



An introduction to the statistical analysis of engineering data

Endorsed by
The Royal Aeronautical Society
The Institution of Mechanical Engineers

ESDU DATA ITEMS

Data Items provide validated information in engineering design and analysis for use by, or under the supervision of, professionally qualified engineers. The data are founded on an evaluation of all the relevant information, both published and unpublished, and are invariably supported by original work of ESDU staff engineers or consultants. The whole process is subject to independent review for which crucial support is provided by industrial companies, government research laboratories, universities and others from around the world through the participation of some of their leading experts on ESDU Technical Committees. This process ensures that the results of much valuable work (theoretical, experimental and operational), which may not be widely available or in a readily usable form, can be communicated concisely and accurately to the engineering community.

We are constantly striving to develop new work and review data already issued. Any comments arising out of your use of our data, or any suggestions for new topics or information that might lead to improvements, will help us to provide a better service.

THE PREPARATION OF THIS DATA ITEM

The work on this particular Data Item was monitored and guided by the Joint Committee for Statistics which has the following membership:

Chairman	
Dr M.S.G. Cullimore	— Independent
Members	
Mr J.K. Bennett	— British Aerospace plc, Space and Communications Division
Dr R.J. Henery	— Strathclyde University
Mr C.A. Hermens*	— Fokker B.V., The Netherlands
Mr J. O'Hara	— British Aerospace (Military Aircraft) Ltd, Brough
Mr A. Triner	— Lloyd's Register of Shipping.

* Corresponding Member

This Item was also accepted for inclusion in the Stress and Strength Series by the Stress Analysis and Strength of Components Committee which has the following membership.

Chairman	
Prof. C.E. Turner	— Imperial College of Science and Technology
Vice-Chairman	
Prof. T.G.F. Gray	— Strathclyde University
Members	
Mr A.J. Batchelor	— Independent
Dr I.J. Bickley	— Mirrlees Blackstone (Stockport) Ltd
Dr M.S.G. Cullimore	— Independent
Dr R.B. Heywood	— Independent
Dr L.C. Laming	— Imperial College of Science and Technology
Mr R.V. Pomeroy	— Lloyd's Register of Shipping
Mr N. Rock	— H.W. Structures Ltd
Mr J.V. Vint	— Independent.

This Item was also accepted for inclusion in the Stress and Strength Series by the Stress Analysis and Strength of Components Committee which has the following membership.

Chairman	
Mr J. K. Bennett	— British Aerospace plc, Space and Communications Division
Vice-Chairman	
Mr K.R. Obee	— Independent
Members	
Dr P. Bartholomew	— Defence Research Agency, Aerospace Division, Royal Aerospace Establishment, Farnborough
Mr K. Fitzsimons	— Westland Helicopters Ltd

(Continued on inside back cover)

AN INTRODUCTION TO THE STATISTICAL ANALYSIS OF ENGINEERING DATA

CONTENTS

	Page
1. PREFACE	1
2. BASIC CONCEPTS	2
2.1 Random Events and Probability	2
2.2 Random Variables and Their Distributions	3
2.2.1 Discrete random variables	3
2.2.2 Continuous random variables	4
2.2.3 Measures of location and dispersion	6
2.2.4 Standard distributions	8
3. SAMPLING	9
3.1 The Random Sample	9
3.2 Presentation of Data	9
3.3 Sample statistics	11
4. STATISTICAL TECHNIQUES	12
4.1 Choice of Distributional Models	12
4.2 Point Estimation	13
4.3 Interval Estimation	14
4.4 Hypothesis Testing	15
4.5 Miscellaneous Techniques	16
4.5.1 Distribution-free (non-parametric) techniques	16
4.5.2 Simulation (Monte Carlo) techniques	16
5. POSTSCRIPT	17
6. REFERENCES	18
7. TERMINOLOGY AND SYMBOLS	19
7.1 Index of statistical terminology	19
7.2 Index of symbols	24

AN INTRODUCTION TO THE STATISTICAL ANALYSIS OF ENGINEERING DATA

1. PREFACE

The aim of the Data Items on statistics is to make statistical methods of data analysis more accessible to the design engineer, and hence to increase the range of analytical tools at his disposal.

Statistical methods may be used by engineers to cater for the uncertainties inherent in many engineering problems. The data on which engineering decisions must be based are usually restricted by economic considerations, and there are many situations in which the use of appropriate statistical techniques will enable the maximum amount of useful information to be extracted from the limited data available. It is essential, however, that the statistical techniques described in this and the related Data Items are used with discrimination and in conjunction with engineering judgement and experience.

This Data Item gives a summary of the basic statistical terms and principles involved and the techniques most commonly used in the analysis of engineering data. The treatment has been restricted to the consideration of a single random variable.

Section 2 covers the fundamental concepts, including probability, random variables and distribution theory, whilst Section 3 introduces sampling from a population. This is followed in Section 4 by an introduction to statistical inference, the process by which deductions are made about a population on the basis of a sample, the main topics covered being point and interval estimation and hypothesis testing.

Since it is not possible in this Data Item to give more than a brief summary of the most important concepts and techniques, references are given in Section 6 to sources from which further information may be obtained if required. Reference 5 gives an elementary treatment of much of the relevant material, and requires only a very basic knowledge of mathematics, whilst References 3, 8 and 9 give a more detailed coverage at a more advanced level. Reference 10 is a good introduction text which covers the subject matter of Section 2 in greater detail. References are also made in the text below to more comprehensive treatments of specific topics; in particular, ESDU 91041* covers the analysis of data from normal distributions and also includes a number of statistical tables.

An index of the statistical terms and symbols used is located for convenient reference at the end of the Data Item in Section 7.

* ESDU 91041 "The statistical analysis of data from normal distributions, with particular reference to small samples".

2. BASIC CONCEPTS

2.1 Random Events and Probability

An **experiment** is defined as the process of making an observation which is capable of being repeated. Occasionally, the outcome of an experiment is completely determined by the conditions under which it is performed, and it could be argued that this would always be the case if the conditions could be completely specified. In practice, however, even when conditions are nominally identical, most engineering experiments are not precisely repeatable, but their outcomes contain an element of chance. This is because there are factors that cannot be measured or controlled, some of which the experimenter may be unaware of. It is this type of engineering experiment with which the Data Items on statistics are concerned.

A **random event** is a subset of all possible outcomes of an experiment involving an element of chance. If the experiment in question is the landing of an aircraft, relevant examples are the event that an aircraft makes a hard landing, and the event that the aircraft bursts one or more tyres on landing.

Probability is a numerical measure of the chance of a random event occurring. It is a number lying between 0 and 1 inclusive, the higher values representing greater chances of occurrence.

The probability of a random event can be determined, in principle, by carrying out a long series of repeated experiments and equating the proportion of times that the event occurs, in the long run, to its probability. In practice, it is often necessary to estimate probabilities from relatively small numbers of experiments.

Let $P(E_i)$ denote the probability that a particular event E_i occurs. Given two random events E_1 and E_2 , the probability that at least one of E_1 and E_2 occurs, or the probability that both of E_1 and E_2 occur, is often of interest.

Let $E_1 \cup E_2$ (read as “ E_1 **union** E_2 ”) denote the event that at least one of the individual events E_1 and E_2 occurs, and let $E_1 \cap E_2$ (read as “ E_1 **intersection** E_2 ”) denote the event that both E_1 and E_2 occur.

The addition rule states that

$$P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2). \quad (2.1)$$

If it is impossible for both E_1 and E_2 to occur, that is,

$$P(E_1 \cap E_2) = 0, \quad (2.2)$$

then E_1 and E_2 are said to be **exclusive events**, and the probability of at least one occurring is just the sum of their individual probabilities.

If at least one of E_1 and E_2 must occur, that is,

$$P(E_1 \cup E_2) = 1 \quad (2.3)$$

then E_1 and E_2 are said to be **exhaustive events**.

The **multiplication rule** states that

$$P(E_1 \cap E_2) = P(E_1) \times P(E_2 | E_1) \quad (2.4)$$

where $P(E_2 | E_1)$ (read as “ E_2 given E_1 ”) is the **conditional probability** that E_2 occurs, given that E_1 has occurred.

E_1 and E_2 are said to be **independent events** if

$$P(E_2 | E_1) = P(E_2); \quad (2.5)$$

otherwise they are said to be **dependent**.

Taking the examples of random events cited above, the probability of an aircraft bursting one or more tyres on landing will be increased if it makes a hard landing; hence the two events are dependent rather than independent.

If the events are independent, the probability of E_2 occurring is the same whether or not E_1 occurs, and the probability of both occurring is just the product of their individual probabilities. In this case, Equation (2.1) simplifies to:

$$P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1) \times P(E_2). \quad (2.1a)$$

When E_1 and E_2 are dependent events, $P(E_2 | E_1)$ may not be known. However, an upper bound for the probability of at least one of E_1 and E_2 occurring can be obtained by dropping the last term in Equation (2.1):

$$P(E_1 \cup E_2) \leq P(E_1) + P(E_2). \quad (2.1b)$$

Equations (2.1) to (2.5) can all be extended to combinations of more than two events (see, for example, Reference 3 Chapter 1).

2.2 Random Variables and Their Distributions

A **random variable** (or variate) is a numerical quantity associated with each outcome of an experiment involving an element of chance. A random variable is, by convention, denoted by a capital letter such as X , Y or Z , and the numerical values that it can take are denoted by the corresponding lower-case letters. Random variables may be either discrete or continuous as defined below.

For example, in the aircraft landing experiment discussed in Section 2.1, the *random variables* of interest are the peak vertical deceleration experienced by the aircraft and the number of tyres burst on landing. (Corresponding *events* of interest are the event that the deceleration exceeds a certain value (defined as a hard landing) and the event that one or more tyres burst.)

2.2.1 Discrete random variables

A **discrete random variable** may take one of a set of distinct values, which are usually integers. For example, the number of defects in a casting is a random variable which can take the values 0, 1, 2, 3, etc.

The probability distribution of a discrete random variable X can be defined by its **probability function**, $p(x)$ (sometimes called a frequency function), which assigns a probability to each possible value of X :

$$p(x) = P(X = x), \quad (2.6)$$

where $0 \leq p(x) \leq 1$ and, summing over all possible values of x ,

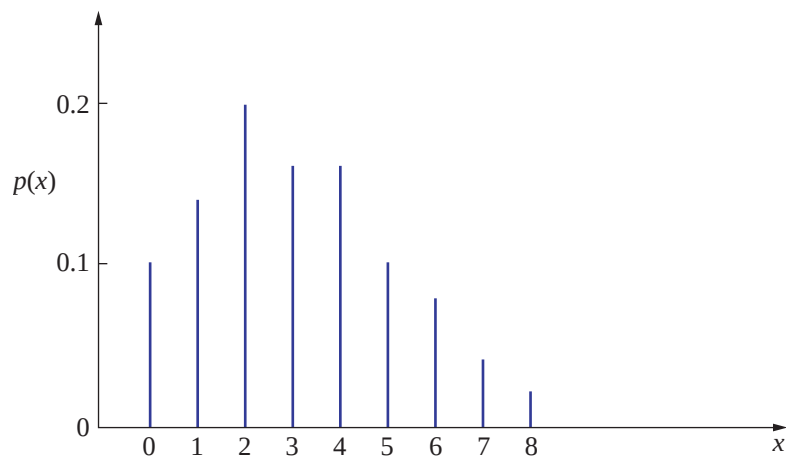
$$\sum_x p(x) = 1.$$

Another way of describing a discrete random variable is by its **cumulative distribution function**, $F(x)$ (often referred to simply as a distribution function), which assigns a probability to the event that the value

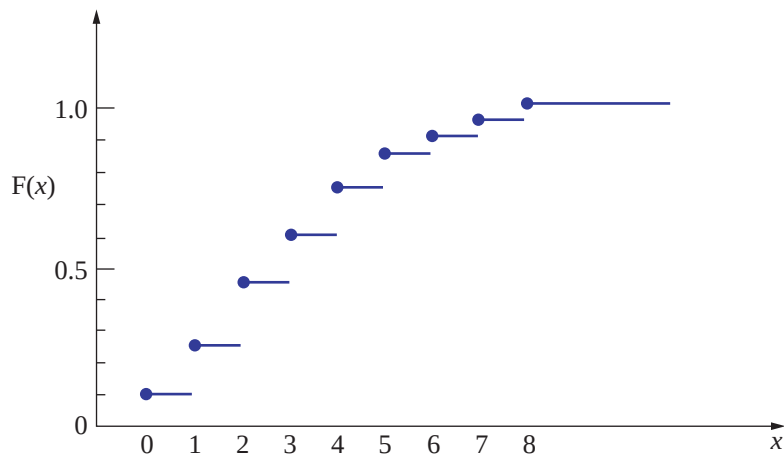
of X is not greater than a given value, x_1 :

$$F(x_1) = P(X \leq x_1) = \sum_{x=-\infty}^{x_1} p(x). \quad (2.7)$$

The probability function and cumulative distribution function can also be represented graphically. Sketches 2.1 and 2.2 illustrate the probability function and cumulative distribution function for the discrete random variable in the example given previously – the number of detectable defects in a particular design of casting. (The probabilities of 8 or more defects are assumed in this case to be negligible for all practical purposes.)



Sketch 2.1 Probability function of a discrete random variable



Sketch 2.2 Cumulative distribution function of a discrete random variable

2.2.2 Continuous random variables

A **continuous random variable** may take any value in a given interval. For example, the time to failure of a component in a fatigue test may take any value in the interval 0 to ∞ . (Although it will only usually be possible to make measurements to a limited number of significant figures, a variable is regarded as continuous when the underlying scale on which it is measured is continuous.)

The probability distribution of a continuous random variable, X , can be defined by its **probability density**

function, $f(x)$, given by:

$$P(x_1 < X < x_2) = \int_{x_1}^{x_2} f(x) dx \quad (2.8)$$

where $f(x) \geq 0$ and $\int_{-\infty}^{\infty} f(x) dx = 1$.

Note that the probability of a continuous random variable taking any given point value is zero, since there are an infinite number of possible values, although there is a finite probability of X falling in a narrow interval centred on that value.

A **cumulative distribution function** can also be specified for a continuous random variable:

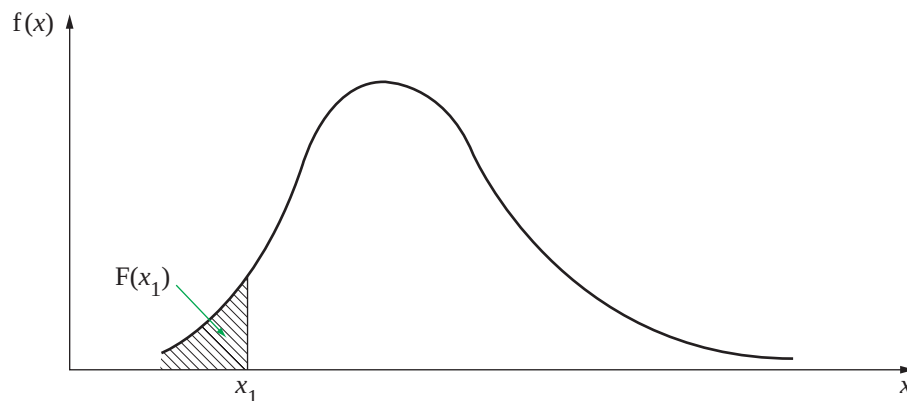
$$F(x_1) = P(X < x_1) = \int_{-\infty}^{x_1} f(x) dx. \quad (2.9)$$

The probability density function and cumulative distribution function are also related by the equation:

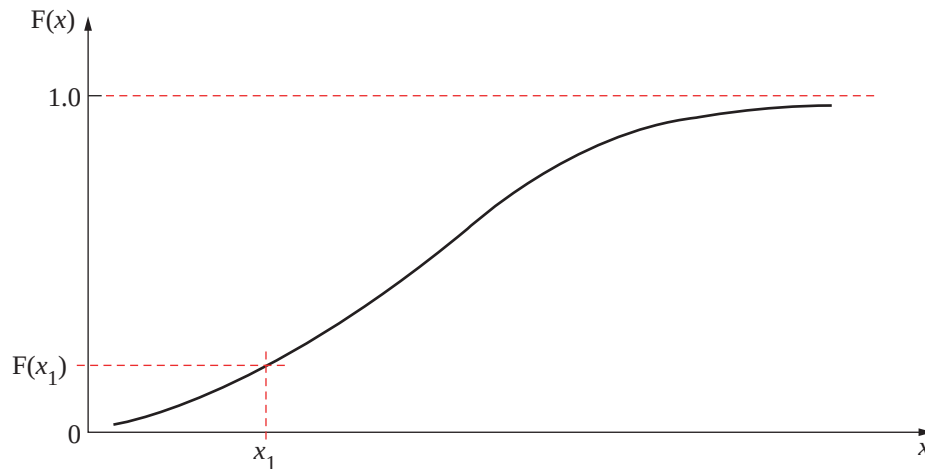
$$f(x) = \frac{dF(x)}{dx} \quad (2.10)$$

for those points x at which $F(x)$ is differentiable.

The probability density function and cumulative distribution function for a continuous random variable, representing the time to failure of a component in a particular fatigue test, are illustrated in Sketches 2.3 and 2.4.



Sketch 2.3 Probability density function for a continuous random variable



Sketch 2.4 Cumulative distribution function for a continuous random variable

2.2.3 Measures of location and dispersion

It is often useful when dealing with random variables and their probability distributions to obtain measures of their central location, and their spread, or dispersion. The location of a random variable is usually expressed in terms of its mean, and less commonly by its median or its mode, whilst the spread of the variable is usually measured by its variance, standard deviation or coefficient of variation. These terms are defined below.

The **mean** or **expected value** of a random variable is denoted by μ or $E(X)$, where E is the **expectation** operator.

For a discrete random variable, X , it is defined to be:

$$\mu = E(X) = \sum_x xp(x) \quad (2.11)$$

and for a continuous random variable it is defined to be:

$$\mu = E(X) = \int_{-\infty}^{\infty} xf(x)dx. \quad (2.12)$$

The **geometric mean** is also occasionally used in statistical analysis; it is only applicable to variables that cannot take negative values. The geometric mean, G , of a discrete or continuous variable, X , can be determined by noting that $\log G$ is the arithmetic mean of $\log X$ (where logarithms may be taken to any convenient base):

$$\log G = E(\log X). \quad (2.13)$$

The **median** of a random variable is denoted by $x_{0.5}$ and is defined to be the smallest value of x that satisfies:

$$F(x) \geq 0.5. \quad (2.14)$$

The median is a special case of a **percentile** (or quantile) of a random variable, the $100p^{th}$ percentile (or p^{th} quantile), denoted by x_p , being defined as the smallest value of x for which $F(x) \geq p$.

For a discrete random variable, x_p is the value at which $F(x)$ “jumps” from a value less than p to a value greater than or equal to p .