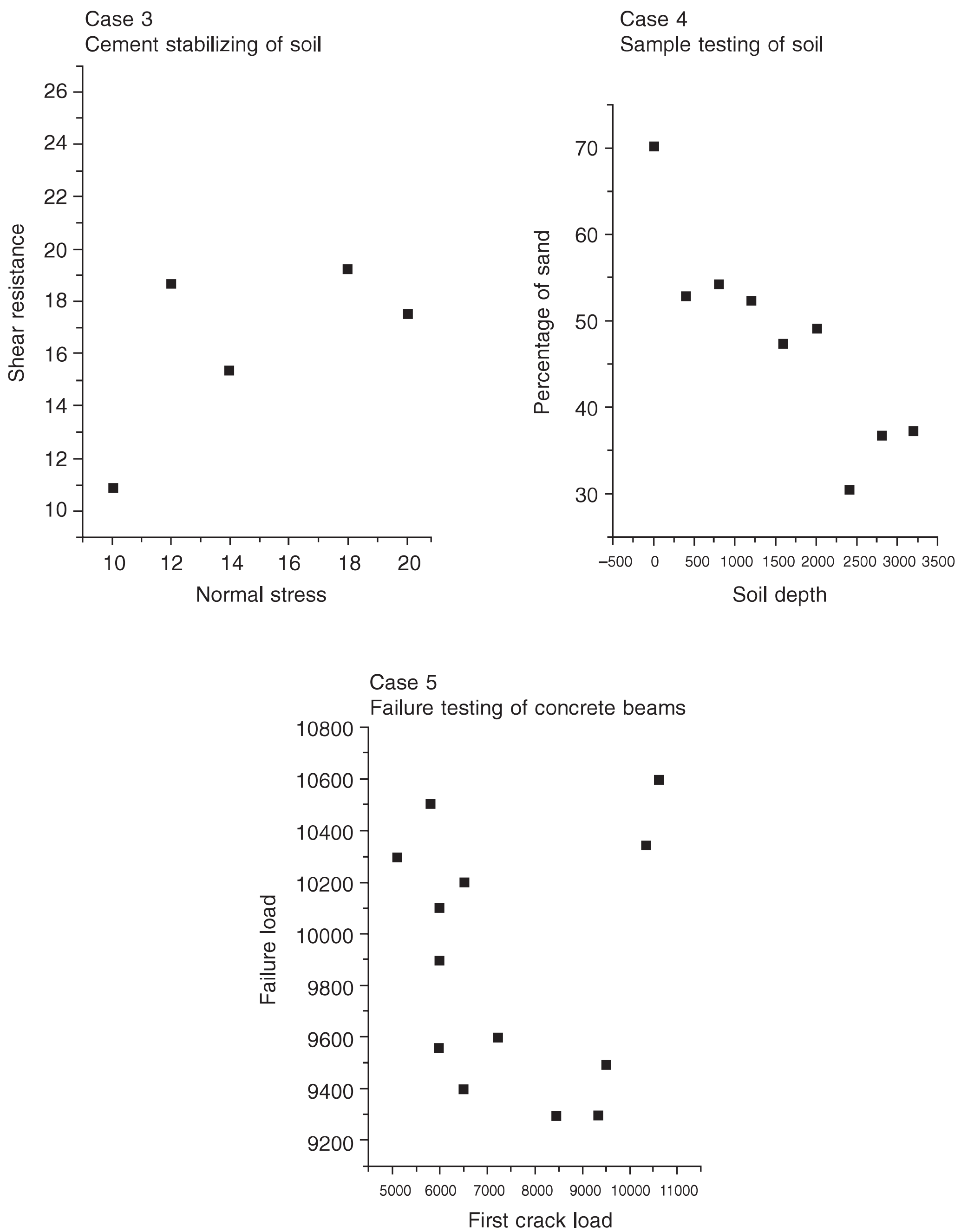


Figure 10.12 Scattergrams of data for Cases 3, 4 and 5.

In *Case* 4 then a trend exists where decreasing values of y are linked to increasing values of x; the data perhaps suggest a linear relationship could exist. Looking at the scattergram for *Case 5* it is difficult to perceive that any meaningful relationship exists between the variables x and y. Data from *Case 3* shows some trend exists in that increasing values of x are associated with increasing values of y, but the relationship is not obviously linear. Clearly, it would not make sense to seek a linear relationship for *Case 5* but that a linear relationship could reasonably be sought for *Cases 2* and *4*. For *Case 3* perhaps some tentative linear relationship could be determined, but some measure of confidence of how this relationship matches with the data values would be desirable. The measure of the association of variables is called correlation. The most widely used association relationship is that of a straight-line (linear) fit to data pair observations between a response variable (y) and an explanatory variable (x). In this section a linear fit approach will be assumed but in practice, care should be exercised to consider the possibility that some other form of relationship might be more appropriate.

10.6.2 Regression – least-squares method

Regression is a general term used in data analysis to mean ‘trend’ or ‘pattern’. Many engineering problems are concerned with determining a relationship between a set of variables. Even for a strongly linear relationship, such as in *Case* 2, in practice all data points are unlikely to align due to random error. Generally, the data points are more scattered and a more realistic aim would be to obtain a ‘best’ curve through the collection of all data points and the most used method is to use a least-squares method.

Given a set of data $(x_1, y_1), \ldots (x_i, y_i) \ldots (x_n, y_n)$ consider fitting a straight line $y = a + bx$ through the data points so as to achieve some form of ‘best fit’. In practice, if data values are plotted as illustrated in Figure 10.13 this means adjusting the slope of the line (parameter b) and the intercept on the y-axis (parameter a) until some form of optimal fit is achieved.

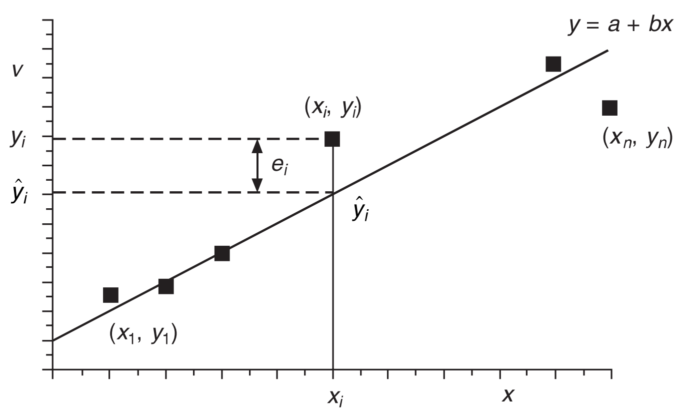


Figure 10.13 Least squares approximation method.

Calculated values for these parameters can be determined provided a measure of fit is defined. At any data value x_i, the data value is specified as y_i and the corresponding regression curve value is $\hat{y}_i = a + bx_i$; this defines an associated error $e_i = y_i - \hat{y}_i = y_i - a - bx_i$.

As the values of a and b are changed, then the error associated with each data point will change; 'best values' for a and b will occur when the errors e_i are minimized. A measure of the global error is given by the sum of squares

$$S = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

$S = S(a, b)$ is a function of the two variables a and b, and can be chosen such that S is minimized.

Applying appropriate calculus for a function of several variables, the result is:

$$b = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n \sum x_i^2 - (\sum x_i)^2}, \quad a = \bar{y} - b\bar{x}$$

where

$$\bar{x} = \frac{1}{n} \sum x_i \quad \text{and} \quad \bar{y} = \frac{1}{n} \sum y_i$$

These formulae determine the parameters associated with a least squares regression line as illustrated in a later example.

10.6.3 Correlation coefficient

A measure of the correlation between data values and a linear fit can be obtained as discussed below, and illustrated in Figure 10.14. At any data value x_i, then both a recorded data value y_i and a value $\hat{y}_i = a + bx_i$ calculated from the regression formula are available. An assessment of the closeness of agreement between the data values is obtained from considering the variations $(y_i - \hat{y}_i)$ between the recorded data value and the regression value at a general data point (x_i, y_i).

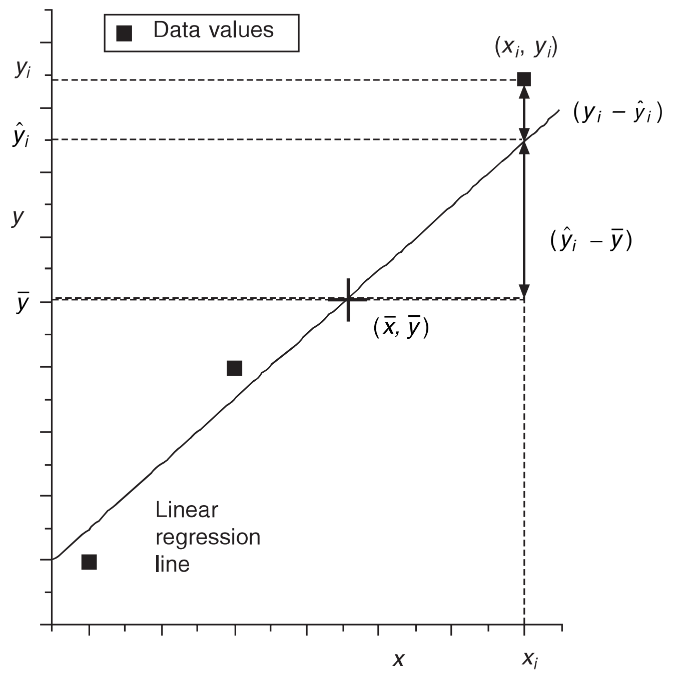


Figure 10.14 Calculation of correlation quantities.

The least-squares method can be shown to have the following properties that are useful in the following analysis:

(i) The regression line will pass through the point with coordinate $(\bar{x}, \bar{y})$, i.e. pass through the sample mean values of both x_i and y_i;

(ii) $\sum_{1}^{n} (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = 0$

It is convenient to use the point $(\bar{x}, \bar{y})$ as a reference point, which from property 1 lies on the regression line. Any data value y_i can then be readily expressed as

$$(y_i - \bar{y}) = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})$$

as illustrated in Figure 10.14.

The above expression is valid for all data points and squaring and summing over all n data values, and using property 2 gives a useful expression

$$\underbrace{\sum_{i=1}^{n} (y_i - \bar{y})^2}_{\text{total variation}} = \underbrace{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}_{\text{unexplained variation}} + \underbrace{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}_{\text{explained variation}}$$

The term $\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$ is called the explained variation and corresponds to a measure if all data values were to lie exactly on the regression line. Discrepancies and variations from data points not lying on the regression line are given by the measure identified as the unexplained variation, and corresponding to the sum $\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$. Clearly the requirement for good correlation is that the unexplained variation is relatively small, or equally, the total variation and the explained variation are almost equal. A useful relative measure of correlation is therefore given by

$$R^2 = \frac{\text{explained variation}}{\text{total variation}} = \frac{\Sigma(\hat{y}_i - \bar{y})^2}{\Sigma(y_i - \bar{y})^2} = \frac{SSR}{CSS}$$

$$= -\frac{\Sigma(y_i - \hat{y}_i)^2}{\Sigma(y_i - \bar{y})^2} = 1 - \frac{SSE}{CSS}$$

In the above, a number of quantities are determined directly from data values as:

$CSS = \sum_{i=1}^{n} (y_i - \bar{y})^2$ – computed sum of squares

$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$ – sum of squares due to regression

$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ – sum of squares due to errors.

The quantity R^2 is called the coefficient of determination and is a useful measure of the association between a linear regression line and the data with $R^2 = 1$ corresponding to a perfect fit and $R^2 = 0$ corresponds to no dependence between x and y.

Values for the Case examples following from fitting a least-squares fit can be readily calculated as:

Case 2: $R^2 = 0.95$ indicates a very strong correlation between data values;
Case 3: $R^2 = 0.26$ indicates that correlation between data values is weak;
Case 4: $R^2 = 0.79$ indicates a fair correlation between data values;
Case 5: $R^2 = 0.00004$ indicates little correlation exists;

10.6.4 Example – Case 2 Beam deflection

The ten data values associated with *Case 2* are used to provide a regression calculation to obtain a least-squares fit, calculation of the correlation coefficient and analysis of residuals. It is convenient to display derived values in a spreadsheet format as follows:

i	x_i	y_i	$x_i - \bar{x}$	$(y_i - \bar{y})$	$\hat{y}_i = a + bx_i$	Residual $y_i - \hat{y}_i$
1	100	45	–45	–21.5	43.56	1.44
2	110	52	–35	–14.5	48.66	3.34
3	120	54	–25	–12.5	53.76	0.24
4	130	54	–15	–12.5	58.85	–4.85
5	140	62	–5	–4.5	63.95	–1.95
6	150	68	5	–1.5	69.05	–1.05
7	160	75	15	8.5	74.15	0.85
8	170	75	25	8.5	79.24	–4.24
9	180	92	35	25.5	84.34	7.66
10	190	88	45	21.5	89.44	–1.44

Derived values associated with the operational formulae used are:

$$\bar{x} = \frac{1}{10}\sum_{i=1}^{10} x_i = 145, \quad \bar{y} = \frac{1}{10}\sum_{i=1}^{10} y_i = 66.5$$

$$\sum_{i=1}^{10}(x_i - \bar{x})(y_i - \bar{y}) = 4205, \ \sum_{i=1}^{10}(x_i - \bar{x})^2 = 8250, \ \sum_{i=1}^{10}(y_i - \bar{y})^2 = 2264.5$$

Using the given formulae, $b = 4205/8250 = 0.510$ and $a = 66.5 – 0.51*145 = – 7.41$.

Hence the least squares linear regression is $y = –7.41 + 0.51x$. Deflection values $\hat{y}_i$ obtained from the least-squares analysis are calculated in the spreadsheet together with residual values $(y_i - \hat{y}_i)$ for information; data values together with the regression curve are displayed in Figure 10.15.

Corresponding values for correlation quantities are:

CSS = 2264.5, SSR = 2143.3, SSE = 121.2, giving a value for the correlation coefficient $R^2 = 0.946$.

10.6.5 Analysis of residuals

The starting point for measuring regression error were the quantities $e_i = y_i - \hat{y}_i$, which are called the residuals. These may be usefully considered following a regression analysis

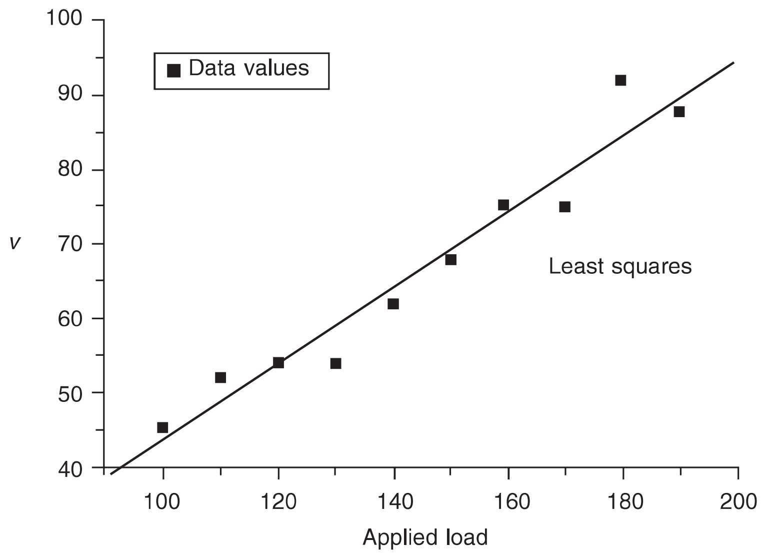


Figure 10.15 Comparison of least-squares linear regression with data values in Case 2 – Beam deflection.

to look at the level of agreement. For example, plotting the residuals for *Case 2* is shown in Figure 10.16. This shows residuals apparently randomly distributed about the zero error line; this is typical of data that is consistent with a linear regression.

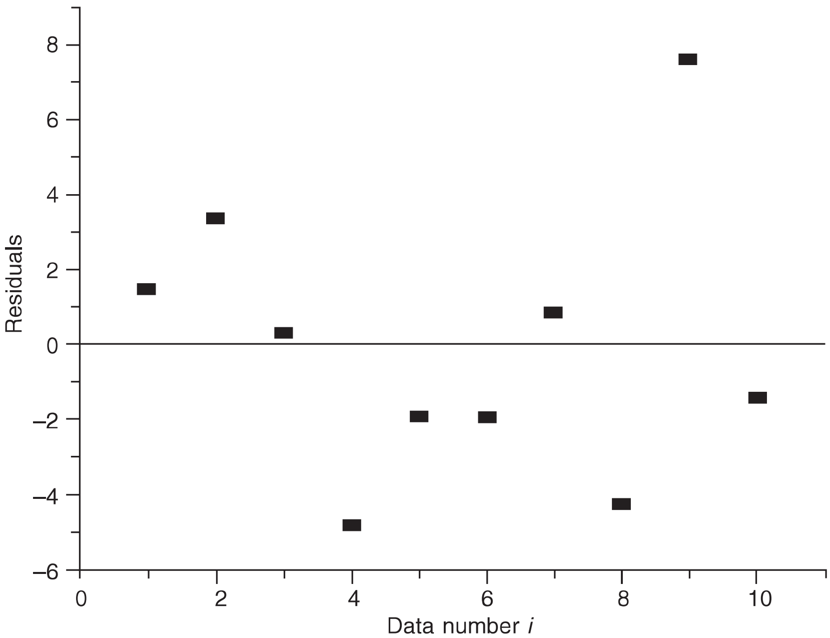


Figure 10.16 Plot of residuals from least-squares fit to data in Case 2.

The value of providing a plot of residuals is to identify if the choice of a linear fit (or other) is appropriate. If residuals follow a distinct trend, then a linear regression curve may not be appropriate, and a quadratic or other curve may be more appropriate. The residual curve can also highlight suspect data points (outliers), i.e. values that are distinct from the rest of the data and may arise from an uncharacteristic operational or measuring error. These outliers may often need special attention to decide whether they should be included within the analysis.

10.6.6 Extension to multivariate

In many practical cases, a dependant variable y may depend upon more than one independent variable, i.e. $y = f(x_1, x_2, \ldots x_m)$. A direct extension to the regression analysis is available to obtain a least squares linear fit of the form $y = \alpha_0 + \alpha_1 x_1 + \ldots \alpha_m x_m$ to the data.

In this case, analysis is best carried out using matrix algebra. Also as drawing a scattergram is not feasible, analysis of the level of fit is conducted through looking at the residuals and the values of the appropriate correlation coefficient. Details will not be covered within this chapter.

10.6.7 Fit of regression curves and confidence lines

Hypothesis testing techniques can also be used to test if a linear regression obtained to a set of data points is acceptable to a given significance. We have already seen that if R^2, the coefficient of determination is close to a value of 1, then a close fit (i.e. good correlation) is expected.

It can be shown that the quantity

$$F = \frac{R^2/v_1}{(1 - R^2)/v_2}, \; v_1 = m \quad \text{and} \quad v_2 = n - (m + 1)$$

is distributed as an F-variate, $F_{[v_1, v_2]}$ where n is the number of data points and $m = 2$ for a linear fit.

Estimation techniques can be applied to linear regression analysis using the t-distribution with respect to confidence limits to the two parameters a and b for a linear fit. This gives rise to associated confidence limit values for each data value and joining these values produces confidence lines.

10.7 Statistical formulae and tables

Selected statistical formulae:

$$\sigma_{\bar{X}} = \frac{\sigma_X}{\sqrt{n}}$$

$$\sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$\hat{\sigma}_e = s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \quad \text{where} \quad s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

$$t = \frac{\text{observed difference in means}}{\hat{\sigma}_e}$$

$$F = \frac{\sigma_1^2}{\sigma_2^2} \quad [\text{where } \sigma_1 > \sigma_2]$$

$$CSS = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$b = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n \sum x_i^2 - (\sum x_i)^2}$$

$$a = \bar{y} - b\bar{x}$$

Statistical Table 10.1 The normal distribution

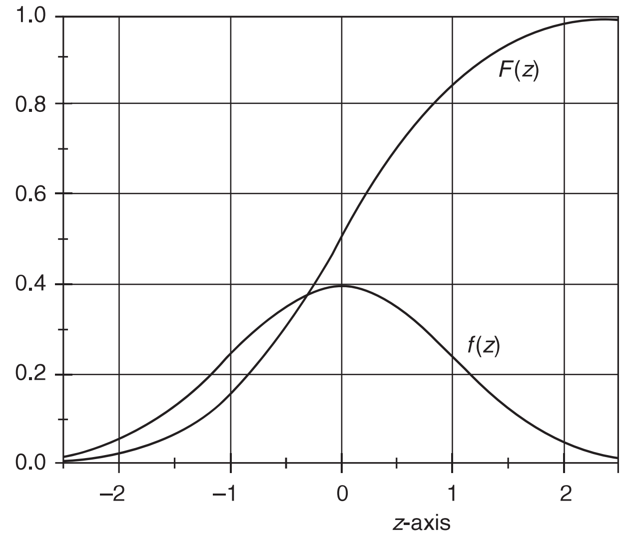
Z is the normalized variate $N(0, 1)$

$f(z)$ is the probability density function (pdf) $= \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$

$F(z)$ is the cumulative distribution function $= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{1}{2}z^2} dz$

Values for $z < 0$ are given by: $f(-z) = f(z)$ and $F(-z) = 1 - F(z)$

Table values given from the Excel function NORMDIS



$f(Z)$	**Z**	$F(Z)$									
		0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.3989	**0.0**	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.3970	**0.1**	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.3910	**0.2**	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3814	**0.3**	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.3683	**0.4**	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.3521	**0.5**	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.3332	**0.6**	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.3123	**0.7**	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.2897	**0.8**	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.2661	**0.9**	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
0.2420	**1.0**	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
0.2179	**1.1**	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
0.1942	**1.2**	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
0.1714	**1.3**	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
0.1497	**1.4**	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319

$f(Z)$	**Z**	$F(Z)$									
		0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.1295	**1.5**	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
0.1109	**1.6**	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
0.0940	**1.7**	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
0.0790	**1.8**	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
0.0656	**1.9**	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
0.0540	**2.0**	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
0.0440	**2.1**	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
0.0355	**2.2**	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
0.0283	**2.3**	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
0.0224	**2.4**	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
0.0175	**2.5**	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
0.0136	**2.6**	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
0.0104	**2.7**	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
0.0079	**2.8**	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
0.0060	**2.9**	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986

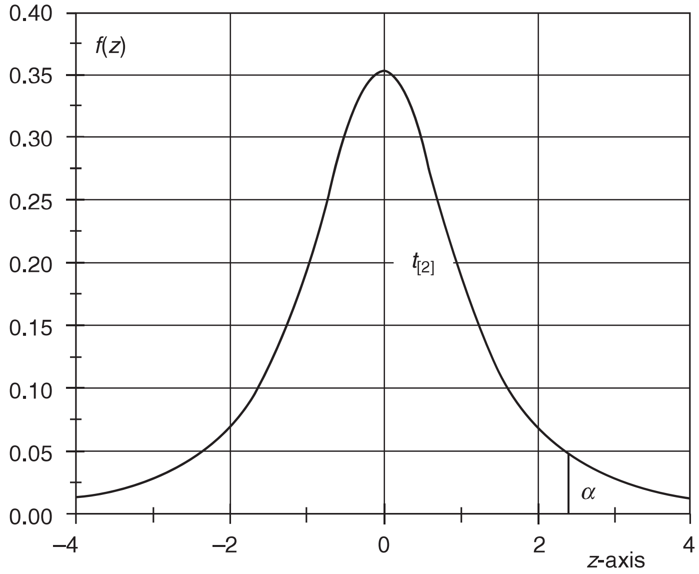
Statistical Table 10.2 Critical values of the t-distribution

The table gives the values of $t_{\alpha;v}$, the critical values for a significance level of α in the upper tail of the distribution. The t-distribution having v degrees of freedom.

Critical values for the lower tail is given by $-t_{\alpha;v}$ (symmetry).

For critical values of $|t|$, corresponding to a two-tailed region, the column headings for α must be doubled.

Table values given from the Excel function TINV



v \ α =	0.1	0.05	0.025	0.01	0.005	0.001	0.0005
1	3.078	6.314	1.000	31.821	636.578	318.29	636.58
2	1.886	2.920	0.816	6.965	31.600	22.328	31.600
3	1.638	2.353	0.765	4.541	12.924	10.214	12.924
4	1.533	2.132	0.741	3.747	8.610	7.173	8.610
5	1.476	2.015	0.727	3.365	6.869	5.894	6.869
6	1.440	1.943	0.718	3.143	5.959	5.208	5.959
7	1.415	1.895	0.711	2.998	5.408	4.785	5.408
8	1.397	1.860	0.706	2.896	5.041	4.501	5.041
9	1.383	1.833	0.703	2.821	4.781	4.297	4.781
10	1.372	1.812	0.700	2.764	4.587	4.144	4.587
11	1.363	1.796	0.697	2.718	4.437	4.025	4.437
12	1.356	1.782	0.695	2.681	4.318	3.930	4.318
13	1.350	1.771	0.694	2.650	4.221	3.852	4.221
14	1.345	1.761	0.692	2.624	4.140	3.787	4.140
15	1.341	1.753	0.691	2.602	4.073	3.733	4.073
16	1.337	1.746	0.690	2.583	4.015	3.686	4.015
17	1.333	1.740	0.689	2.567	3.965	3.646	3.965
18	1.330	1.734	0.688	2.552	3.922	3.610	3.922
19	1.328	1.729	0.688	2.539	3.883	3.579	3.883
20	1.325	1.725	0.687	2.528	3.850	3.552	3.850
21	1.323	1.721	0.686	2.518	3.819	3.527	3.819
22	1.321	1.717	0.686	2.508	3.792	3.505	3.792
23	1.319	1.714	0.685	2.500	3.768	3.485	3.768
24	1.318	1.711	0.685	2.492	3.745	3.467	3.745
25	1.316	1.708	0.684	2.485	3.725	3.450	3.725
26	1.315	1.706	0.684	2.479	3.707	3.435	3.707
27	1.314	1.703	0.684	2.473	3.689	3.421	3.689
28	1.313	1.701	0.683	2.467	3.674	3.408	3.674
29	1.311	1.699	0.683	2.462	3.660	3.396	3.660
30	1.310	1.697	0.683	2.457	3.646	3.385	3.646
40	1.303	1.684	0.681	2.423	3.551	3.307	3.551
60	1.296	1.671	0.679	2.390	3.460	3.232	3.460
120	1.289	1.658	0.677	2.358	3.373	3.160	3.373
∞	1.282	1.645	0.675	2.327	3.291	3.091	3.291

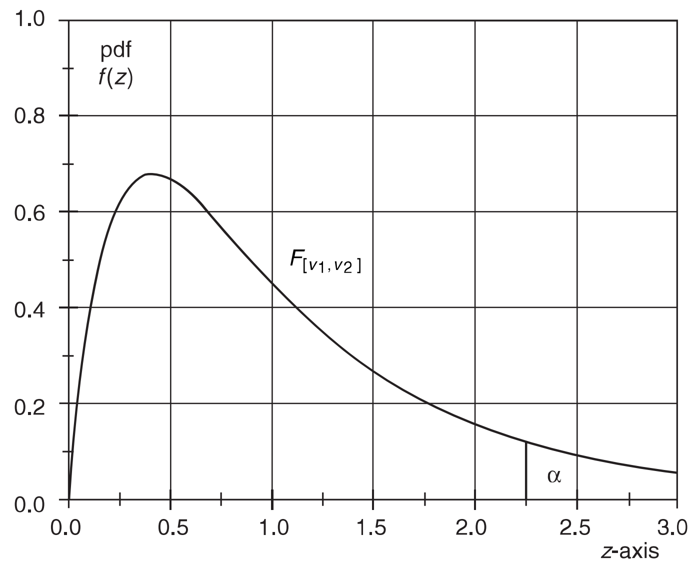
Statistical Table 10.3 Critical values of the *F*-distribution

The table gives the values of $F_{\alpha;v_1,v_2}$, the critical values for a significance level of α in the upper tail of the distribution, the *F*-distribution having v_1 degrees of freedom in the numerator and v_2 degrees of freedom in the denominator.

For each pair of degrees of freedom of v_1 and v_2, the upper values $F_{\alpha;v_1,v_2}$, is tabulated for $\alpha = 0.05$ and 0.025 (bracketed).

Critical values corresponding to the lower tail of the distribution are obtained by the relation $F_{1-\alpha;v_1,v_2} = 1/F_{\alpha;v_2,v_1}$

Table values given from the Excel function FINV



v_2 \ v_1	1	2	3	4	5	6	7	8	10	12	24	∞
1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	241.9	243.9	249.1	254.3
	(648)	(799)	(864)	(900)	(922)	(937)	(948)	(957)	(969)	(977)	(997)	(1018)
2	18.5	19.0	19.2	19.2	19.3	19.3	19.4	19.4	19.4	19.4	19.5	19.5
	(38.5)	(39.0)	(39.2)	(39.2)	(39.3)	(39.3)	(39.4)	(39.4)	(39.4)	(39.4)	(39.5)	(39.5)
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.79	8.74	8.64	8.53
	(17.44)	(16.04)	(15.44)	(15.10)	(14.88)	(14.73)	(14.62)	(14.54)	(14.42)	(14.34)	(14.12)	(13.90)
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	5.96	5.91	5.77	5.63
	(12.22)	(10.65)	(9.98)	(9.60)	(9.36)	(9.20)	(9.07)	(8.98)	(8.84)	(8.75)	(8.51)	(8.26)
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.74	4.68	4.53	4.37
	(10.01)	(8.43)	(7.76)	(7.39)	(7.15)	(6.98)	(6.85)	(6.76)	(6.62)	(6.52)	(6.28)	(6.02)
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.06	4.00	3.84	3.67
	(8.81)	(7.26)	(6.60)	(6.23)	(5.99)	(5.82)	(5.70)	(5.60)	(5.46)	(5.37)	(5.12)	(4.85)
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.64	3.57	3.41	3.23
	(8.07)	(6.54)	(5.89)	(5.52)	(5.29)	(5.12)	(4.99)	(4.90)	(4.76)	(4.67)	(4.41)	(4.14)

v_2 \ v_1	1	2	3	4	5	6	7	8	10	12	24	∞
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.35	3.28	3.12	2.93
	(7.57)	(6.06)	(5.42)	(5.05)	(4.82)	(4.65)	(4.53)	(4.43)	(4.30)	(4.20)	(3.95)	(3.67)
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.14	3.07	2.90	2.71
	(7.21)	(5.71)	(5.08)	(4.72)	(4.48)	(4.32)	(4.20)	(4.10)	(3.96)	(3.87)	(3.61)	(3.33)
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	2.98	2.91	2.74	2.54
	(6.94)	(5.46)	(4.83)	(4.47)	(4.24)	(4.07)	(3.95)	(3.85)	(3.72)	(3.62)	(3.37)	(3.08)
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.85	2.79	2.61	2.40
	(6.72)	(5.26)	(4.63)	(4.28)	(4.04)	(3.88)	(3.76)	(3.66)	(3.53)	(3.43)	(3.17)	(2.88)
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.75	2.69	2.51	2.30
	(6.55)	(5.10)	(4.47)	(4.12)	(3.89)	(3.73)	(3.61)	(3.51)	(3.37)	(3.28)	(3.02)	(2.72)
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.67	2.60	2.42	2.21
	(6.41)	(4.97)	(4.35)	(4.00)	(3.77)	(3.60)	(3.48)	(3.39)	(3.25)	(3.15)	(2.89)	(2.60)
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.60	2.53	2.35	2.13
	(6.30)	(4.86)	(4.24)	(3.89)	(3.66)	(3.50)	(3.38)	(3.29)	(3.15)	(3.05)	(2.79)	(2.49)
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.49	2.42	2.24	2.01
	(6.12)	(4.69)	(4.08)	(3.73)	(3.50)	(3.34)	(3.22)	(3.12)	(2.99)	(2.89)	(2.63)	(2.32)
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.41	2.34	2.15	1.92
	(5.98)	(4.56)	(3.95)	(3.61)	(3.38)	(3.22)	(3.10)	(3.01)	(2.87)	(2.77)	(2.50)	(2.19)
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.35	2.28	2.08	1.84
	(5.87)	(4.46)	(3.86)	(3.51)	(3.29)	(3.13)	(3.01)	(2.91)	(2.77)	(2.68)	(2.41)	(2.09)
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.30	2.23	2.03	1.78
	(5.79)	(4.38)	(3.78)	(3.44)	(3.22)	(3.05)	(2.93)	(2.84)	(2.70)	(2.60)	(2.33)	(2.00)
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.25	2.18	1.98	1.73
	(5.72)	(4.32)	(3.72)	(3.38)	(3.15)	(2.99)	(2.87)	(2.78)	(2.64)	(2.54)	(2.27)	(1.94)
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.22	2.15	1.95	1.69
	(5.66)	(4.27)	(3.67)	(3.33)	(3.10)	(2.94)	(2.82)	(2.73)	(2.59)	(2.49)	(2.22)	(1.88)
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.19	2.12	1.91	1.65
	(5.61)	(4.22)	(3.63)	(3.29)	(3.06)	(2.90)	(2.78)	(2.69)	(2.55)	(2.45)	(2.17)	(1.83)
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.16	2.09	1.89	1.62
	(5.57)	(4.18)	(3.59)	(3.25)	(3.03)	(2.87)	(2.75)	(2.65)	(2.51)	(2.41)	(2.14)	(1.79)
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.08	2.00	1.79	1.51
	(5.42)	(4.05)	(3.46)	(3.13)	(2.90)	(2.74)	(2.62)	(2.53)	(2.39)	(2.29)	(2.01)	(1.64)
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	1.99	1.92	1.70	1.39
	(5.29)	(3.93)	(3.34)	(3.01)	(2.79)	(2.63)	(2.51)	(2.41)	(2.27)	(2.17)	(1.88)	(1.48)
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.91	1.83	1.61	1.25
	(5.15)	(3.80)	(3.23)	(2.89)	(2.67)	(2.52)	(2.39)	(2.30)	(2.16)	(2.05)	(1.76)	(1.31)
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.83	1.75	1.52	1.00
	(5.02)	(3.69)	(3.12)	(2.79)	(2.57)	(2.41)	(2.29)	(2.19)	(2.05)	(1.94)	(1.64)	(1.00)

Further reading

Benjamin, J.R. and Cornell C.A. *Probability, Statistics and Decision for Civil Engineers*. McGraw-Hill, New York.
Bowker, A.H. and Lieberman G.J. *Engineering Statistics*. Prentice Hall, Engleword Cliffs, NJ.
Chatfield, C. *Statistics for Technology*. Penguin Books, London.
Kennedy, J.B. and Neville, A.M. *Basic Statistical Methods for Engineers and Scientists,* 3rd edn. Harper & Row, New York.
Metcalfe, A.V. *Statistics in Engineering – A practical approach*. Chapman and Hall, London.
Metcalfe, A.V. *Statistics in Civil Engineering*. Arnold, London.
Miller, I. and Freund, J.E. *Probability and Statistics for Engineers*, 2nd edn. Prentice Hall, Englewood Cliffs, NJ.
Ross, S.M. *Introduction to Probability and Statistics for Engineers and Scientists*. Wiley, Chichester.
Smith, G.N. *Probability and Statistics in Civil Engineering – An Introduction*. Collins, London.
Walpole, R.E. and Myers, R.H. *Probability and Statistics for Engineers and Scientists*, 4th edn. Macmillan, London.