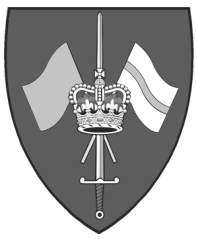


THE ROYAL SCHOOL OF SIGNALS

TRAINING PAMPHLET NO: **363**

DISTANCE LEARNING PACKAGE
CISM COURSE 2000
MODULE 4 - STATISTICS

Prepared by:

Technology Wing
CISM Group

Issue date: 29 September 2000

Training & Recruiting
ARMY TRAINING AND RECRUITING AGENCY

DP Bureau

This publication is for training purposes only. It is not for general external use. It is not subject to amendment and must not be quoted as an authority.

Authority: _____ **(Name in blocks)**

Signature: _____

Date: _____

Review Date	Review Date	Review Date
Signature	**Signature**	**Signature**

Authority: _____ **(Name in blocks)**

Signature: _____

Date: _____

Review Date	Review Date	Review Date
Signature	**Signature**	**Signature**

Authority: _____ **(Name in blocks)**

Signature: _____

Date: _____

Review Date	Review Date	Review Date
Signature	**Signature**	**Signature**

CONTENTS

STATISTICS

CHAPTER 1
INTRODUCTION

101. Outline of Statistics.

a. Statistics is an increasingly important subject used in many types of engineering and scientific investigation. It is the science of collecting, analysing and interpreting observed data, using the theory of probability, and making valid conclusions and reasonable decisions and predictions on the basis of the analysis.

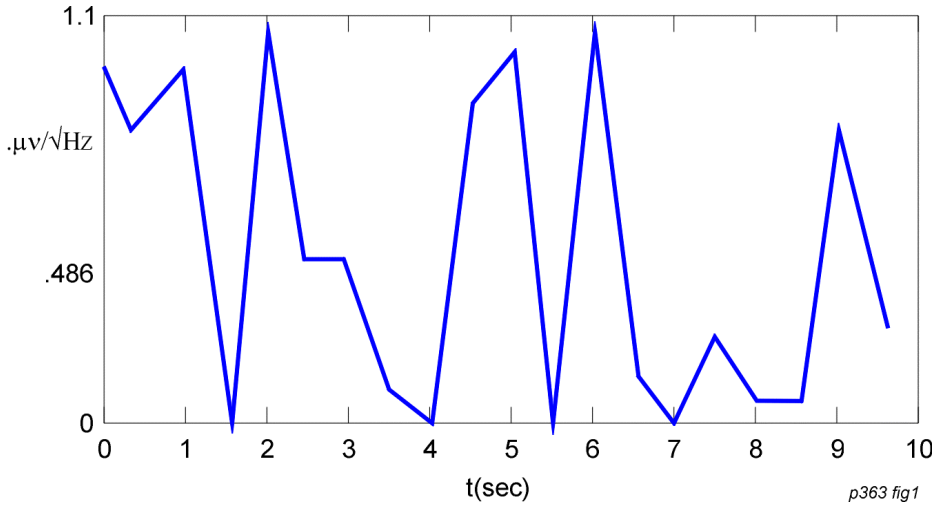
b. Statistics originally referred to simply the collection, organisation, and representation of data and dates back to ancient Rome. Increasingly, during the nineteenth and twentieth centuries, mathematics and probability theory were applied to the analysis. The word "statistics" is also used to mean the data itself, for example in the sense of "vital statistics", a set of measurements. Hence a "statistic" is an item of data or something calculated from data.

c. Statistics is particularly useful in situations where there is an *experimental uncertainty* and has particular applications to **information theory** and the theory of **noise** in communication channels.

EXAMPLE

The noise voltage generated by an operational amplifier at a constant temperature was measured over a period of 10 seconds and the following results obtained.

μV/√Hz	1·021	0·772	0·906	0·000	1·050	0·518	0·517	0·145	0·020	0·815
	0·983	0·002	0·985	0·144	0·002	0·296	0·060	0·058	0·773	0·295



These values vary from 0 to 1·05 with an average value of about 0·468. The variations do not seem to be systematic in any way.

A similar measurement made over another 10 second interval produced a completely different set of figures but with approximately the same average value and a similar spread. There is no apparent reason for the variations about this average value, nor do the variations appear systematic in any way.

Any variation in which there is no consistent pattern or regularity is called a ***random variation***. There is no way in which we can predict what the next value will be. We can measure the *average* noise voltage and we can estimate the *probability* that a measurement will be within a certain range, but we can never predict the magnitude of any particular value.

All naturally occurring phenomena are subject to a random variation, although sometimes this variation may be so small that we are not aware of it.

In this distance learning package, we shall mainly be concerned with *descriptive statistics*, i.e. collection and organisation of data. The study of probability theory and mathematical statistics will form part of the course at the Royal School of Signals.

102. Probability.

Although we shall not be discussing probability theory yet, the term may be mentioned from time to time. For the time being you may think of probability as being the chance or likelihood that something will happen. For example; you throw a dice. The chance of its coming up with a particular face, e.g. a six, is 1 in 6 or $\frac{1}{6}$. **Probability** is a number between 0 and 1 (inclusive).

103. Calculators.

For this course you are recommended to use a scientific calculator. If you do not have one already you should obtain a scientific calculator which has

a. Statistical functions.

b. Complex numbers.

Even the most basic, cheapest scientific calculator should have these and other functions.

In addition, if you have a PC, many of the statistical examples may be handled using Microsoft Excel or similar packages.

CHAPTER 2
ACCURACY AND APPROXIMATION

201. Variables.

a. In statistics, numbers are used to measure characteristics, e.g. length, height, time, voltage, the number of children in a family. A characteristic that is being measured is called a *variable*. In pure mathematics a variable is a quantity which takes some exact value according to a deterministic formula. For example, $y = x^2$. x is called the independent variable and y is called the dependent variable. For any value of x there corresponds an exact value of y, e.g. if $x = -3$ then $y=9$. In traditional physics we are accustomed to deterministic relationships such as Newton's law, $\mathsf{F = ma}$. In reality everything that we measure has some random element, even in physics and especially in nuclear physics and quantum theory.

b. In statistics we are usually dealing with variables that can vary considerably from some predictable path and so we call them ***random variables.*** A random variable is a variable which can take values in a particular range with various probabilities.

202. Discrete and Continuous Variables.

a. A discrete variable is one which takes only fixed discrete values in an interval. For example, the number of children in a family is a discrete variable which takes the discrete values 0, 1, 2, 3, 4, . . . The most probable value is 2, but if we select a family at random, the value could be anything between 0 and some large number, therefore it is a random variable. We toss a coin 6 times and count the number of heads, N that we get. N is a discrete random variable which can take the values 0, 1, 2, 3, 4, 5, 6 with different probabilities. (The most likely value of N is 3 but we cannot predict what it will be). We cannot get 1½ heads or 7 heads, therefore N is a random variable defined over the integers 0 to 6, i.e. $0 \leq \mathsf{N} \leq 6$.
In general, discrete variables arise from countings and enumerations.

b. A continuous variable is one which can take any value in an interval. For example an adult's height can take any value between say 1m and 3m. The time taken to run one mile is a continuous variable which can take any value within a certain range. In general, continuous variables arise from measurements.

203. a. Accuracy.

(1) Theoretically a continuous variable can take *any* value in an interval. The heights of adults form a continuum from approximately 1m to 3m. There are an infinite number of points in that interval. However, in practice we measure people's height to the nearest cm or the nearest ½ inch. We would not attempt to measure it to the nearest micron. Time is a continuous variable but we often quote the time to the nearest minute or the nearest 5 minutes. The time taken to run a race used to be measured to the nearest second but now to the nearest 0·1 second. We do not measure it to the nearest μs although we can measure time extremely accurately.

(2) All measurements are made and recorded to a given degree of accuracy. Firstly our measuring instrument has a certain degree of accuracy and a certain resolution. Secondly, the measurement may be rounded to some convenient level.

b. Resolution.

The resolution of a measuring device is the smallest change in the quantity which it can accurately detect. Ideally, an instrument should have a degree of accuracy better that its resolution.

c. Spurious Accuracy.

(1) How long is a piece of string?

(a) Measure it with a tape measure. The smallest division on the tape measure is 1 mm. Therefore our measuring instrument, the tape measure, has a *resolution* of 1 mm. We should not attempt to express the measurement to a greater degree of accuracy. (We assume that our tape measure is actually accurate to rather less than 1mm.)

(b) Measure it with a travelling microscope, (an extremely accurate measuring instrument). We obtain a reading to the nearest nanometer, however, at each measurement attempt the string stretches by 0·1 mm and so this degree of accuracy is meaningless.

(2) We measure a voltage on a digital voltmeter. The display is in steps of 0·001 V. Therefore the resolution is 1mV. However, we know the meter has an error of $\leq \pm 0{\cdot}01$ V and so our measurement is only accurate to the nearest 0·01 V.

(3) My digital watch displays the time to the nearest minute, therefore it has a resolution of 1 minute. However, it is approximately two minutes slow and therefore the reading is not accurate to the nearest minute. We should beware of attributing an accuracy to a measurement just because our instrument displays it. The accuracy may be worse than its resolution.

(4) We weigh 6 apples to the nearest gram. The weights are:

71 g, 64 g, 69 g, 58 g, 62 g, 73 g.

The mean weight is therefore $(71 + 64 + 69 + 58 + 62 + 73) \div 6$ grams
$= 66{\cdot}16666667$ grams.

This statistic suffers from spurious accuracy. Just because my calculator gives 10 digits does not justify this answer. As the original data was measured to the nearest gram, it is only valid to give the mean to the same degree of accuracy, i.e. 66 g.

204. Rounding.

Data may be rounded if:

a. the accuracy is in doubt

b. a lesser degree of accuracy is sufficient.

Examples : Round to two places of decimals:

21.2312 becomes 21·23. We round *down* because it is closer to 21·23 than to 21·24.

6·328 becomes 6·33. We round *up* because it is closer to 6·33 than to 6·32.

If the digit is a 5 do we round up or down ?

64·135 is equidistant from 64·13 and 64·14. Common practice is to round up but there is a school of thought which says that consistent rounding up causes systematic bias, and so adopt the following rule:

If the previous digit is even, round down.
If the previous digit is odd, round up.
Therefore 64·135 becomes 64·14
64·125 becomes 64·12

but note : 64·1251 becomes 64·13

205. Significant Figures.

Significant figures are the digits which carry information, provided that they are free from spurious accuracy. The numbers of digits which are significant may not be apparent from the data because it will partly depend upon how much rounding has occurred previously.

	ROUNDED TO		
Calculated figure	4 significant figures	3 significant figures	2 significant figures
512·76	512·8	513	510
0·0036162	0·003616	0·00362	0·0036
5 142 936	5 143 000	5 140 000	5 100 000
25·003	25·00	25.0	25
$1{\cdot}5837 \times 10^{4}$	$1{\cdot}584 \times 10^{4}$	$1{\cdot}58 \times 10^{4}$	$1{\cdot}6 \times 10^{4}$
$2{\cdot}6365 \times 10^{-3}$	$2{\cdot}636 \times 10^{-3}$	$2{\cdot}64 \times 10^{-3}$	$2{\cdot}6 \times 10^{-3}$
64 000	64 000	64 000	64 000

In the last case, it is not apparent by inspection, which of the trailing zeros are significant. We would have to know how much the figure had been rounded.

SAQ 4-1

Round the data to 2 decimal places.

a. 26·236

b. 59·344

c. 74·215

d. 86·225

e. 39·3252

f. 71·475

g. 0·063

h. 0·016

i. 0·002

j. 0·007

SAQ 4-2

Complete the following table.

	ROUNDED TO		
Calculated figure	4 significant figures	3 significant figures	2 significant figures
232·85			
6 743 817			
0·000252371			
69·006			
$2{\cdot}3485\times10^{6}$			
$3{\cdot}6472\times10^{-3}$			

CHAPTER 3
SAMPLING

301. Population.

a. A population is the total set of all possible measurements of the particular characteristic under consideration. It is a collection of *statistics*. For example the heights of all adult males in the U.K. In this sense, **population** means the set of measurements; not the people who are being measured.

b. A population may be finite; small or large, or practically infinite. For example; the heights of all the soldiers in one regiment is a relatively small finite population. The heights of all adult males in the U.K. is a large finite population. The weights of all of one species of ant, is for all practical purposes an infinite population. When a population is infinite or very large, it is not possible to record it. Therefore we must take a **sample.**

3.2 Sample.

a. A sample is the subset of the population which is actually measured.

b. **A random sample** is one in which every item in the population has the same probability of being selected.

c. Sampling is necessary because in most cases it is impossible or very costly to measure the complete population, because:

(1) the population is very large or almost infinite.

(2) the sampling process alters the population in some way. For example; destructive testing of components. Killing of plants or animals before performing a test.

(3) The population does not actually exist. For example; initial sampling off the first run on a production line.

d. In practice, a sample gives sufficient information about a population, provided that the sample is taken properly and that it is sufficiently large. From an almost infinite population a sample of 1000 would be quite adequate. Practically, it can be very difficult to obtain a truly random sample.

CHAPTER 4
MEASURES OF CENTRAL TENDENCY

One of the statistics that we often wish to find is a central or average value.

401. a. The Arithmetic Mean (common average).

This is the most commonly used average.

b. POPULATION MEAN, μ.

The mean of a population of size N is : $\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$

where $x_1, x_2, x_3, \cdots, x_N$ are the population.

You will recall that this Σ notation was used in the calculus section. Σ stands for the **sum of**.

$\sum_{i=1}^{N} x_i$ is simply a shorthand way of writing $x_1 + x_2 + x_3 + \cdots + x_N$

i.e. to calculate the mean, add up all the values and divide by the number of them.

c. SAMPLE MEAN, $\bar{x}$.

In practice it is usually impossible to determine the true population mean and therefore we *estimate* it by calculating the sample mean:

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

where $x_1, x_2, x_3, \cdots, x_n$ is the sample of size n.

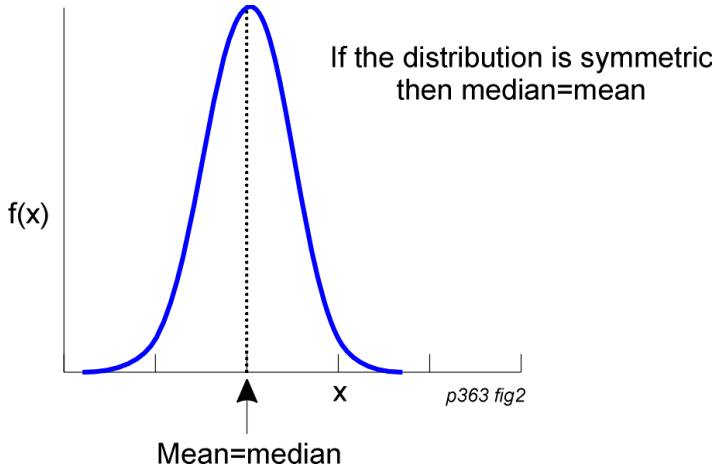
$\bar{x}$ is said to be an *estimate* or *estimator* of μ.

$\bar{x}$ is a *random* var*iable* and therefore also has a probability distribution.
In general we can say that the larger the sample, the better estimate it is of μ.

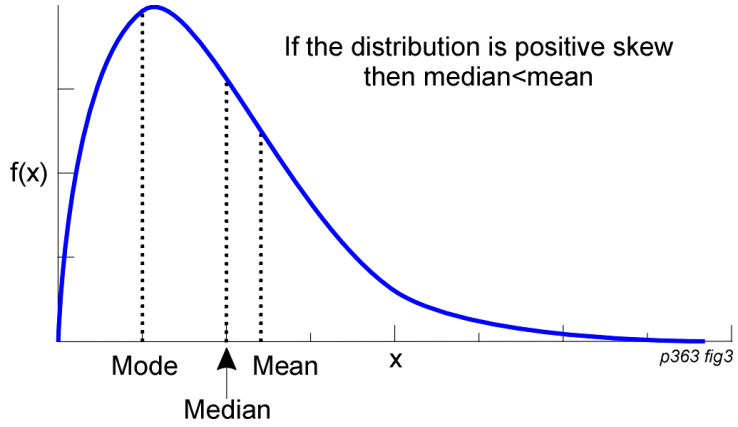
402. Other Measures of Central Tendency.

The median. This is the "middle value". Half the population is less than the median and half the population is greater than the median. If m is the median then

probability($x \le$ m) = 0·5 and probability($x \ge$ m) = 0·5

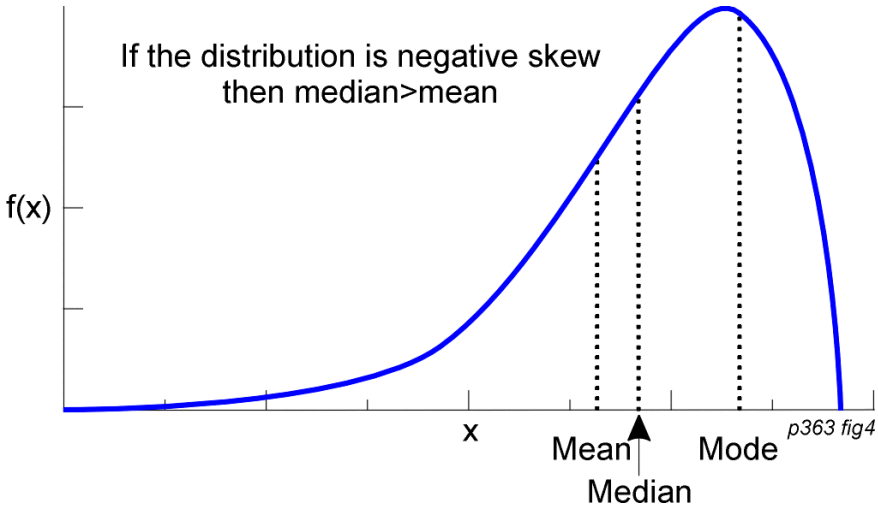


If the distribution is symmetric then median = mean



If the distribution is positive skew then median < mean

If the distribution is negative skew then median > mean



403. The Mode.

This is the most "popular" value, i.e. the one with the greatest frequency or the greatest probability. In the case of the Gaussian distribution which is symmetric and only has one peak, mean = median = mode.

SHORTCUTS IN COMPUTING THE MEAN

404. Consider the data:

5·1, 5·2, 5·5, 5·7, 5·3

If we subtract 5 from each item of data, and find the mean of 0·1, 0·2, 0·5, 0·7, 0·3, then the result is obviously equal to $\bar{x} - 5$. Therefore we simply add back 5 to the result.

CHAPTER 5
MEASURES OF SPREAD

501. The average value on its own does not give all the information about the distribution. It is necessary to know the spread.

502. The range.

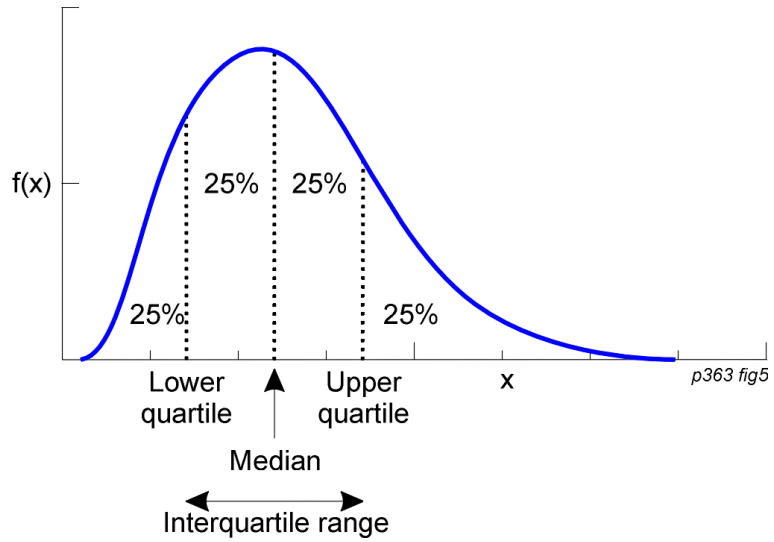
The difference between the greatest and smallest values. While this gives indication of the overall spread, as a measure it is unreliable because it is affected by freak extreme readings.

503. The interquartile range.

The difference between the upper and lower quartiles.

(25% of the population is less than the lower quartile and 75% of the population is less than the upper quartile. Other commonly quoted percentiles are the 10th and the 90th percentiles.)

Sometimes the semi-interquartile range is used instead. This is simply half the interquartile range.



504. The mean deviation.

The mean of the absolute values of the deviations from the mean, given by $\frac{1}{N}\sum_{i=1}^{N}|x_i - \mu|$

If the absolute value were not taken, the mean deviation would, of course, be zero. This statistic is more correctly known as the ***mean absolute*** deviation. This measure is rarely used.

505. The standard deviation.

This is the root-mean-square of the deviations from the mean. It is the most commonly used measure of spread.

506. THE POPULATION STANDARD DEVIATION, σ

This is defined as:

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

where $x_1, x_2, x_3, \cdots\cdots, x_N$ are the population.

σ^2 is known as the variance.

507. THE SAMPLE STANDARD DEVIATION, s.

In practice, it is not possible to measure the population standard deviation and so we use the sample standard deviation. Note that the mean will generally also be the sample value in this case. The sample standard deviation is defined as:

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

where $x_1, x_2, x_3, \cdots x_n$ is the sample of size n.

s^2 is known as the sample variance.

508. However, for **small samples** it is usual to use the value of s which gives an unbiased estimate of s, dividing by $n-1$ instead of n, i.e.

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

Obviously, if n is large, it makes almost no difference. The reason for applying this slightly different statistic to small samples is beyond the scope of this lesson and will be explained during the course at the Royal School of Signals.

509. Computing the standard deviation.

A shortcut in computing the standard deviation is to use the identity

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

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

The statistic $\sum_{i=1}^{n}(x_i - \bar{x})^2$ is known as S_{xx}.

It is evident from the formula for standard deviation that if we transform the data by adding or subtracting a constant, that the standard deviation is the same, i.e. the transformed data has the same **spread**.

510. For Example.

Consider the following data:

$$2{\cdot}125,\ 2{\cdot}106,\ 2{\cdot}048,\ 2{\cdot}225,\ 2{\cdot}349$$

If we subtract 2 from each item of data this may make the figures more manageable and so we find the standard deviation of

$$0{\cdot}125,\ 0{\cdot}106,\ 0{\cdot}048,\ 0{\cdot}225,\ 0{\cdot}349\ ,\ \text{which is of course the same.}$$

NUMERICAL EXAMPLES

511. The following table contains values of the resistance in ohms of a sample of 1kΩ resistors.

$$1015,\ 1066,\ 1100,\ 995,\ 915,\ 1082,\ 927,\ 933,\ 1077,\ 973.$$

Calculate the mean and standard deviation using the formula for small samples.

Even when using a calculator, it is often better to use a shortcut. If we transform the data by subtracting 1000 from each item, then we simply add 1000 back to the computed mean value. The standard deviation is unaffected.

	x	x^2
	15	225
	66	4356
	100	10000
	–5	25
	–85	7225
	82	6724
	–73	5329
	–67	4489
	77	5929
	–27	729
SUM	83	45031

$\Sigma x = 83 \qquad \Sigma x^2 = 45031$

Mean of transformed data = $83 \div 10 = 8{\cdot}3$

$\therefore \quad \bar{x} = 1000 + 8{\cdot}3 = 1008{\cdot}3\ \Omega$

$$\sum_{i=1}^{n}(x_i-\bar{x})^2 = \sum_{i=1}^{n}x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n}x_i\right)^2 = 45031 - \tfrac{1}{10}(83)^2 = 44342\cdot1$$

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2} = \sqrt{\frac{1}{9}\times 44342\cdot1} = 70\cdot2\,\Omega$$

SAQ 4-3

Ten measurements of a current in mA were made and were recorded as:

38·8, 40·9, 39·2, 39·7, 40·2, 39·5, 40·3, 39·2, 39·8, 40·6

Transform the data by subtracting 40 from each and hence find the mean value of current.

SAQ 4-4

A sample of 10 resistors of nominal value 2200 Ω were measured with the following results:

R in ohms : 2182, 2220, 2125, 2221, 2218, 2175, 2188, 2218, 2195, 2200

a. Transform the data by subtracting 2200 from each and find the mean resistance.

b. Using the formula for small samples, find the standard deviation.

CHAPTER 6
FREQUENCY DISTRIBUTIONS

601. When summarising large amounts of data it is often useful to distribute the data into *classes* or *categories*. For example: The following data are the heights in cm of 20 soldiers.

182, 175, 183, 179, 176, 175, 182, 180, 180, 182, 179, 179, 177, 177, 178, 178, 175, 180, 176, 183

This data may be summarised as follows:

Height (cm)	Frequency
175	3
176	2
177	2
178	2
179	3
180	3
182	3
183	2

It is obvious that the sum of all the heights of 175 cm is 3×175, the sum of all the height of 176 cm is 2×176, etc.

The total of all the heights can be calculated by multiplying each class value by its frequency and summing these products.

It is also evident that the sum of all the frequencies is equal to the total number of soldiers.

Height (x)	frequency (f)	fx
175	3	525
176	2	352
177	2	354
178	2	356
179	3	537
180	3	540
182	3	546
183	2	366
	20	3576

The mean height, $\bar{x} = \dfrac{\text{sum of all the heights}}{\text{total No of soldiers}}$

$= \dfrac{3576}{20}$

$= 178 \cdot 8 \text{ cm}$

Since the measurements were presumably made to the nearest cm we should round this off to 179 cm.

602. For a grouped frequency table, the formula for the sample mean is
:

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

where m is the number of groups (=8 in the above example)

STANDARD DEVIATION OF GROUPED DATA.

603. The formula for standard deviation from a grouped frequency table is:

$$s = \sqrt{\frac{\sum_{i=1}^{m} f_i (x - \bar{x})^2}{\sum_{i=1}^{m} f_i}}$$

$$= \sqrt{\frac{\sum_{i=1}^{m} f_i x_i^2 - \frac{1}{\sum_{i=1}^{m} f_i}\left(\sum_{i=1}^{m} f_i x_i\right)^2}{\sum_{i=1}^{m} f_i}}$$

where m is the number of groups.

In this formula, it can be seen that Σx^2 is replaced by $\Sigma f x^2$ and $(\Sigma x)^2$ is replaced by $(\Sigma f x)^2$.

n is replaced by Σf. For grouped frequency data, n is usually large, but if the sample is small we may use $n - 1$ instead.

Using the above figures for heights of soldiers;

Height (x)	frequency (f)	fx	x^2	fx^2
175	3	525	30625	91875
176	2	352	30976	61952
177	2	354	31329	62658
178	2	356	31684	63368
179	3	537	32041	96123
180	3	540	32400	97200
182	3	546	33124	99372
183	2	366	33489	66978
SUM	20	3576		639526

$$s = \sqrt{\frac{\sum_{i=1}^{m} f_i x_i^2 - \frac{1}{\sum_{i=1}^{m} f_i}\left(\sum_{i=1}^{m} f_i x_i\right)^2}{\sum_{i=1}^{m} f_i}} = \sqrt{\frac{639526 - \frac{1}{20}(3576)^2}{20}} = 2 \cdot 62 \text{ cm.}$$

As this sample is small, we shall use 19 instead of 20 (Note: in the denominator only)

$$\sqrt{\frac{639526 - \frac{1}{20}(3576)^2}{19}} = 2 \cdot 69 \text{ cm.}$$

FURTHER CLASSIFICATION INTO GROUPS

604. In order to make data more presentable and manageable, we may distribute the data into wider groups. For example:

100 soldiers were weighed and their mass recorded to the nearest kilogramme. The data is presented as follows:

Mass (kg)	No of soldiers
60 - 62	5
63 - 65	18
66 - 68	42
69 - 71	27
72 - 74	8

605. Although the measurements were made to the nearest kilogramme, we have grouped the data into wider intervals. Although grouping in this way may lose some of the original detail of the data, it may present a clearer overall picture.

CLASS INTERVALS AND CLASS LIMITS

606. A symbol defining a class such as 60 - 62 is called a *class interval* or sometimes simply the ***class***. The end numbers 60 and 62 are called *class limits* (the *lower class limit* and the *upper class limit*, respectively.)

607. Class Boundaries.

If masses are measured to the nearest kg, the class interval 60 - 62 includes all measurements from 59·5 to 62·5 kg. These numbers are called ***class boundaries*** or ***true class limits.*** Sometimes class boundaries are used to symbolise classes, e.g. the table could be written:

Mass (kg)	No of soldiers
59·5 - 62·5	5
62·5 - 65·5	18
65·5 - 68·5	42
68·5 - 71·5	27
71·5 - 74·5	8

608. To avoid ambiguity, such class limits should not coincide with actual observations. For example if a measurement were 62·5 kg, it would not be clear whether it belonged to the first or second class. In this case, since we recorded data to the nearest kg, the problem does not arise.

THE SIZE OR WIDTH OF A CLASS INTERVAL

609. The ***class width*** is the difference between the upper and lower class boundaries. In the above example, the class width is **3 kg**. Note that in general, the classes may not necessarily have the same width, although it is common practice to make them equal. In the above example, the class width is constant.

THE CLASS MARK OR CLASS MID-POINT

610. a. The class mark is the mid-point of the class interval and is the average of the lower and upper class limits or the average of the lower and upper class boundaries.

b. All observations within a class interval are represented by the class mark and this value is used for calculations. In the above example, the class marks are:

61, 64, 67, 70, 73.

These values would be used to calculate the mean or standard deviation.

HISTOGRAMS

611. A histogram is a graphical representation of a frequency distribution.

612. A histogram is similar to a vertical bar chart, but the rectangles are continuous.

a. It consists of a series of rectangles with their bases on the horizontal axis, their centres at the class marks, and their widths equal to the class interval sizes.

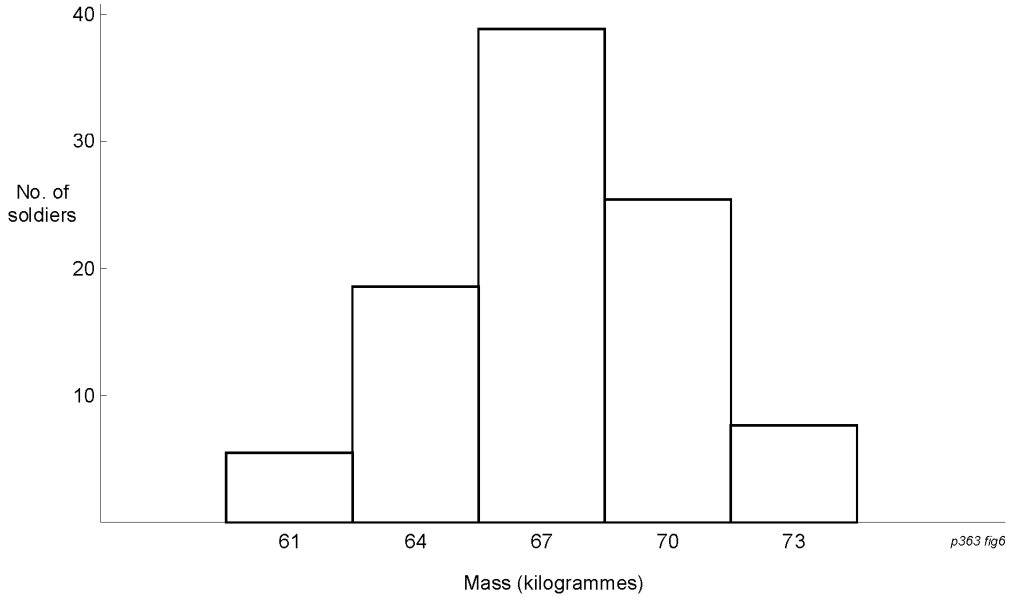
b. The area of a rectangle is proportional to the class frequency.

613. If the class intervals all have equal sizes, the heights of the rectangles are proportional to the class frequencies and it is then customary to make the height **equal** to the class frequency. If class intervals do not have equal sizes, then the heights must be adjusted so that the areas are proportional to class frequencies.

614. Consider the previous example.

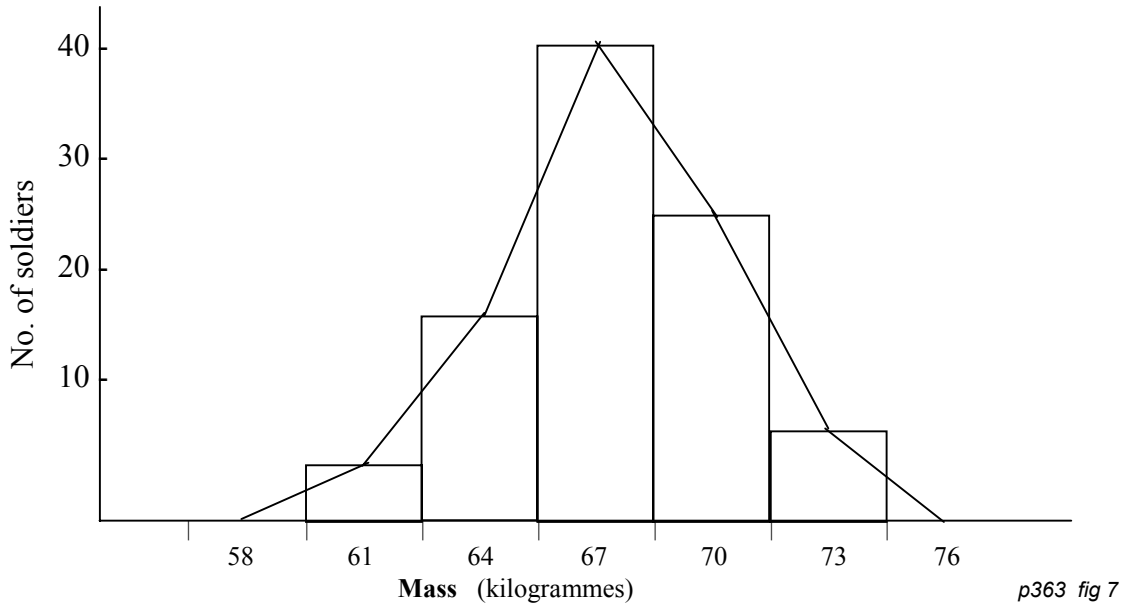
Mass (kg)	Class Mark	No of soldiers
59·5 - 62·5	61	5
62·5 - 65·5	64	18
65·5 - 68·5	67	42
68·5 - 71·5	70	27
71·5 - 74·5	73	8

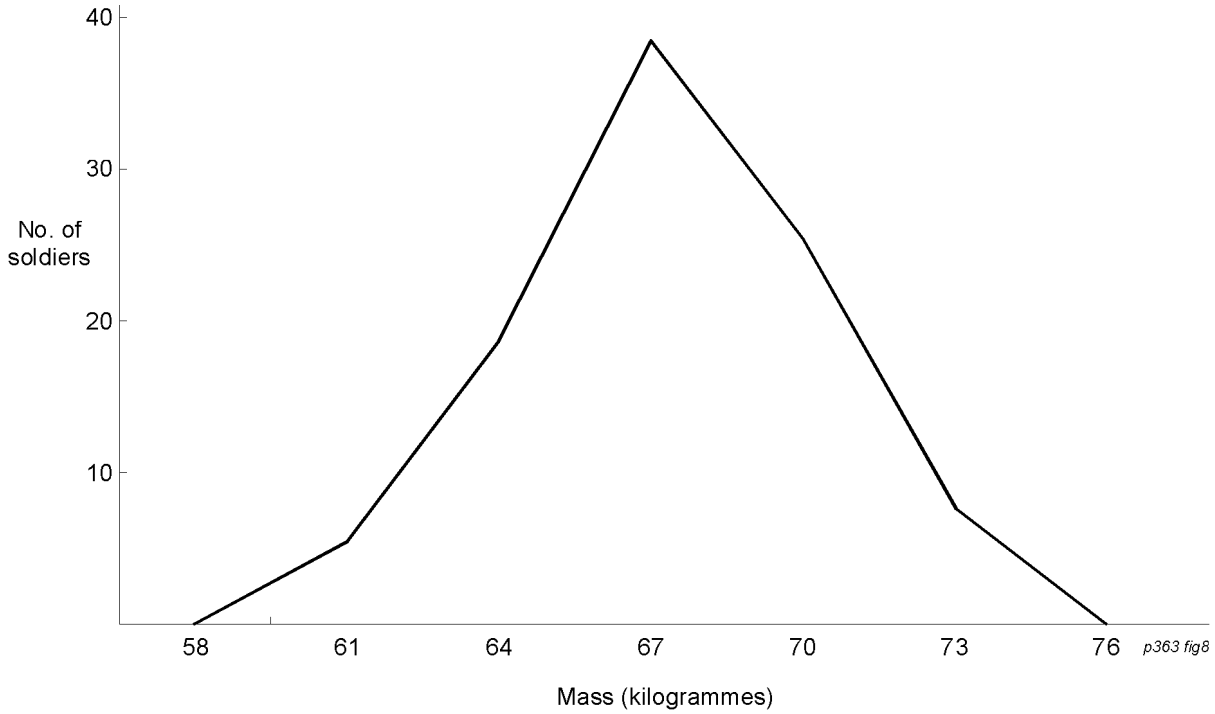
Below is the histogram drawn from this data.



THE FREQUENCY POLYGON

615. A frequency polygon is a line graph of class frequency plotted against class mark. It can be obtained by connecting the midpoints of the tops of the rectangles in the histogram. Note the extensions to the left and right indicating frequencies of zero.





The polygon is shown here on its own. The total area under the polygon will be equal to the area under the histogram and is therefore proportional to the total number of observations.

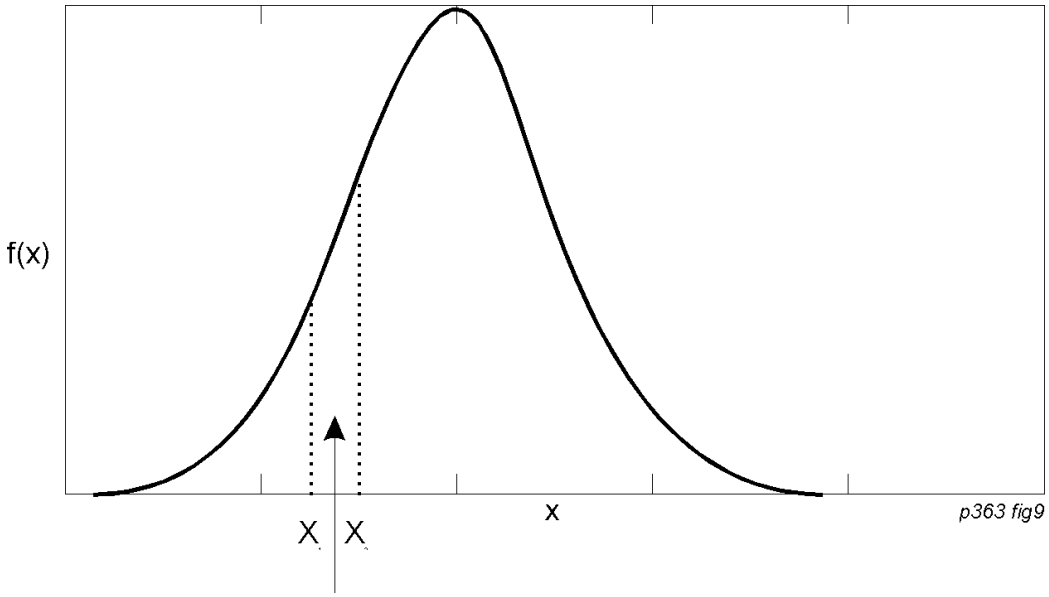
RELATIVE FREQUENCY DISTRIBUTION

616. The ***relative frequency*** of a class is the frequency of the class divided by the total frequency of all classes (the number of observations). For example, the relative frequency of the class with mark 67 is $42 \div 100 = 0{\cdot}42$. The relative frequency is an estimate of the **probability** of obtaining a measurement within this class range.

617. Graphical representation of a relative frequency distribution is obtained by simply changing the vertical scale of the histogram or frequency polygon from frequency to relative frequency. In this case, the total area under the diagram will be equal to 1.

618. Now imagine that we could measure the heights of all adult males in the U.K. (a practical impossibility) and plot the relative frequency polygon. Height is actually a **continuous** variable, so suppose that we make our measurements to the nearest 0·1 mm. The relative frequency polygon would consist of a very large number of lines, so close that they would merge into a continuous curve. The total area under the curve would be equal to 1. The proportion of men with height in a certain range would be given by the area under the curve between the two ordinates.

619. The frequency polygon is in fact tending in the limit towards the ***probability density function*** which would very likely be a Normal or Gaussian distribution.



The theory of probability density functions is very important in communications, and will be taught later in your course at the Royal School of Signals.

SAQ 4-5

The following table gives the breaking strain of a sample of 61 antenna stays.

a. Draw the histogram for this data.

b. Using the class marks, calculate the mean.

c. Using the class marks, calculate the standard deviation.

Breaking strain (kN)	No. of stays
93 - 97	2
98 - 102	5
103 - 107	12
108 - 112	17
113 - 117	14
118 - 122	6
123 - 127	4
128 - 132	1
	TOTAL 61

CHAPTER 7
ANSWERS TO SAQS

SAQ 4-1

a. 26·24

b. 59·34

c. 74·22

d. 86·22

e. 39·33

f. 71·48

g. 0·06

h. 0·02

i. 0·00

j. 0·01

SAQ 4-2

	ROUNDED TO		
Calculated figure	4 significant figures	3 significant figures	2 significant figures
232·85	232·8	233	230
6 743 817	6 744 000	6 740 000	6 700 000
0·000252371	0·0002524	0·000252	0·00025
69·006	69·01	69·0	69
$2{\cdot}3485 \times 10^{6}$	$2{\cdot}348 \times 10^{6}$	$2{\cdot}35 \times 10^{6}$	$2{\cdot}3 \times 10^{6}$
$3{\cdot}6472 \times 10^{-3}$	$3{\cdot}647 \times 10^{-3}$	$3{\cdot}65 \times 10^{-3}$	$3{\cdot}6 \times 10^{-3}$

SAQ 4-3

x	$x - 40$
38·8	–1·2
40·9	0·9
39·2	–0·8
39·7	–0·3
40·2	0·2
39·5	–0·5
40·3	0·3
39·2	–0·8
39·8	–0·2
40·6	0·6
TOTAL	–1·8

Mean of transformed data = –1·8 ÷ 10 = –0·18

= 40 – 0·18 = 39·82 mA

Since the measurements were made to the nearest 0·1 mA, we shall round this to 39·8 mA

SAQ 4-4

X	$x = X - 2200$	x^2
2182	–18	324
2200	0	0
2125	–75	5625
2221	21	441
2218	18	324
2175	–25	625
2188	–12	144
2218	18	324
2195	–5	25
2200	0	0
TOTALS	– 78	7832

a. $\text{Mean} = \dfrac{-78}{10} + 2200 = 2192\,\Omega$

rounded to the nearest ohm.

b. $$\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2$$

$$= 7832 - \tfrac{1}{10}(-78)^2 = 7223 \cdot 6$$

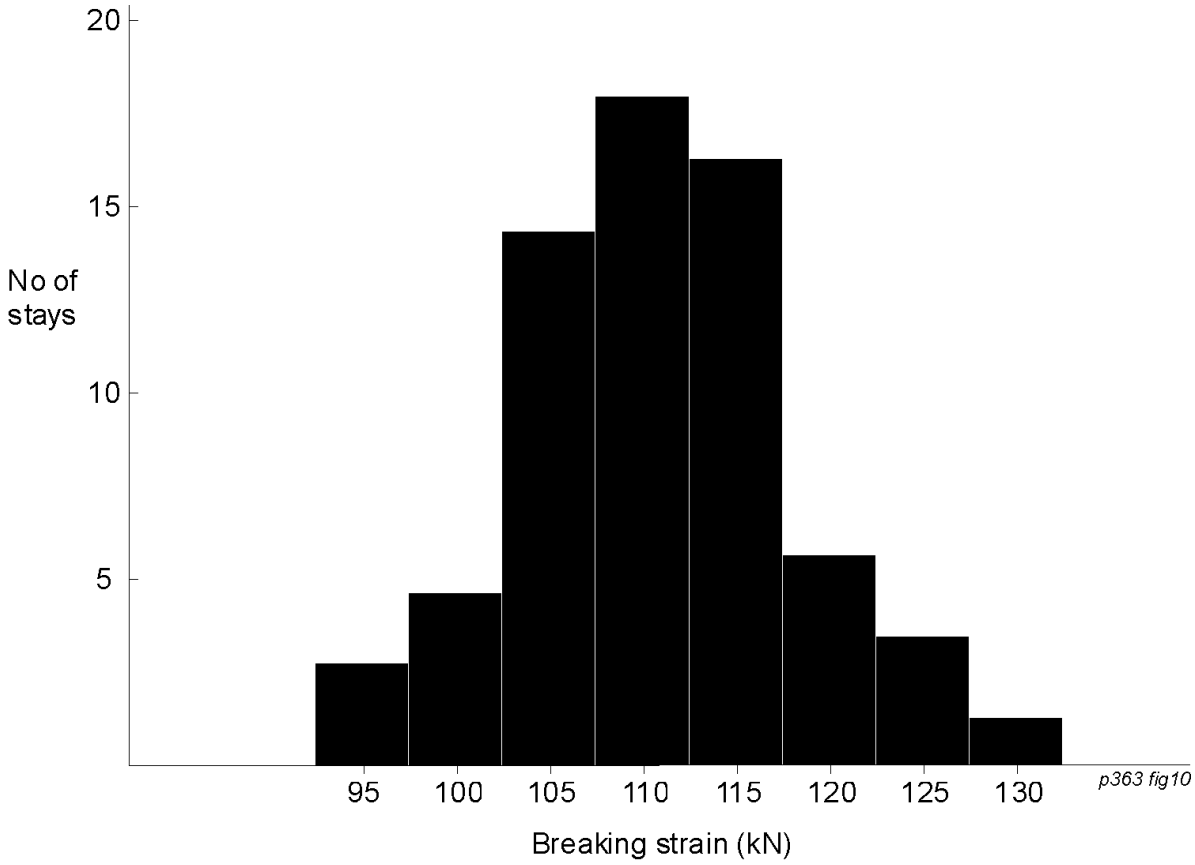
$$\therefore s = \sqrt{\frac{7223 \cdot 6}{9}} = 28 \cdot 33\ \Omega$$

Since the measurements were made to the nearest ohm we shall round this to 28Ω.

SAQ 4-5

a.

Class boundaries	Class mark	Frequency
92·5 - 97·5	95	2
97·5 - 102·5	100	5
102·5 - 107·5	105	12
107·5 - 112·5	110	17
112·5 - 117·5	115	14
117·5 - 122·5	120	6
122·5 - 127·5	125	4
127·5 - 132·5	130	1



b.

Class mark, x	Frequency, f	fx	x^2	fx^2
95	2	190	9025	18050
100	5	500	10000	50000
105	12	1260	11025	132300
110	17	1870	12100	205700
115	14	1610	13225	185150
120	6	720	14400	86400
125	4	500	15625	62500
130	1	130	16900	16900
TOTALS	61	6780		757000

$$\bar{x} = \frac{\sum_{i=1}^{m} f_i x_i}{\sum_{i=1}^{m} f_i} = \frac{6780}{61} = 111{\cdot}1 \text{ kN}$$

c.

$$s = \sqrt{\frac{\sum_{i=1}^{m} f_i x_i^2 - \frac{1}{\sum_{i=1}^{m} f_i}\left(\sum_{i=1}^{m} f_i x_i\right)^2}{\sum_{i=1}^{m} f_i}} = \sqrt{\frac{757000 - \frac{1}{61}(6780)^2}{61}} = 7{\cdot}5 \text{ kN}$$