Quality Assurance in Measurement

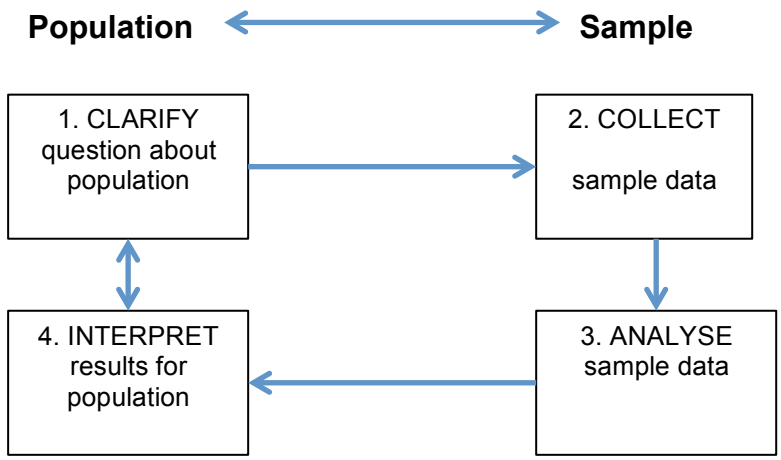
Module 3 - Part I - Statistical Methods in Measurement

Developed by Professor A Furness and Dr P Worthington © 2017

Contents

1.0 Statistical Populations

The population is the collection of things about which our investigation and analysis are designed to provide some insight and information. More specifically, we may be concerned with the way in which a particular quantity or characteristic varies i.e. the population (for example, weights or other measurements). We attempt to obtain information about such distributions by analysing samples drawn from these populations; this idea is summarized in the following diagram - we shall refer back to this at various points.



For our purposes there are two meanings to the term "population". First, we have populations that arise directly from practical situations, such as the one discussed below. Such populations may be called "real-life" populations. We also need "ideal" or "model" populations, whose distributions do not really exist in real life but which nonetheless are used to model real-life populations. Later in this course we shall examine the most important features of those ideal distributions most relevant to metrology.

1.1 "Real-Life" Populations

To a layperson, the word "population" usually means the inhabitants of a region or country. However, statisticians impart a much broader meaning to the word so that it does not relate just to a group of humans (or animals). In its most general form, a population is the collection of all items under investigation. More specifically, it is the collection of the values of those characteristics we wish to study.

Examples of such populations could be:

(a) The average petrol consumption of Ford automobiles manufactured in Great Britain last year.
(b) The average depth of tread of new ZX16 tyres.

(c) The weight of contents of nominally 750 g packets of Corn Flakes filled by a particular machine.

At this point, several words of caution are appropriate. As we indicated above, a population is ideally a specific collection of "things" on which unambiguous and relevant measurements or counts can be made on the characteristics of interest. It is also explicitly or implicitly assumed that so-called "random samples" can be taken of these things or their values. Although many teachers of statistics and books on the subject seem to assume (apparently, with scarcely a second thought) that the world is like that, a few moments of considered thought will reveal that this is often not the case. For instance, consider the above three examples.

(a) "The average petrol consumptions of Ford automobiles manufactured in Great Britain last year." How is "average petrol consumption" measured? What is meant by that expression? Perhaps we are referring to one of the standard Government or consumer organisation tests; but which one? Is the consumption at a steady 56 mph or at a steady 70 mph or under "simulated city driving" (whatever that is)? How about some real consumption figures from actual driving? But that depends very much on who is driving, how they are driving and where they are driving. Maybe we want to take such features into account and maybe we do not.

Let us turn to the other part of the definition of this population: "manufactured in Great Britain last year". This sounds simple enough - but wait! These days, different parts of cars are made in different companies and countries. Also, assembly may take place in still further locations. Thus it is difficult to define exactly what is meant by "Made in Great Britain".

Even "manufactured ... last year" is hard to define. Different components are made at different times and different sections are assembled at different times. Perhaps we mean, "completed last year". But when is a car "complete"? Is it when it is painted or when its accessories are incorporated or when it is delivered to a distributor or when it is sold or when it is registered?

All these may be irritating questions, but until they are answered, how can we know which vehicles we should regard as candidates for our sample?

(b) "The average depth of tread of new ZX16 tyres" In some respects, this is even more difficult to pin down than the Ford cars in (a). At least in that case there was an attempt (albeit rather unsuccessful) to define the population in terms of time and location. Here there is no such attempt. For one thing, this "population" is forever increasing, at least until this type of tyre ceases to be made. Also the depth may change over time because of

changes in manufacturing techniques. To put it mildly, it will be rather difficult to define what we mean by the selection of a random sample from this "population", let alone to actually take it!

(c) The weights of contents of nominally 750 g packets of Corn Flakes filled by a particular machine. The particular problems here are that (as in (b)) the number of things (packets) to be included in the population keeps on growing, and that the contents may depend very much on factors such as who is operating the machine, the machine settings used, and so on.

So, in summary, there can be (and usually are) many problems associated with the basic conditions assumed for statistical sampling, estimation and inference procedures. The population of "things" is rarely static being continually changing, with extra items being created and, possibly, others dropping out; consequently, in practice, a "random sample" is difficult to define, let alone take. Further, the characteristic(s) being measured may be continually changing for any given member of the population. Also, they might be very dependent on extraneous factors, as opposed to being a property of the population member itself.

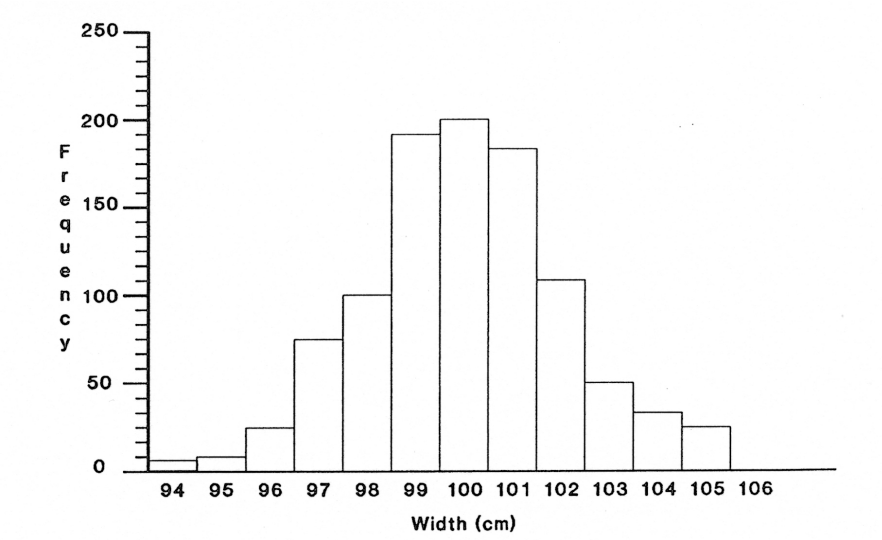
These problems are usually impossible to solve completely. Therefore, the best that we can do, if we are going to use standard statistical techniques, is to be continually aware of such problems, to evaluate their seriousness, and to alleviate their effects as far as possible. Also, we should take particular care to reduce looseness in the definitions as far as possible, and to define measurements on population items that are meaningful and informative. Further, if populations are to be compared with each other, we should try to ensure that these comparisons are as "fair" as is possible. We should also not be afraid of saying that standard statistical techniques are inappropriate and should not be carried out when conditions are too far from what the theory assumes. In particular, when either the population, or the measurements of the characteristics of interest are varying in an unstable and unpredictable way, a sampling experiment carried out at the present time may have little bearing on the future, and may mislead. Indeed, the concept of stability is paramount for the use of the techniques covered in this course and for most other statistical procedures as well; after all, if the subjects of our study are unstable over time, how can the "now" tell us anything about what is to come? It is for this reason that we make a point of looking at the stability of data later in the course.

All the examples and illustrations in this course are to be viewed with the understanding that these matters have been fully considered, and the researchers are satisfied that the situations are sufficiently stable for the statistical techniques to be meaningful.

To illustrate a few ideas about populations, we have chosen data collected during June 2015 related to the width (in cm) of a fabric product. The table below gives information on the distribution of widths.

Width (cm)			Frequency
93.5	to	94.5	4
94.5	to	95.5	8
95.5	to	96.5	27
96.5	to	97.5	78
97.5	to	98.5	108
98.5	to	99.5	195
99.5	to	100.5	202
100.5	to	101.5	83
101.5	to	102.5	118
102.5	to	103.5	55
103.5	to	104.5	34
104.5	to	105.5	12
105.5	to	106.5	1
			1025 = Total

Such a table is often called a *frequency distribution* table since, as in this case, it shows how the frequencies of widths are distributed according to the various groups. From it we can glean a lot of information about the population. For instance, only about 1% (i.e. 12/1025 + 1/1025 = 0.0127) of all measurements were greater than 104.5 cm. About 11% (117/1025 = 0.1141) of the fabrics had a width less than 97.5 cm. Also the bulk of the measurements are centred on 100 cm (the target value). This type of information can be nicely illustrated in a *histogram* in which the frequencies are represented diagrammatically by rectangular blocks. A histogram for the fabric data is shown in Figure 1.1



Our statement about the bulk of the measurements being centred on the target value of 100 cm can now be seen "at a glance" from the histogram. A particularly important feature of the distribution is also immediately revealed by the histogram, namely its symmetry about the target value. This symmetry is what we would expect from such data where there have been no trends (up or down) throughout June that could have destroyed the symmetry and where we are expecting the majority of the values to centred on the target value.

In addition to extracting information about proportions and shape of the distribution, it is also useful to obtain some numerical measures that summarize main features of the population. The two typical numerical measures of greatest importance are (a) averages which, in some sense, locate the "centre" of the distribution - hence such averages are often called measures of location; and (b) measures of dispersion which indicate, for example, whether the majority of the population are closely clustered around the average or are more widely dispersed.

1.2 Measures of Location

There are many ways of measuring the "centre" or "average" of a population. The two most popular ones are the *mean* and the *median*.

The *mean* of a population (usually denoted by the Greek letter μ, pronounced "mew") is equal to the sum of all the values divided by the size of the population. For the fabrics data we obtain

$$\mu = \frac{102475}{1025} = 99.98 \text{ cm}$$

Although the mean is a much-used measure of location, it does suffer some disadvantages For example, unusually high or low values can make the mean unrepresentative of the population as a whole.

The median, which we will denote by the Greek letter phi, ø, (we have chosen to represent the median of a population by ø as this represents a population split into two equal parts - which is just what the median does!) is very simple in concept: it is just the middle value of the population in the sense that half the population lies below it and half above it. This is, of course, an eminently reasonable measure of location; in fact, many would say that it is a more natural measure than the mean It certainly satisfies the criterion of being "central" more obviously than does the mean.

If the population is not too large then the median is easily obtained after simply arranging the numbers into ascending order; for example the numbers

4 8 7 5 8 5 6

are arranged as

4 5 5 6 7 8 8

from which we clearly see that the median, ø, is equal to 6.

If there is an even number of observations, for example,

5 5 7 7 8 9 12 16

then the median is taken to be halfway between the middle pair, i.e. ø = (7 + 8)/2 = 7.5

However, most populations are too large for this simple technique to be suitable and a more appropriate method of obtaining the median is from the so-called *cumulative distribution function* (cdf). Firstly the cumulative frequencies are calculated by simply summing successive frequencies in the frequency table. A table of cumulative frequencies for the above data is given below.

Width (cm)			Frequency	Cumulative Frequency	% Cumulative Frequency
93.5	to	94.5	4	4	0.39
94.5	to	95.5	8	12	1.17
95.5	to	96.5	27	39	3.80
96.5	to	97.5	78	117	11.41
97.5	to	98.5	108	225	21.95
98.5	to	99.5	195	420	40.98
99.5	to	100.5	202	622	60.68
100.5	to	101.5	83	805	78.54
101.5	to	102.5	118	923	90.05
102.5	to	103.5	55	978	95.41
103.5	to	104.5	34	1012	98.73
104.5	to	105.5	12	1024	99.90
105.5	to	106.5	1	1025	100
			1025 = Total		

You will see that we have added a final column of percentage frequencies; these will make the identification of the middle value (the 50% value) easier. Once the cumulative distribution table has been calculated, a close approximation to the median can be easily obtained by drawing a graph of the cdf. Figure 1.2 shows the correct representation of the cdf of this population while Figure 1.3 shows the sort of graph that is usually drawn when one wants to obtain an estimate of ø by locating the 50% point.

Figure 1.2

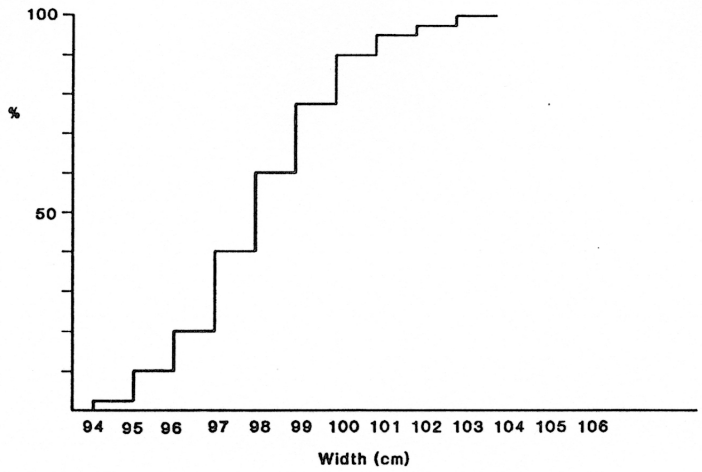
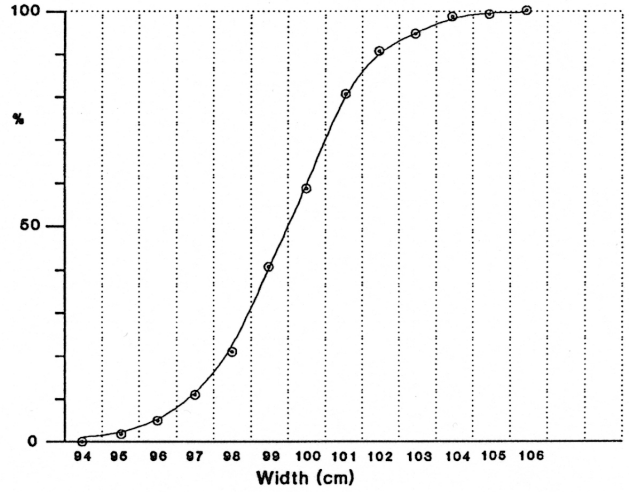


Figure 1.3



In Figure 1.3, the dotted line "shows the way" and gives the median value for the population of about 99.5 cm (incidentally, the true value is 99.97 cm). Of course, this estimate of the median is not absolutely accurate, but it will generally be good enough for "most practical purposes".

1.3 Measures of Dispersion

The second vitally important characteristic of a population is the dispersion (or spread, or variability) of its members. It might be that the members of a population are tightly bunched, in which case there is little dispersion in the population (e.g. see Figure 1.4). Low dispersion is most desirable in many populations; for instance, those relating to the manufacture of scientific instruments, the weights of foodstuffs, filling operations. Indeed, virtually every kind of measurement on virtually every production process. On the other hand, large variability, or dispersion, between items in any such situations (see Figure 1.5) may indicate faulty machinery, slap-dash methods, erratic quality of raw material etc., - which certainly would cause displeasure to the customer. In other contexts, high dispersion of a population might be an entirely expected and natural phenomenon, such as, for example, with the weights or heights of a human population.

Let us briefly say something about the standard deviation of a population, which is the quantity most commonly used in classical statistics to measure dispersion.

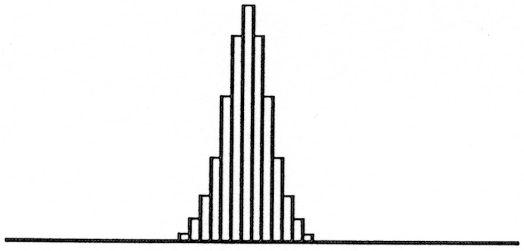


Figure 1.4

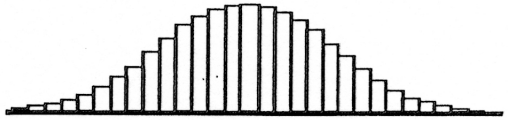


Figure 1.5

As we shall see shortly, the standard deviation does not arise in a particularly obvious manner. One "natural" measure of dispersion is the average of the differences between the values of the population and the "centre" of the population. Now, in classical statistics, "averages" and "centres" are usually described in terms of means and so, if a population has N members (denoted by X_1, X_2, X_3 ..... , X_N), a "natural" measure of dispersion is

$$\frac{1}{N}\{ | X_1 - \mu | + | X_2 - \mu | + | X_3 - \mu | + ... + | X_N - \mu | \}$$

where the quantities $|X_1 - \mu|$, $|X_2 - \mu|$ etc. are the distances of each member from the mean of the population. (The vertical bars denote what are called "absolute values" or "sizes".) So for example, if $\mu = 10$ and $X_1 = 13$ then $|X_1 - \mu| = |13-10| = 3$, and if $X_2 = 7$ then $|X_2 - \mu| = |7-10| = 3$

Both of these sizes are 3 since 13 and 7 are each 3 units from $\mu = 10$.

The above measure of dispersion is sometimes referred to as the ***Mean Absolute Deviation*** and is often abbreviated by MAD. By using some standard mathematical notation, the above expression for the MAD can be abbreviated to:

$$\text{MAD} = \frac{1}{N} \sum_{i=1}^{N} |X_i - \mu|$$

where $\sum_{i=1}^{N}$ is a summation notation that stands for "in the expression which follows, set I equal to 1, then 2, then 3, and so on1 up, to N, and then add up (sum) all those expressions so obtained.

Unfortunately, mathematical statisticians have never been too fond of the MAD as the absolute values $|X_1 - \mu|$ are a bit awkward to handle algebraically. But the use of absolute values in the MAD expression seems necessary in order to prevent positive and negative differences (deviations) cancelling each other out.

However, there is an alternative way of preventing this; and that is to square the deviations. The resulting measure is:

$$\frac{1}{N}\{(X_1 - \mu)^2 + (X_2 - \mu)^2 + ... + (X_N - \mu)^2\} = \frac{1}{N}\sum_{i=1}^{N}(X_i - \mu)^2$$

This measure is known as the ***variance*** of the population and is denoted by σ^2. However, this is really not at all a "proper" measure of dispersion (although we shall certainly need to use the variance when we come to combine errors). For example, if you are comparing two populations where the deviations in one are three times as big as in the other, then you will find that the variance of that population is 9 times as big as the other.

Similarly, if the deviations were 4 times as big, the variance would be 16 times that of the other. The remedy is simple, we take the square root of the variance to give a valid measure of dispersion; this overcomes the problem just mentioned.

It is this measure that is known as the ***standard deviation*** of the population. The standard deviation is usually denoted by the Greek letter σ (sigma). Thus we have

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

The standard deviation of the widths of the fabric is equal to 1.99 cm. A much smaller value would have indicated that most of the widths were closer to the mean value (so there would not be much dispersion i.e. a consistent product); while a much greater value (for example, $\sigma = 10$) would have indicated that the widths were spread more broadly over the possible range of 70 to 130 cm, for example (i.e. the product would be quite variable.)

There is another measure of dispersion, namely the range (R). This is simply equal to:

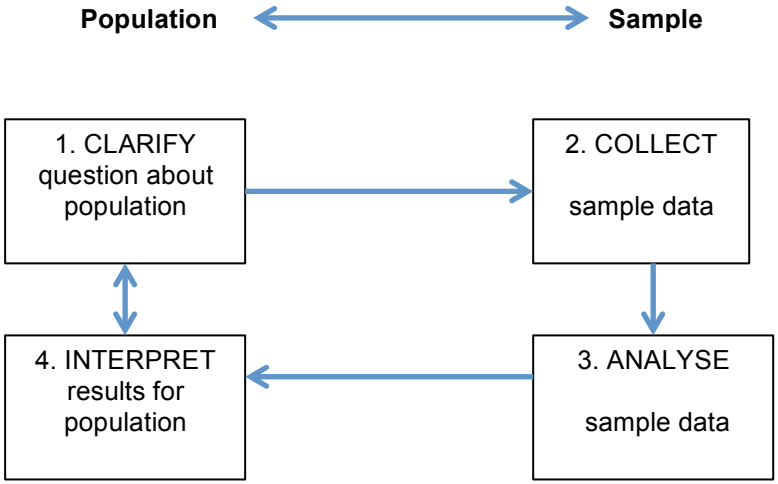
biggest value - smallest value.

For the fabric data, this is equal to 105.82 - 93.56 = 12.26 cm.

The range is most useful when dealing with samples where we wish to estimate the population standard deviation; there is a particularly quick way of doing this using the sample range - but more of that later!

2.0 Samples

More often than not, it is impractical for us to obtain the various population features we have mentioned: the histogram, the median, the mean and the standard deviation. This is likely to be because of the high cost or sheer impossibility of measuring or observing every member of a large population. So we are obliged to take a sample from the population in the hope that the sample characteristics (sample median, mean etc.) will be close to the corresponding population characteristics. Indeed this is really what much of the subject of statistics is all about: we are finding out things about populations by investigating samples taken from them. The modelling diagram again illustrates this point:



To illustrate a few of these sample characteristics we have taken a random sample of 40 from the population of fabric widths. (Here, the word "random" implies that all possible selections from the population are equally likely). The widths of the 40 fabrics were:

99.39	99.87	100.83	95.45	101.67
100.02	101.09	100.20	100.21	100.14
99.54	101.29	103.55	101.77	96.27
100.07	102.22	103.60	99.89	102.43
99.05	100.43	95.77	100.91	103.23
102.04	98.88	99.03	101.25	100.74
100.26	102.30	103.18	101.35	97.22
99.03	98.86	99.55	98.49	97.76

It is particularly interesting to calculate the various sample characteristics in this case, since we will be able to compare them with the known corresponding population values! First let us construct the sample histogram. To help with this we have formed the frequency table using the same intervals as for the population:

Width (cm)			Frequency	Cumulative Frequency	% Cumulative Frequency
93.5	to	94.5	0	0	0
94.5	to	95.5	1	1	2.5
95.5	to	96.5	2	3	7.5
96.5	to	97.5	1	4	10.0
97.5	to	98.5	2	6	15.0
98.5	to	99.5	6	12	30.0
99.5	to	100.5	11	23	57.5
100.5	to	101.5	7	30	75.0
101.5	to	102.5	6	36	90.0
102.5	to	103.5	2	38	95.0
103.5	to	104.5	2	40	100
104.5	to	105.5	0	40	100
105.5	to	106.5	0	40	100
			40 = Total		

Figure 2.1 shows the sample-histogram superimposed on the population histogram.

Similarly, Figure 2.2 shows a graph of the sample-cdf superimposed on a graph of the population cdf (note that in both cases we have drawn the true graphs of the cdfs rather than the smoothed versions such as we drew before in order to obtain a value for ø).

Figure 2.1

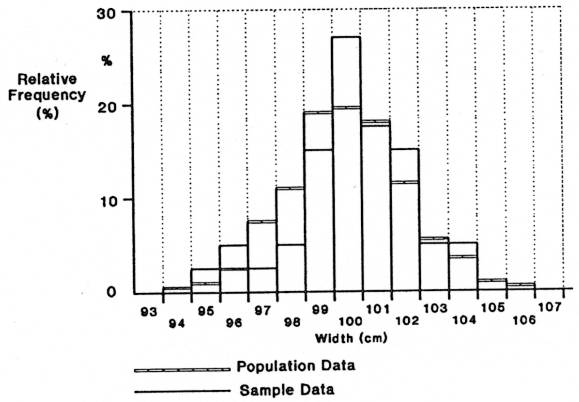
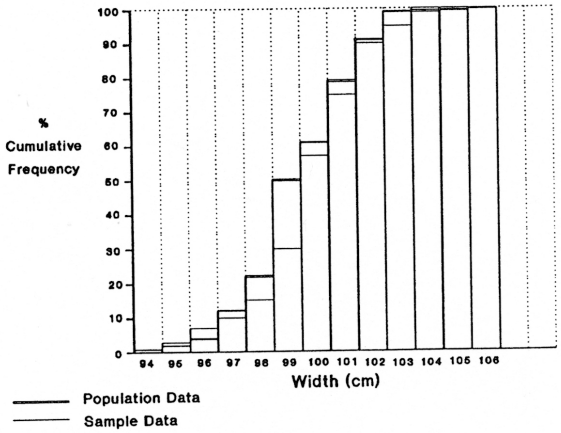


Figure 2.2



The size of the population (40 compared to 1025) is how well the pictures for the sample reflect the corresponding population pictures. Let us see whether the numerical measures for the samples similarly bear any resemblance to those for the population. The sample mean, which we will denote by $\bar{x}$, is calculated using the analogous formula to that for the population, namely:

$$\text{sample mean } \bar{x}, \quad = \frac{\text{sum of all the values}}{\text{number of values}}$$

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

where we have denoted then values in the sample by x_1, x_2, x_3, ... , x_n.

Using this formula on our sample gives

$$\bar{x} = \frac{4008.83}{40} = 100.22$$

The sample median, which we will denote by $\tilde{x}$, is the halfway point in the sample which here, with a sample of 40, is halfway between the 20th and 21st members when the data have been ordered.

A clever way to reduce the effort involved is to note from the cumulative frequency table that these two values must be in the width group "99.5 to 100.5" and so in fact we need only sort through the 11 values in this group. Doing this, we find that the 20th and 21st values are 100.2 and 100.21 respectively, giving

$\tilde{x}$ = (100.2 + 100.21)/2 = 100.21 (to 2 decimal places)

The standard deviation, sd, of the sample is also calculated in a similar fashion to its population counterpart. The formula for sd is

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

and for our sample this produces sd = 1.95 cm.

These sample statistics are remarkably close to their population counterparts as the following table shows.

	Mean	Median	Standard Deviation
Population	μ= 99.98	ø = 99.97	σ = 1.99
Sample	$\bar{x}$ = 100.22	$\tilde{x}$ = 100.21	sd = 1.95

Of course, one cannot depend on the sample statistics always being as close as this to their population counterparts (you may well have discovered this fact in doing the above Activities!). However, with a sample as large as 40, it is quite usual for them to be fairly close. Indeed, classical statistical methods very largely depend on this being the case.

We have calculated the sample statistics in this instance for interest only, since we already know the population equivalents. In real-life situations, these population "parameters" are unknown, and the sample statistics are used to provide information about them - this example shows they can be quite effective in so doing.

2.1 A necessary modification of the standard deviation

Although in classical statistics the sample mean $\bar{x}$ is used to estimate the population mean μ, the sample standard deviation, sd, is not generally used to estimate the population standard deviation σ. The reason for this is that sd tends to under estimate σ, particularly with small samples. The remedy for this problem is to use a modified version of the sample deviation that is defined by:

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

i.e. s is the same as sd but with the n in the denominator replaced by n-1. So whenever we need to estimate σ we shall invariably use s.

3.0 Exploratory Data Analyses (EDA)

In the late 1970's Professor John Tukey presented a number of simple techniques designed for examining data in its "raw state" i.e. for getting an initial feel for the data. These techniques easily highlight outliers (see "A Note on Outliers" below), and inform us, in a simple manner, of the main characteristics of the sample and enable quick and very effective comparisons to be made between two or more data sets. The success and effectiveness of these techniques is such that very often instead of giving us just initial thoughts about our data, they supply us with the final conclusions about the situation! So rather than having to involve ourselves with complicated statistical and mathematical theory we often "get away" with using one or more of Tukey's ideas (though some of the techniques we examine in this unit are older than Professor Tukey!).

The techniques we will look at are

(a) run charts

(b) dot plots

(c) boxplots (sometimes called box-and-whisker plots).

In dealing with a real situation we will usually find ourselves using more than one of the above techniques in an effort to squeeze as much information as we can from the data. We could have included histograms in this list; however we excluded them since we have already mentioned them in the unit on samples and will meet them again in the unit on normality. (Actually, histograms are a little more tedious to produce than many of the other techniques in this unit and so they would have been "out of place".) Of course, you might ask: "Why EDA in metrology?" Well, there is very often a need to be able to sense quickly what is going on in a given situation. These simple "pencil and paper" techniques are ideal for enabling us to do just that at the place where the measurements are being taken! Thus there is no need for us to go in search of a computer to analyse our data (though we will have the facility to illustrate these ideas on our computer!).

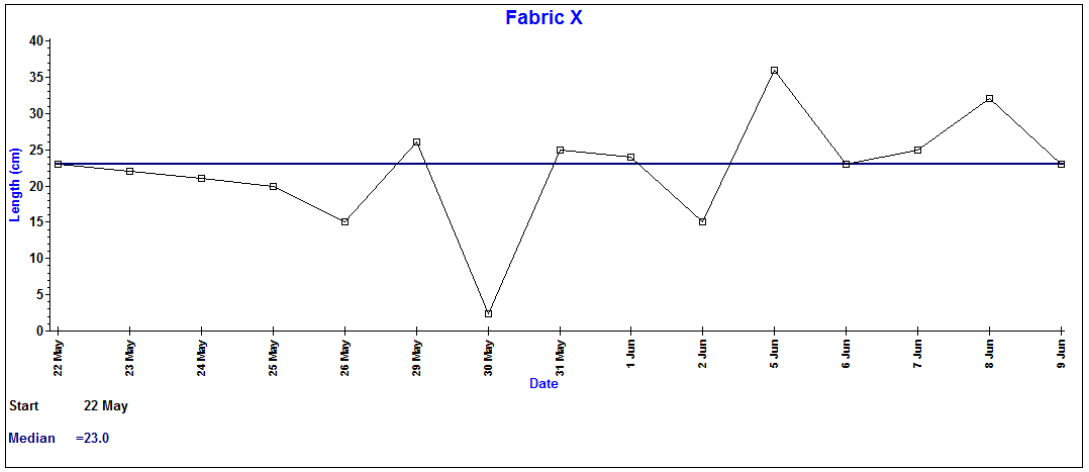
A Note on Outliers

An observation(s) lying some distance away from the bulk of the observations should always be regarded with some suspicion. Naturally, we want to know whether it is a rogue value (an outlier in the true sense) or there by virtue that its parent distribution is skewed.

To classify a "funny" value as being an outlier or there "by right", we need some understanding of the physical process being measured. However, most measurement processes are inherently symmetrical about a central value and so any "funny" values could be regarded as potential outliers. Even so that does not mean that such values should be discarded without some kind of investigation as to their cause. There is often a simple explanation such as incorrectly recording the result, carelessness in performing the measurement, not obeying instructions, sudden changes in environmental conditions, etc. Thus it should be a matter of routine for you to record the order in which the measurements were taken, the technician, environmental conditions and so on.

3.1 Run Charts

The run chart is a graph of the type sometimes called a "time-series", i.e. a graph in which the time is measured along the horizontal axis and a quantity of interest on the vertical axis. Figure 3.1 illustrates a "typical" **run chart**.



Simple though they are, run charts play a crucial role in our study of measurement data. Why? Well, all our attempts at analysing our data with the simple techniques here, or even using more sophisticated ones, are all negated if the data are not stable over time. The run chart provides us with a simple but very effective method of checking for stability over time and we should certainly always use it before progressing on to any other type of analysis of the data.

Activity 3.1 *The specimens used in the experiment had a mass of about 5 gm and so the laboratory decided to use a method based on Archimedes principle. Fortunately, the laboratory recorded the results in the order in which they were taken. The results (in gm/cc) are given below.*

Order:	1	2	3	4	5	6
Density:	1.1566	1.1569	1.1563	1.1566	1.1569	1.1568
Order:	7	8	9	10	11	12
Density:	1.1574	1.1569	1.1569	1.1570	1.1571	1.1576

Construct a run chart and hence comment on the stability of the measurement process. Offer an explanation for what you observe.

3.2 Dot Plots

A dot plot is very simply constructed by plotting the value of each observation on a linear scale (the scale need not be drawn to great accuracy - roughly linear usually suffices). Thus we build up a picture that resembles a histogram but which has far more detail than a histogram since every observation is recorded individually. As an example, the dot plot in Figure 3.2 is for the density data given in the above Activity.

Figure 3.2

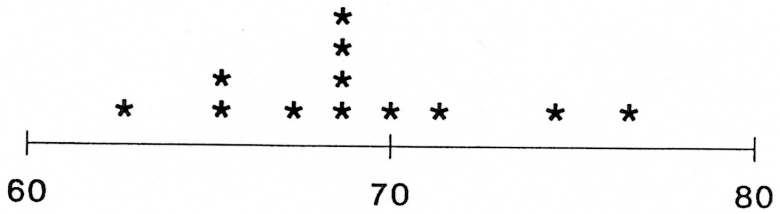


Figure 3.2 reveals a centering around 69 (1.1569 actually) and also a large spread due to the trend we discovered in these data.

Because dot plots are so easy to construct, we should always use them as second nature when we wish to examine data for spread, shape, location and "funny" values.

Activity 3.2 *An important check on the final quality of a product is its vibration characteristics. The following data are vibration measurements of the product from one machine.*

8.1	13.7	25.1	21.1	11.3	19.7	11.9	26.4	26.9	72.2	55.6	30.1
30.4	55.5	53.9	30.4	48.1	38.7	30.7	31.4	32.8			

Construct a dot plot of these data and comment on your findings.

In the next Activity we see the usefulness of dot plot in comparing two data sets.

Activity 3.3 *As in the previous Activity, we are dealing with vibration measurements but this time we are examining samples of products taken from two different machines. Construct a dot plot of the data, using the same axis for each sample but plotting Machine 1 above and Machine 2 below the axis.*

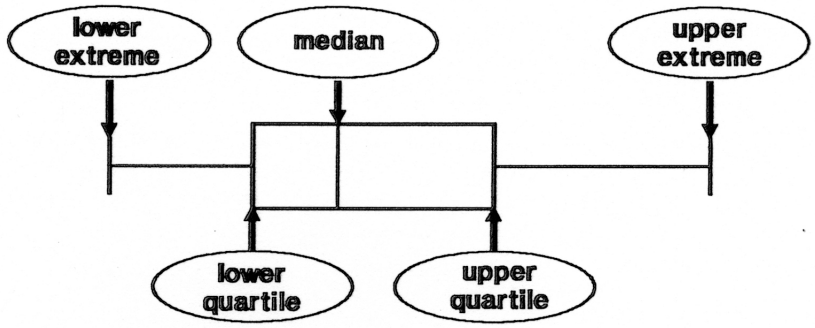
Machine 1:	6. 7	20.5	25.3	30. 7	33.9	50.1	56.8	58.9
Machine 2:	6.9	21.5	25.2	37.8	40.3	42.0	50.0	53.4
	53.6	57.2	57.5	67.0				

This is a particularly interesting case study. There is a difference between the sample means of just over 7 units and the commonly held view by the technicians working in this area was that if there is a difference of more than 7 between the sample means then there is a significant difference between the machines. Use the dot plots to give your opinion on the situation.

3.3 Box-and-Whisker Plots (Boxplots)

These strange sounding devices have become very popular for examining data with respect to their shape, average, spread and outlier values. In particular, we shall find them very useful in comparing two or more data sets (for example, they can assist us in answering questions such as "do these three instruments give consistent results?") without having to delve deeply into the intricacies of topics such as "Analysis of Variance". A boxplot, in its simplest form, is shown in Figure 3.3.

Figure 3.3



Now you can guess why they are called "box-and-whisker" plots!

Unfortunately, the type of boxplot shown in Figure 3.3 masks any outliers that may exist in the sample. We can overcome this by only drawing the whiskers to points that just lie within what are called fences. Fences are imaginary lines at a distance of 1.5 times the quartile spread from the actual quartiles; we then reason that any points outside the fences are sufficiently far removed from the bulk of the data as to be deemed "possible outliers". Figure 3.4(a) shows the fences for the example in Figure 3.3 (with the actual observations outside the box shown by a spot) while Figure 3.4(b) shows the completed boxplot of the type we shall use from now on

Figure 3.4(a)

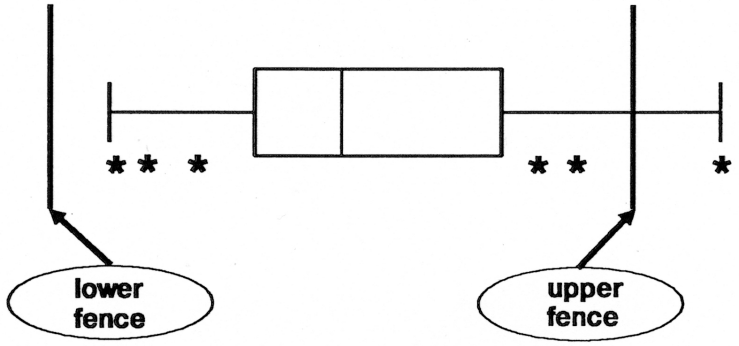
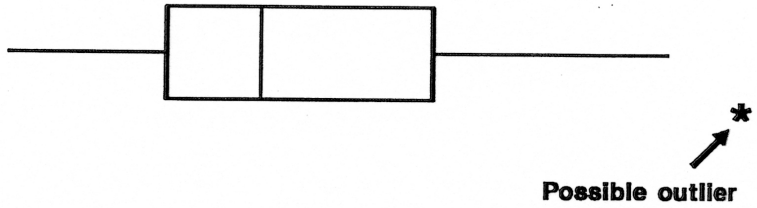


Figure 3.4(b)



Activity 3.4 *Sketch the kind of boxplots you would expect with samples taken from (i) symmetrical, (ii) left skew and (iii) right skew distributions.*

Activity 3.5 *Construct boxplots for the data sets given in Activity 3 and comment on the comparison between the two.*

Activity 3.6 *The two machines referred to in Activity 3 later gave the following sets of results. Use boxplots to compare the two machines.*

Machine 1:	*6. 7*	*20.5*	*25.3*	*30. 7*	*33.9*	*50.1*	*56.8*	*58.9*
Machine 2:	*6.9*	*21.5*	*25.2*	*37.8*	*40.3*	*42.0*	*50.0*	*53.4*
	53.6	*57.2*	*57.5*	*67.0*				

4.0 Probability

In discussing the theory of errors and other problems involving random variables, we need knowledge of some common theoretical distributions that have proven to be good at modelling "real-life". Before looking at these theoretical distributions, we shall use .the distribution of fabric widths to help introduce us to some notions of probability that will be important later and also to some of the underlying concepts of theoretical distributions.

4.1 What information do probabilities give us?

The probabilities of events are numerical assessments of the chance or likelihood of these events occurring. To begin with extreme cases, an impossible event is said to have the probability of 0 of occurring, while an event which is certain to occur is said to have a probability of 1

Usually, the probability of an event occurring lies somewhere between these extremes; a probability close to 0 (say, less than 0.05) indicates there is only a small chance of the event happening and a probability close to 1 (say, greater than 0.95) indicates there is a very strong chance of the event occurring. A probability of ½ indicates an even (or "50-50") chance of the event happening; a probability below ½ indicates that the odds are against the event happening, while a probability above ½ indicates that the event is "odds-on".

A probability of ⅔ can be interpreted as, in the long-run the event happens "2 times out of 3", while a probability of 0.9 indicates that in the long-run the event happens "9 times out of 10".

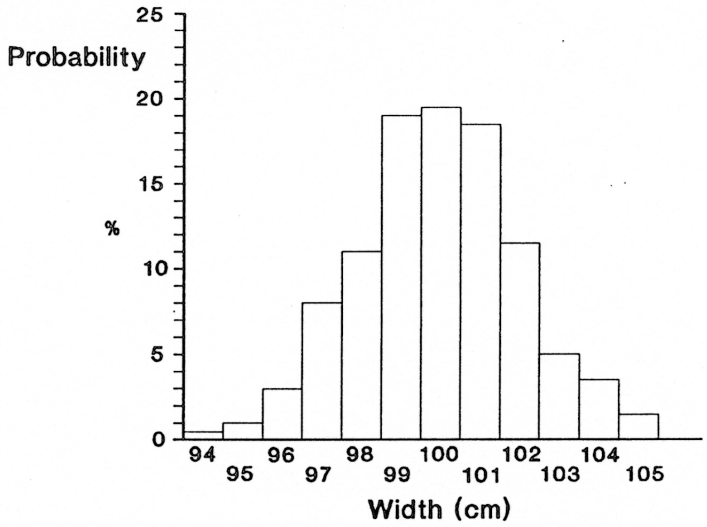
Activity 4.1 *A certain measurement is seen to exceed 4.5 cm on just 3 occasions out of 42. What is the probability of such a measurement exceeding 4.5 cm? What are the underlying assumptions behind your answer?*

With these thoughts in mind, let us now look again at the distribution of fabric measurements.

For the moment, our ideas will be based on the histogram for this population; we reproduce this in Figure 4.1 with the frequency scale replaced by a relative (percentage) frequency scale (which effectively goes from 0 to 1 instead of from 0 to N = 1025), since this is the most helpful scale for our purposes. For convenience, we have also indicated the actual percentage frequency on each block.

What kind of information can be gleaned from this? Well we see that, for example, the proportion of widths over 104.5 cm is 1.2 + 0.1 = 1.3%. We can interpret this result as being the chance that when a measurement is selected at random from the population, it is at least 104.5 cm (obviously, quote a small chance).

Figure 4.1



As a further example, consider the probability of selecting a width between 98.5 cm and 101.5 cm inclusive. From the histogram in Figure 4.1 we see that the proportions for the three blocks covering this range are about 0.19, 0.197 and 0.179 and so the probability of selecting a width in this range is (about) 0.19 + 0.197 + 0.179 = 0.566, which indicates just over a "50-50" chance of having a measurement in this range.

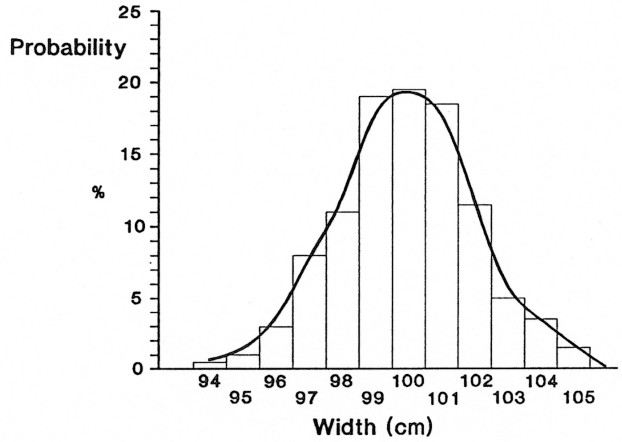
Activity 4.2 *Estimate the probability of a measurement being less than 95.5 cm.*

Now, of course, the sum of the proportions for all the individual classes is 1, and so the histogram in Figure 4.1 is really showing us how the total probability of 1 is distributed over the various measurement groups. This immediately gives us what is known as the probability distribution of the widths. Sometimes it is more convenient to display such a probability distribution as a table:

Width (cm)	**Probability**
93.5 to 94.5	0.004
94.5 to 95.5	0.008
95.5 to 96.5	0.026
96.5 to 97.5	0.076
97.5 to98.5	0.105
98.5 to 99.5	0.190
99.5 to 100.5	0.197
100.5 to 101.5	0.179
101.5 to 102.5	0.115
102.5 to 103.5	0.054
103.5 to 104.5	0.033
104.5 to 105.5	0.012
105.5 to 106.5	0.001
	Sum = 1

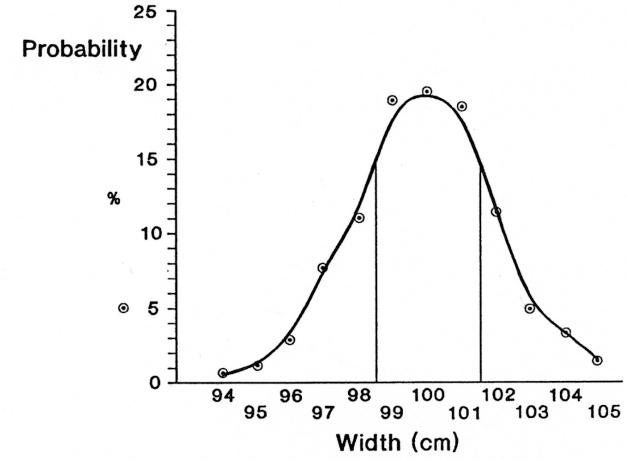
Actually, in this example, we are combining ideas of both discrete and continuous distributions. The width of a fabric is a continuous variable (it is a measurement of length) and so the above distribution is really a discrete approximation to what is essentially an underlying continuous distribution. In the case of a continuous variable, it is usual to have some kind of algebraic description. Clearly, in this instance this is not possible. However, we can supply a graphical indication of the behaviour of the widths when they are viewed as a continuous variable. By a judicious use of freehand on Figure 4.1, we obtain the sketch shown in Figure 4.2.

Figure 4.2



If we use such a curve to represent the probability distribution then the probability of a measurement being between 98.5 and 101.5 cm will now correspond to the area indicated on the sketch in Figure 4.3.

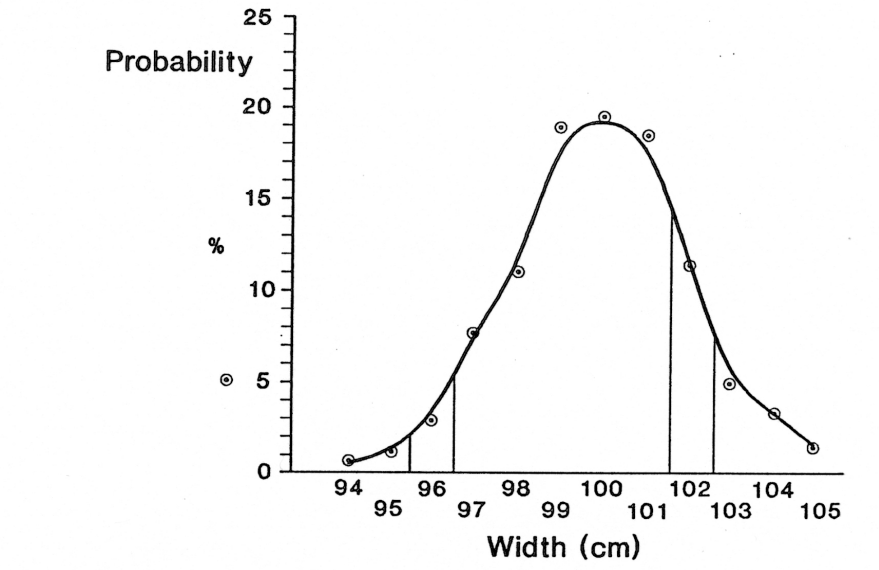
Figure 4.3



This is just a generalisation of what happens in the histogram; if you think about it, you will realise that areas represent probabilities there as well.

Now consider two groups of equal "lengths", say 95.5 - 96.5 cm and 101.5 - 102.5 cm. The probabilities associated with these groups are represented in Figure 4.4.

Figure 4.4



Just by looking at the two areas, we see that the probability of selecting a measurement in the interval 101.5 - 102.5 is of course far greater than for the other group. This feature is often expressed by saying that there is a greater **density** of probability in the second group than in the first. Because a curve such as that in Figure 1.9 indicates this relative density of probability, it is called a ***probability density function*** (or, pdf for short).

If the distribution is derived form theoretical arguments (instead of empirically as here) then we are usually fortunate enough to find an algebraic representation of the pdf; this often allows us to obtain probabilities quite easily and more accurately.

Activity 4.3 *Estimate the probability of a measurement being (i) greater than 103cm, (ii) less than 96.5 cm, and (iii) less than 96 cm.*

Activity 4.4 *Write down that probability distribution that gives the probabilities of the scores 1 to 6 on a "fair" die. (Give the distribution as a list and as a formula.)*

Activity 4.5 *Write down that probability distribution which gives the probabilities of the tossing a "head" and a "tail" on an unbaised coin.*

4.2 Probability Laws

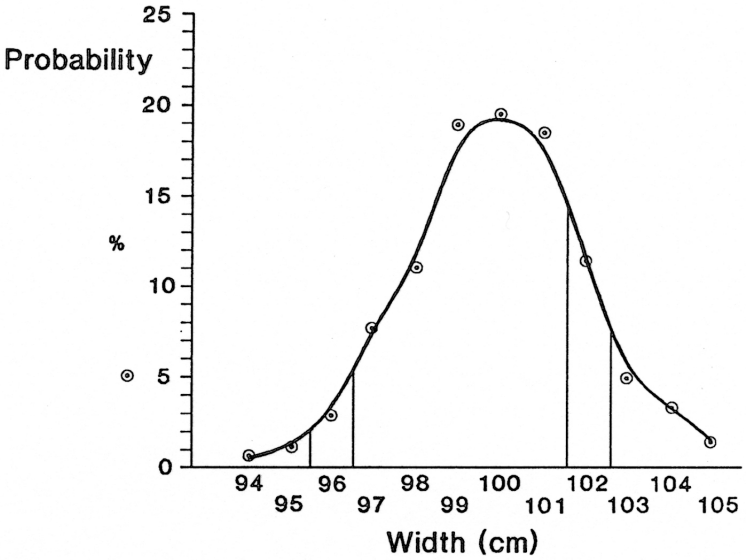
Once we have available the probabilities of some events (either by empirical or theoretical means), there are a number of Probability Laws that can be used to find probabilities of other events in which we might be interested. We shall only mention the few we need for our purposes of in measurement.

First, consider again the two events, and their probabilities, highlighted in Figure 4.5. They are,

Prob (width is between 95.5 and 96.5) = 0.026
and Prob (width is between 101.5 and 192.5)= 0.115

Suppose that we were now interested in the probability of the measured width being in either of these two groups. Since the area representing this probability is obviously obtained by adding the two individual areas in Figure 4.5, we have,

Figure 4.5



Prob {(width is between 95.5 and 96.5cm) **or** (width is between 101.5 and 102.5cm)} =

Prob (width is between 95.5 and 96.5cm) **+** Prob (width is between 101.5 and 102.5cm) = 0.026 + 0.111 = 0.137

Activity 4.6 *Write down the probability of getting (i) a score of 1 or 6, (ii) an even score when a fair die is tossed.*

4.1.1 Addition Law:

This is a direct application of the simple form of the so-called addition law of probability that states that if two events, E_1 and E_2 say, are mutually exclusive (that is, they cannot occur simultaneously, which is indeed the case in our examples so far), then

$$\text{Prob}(E_1 \text{ or } E_2) = \text{Prob}(E_1) + \text{Prob}(E_2)$$

However, if E_1 and E_2 are not mutually exclusive then we need a more general addition law; this is:

$$\text{Prob}(E_1 \text{ or } E_2) = \text{Prob}(E_1) + \text{Prob}(E_2) - \text{Prob}(E_1 \text{ and } E_2)$$

where Prob (E_1 and E_2) is the probability of both E_1 and E_2 occurring.

4.1.2 Multipication Law:

The addition law is one of the two most-used probability laws. The other is the multiplication law. As the name implies, the independence of two events E1 and E2 states that the occurrence or non-occurrence of the either event does not, in any way, affect the probability of E1 (say) occurring and does not depend at all on whether E2 does or does not occur.

When we have two events that are independent in this sense, the multiplication law provides us with a way of calculating the probability of both events occurring:

$$\text{Prob}(E_1 \textbf{ and } E_2) = \text{Prob}(E_1) \times \text{Prob}(E_2)$$

Activity 4.7 *Calculate the probability of getting a "head" when a fair coin is tossed and a score of 6 when a fair die is thrown. Give an intuitive reason for your result.*

If we have more than two events, say, E_1, E_2……. E_n, which are all mutually independent then a similar result is:

$$\text{Prob}(E_1 \textbf{ and } E_2 \textbf{ and } \ldots\ldots\ldots \textbf{ and } E_n) = \text{Prob}(E_1) \times \text{Prob}(E_2) \times .. \times \text{prob}(E_n)$$

To illustrate the multiplication law further, suppose that in the population of fabric measurements, we have information on the colour of the fabric. In particular, suppose that 40% of the fabrics are black, so that the probability of selecting a black fabric is 0.40. Then, making the reasonable

assumption that fabric colour is independent of width, we can calculate the probability of (for example) selecting a measurement that is both between 102.5 and 102.5 cm **and** black by the following:

Prob (width between 101.5 & 102.5 **and** black)
= Prob (width between 101.5 & 102.5) **x** Prob (black)
= 0.5 x 0.40
= 0.046.

- Quite a small chance!

Having sown the seeds of a few of the main concepts of theoretical distributions, we can now progress to a more formal view.

5.0 Random Variables

Now we shall look at some concepts of random variables. We will base the discussion around the familiar theme of dice throwing! Clearly, this is an artificial situation for a metrology course; nonetheless it does have the distinct advantage of enabling us to concentrate on the concept of random variables without clouding the issues with other ideas. Why should we study random variables? Well, as we shall see, they form the basis of all our later work with the theory of errors.

Probably, all of us at some time or other have thrown a die (or dice) in playing some game - Ludo, Monopoly, etc. Now before we throw a die, we have no idea as to which of the scores 1, 2, 3, 4, 5 or 6 will occur; all we know is that, if the die is fair, then each score has an equal probability (1 chance in 6, or ⅙) of occurring. So we have a variable quantity (the score) whose values 1, 2, 3, 4, 5 or 6 occur according to a rule of chance, viz. associated with each score is the probability of ⅙. Other examples of variables whose outcomes depend on some rule of chance are:

- the number of defectives in a sample of, say 20 electronic components,
- the number of machine breakdowns per shift,
- the actual time (as opposed to intended) of arrival of buses at a particular stop,
- and the total score on two dice in a game of Monopoly.

However, we note that the rules of chance for these variables are not easily predetermined as with the score on a die! In each case, we are dealing with variable quantities whose values (often called realisations) occur according to a law of chance, i.e. its probability distribution.

Such variables are called **random variables** to distinguish them from the sort of variables we get in algebra which are determined by definite formula (e.g. y = 2x + 6) rather than by chance behaviour.

Random variables can be classified as being either discrete or continuous. A discrete random variable can take on only certain values usually (but not always) just integers, as with the scores 1, 2, 3, 4, 5 or 6 of a die. On the other hand, a continuous random variable can take any value over its possible range, as with the times of arrivals of buses at a stop. More often, the values of a discrete random variable are obtained by a counting process, while those of a continuous random variable are derived by a measurement process.

It is usual to denote random variables by upper case (capital) letters and their realisations by lower case letters. So, for example, if we were interested in the scores from throwing a die we would perhaps denote "score on die" by D and then write its probability distribution as:

d:	1	2	3	4	5	6
Prob (D = d):	⅙	⅙	⅙	⅙	⅙	⅙

To illustrate these ideas further let us consider the situation which arises in the game of Monopoly: a player throws two dice and then moves the marker the number of squares indicated by the total score of the dice. Let us use T to denote this random variable, i.e. T = total score on the two dice

Observe first that when two dice are thrown, 36 equally likely realisations are possible:

(1, 1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
(2, 1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
(3, 1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
(4, 1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
(5, 1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
(6, 1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

Where, say, the dice are of different colours (blue and red), and the first number in each pair represents the outcome on the blue die and the second number the outcome on the red die. We can then immediately see, for example, that a total score of 2 can be achieved in only one way, by (1, 1) and so Prob (T = 2) = 1 /36.

Similarly there are two ways out of 36 of scoring a total of 3, and so Prob (T = 3) = 2/36. Continuing in this way, we obtain the distribution of T as:

t:	2	3	4	5	6	7	8	9	10	11	12
Prob (T = t):	1/36	2/36	3/36	4/36	5/36	6/36	5/37	4/36	3/36	2/36	1/36

These probabilities should be consistent with the experience of avid players of Monopoly; for example, the most likely score is 7 (but it still only occurs 1 time in 6 on average) while the scores of 2 and 12 each only occur 1 time in 36 in the long run.

5.1 Cumulative Distribution Function (cdf)

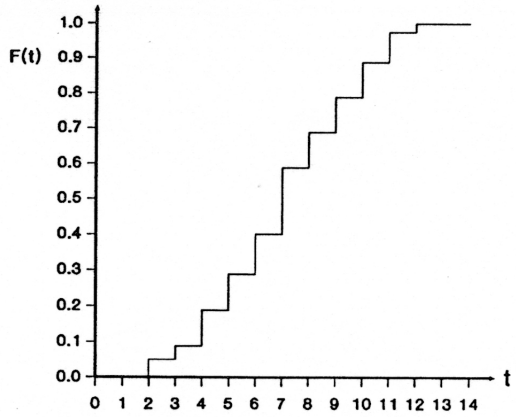
Once the probability distribution of a random variable, which for generality we call x, has been obtained we can derive its *cumulative distribution function* (cdf). This function, which is usually denoted by F(x), gives probabilities of events of the type: $X \leq x$ (i.e. the event that X takes a value of which is less than or equal to x). The cdf can be obtained in a manner similar to that for the distribution of widths, i.e. we simply add successive probabilities. Doing this for the random variable T (the total score from two dice), we obtain:

F (t) = Prob ($T \leq t$)

2/36	$2 \leq t$
3/36	$3 \leq t$
6/36	$4 \leq t$
10/36	$5 \leq t$
15/36	$6 \leq t$
21/36	$7 \leq t$
26/36	$8 \leq t$
30/36	$9 \leq t$
33/36	$10 \leq t$
35/36	$11 \leq t$
1	$t \leq 12$

Figure 5.1 shows a graph of this cdf.

Figure 5.1



Notice that the cdf is specified for all real numbers and not just the individual scores 2, 3, , 12. Thus (if one wished!) we could say that Prob ($T \leq 4.5$) = 6/36 and Prob ($T \leq 20$) = 1; these results are undoubtedly true, though they are hardly likely to be very useful. However, the cdf does help us to find probabilities of more sensible events such as:

Prob ($T \leq 7$) = F(7) = 21/26,

Prob ($T \geq 10$) = 1 - Prob ($T \leq 10$) = 1 - F(7) = 1 - 33/36 = 3/36,

Prob($T \geq 8$) = 1 - Prob($T \leq 7$) = 1-F(7) = 1 - 21/36 = 15/36,

Although our discussion on probabilities has so far been concerned with discrete random variables, a similar analysis can be performed with continuous random variables (though not always as easily!) In the next section we discuss two continuous distributions, both of which are of great importance in metrology.

6.0 Some Standard Theoretical Distributions

6.1 Discrete Distributions

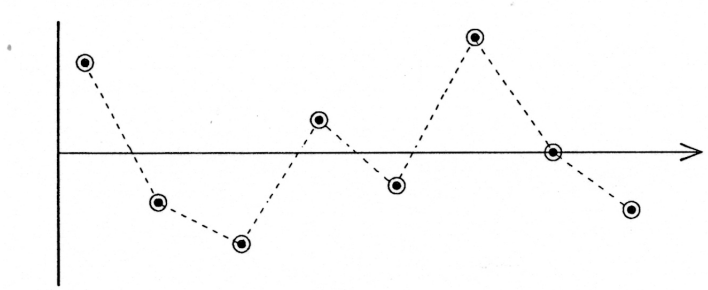
It is fortunate that the behaviour of many unpredictable events in real-life can often be modelled by well-known theoretical distributions whose properties are well documented. However, we should always bear in mind that any model is only an approximation to what is happening in real-life! In this section we briefly look at just one of these standard distributions: the binomial distribution.

6.1.2 The Binomial Distribution

One of the remarkable things about the binomial distribution is the large number of situations to which it can be applied. To quote just two situations we have:

Example 1

Let us consider a run chart where the data are plotted in order to see whether there exists patterns of any kind. Now if the process is "random" then there is a 50-50 chance of each value being above or below the mean line.



Now suppose we have 20 values in total. Assuming that the one value is independent of any other value (i.e.no trends or other “nasties”) the probability of a particular value being above the mean line is 0.5 (and of course there is a probability of 0.5 of a point occurring below the line).

Here we have a situation where each outcome is one of two possibilities (above or below) and, unless there is a trend or some other nasty, one outcome is independent of all the others.

Example 2

It is known that 10% of all males suffer to some extent from colour-blindness. If a random sample of, say, 10 males were selected then, for instance, we could calculate the probability of at least 2 males in the sample being colour-blind.

As in Example 1, the outcome of each trial or inspection is one of two possibilities, namely, colour-blindness or not. Also, whether or not one person is colour-blind is quite independent of the condition of other people in the sample. (This is still effectively true if colour-blindness tends to run in families, since the members of our sample are randomly selected from the population.)

Both of these situations have the common features that are characteristic of ***binomial distributions***. In each case we have:

(i) A number, n, of independent similar trials (n = 20 in the first example and n = 10 in the second)

(ii) The outcome of each trial is one of just two possibilities. In general, let us refer to these two possibilities as success (S) and failure (F). The probability of a success in any one trial is commonly denoted by p and remains the same in every trial; similarly the probability of a failure at any trial is denoted by q. Of course, since success and failure are the only two possible outcomes, it is clear that p + q = 1.

To fix our ideas, consider an experiment in which n = 5 independent trials are conducted. Each outcome will be either a success (with probability p) or a failure (with probability q). Suppose we want to find the probability of getting just 2 successes out of the 5 tries.

It might happen that the first two tries were successes followed by three failures. We can represent this realisation by

S S F F F

Now, since it is assumed that the outcomes of all trials are independent of each other, the probability of this particular arrangement or configuration is obtained by using the multiplication law:

$$p \times p \times q \times q \times q = p^2q^3$$

But, since we want the probability of 2 successes out of 5, irrespective of whether they are the first two, or the first and third, or the second and the fifth etc., we need to take into account every configuration which can give rise to 2 successes and 3 failures. There are in fact ten such arrangements. We display them below with the probabilities computed alongside.

Configuration	**Probability**
S S F F F	$p \times p \times q \times q \times q = p^2q^3$
F S S F F	$q \times p \times p \times q \times q = p^2q^3$
F F S S F	$q \times q \times p \times p \times q = p^2q^3$
F F F S S	$q \times q \times q \times p \times p = p^2q^3$
S F S F F	$p \times q \times p \times p \times q = p^2q^3$
F S F S F	$q \times p \times q \times p \times q = p^2q^3$
F F S F S	$q \times q \times p \times q \times p = p^2q^3$
S F F S F	$p \times q \times q \times p \times q = p^2q^3$
S F F F S	$p \times q \times q \times q \times p = p^2q^3$

So, each such configuration has the same probability, p^2q^3, of occurring. As there are 10 of them, the total probability of having precisely 2 successes out of 5 is 10 p^2q^3.

This, and all similar binomial probability computations are composed of two parts. The p^2q^3 part is easy; it comes straight from the multiplication rule. The "10" part, which counts the number of possible configurations, is not so easy. In practice it would certainly be extremely tedious to have to resort to listing every arrangement to determine this number! Luckily, there is a well-known formula that gives us the number without all that tedious work. To understand the formula, you will need to know what is meant by "factorials". For any positive integer n, "factorial n", which is denoted by n! is computed as

$$n! = n \times (n-1) \times (n-2) \times \ldots \times 3 \times 2 \times 1$$

so, that, for example, 5! = 5 x 4 x 3 x 2 x 1 = 120. Other examples you might like to check are: 2! =2, 3! =6, 4! =24 and 6! =720. An exception to this rule is 0! which is simply taken to have the value of 1.

Now we can state our important result:

The number of configurations of exactly x successes and n-x failures in n trials is given by

$$\frac{n!}{x!\,(n-x)!}$$

This quantity is called a **binomial coefficient**, and is usually denoted by either $\binom{n}{x}$ or ${}^{n}C_{x}$

In our problem above, the number of configurations of 2 (= x) successes and 3 (= n - x) failures in 5 (= n) trials, is computed by the formula as:

$$\frac{5!}{2!3!} \qquad \frac{5x4x3x2x1}{12} = \qquad \frac{120}{12} \; = \; 10$$

which is in agreement with our previous method of listing all possible arrangements

Note that we can now write the probability of getting 2 successes out of 5 as

$$^5C_2\, p^2q^3$$

It is therefore straightforward to see how to generalise the result to any case where we have, say, n trials and want the probability of getting precisely x successes. Writing X for the number of successes in n trials, we have

$$\text{Prob}(X = x) = \text{Prob (x successes out of n trials)}$$

$$= {}^nC_x\, p^x q^{n-x},\ (x = 0,1,2,3, \ldots.. n)$$

This general result gives the probabilities of every possible number x of successes (x = 0, 1,2, ... ,n) in a binomial situation, and is called the ***binomial distribution***.

To further illustrate this important distribution let us consider Example 1. In this case, n = 20 and the probability of a point being above the line (i.e. a "success") is ½ i.e. p = ½ (and so, also q = ½). If we let X denote the total number of successes out of the 20 we find, by substituting p = q = ½ and n = 20 in the above formula that

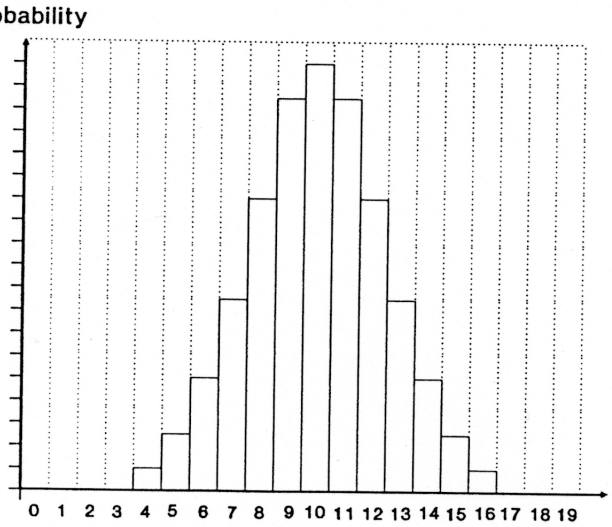
$$\text{Prob}(X = x) = {}^{20}C_x\, ½^x\, ½^{20-x},\ (x = 0, 1, 2 \ldots 20)$$

So, for instance, if we wanted the probability of having just 4 points above the line, we simply put x = 4 in the above formula. This gives:

$$\text{Prob}(X = 4) = {}^{20}C_4\, ½^4\, ½^{16} = 0.0046 \text{ (to 4 dp)}.$$

It is interesting to look at the histogram (shown in Figure 6.1) of this distribution which, because p = q = ½, is symmetrical about x = 10 (the mean number of successes).

Figure 6.1



For later use, we have also give below the actual probabilities associated with this distribution.

x	Prob (X = x)	x	Prob(X = x)	x	Prob(X = x)
0	0.0000	7	0.0739	14	0.0370
1	0.000	8	0.1201	15	0.0148
2	0.0002	9	0.1602	16	0.0046
3	0.0011	10	0.1762	17	0.0011
4	0.0046	11	0.1602	18	0.0002
5	0.0148	12	0.1201	19	0.0000
6	0.0370	13	0.0739	20	0.0000

Even for quite small values of n, calculating the binomial probabilities can be very tedious, particularly probabilities of ranges of values like X ≤ 12. However, for n up to 20 and a wide selection of values of p, binomial probabilities can be read directly from tables. (see “Elementary Statistics Tables” by Henry Neave)

For some purposes we need to know the mean and standard deviation of the number of successes. We state the results without proof (although the result for the mean is fairly obvious):

The mean (or "expected") number of successes is μ = np, and the standard deviation of the number of successes is σ = $\sqrt{npq}$

In the case of Example 1 where n = 20 and p = ½, the expected number of points above the line is μ = np = 20 x ½= 10 (as you surely "expect").

The standard deviation of the number of points is σ = $\sqrt{npq}$ = √(20 x ½ x ½) = √5 = 2.24.

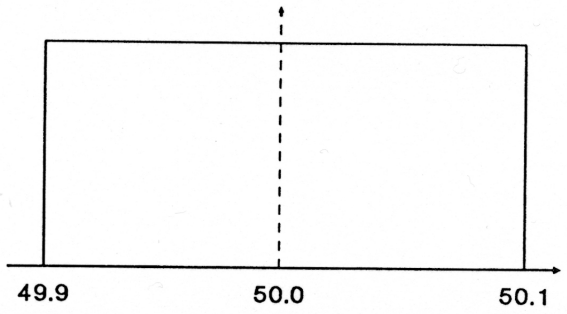
It is interesting to note that this standard deviation can be quite useful in the "3σ-rule". Particularly with symmetrical distributions, it is the case that about 99% of the points will lie within 3 standard deviations of the mean.

7.0 Continuous Distributions

7.1 Uniform Distribution

As an example, the distribution hat best describes an elementary systematic error specified by 50 ± 0.1 is something like the following:

Figure 7.1

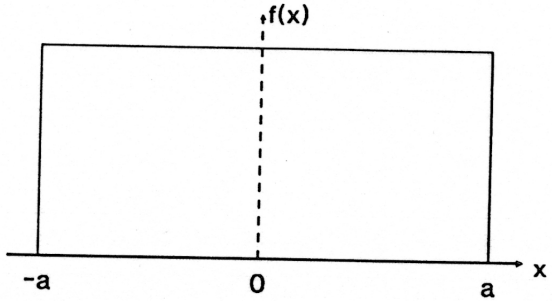


Such a distribution is said to be ***uniform*** over the range (49.9, 50.1) and is such that values have an equal chance of being **anywhere** within this range. (This feature is certainly characterized by the shape of its pdf.)

So, when a manufacturer or calibration laboratory quotes an "accuracy" figure of 50 ± 0.1 they are (often) expressing the view that the true value can be anywhere, with equal chance, within 0.1 of 50.0.

For a uniform distribution over (-a, a) the graph of its pdf is:

Figure 7.2



The formula for the pdf is simply:

$$f(x) = \begin{cases} 1/2a, -a < x < a \\ 0, \text{elsewhere} \end{cases}$$

Obviously, the mean of such a distribution is 0 and it can be shown that the standard deviation is given by $\sigma = \frac{a}{\sqrt{3}}$

The interpretation of this last result is a bit difficult with a distribution where every value is equally likely to occur. Nonetheless, we shall have need of this result when we discuss the problem of combining errors to obtain an "overall" uncertainty of a measurement.

7 .2 Normal Distribution

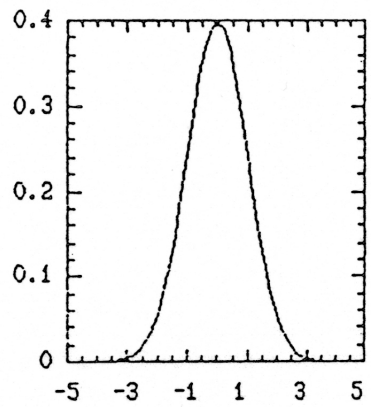
The normal distribution is probably the most important of all continuous probability distributions. But what is it? It is none other than a mathematical model that has been found very useful in describing the behaviour of results of many measuring processes. Typical examples are:

1. The heights of females (or males) in a community.
2. The weight of cans of pet-food.
3. The lengths of mass-produced components.

Also, it is generally assumed that random and repeatability errors in measurements "follow" a normal distribution.

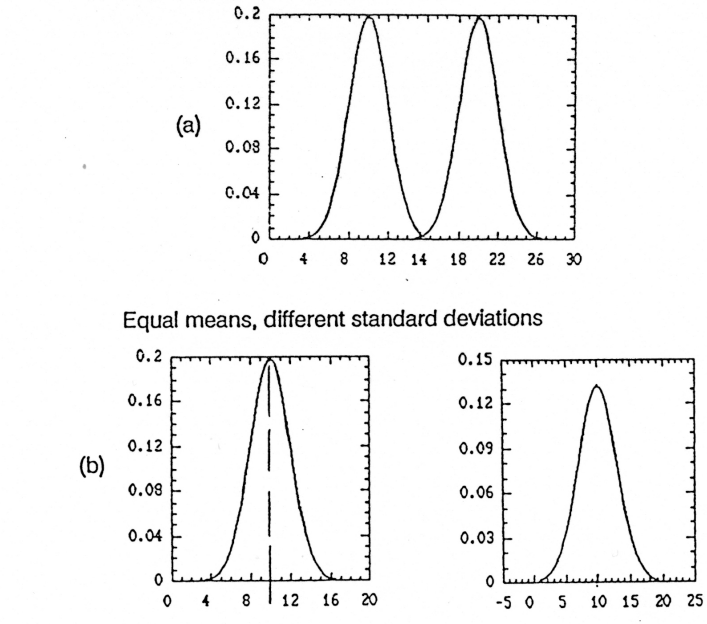
All have histograms that first gradually rise in height to a maximum and then similarly decrease to give an approximately symmetrical shape. The normal curve (or density function) is a continuous curve that behaves in a similar way and has a perfectly symmetrical shape. A typical normal curve is shown in Figure 7:3

Figure 7.3



Such a curve is often described as being "bell-shaped". Whereabouts the curve is centred, and how wide or narrow it is, depend on its values of μ and σ. What would be the effect on the shape of the normal curve if the mean and/or standard deviation were different? Figure 7.4 below illustrates normal distributions for varying means and standard deviations.

Figure 7.4 Equal standard deviations



So we see that the shape of a normal distribution describes very well its characteristics where values are tending to be "targeted" on a central value (the mean μ) giving more chance of being in the middle rather than the tails.

Incidentally, the equation of the normal distribution curve is given by

$$\frac{1}{\sqrt{2\pi\sigma}} e^{-½\left(\frac{x-\mu}{\sigma}\right)2}$$

Where μ is the mean and σ is the standard deviation of the distribution. The equation was discovered by Abraham de Moirve while attempting to devise better strategies for London punters in the eighteenth century!

Let us not forget that since this is only a mathematical model used to describe real-life distributions, we can only speak of “real-life” distributions as being approximately normal (note there are many bell-shaped theoretical distributions which are not normal!).

It is common to refer to a normal distribution whose mean is μ and standard deviation is σ (i.e. variance is σ^2) by the notation N(μ, σ^2). Thus the two distributions in Figure 7.2(a) could be denoted by N(10, 4) and N(20, 4) and those in Figure 7.2(b) by N (10, 4) and N(10, 9) and those in Figure 7.2(b) by N(10, 4} and N(10, 9).

In each case, we see that the basic shape is retained. Changing the mean does not alter the shape of the curve at all; it simply shifts the position. Increasing the standard deviation flattens the shape that is to be expected since the total area of 1 has to be spread out over a greater range of values). On the other hand, a small standard deviation results in the curve being narrow and more pointed.

To illustrate how well the normal curve describes real data we have collected measurements of 200 ½-metre samples of a fabric. Figure 7.5 shows the histogram produced from this sample.

Figure 7.5

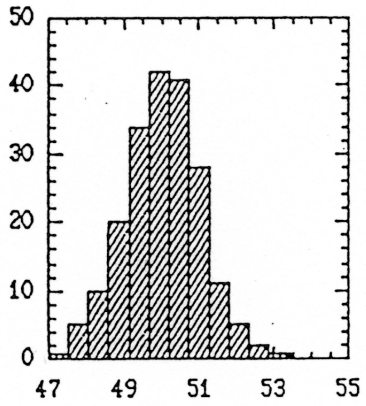
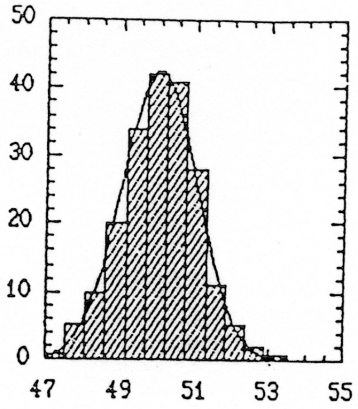


Figure 7.6

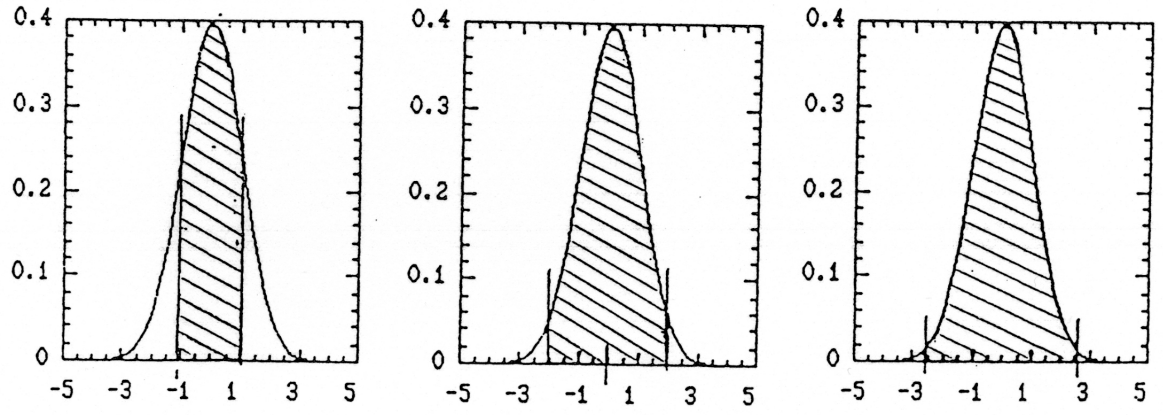


This histogram is reproduced in Figure 7.6, and superimposed on it is the normal curve with the mean and standard deviation the same as those of the sample.

A important feature of normal distributions is that, regardless of their means and standard deviations, results such as the following are always true:

(a) 68% of its area lies within 1 standard deviation of μ,
(b) 95% of its area within 1.96 (i.e. roughly 2) standard deviations of μ,
(c) 99% of its area lies within 2.58 standard deviations of μ.

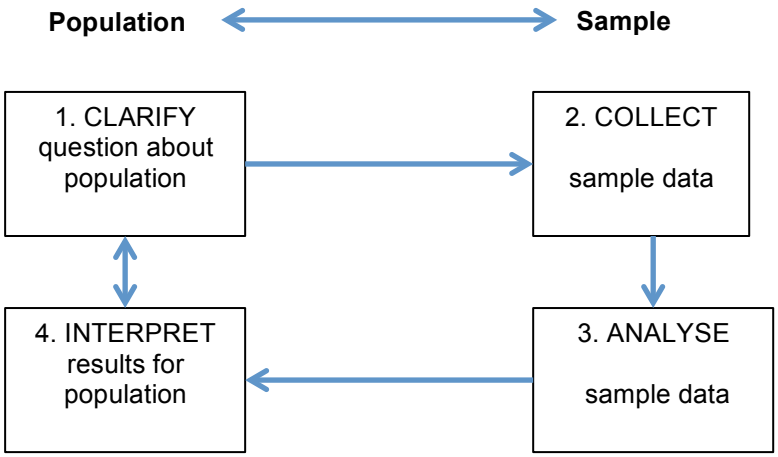
Figure 7.7



A useful interpretation of this feature is that the probability of obtaining an observation more than 2 standard deviations away from the mean is about 5%, more than 2.5 standard deviations away is about 1%. Thus if we observed a value which is more than, say, 2 standard deviations away from the mean then we might classify it as a somewhat unusual event which might cause doubt as to whether the observation even comes from this distribution.

8.0 Hypothesis Tests

Statistical tests (usually called hypothesis tests or significance tests) can often play a major part in statistical investigations. The basic idea of most statistical techniques is to increase our knowledge about populations using information in samples taken from them. The model diagram below illustrates the stage we have reached with our statistical explorations.



In statistical testing, we are concerned with examining the truth, or otherwise, of hypotheses (suppositions, proposals, claims, guesses, etc.) about some feature(s) of one or more populations. For instance, we might find ourselves considering claims such as:

(a) "There are no differences between the population means from four different types of weighing instruments."

(b) "A new digital depth gauge is more effective than the existing instrument."

Whether or not particular hypotheses are eventually assessed to be reasonable will depend on the amount of evidence contained in data from the relevant population(s). Actually, we never claim to completely prove anything by means of a statistical test; we simply pronounce a judgement based on the available evidence, and give an assessment of the strength of that evidence. Of course, sometimes the evidence may be so overwhelming that a hypothesis may be regarded as proved (or disproved) "for all practical purposes".

With statistical testing, we should always pay considerable attention to each of the following matters:

(a) understanding the logical structure of the statistical testing procedure;
(b) choosing the most appropriate hypotheses;

(c) stating the conclusions as accurately and as informatively as possible;
(d) interpreting these conclusions "in English" rather than merely using mathematical or statistical jargon.

To these ends, in the next section we present five main steps which should be used to govern the formulation and execution of any statistical test. The use of these five steps virtually guarantees that the important points listed above are all properly covered. We shall present that section as if we were designing a test "from scratch".

9.1 The Concepts Of Statistical Testing

Rather than presenting the underlying ideas of statistical testing "in the abstract", it seems more sensible to discuss them with reference to a specific problem. So, consider the following situation.

> On a control chart, we usually expect the number of points above and below the overall mean line to be about equal. More points above/below the line than expected would cause us to suspect there had been a downward/upward shift in the mean of the process. We shall investigate whether there is evidence of such a shift (up or down) on a control chart containing 20points.

Let us now describe our five steps in terms of this experiment.

STEP 1 Formulation Of The Practical Problem In Terms Of Hypotheses

We begin by attempting to clarify exactly what is meant by the word "hypothesis". According to Concise Oxford Dictionary, a hypothesis is "a supposition made as basis for reasoning, without reference to its truth, or as a starting-point for investigation" and then, maybe rather pessimistically, as a "groundless assumption". The dictionary's definition is certainly quite appropriate for one of the hypotheses we use when setting up a test; this is the so-called null hypothesis, which is usually denoted by H_0.

This null hypothesis usually represents the proposition that there is no difference between two or more populations (at least in respect of the feature(s) in which we are interested); or, when we are studying just one population, it might represent maintenance of the "status quo", i.e. no difference now from what has been the case before, or no difference from the manufacturer's specifications, and so on. However, as we shall see, any test always involves (at least) two hypotheses, namely

(a) the null hypothesis, H_0 and
(b) an alternative hypothesis, H_1

A positive diagnosis will indicate that we should perhaps take action of some kind (for example. to adopt a new instrument if the test indicates it to be better than the existing one, or to recommend a new process if it is judged better than the rest), or at least to take steps to investigate the situation further.

Although of less practical interest, the null hypothesis plays a fundamental role in the formal details of the test procedure. Remember that, in situations where we have two or more samples, the null hypothesis usually states that there is no difference between the populations being

compared; while in a situation involving only one population it expresses something like the maintenance of the status quo. Thus H_0 is basically a standard or control against which the strength of evidence in favour of the type of difference described by H_1 can be measured.

The hypotheses appropriate for our problem can now be immediately identified as:

H_0 : There is an equal number of points above as below the overall mean line.
H_1 : There is an unequal number of points above as below the line.

Once the hypotheses have been selected, we can proceed to the crux of the formal argument in statistical testing. This may be stated in the following terms:

At the outset, we tentatively assume the truth of H_0 (whether or not we actually believe it!). Then, with this assumption, if the data lead to a result that is relatively unlikely and would be much better explained if H_1 were true, then we have reason to reject H_0 in favour of H_1 by what is essentially an "argument by contradiction".

It is this argument that we pursue in Steps 3 to 5.

STEP 2 Choice Of The Test Statistic

We ask: How can the data be used in some reasonable and convenient way to test the hypotheses?

The first point to make is that, in general, it would be too complicated and cumbersome to use each piece of available data explicitly in the test procedure. Instead, what we usually do is to carry out some calculation on the data so as to summarise them in terms of a single representative figure. Any such quantity calculated from the data is called a test statistic and, for the time being, we shall denote our chosen statistic by T. Obviously we want T to relate to those aspects of the data that are relevant to the particular problem we are trying to solve. There are two important properties that T must satisfy in order to be of use in testing the hypotheses:

1. Since T is to be used to indicate the weight of evidence pointing to the truth of H_1 as opposed to H_0, then T must behave differently when H_1 is true from when H_0 is true; and, in general, the greater the difference in the real situation from that expressed by H_0, the greater should be the difference in the behaviour of T.

2. The probability distribution of T must be calculable (at least approximately) under the assumption that H_0 is true. This distribution is called the null distribution and provides us with the guide as to what are or are not reasonable values of T under H_0 (i.e. whether or not the value of T can be regarded as outrageous!).

So how can we devise an appropriate test statistic T for our problem? Let us consider the following simple procedure:

Suppose we denote by "+" or "-" whether the point is above or below the overall mean line. Let us define T as the total number of "+ "s, i.e. T is the number of points above the line.

Is this test statistic T appropriate for the job?

The answer is clearly "yes". It certainly satisfies the former of the above two criteria; for if there has been a shift in the process mean (i.e. H_1 is true) then we can expect T to have a high or a low value, whereas if the process mean has not shifted (i.e. H_0 is true) we would expect the numbers of "+ "s and "-"s to be fairly similar, leading to a more moderate value of T. Thus the behaviour of T depends on whether H_0 or H_1 is true, as required.

The second criterion is also satisfied since, if H_0 is true, whether a point is above or below the line will be rather like tossing a coin, i.e. a "+" or a "-" will be produced with a probability of each; consequently the null distribution of T is binomial with n=20 and p= ½.
Once we have discovered the null distribution of T (i.e. its probability distribution when Ho is true) we can discriminate between what are and are not reasonable values of T under Ho.

STEP 3 The Critical Region And How It Is Used

Once the test statistic T has been selected we ask:

"What kind of values of Twill most strongly point to H_1 being true rather than H_0 being true?"

Such a collection of values of T is called a critical region. Then, if we find that the value of T calculated from our data lies in such a critical region, our conclusion is that we may "reject Ho in favour of H_1"; when this happens, the value of T is said to be significant. But if T does not lie in the critical region then we do not have sufficient evidence to reject Ho in favour of H_1 and the value of T is then said to be not significant.

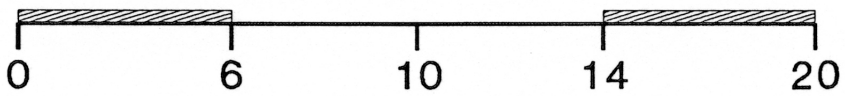
Notice that we avoid talking about "accepting H_0" in this latter case. The reason is that there is invariably a wide range of possible situations under which any particular value of T can be regarded as reasonable and, since we cannot discriminate between such situations, then there is no reason why we should believe (i.e. "accept") H_0 rather than any of the other (possibly countless) possibilities under which our value of T may be regarded as reasonable. For example, suppose an investigation involves the hypotheses

H_0: Mean length = 60 cm

H_1: Mean length > 60 cm.

To speak of accepting H_0 in this case seems to imply that we believe the average length is definitely 60 cm and not 59.7 cm or 60.1 cm or any other possible value for which the same value of T is still unsurprising.

In our problem, remember that if H_1 is true then we expect either a high value of T or a low value of T; otherwise if H_0 is true we expect similar numbers of "+ "s and "-"s thereby giving a moderate value of T. So we should reject H_0 in favour of H_1 only if T is large or if T is small. Now under H_0, we know that the "expected" value of T is 10 and that its distribution is symmetrical (being binomial with p = ½ and n = 20). So we need a critical region of the form "distance between T and 10 ≥ c, i.e. |T - 10 ≥ c where c is some appropriately chosen critical value. For example, if we choose c = 4, the critical region would contain all possible values of T satisfying T ≤ 6 or T ≥ 14.



The selection of a suitable critical value from these possibilities is the subject of STEP 4.

STEP 4 Deciding On A Significance Level And The Critical Value

We have now decided on the type of critical region appropriate for our test, but how do we decide on precisely how large the critical region should be? To help us with this, we need to consider what are the acceptable levels of risk of the test giving incorrect conclusions from the data; for it is a sad fact of life that, however we decide on the criteria for rejecting or not rejecting H_0, incorrect conclusions will always be possible.

If, for example, the critical region is relatively large, then H_0 may often be rejected when it should not be (an injustice!). On the other hand if, in order to avoid this type of error, we make the critical region very small, then we will find ourselves frequently failing to reject H_0 when we should be doing so (another injustice!). In practice, the decision on whether or not to reject H_0 depends on what is known as the significance level (or the size) of the test. This is defined as the risk we are prepared to take of committing the error of rejecting H_0 when it is in fact true. (This kind of risk is sometimes referred to as a Type I error. A Type II error refers to the other risk i.e. of not rejecting H_0 when it is in fact false.) In other words, it is the probability of obtaining a significant value of T when Ho is true.

The larger or smaller the significance level, the larger or smaller will be the critical region.

We use significance levels of between about 0.1% and 10%, depending on the practical seriousness of making this kind of error. In fact, in our decision rules we shall usually use a 5% significance level as this level represents what we regard as the beginning of "rare events".

So for our problem we decide on a significance level of about 5%. We say "about 5%" since it is usually impossible with a discrete null distribution to obtain critical regions which correspond exactly to a given nominal probability such as 5%. So, for this approximate significance level, what is the appropriate critical region? Now, as we have already seen, the null distribution of the test statistic T is binomial with n = 20 and p = ½, and we seek a value c such that Prob ($|T - 10| \leq c$) = 0.05 when H_0 is true.

Activity *Use the binomial distribution for n = 20, p = ½ to work out the probability of having:*
(a) 0, 1, 19 or 20,
(b) 0, 1,2, 18, 19 or 20,
(c) 0, 1,2,3, 17, 18, 19 or 20 "+ "s.

What is the most appropriate critical region for a 5% significance level for this test?

Thus, we may state the test's decision rule as follows:

"If $T \leq 5$ or $T \geq 15$ we shall reject H_0 in favour of H_1 at a significance level of about 5%, otherwise we shall not reject H_0".

In other words, we are saying that such values of T are so much more likely under H_1 than under H_0 that we claim positive support for the truth of H_1. Otherwise, if $6 \leq T \leq 5$ 14, we cannot reject H_0 in favour of H_1, for then the value of T is "middling" and so entirely reasonable under H_0.

Calculation of T.
The results from the control chart were:

+ - + + + - + + + + + + + - + + + + + +

giving T = 17.

STEP 5 Conclusion

Referring to the above decision rule, we see that there is sufficient evidence (at the 5% level of significance) to reject the null hypothesis. (In fact, it seems there has been a shift upwards!)

We could leave the issue there. However, a conclusion like this, which is based merely on whether we are or are not able to reject H_0 in favour of H_1 at some prescribed significance level (5% in this case) can be very wasteful of the information in the data.

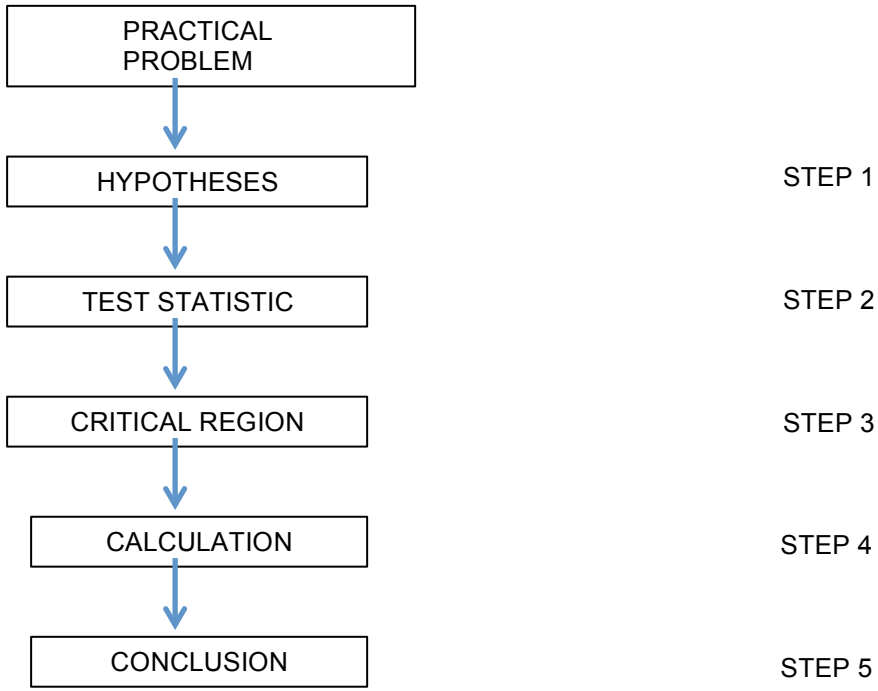
Suppose the investigation had produced the even more extreme result that all 20 signs were "+ "s or were all "-"s. Then, of course, again we would have rejected H_0 in favour of H_1 at the 5% significance level. But this conclusion would also have been reached had there been 19, 18, 17, 16, 15,5, 4, 3, 2, 1 or o "+ "s! In practice, a much more informative conclusion maybe reached by considering various significance levels and determining how big the critical region needs to be in order for the observed value of T to just fall into it. How accurately one can do this depends of course on how much detail is available on the null distribution - we do not always have the complete information available as we do in our current example.

In practice though, no great accuracy is necessary; all that is really needed is a rough assessment of the weight of evidence supporting H_1. Even a fairly approximate idea of the "borderline" significance level leads to much more useful and meaningful interpretations. Typical conclusions could be that there is:

Conclusion	**Significance Level**
very little, if any, evidence	exceeding 10%
some suspicion	between 5% and 10%
considerable evidence	between 1% and 5%
very strong evidence	between 0.1% and 1%
practically conclusive evidence	less than 0.1%

Summary

The above ideas on statistical testing are conveniently summarised in the following flowchart.



Incidentally, the test we have devised for the control chart problem is a standard procedure known as the "sign-test".

8.1 One-Sided and Two-Sided Tests

We have already seen the fundamental role played by the alternative hypothesis H_1 in determining the critical region; for, if you recall, the critical region represents the types of samples that, if they occur, would most strongly influence us to believe in H_1 rather than H_0. In the above problem, this led to critical regions of the form $|T - 10| \geq c$. We shall now see that H_1 is responsible for classifying a test as being either one-sided or two-sided.

In our problem it seemed that the most appropriate alternative hypothesis for the situation was

> H_1: There has been an upward/downward shift.

But suppose that we were interested in finding out whether there had been an upwards shift in the mean of the process. The appropriate alternative hypothesis would be:

> H_1: There has been an upward shift.

Now, unlike the original alternative, this H_1 only allows for the possibility of an upward shift (a large number of "+ "s). Because of the "one-sidedness" of this H_1, the statistical test for this situation is referred to as a one-sided test, as opposed to the test based on the original "two-sided" H_1 that is called a two-sided test.

The critical region for such a one-sided test is quite different from that for our original two-sided test.

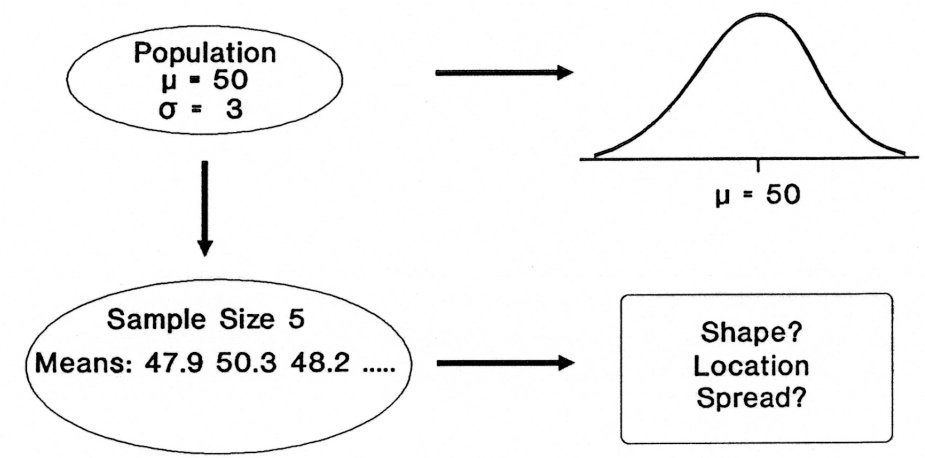
Activity *Obtain an appropriate 5% critical region for the one-sided test specified by the above H_1. Hence write down the decision rule for this test.*

The question of which form of alternative hypothesis is relevant for the particular task on hand, and so whether a one-sided or two-sided test should be used, should always be decided before the data are inspected (or preferably even collected). The deciding factor is whether one is concerned in practice with diagnosing a difference in either direction or in just one direction, should H_0 be untrue; and this is ruled out by the practical circumstances of the problem being investigated. For instance, a company might be interested in discovering whether one of its products differs in quality from its rivals' (giving a two-sided test); or on the other hand, it might only be interested in testing whether its product is inferior (giving a one-sided test) since evidence of this might necessitate action involving redesign, price-cutting, etc. A one-sided test is more sensitive (more likely to reject H_0) if there is actually a difference in the prescribed direction; but this is at the expense of having no sensitivity at all for recognising differences in the opposite direction.

The choice between one- and two-sided tests is usually easier to make when dealing with actual practical problems (when the details of the problem are well understood) than with some textbook problems (when such details may be lacking).

9.0 The Distribution Of Sample Means

What is meant by "The Distribution of Sample Means"? All we are seeking is information on how the sample means behave - is their spread the same as their parent population? Is the shape of their distribution the same as the parent population? Is their centre (mean) the same as the parent population?



BUT WAIT! Why do we want such information?

Well, it's certainty not for idle curiosity! If you reflect for a moment you will realise that most of your previous (and future!) work in statistics deals with sample means - for estimating a population mean, for summarizing data etc. So really, it is most necessary for us to know more about the behaviour of sample means if we are to progress further.

To explore the characteristics of the sampling distribution of means we shall first consider an artificial population and then afterwards look at what happens in a real-life population.

Consider the small population consisting of just the numbers 3, 4, 5, 6 and 7. The shape of its distribution is shown below in Figure 10.1.

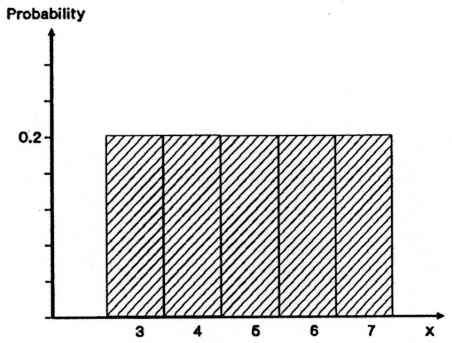


Figure 10.1

Activity 10.1 *For this distribution, show that μ = 5 and σ^2 = 2.*

NOTE: We shall use the variance as our measure of spread (it is actually more convenient, just for the moment, than the standard deviation).

Now that we have the facts about this population distribution let us take a look at the sampling distribution of means for a sample size n = 2. We sample in such a manner as to allow repeats ("sampling with replacement"); this gives 25 samples in all:

(3,3) (3,4) (3,5) (3,6) (3,7)
(4,3) (4,4) (4,5) (4,6) (4,7)
(5,3) (5,4) (5,5) (5,6) (5,7)
(6,3) (6,4) (6,5) (6,6) (6,7)
(7,3) (7,4) (7,5) (7,6) (7,7)

The corresponding 25 sample means are:

3.0, 3.5, 4.0, 4.5, 5.0
3.5, 4.0, 4.5, 5.0, 5.5
4.0, 4.5, 5.0, 5.5, 6.0
4.5, 5.0, 5.5, 6.0, 6.5
5.0, 5.5, 6.0, 6.5, 7.0

Now, the mean of these sample means is 5.0 – the same value as the mean of the population!

How does the variance of these sample means compare with the variance ($\sigma^2 = 2$) of the population. Well, calculating the variance of the sample means produces a value of 0.4 or 2/5 (i.e. $\sigma^2/n = 2/5$).

Activity 10.2 *Sketch the histogram of the distribution of sample means and compare it with that for the population. Give an intuitive reason for the spread of the means being less than that of the population.*

We have illustrated the following features about the distribution of sample means:

1. The means have less spread than the parent population. In fact, the variance of means = σ^2/n [and thus the standard deviation of the means = $\sigma/\sqrt{n}$].

2. The centre (i.e. mean) of the sample means equals the population mean.

3. The shape of the distribution of sample means is symmetrical and "bell-shaped".

9.1 The Central Limit Theorem

The so-called Central Limit Theorem gives us insight into another aspect the behaviour of the sampling distribution of means.

Summary of the Central Limit theorem:

If the sample size n is "big enough" then the distribution of the sample means is approximately normal, irrespective of the shape of the parent population. The mean and standard deviation of this distribution are respectively μ and $\sigma/\sqrt{n}$, where μ and σ are the mean and the standard deviation of the parent population.

Summary

If we want to estimate the population we should use the sample mean $\bar{x}$. The precision of this estimate is measured by the standard deviation (often called the standard error) of the means as given by $s/\sqrt{n}$.

Furthermore, if the sample size n is "big enough" (often as small as 5 is fine!) then the sampling distribution of means is approximately normal.

10.0 Confidence Intervals for the Population Mean

The best estimator of the population mean is the sample mean $\bar{x}$. The implication of this for us is, for example, that if we wanted to estimate the true value of some physical quantity then we should use the mean of a sample of observations.

What if two or three people each took a random sample of, say 10, values from the same population and each calculated the sample mean, $\bar{x}$. No doubt these sample means would be quite different from each other and *yet they are all estimates of μ*. So, who is right? The answer is simply "all of them"! You now realise the limitations of using just the sample mean to estimate μ. What would be much more preferable would be to obtain an interval which expresses the thought "we do not know the true value of μ but we are (say) 95% confident that it lies "between …….. and ……..." i.e. we really need is a so-called 95% confidence interval.

To obtain such an interval we simply make use of our knowledge of the behaviour of the sample means:

(a) the standard deviation of the sample means is $\sigma/\sqrt{n}$

and

(b) if n is large then the distribution of the sample means is approximately normal.

For the moment we will use the true value of a for our population, though of course in real-life this quantity would have to be estimated; we shall deal with this problem shortly when we have illustrated the principles with the "simple" case.

Now we recall that with a normal distribution, 95% of all the values lie within 1.96 standard deviations (= $\sigma/\sqrt{n}$, here) of the mean μ. Thus it seems reasonable to claim that a 95% confidence interval for is:

$\bar{x}$ - 1.96 $\sigma/\sqrt{n}$ to $\bar{x}$ + 1.96 $\sigma/\sqrt{n}$ i.e. we can be 95% confident that this interval contains μ.

Activity *What value would we use in place of 1.96 in calculating a 99% confidence interval for μ?*

We now ask "How can we calculate a confidence interval for μ when the population standard deviation σ is unknown? Unfortunately, it is not just a case of replacing σ in the calculations by its estimate s. At the turn of 20th century W. S. Gossett, writing under the pen-name of Student, discovered that the quantity

$$t = (\bar{x} - \mu)/(s/\sqrt{n})$$

does not follow a normal distribution as does the quantity $Z = \frac{(\bar{x}-\mu)}{\left(\frac{\sigma}{\sqrt{n}}\right)}$.

In fact, Gosset demonstrated that if you repeatedly took samples of size n from a normal distribution then the quantity t has a distribution that is symmetrical but with tails "fatter" than those of a normal distribution. Also, the precise shape of this so-called t-distribution depended on a quantity called degrees of freedom ν, which in this case is equal to n-1. Figure 11.1 illustrates the essential differences between a normal distribution and a t-distribution while Figure 11.2 shows several t-distributions for various numbers of degrees of freedom.

Figure 11.1

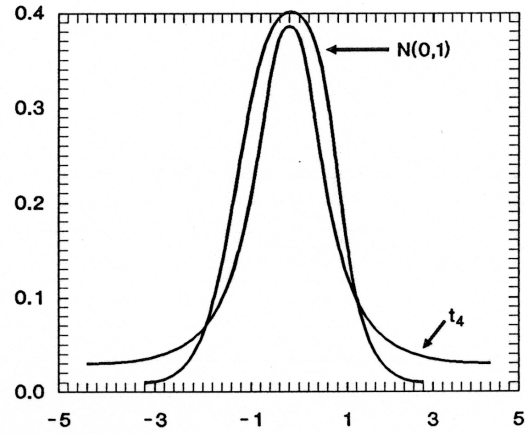
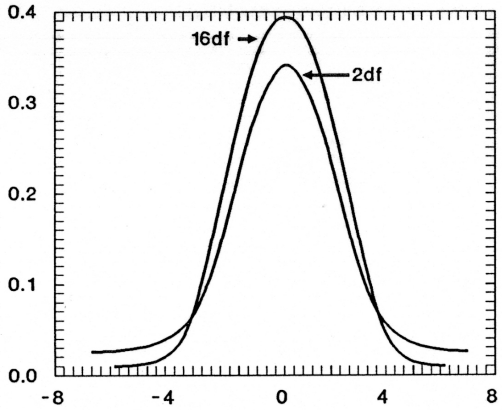


Figure 11.2



In statistical tables such as Neave's "Elementary Statistics Tables", t-tables list the appropriate critical values for the number of degrees of freedom ranging from 1 to ∞ (note: it can be shown that the t-distribution approaches the normal distribution as $\nu \rightarrow \infty$). Thus for example, in the above Activities we have n = 9 giving ν = n-1 = 8 and so a critical value for a 95% confidence interval is 2.306 (compared with 1.96 when we could use the true value of σ).

10.1 Confidence Intervals for the Population Variance

Obtaining a confidence interval for the population variance σ^2 is just as important (and sometimes more so!) as calculating an interval for the population mean μ.

As with the mean, we base the confidence interval for σ^2 on the equivalent sample statistic s^2; however (unfortunately) the distribution of s^2 is not normal but is chi-squared with n-1 degrees of freedom. We thus produce an interval that is not symmetric about the sample value as was the case with the mean. So, how can we construct a γ = 95% confidence interval for σ^2 ? Well, if C_l and C_u are the lower and upper critical values (corresponding to γ = 95%) from a chi-squared distribution with n-1 degrees of freedom then the required interval is given by:

$$\frac{(n-1)s^2}{C_u}, \qquad \frac{(n-1)s^2}{C_l}$$

Example: An important check on the final quality of a product is its vibration characteristics. The following data are vibration measurements of the product from one machine.

8.1 13. 7 25.1 11.1 11.3 19.7 11.9 26.4 26.9 72.2 55.6 30.1 30.4 55.5 53.9 30.4 48.1
38.7 30.7 31.4 32.8

Here the relevant sample statistics are: s^2 = 281. 7 and n = 21 giving 20 degrees of freedom. Now for a 95% confidence interval the critical values are given in the chi-squared tables as C_l = 9.591 and C_u= 34.170. Thus a few moments' work with the calculator gives the desired confidence for σ^2 as:

(164.9, 587.4)

and, by square rooting these limits we obtain a 95% confidence interval for the population standard deviation, σ, of (12.8, 24.2).

11.0 Correlation

The situation that confronts us here is one of deciding whether two random variables are related (i.e. correlated) in the sense of both variables increasing (or decreasing) together (positive correlation) or one variable increasing while the other correspondingly decreases (negative correlation). Furthermore, if there is evidence of such correlation, what is its strength?

We have a double motive for studying correlation in a metrology course. First, we shall shortly be discussing the problem of obtaining an overall uncertainty when two or more uncertainties are combined; just how we go about this will depends on whether or not the component uncertainties are independent or are correlated. Second, a variable representing an important product characteristic is often difficult or expensive to measure; in which case we look for another related variable that is easier or cheaper to measure and which will give us equivalent information about the process as the original variable. In fact, the example we have chosen to develop the ideas of correlation is just such a case.

The measure of correlation we will evolve is Pearson's linear correlation coefficient (or the product moment correlation coefficient, as it is sometimes called). In deriving this coefficient we shall encounter the notion of the covariance between random variables. Although this is itself a (crude) measure of association it is more important to us as the "correction factor" when determining an overall uncertainty from the sum (or difference) of two or more dependent uncertainties.

The example we now consider comes from a study of an extruder. Currently, the process is monitored by weighing an ½-metre sample of the product. During a production run some six or seven samples are taken at various intervals and, naturally, the taking of each sample interrupts the process. If some other variable, related in some way to the weight, could be found which could be equally as good (or better since few people have faith in the quality of the weight measurements!) in monitoring the process without interrupting it, then clearly much cost would be saved. After some preliminary investigations involving several possible variables, an experiment was conducted into the relationship between the operating pressure of the extruder and the weight of the ½-metre sample. The results of this experiment are given below, where the weights are expressed in deviations from some target value.

Pressure (X):	226	211	221	222	215	217	208
Weight (kg) (Y):	4	-11	-1	7	-3	-12	-21

Whenever we wish to see whether there is a relationship between two variables we always construct a scatter graph; there really is no other better way for assessing the situation. Figure 12.1 below shows the scatter graph for the above data.

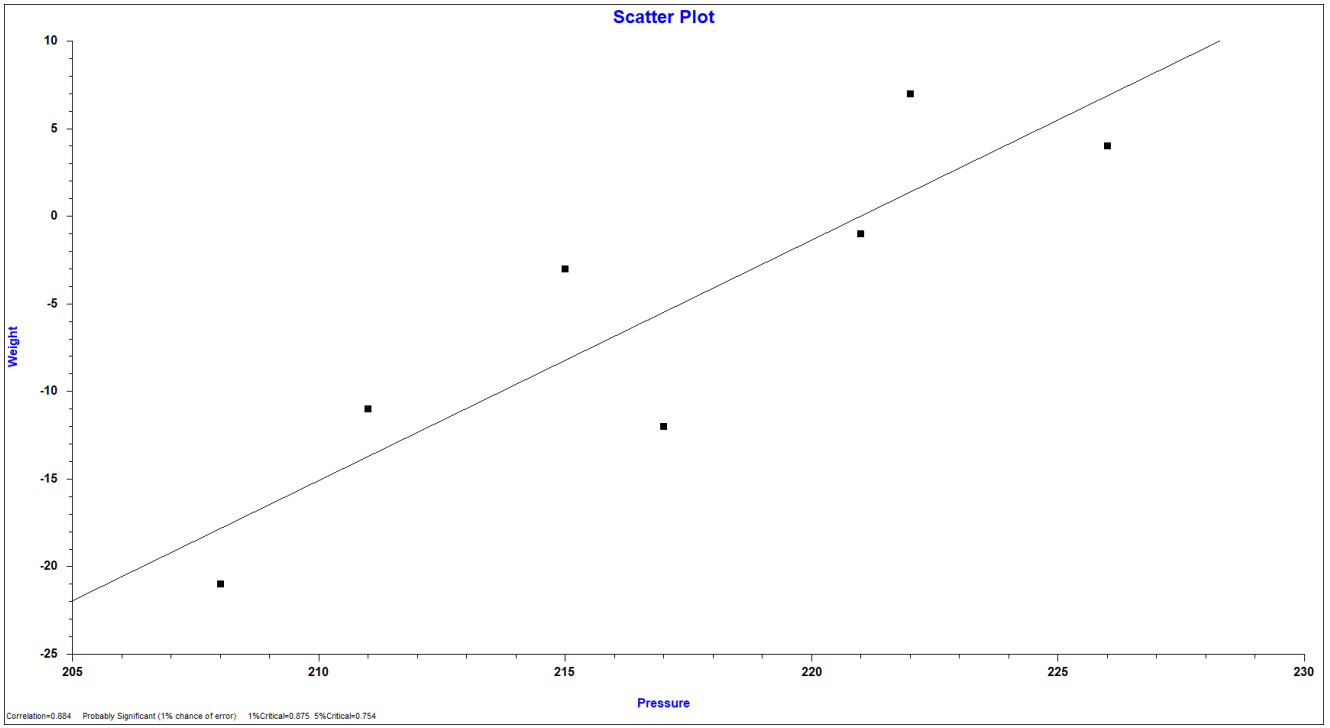
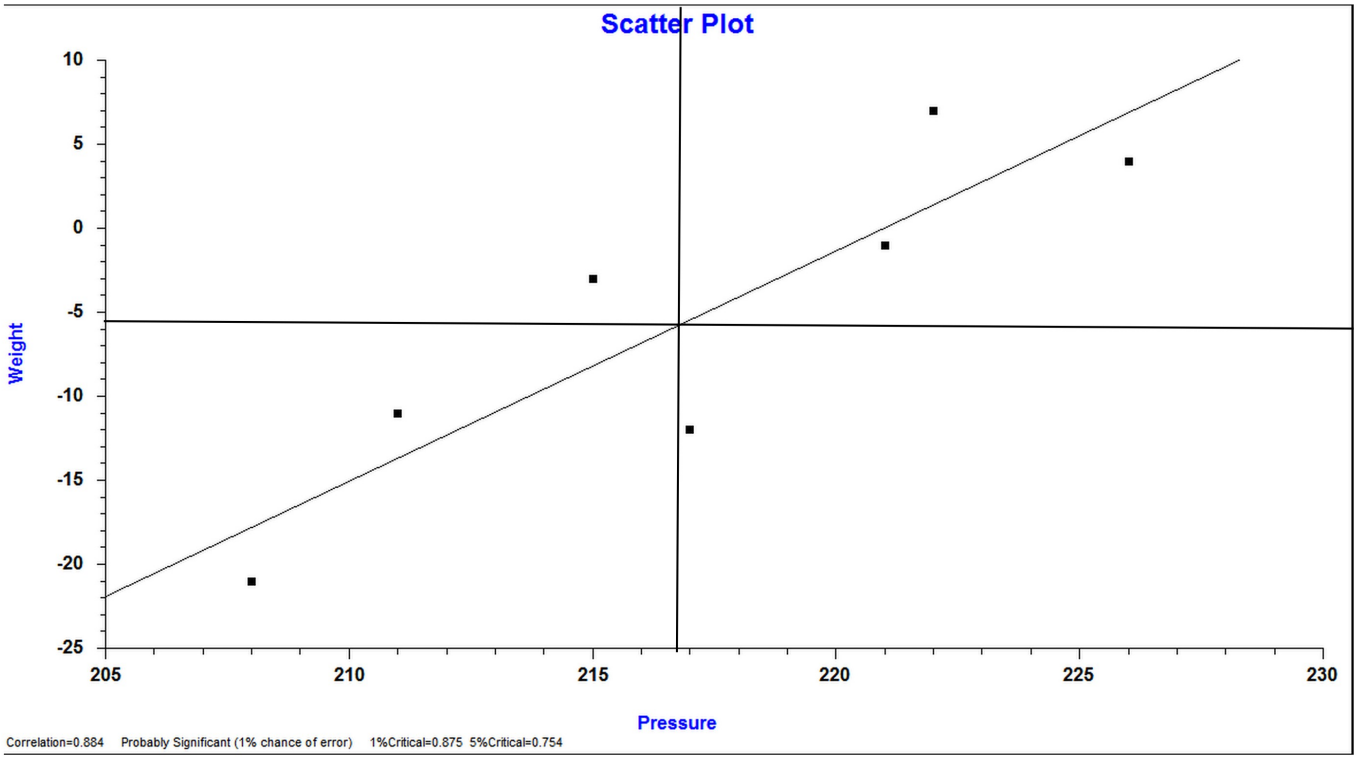


Figure 12.1

Immediately, we see strong evidence of a positive (in the sense of both variables increasing together) relationship i.e. of positive correlation. Furthermore, it is clear this relationship is fairly linear.

Activity 12.1 *Sketch graphs illustrating (a) negative correlation and (b) no correlation between two variables.*

Our next task is to develop a measure of the strength of association between our two variables. We start by considering Figure 12.2 where we have superimposed on Figure 12.1 the lines X = $\bar{X}$ and Y = $\bar{Y}$.


Scatter Plot
10
5
0
-5
-10
-15
-20
-25
Weight
205
210
215
220
225
230
Pressure
Correlation=0.884 Probably Significant (1% chance of error) 1%Critical=0.875 5%Critical=0.754

$$\bar{X} = 217.1$$

$$\bar{Y} = -5.3$$

From Figure 12.2 we see that because of the positive correlation between our two variables, most of the points lie in the two quadrants where (X - $\bar{X}$) and (Y - $\bar{Y}$) are either both positive or both negative. Consequently, if we calculate the sum

$$\sum(X - \bar{X})(Y - \bar{Y})$$

we should obtain a large positive quantity.

If we now "average" value this sum by dividing it by (n-1) we obtain the quantity known as the covariance, denoted by cov(X,Y). Thus we have:

$$\text{cov}(X,Y) = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{n - 1}$$

For our problem, the covariance between pressure and weight is cov(X, Y) = 333.29/6 = 55.548.

Now, although the covariance plays an important role in calculating an overall uncertainty of a measurement, it is a rather poor measure of correlation. The reason is that by changing the measurement units (e.g. inches to cm) will give a totally different value of covariance - and yet the strength of the relationship between our variables has not changed!

This problem is overcome by standardizing the quantities (X - $\bar{X}$) and (Y - $\bar{Y}$) in the expression for the covariance by dividing each by their respective standard deviation. This produces the measure known as the *linear correlation coefficient* (linear because it is designed specifically to measure the amount of linear association between variables, being based on the linear quantities (X - $\bar{X}$) and (Y - $\bar{Y}$). The formula for this correlation coefficient is:

$$r = \frac{\text{cov}(X,Y)}{s_x s_y}$$

where s_x and s_v are the standard deviations of X and Y respectively.

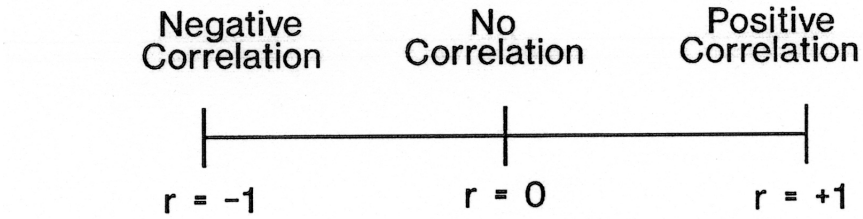
An alternative version of this formula is

$$r = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum(X - \bar{X})^2(Y - \bar{Y})^2}}$$

For our pressure and weight data, the value of r is 0.884.

The maximum value of r is + 1 and occurs when there is a perfect positive linear relation between the variables (all the values lie on a straight line with positive gradient) while the smallest value that r can have is -1 and occurs for a perfect negative linear relationship (all the values lie on a straight line with negative gradient). This is summarised in Figure 12.3.

Figure 12.3



Of course, we now ask "How indicative of the strength of the relationship is this value of r = 0.884?". A subjective opinion, guided by Figure 12.3, is that the value r = 0.884, being fairly close to + 1, indicates strong positive correlation (certainly strong enough for there to be a case for monitoring the process using easily read pressure rather than the costly weight measurement).

Usually, this judgment, particularly when backed up by the scatter diagram in Figure 12.1, is quite sufficient. However, we can use a hypothesis test to assess the real strength of evidence supporting a hypothesis of positive correlation. The Table in Neave's *"Elementary Statistics Tables"* gives critical values of r for n = 3 to 100.

Activity 12.2 *Test the null hypothesis "H_0: There is no correlation between pressure and weight." against the alternative hypothesis "H_1: There is positive correlation between pressure and weight."*