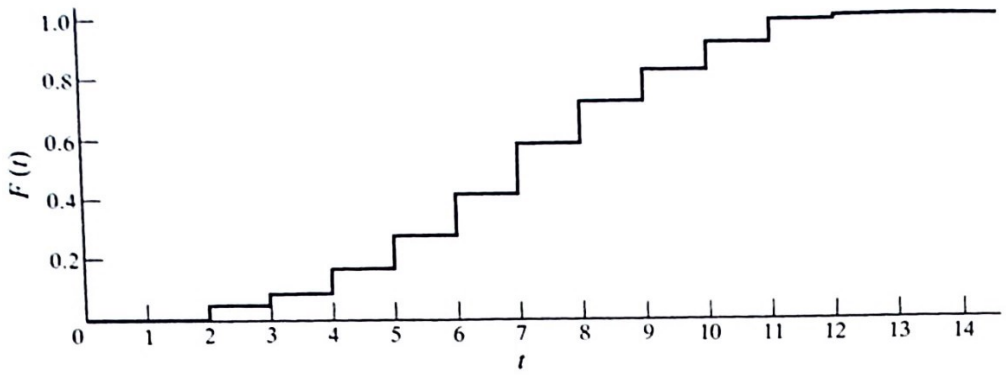


Figure 1.10

$$\text{Prob}(T \geqslant 8) = 1 - \text{Prob}(T \leqslant 7) = 1 - F(7) = 1 - 21/36 = 15/36$$

and, slightly more involved,

$$\begin{aligned}\text{Prob}(6 \leqslant T \leqslant 9) &= \text{Prob}(T \leqslant 9) - \text{Prob}(T < 6) \\ &\quad \text{Prob}(T \leqslant 9) - \text{Prob}(T \leqslant 5) \\ &= F(9) - F(5) = 30/36 - 10/36 = 20/36\end{aligned}$$

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). As an example, consider a random variable, U, whose realizations (expressed to any prescribed number of decimal places) are equally likely to take *any* value over the range 0 to 1. This random variable may not be as new to some of you as you may imagine, for it is this random variable that creates the unpredictability in most computer games*.

The graph (see Fig. 1.11) of its probability density function, which we denote by $f(u)$, reflects this property of equal chances very well. Also note that the area is equal to 1. The mathematical formula for its pdf is simply

$$f(u) = \begin{cases} 1 & 0 \leqslant u \leqslant 1 \\ 0 & \text{otherwise} \end{cases}$$

In fact, since this is such a useful distribution, it has been given a name: it is 'the *uniform* distribution over (0,1)', a name that clearly reflects its characteristics.

* The BBC Micro function RND(1) gives a realization of this random variable, as do buttons marked RND or RAND, etc., on many pocket calculators.

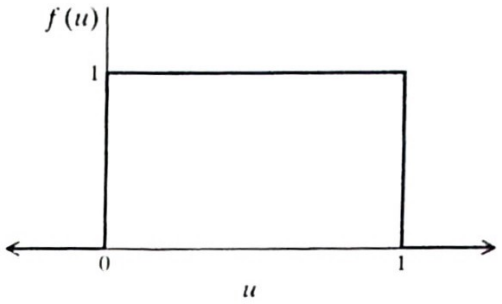


Figure 1.11 Graph of the pdf of the uniform distribution over (0,1).

Probabilities relating to this distribution are easily obtained. For instance

$$\begin{aligned}\text{Prob}(0.5 \leqslant u \leqslant 0.75) &= (\text{area under the pdf curve between } u=0.5 \text{ and } u=0.75) \\ &= 0.25 \times 1 = 0.25\end{aligned}$$

and

$$\text{Prob}(u \geqslant 0.8) = (\text{area between } u=0.8 \text{ and } u=1) = 0.2 \times 1 = 0.2$$

These areas are illustrated in Figures 1.12a and b respectively.

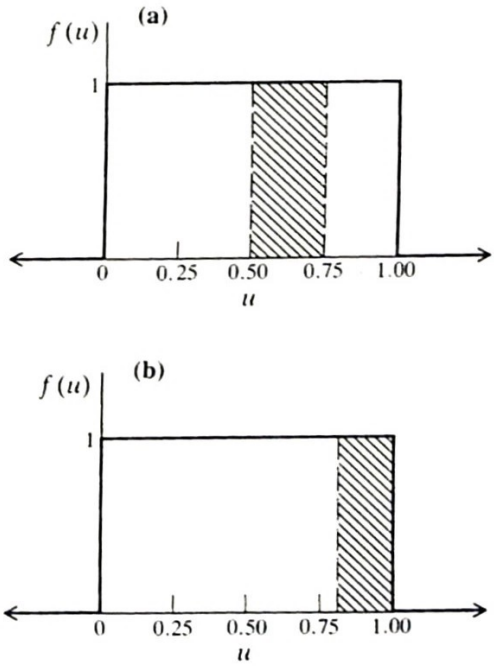


Figure 1.12 (a) Area between $u = 0.5$ and $u = 0.75$. (b) Area between $u = 0.8$ and $u = 1$.

The cdf of U is just as easy to derive. We know that $F(U) =$

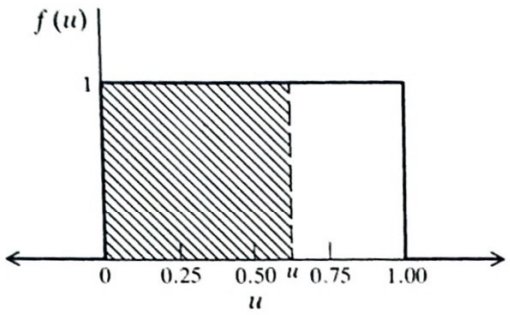


Figure 1.13 Area between $U = 0$ and $U = u$.

Prob($U \leqslant u$) = (area between $U=0$ and $U=u$) = $u \times 1$ (see Fig. 1.13). Thus a complete description of its cdf is

$$F(U) = \text{Prob}(U \leqslant u) = \begin{cases} 0 & u<0 \\ u & 0 \leqslant u \leqslant 1 \\ 1 & u>1 \end{cases}$$

The graph of the cdf also has a simple form and is shown in Figure 1.14.

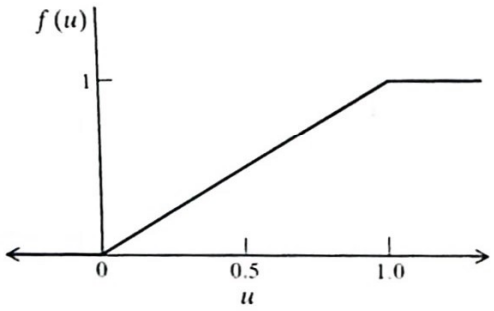


Figure 1.14 Graph of the cdf of the uniform distribution over (0,1).

SOME STANDARD THEORETICAL 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. In this section we briefly look at just three of these standard distributions – the binomial, the normal and the chi-squared distributions – each of which, in one way or another, is needed in various parts of the book.

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

A multiple-choice examination paper contains 18 questions, each having just two possible answers. Now, if a candidate were to choose each answer solely by guesswork, then the probability of choosing the correct answer is $\frac{1}{2}$ in each case. If the candidate needs 12 or more correct answers to pass, then by using the binomial distribution we can calculate the probability of a pass being obtained purely by such guesswork!

Here we have a situation where each guess (outcome) is one of two possibilities (right or wrong) and, unless the candidate is operating a 'scheme', the outcome of a particular guess is independent of all the others.

EXAMPLE 2

Suppose it is known that 25% of all males suffer to some extent from colour-blindness. If a random sample of, say, 20 males were selected then, for instance, we could calculate the probability of at least two 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-blind or not colour-blind. 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 even 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 characteristic of binomial distributions. In each case we have:

(a) A number, n, of independent similar trials ($n=18$ in the first example and $n=20$ in the second).
(b) 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 two successes out of the five tries.

It might happen that the first two tries were successes followed by three failures. We can represent this realization 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 two successes out of five, *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 that can give rise to two successes and three failures. There are in fact 10 such arrangements. We display them below, and their probabilities are 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	$p \times q \times p \times q \times q = p^2q^3$
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	$p \times q \times q \times p \times q = p^2q^3$
F S F F S	$q \times p \times q \times q \times p = p^2q^3$
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 occurrence. As there are 10 of them, the total probability of having precisely two successes out of five is $10p^2q^3$.

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 which 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

$$n! = n\times(n-1)\times(n-2) \times \ldots \times 3\times2\times1$$

so that, for example, $5!=5\times4\times3\times2\times1=120$. Other examples that 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 ${}^nC_x$.

In our problem above, the number of configurations of two $(=x)$ successes and three $(=n-x)$ failures in five $(=n)$ trials is computed by the formula as

$$\binom{5}{2} = \frac{5!}{2!3!} = \frac{5\times4\times3\times2\times1}{2\times1\times3\times2\times1} = 10$$

which is in agreement with our previous method.

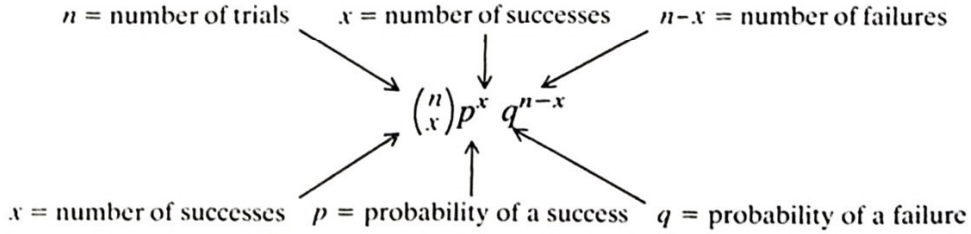
Note that we can write the probability of getting two successes out of five as

$$\binom{5}{2} p^2q^3$$

It is therefore easy to see how to generalize 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

$$\begin{aligned}\text{Prob}(X=x) &= \text{Prob}(x \text{ successes out of } n \text{ trials})\\ &= \binom{n}{x}p^xq^{n-x} \qquad (x = 0,1,2, \ldots ,n)\end{aligned}$$

Notice what each part of this expression represents:



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

To illustrate this important distribution further, let us consider the student in Example 1. In this case, $n=18$ and the probability of correctly answering a question is $\frac{1}{2}$, i.e. $p=\frac{1}{2}$ (and so also $q=\frac{1}{2}$). If we let X denote his total number of correct answers out of the 18, we find, by substituting $p=q=\frac{1}{2}$ and $n=18$ in the above formula, that

$$\text{Prob}(X=x) = \binom{18}{x}\left(\frac{1}{2}\right)^x\left(\frac{1}{2}\right)^{18-x} \qquad (x = 0,1,2, \ldots ,18)$$

So, for instance, if we wanted the probability of the student only guessing four answers correctly, we simply put $x=4$ in the above formula. This gives

$$\text{Prob}(X = 4) = \binom{18}{4}\left(\frac{1}{2}\right)^4\left(\frac{1}{2}\right)^{14}$$

Now

$$\binom{18}{4} = \frac{18!}{4!14!} = \frac{18\times17\times16\times15}{4\times3\times2\times1} = 3060$$

and so

$$\text{Prob}(X = 4) = 3060 \times \left(\frac{1}{2}\right)^4 \times \left(\frac{1}{2}\right)^{14}$$
$$= 0.0117 \quad \text{(to four decimal places)}$$

In the same way we could calculate the probability of his guessing correctly any number from 0 to 18. The results for this are shown in the following table:

x	0	1	2	4	4	5	6
Prob($X = x$)	0.0000	0.0001	0.0006	0.0031	0.0117	0.0327	0.0708
x	7	8	9	10	11	12	
Prob($X = x$)	0.1214	0.1669	0.1855	0.1669	0.1214	0.0708	
x	13	14	15	16	17	18	
Prob($X = x$)	0.0327	0.0117	0.0031	0.0006	0.0001	0.0000	

It is interesting to look at the histogram (shown in Fig. 1.15) of this

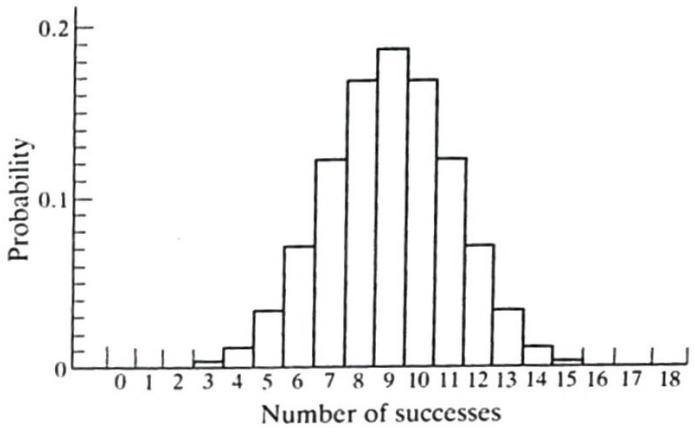


Figure 1.15

distribution which, because $p=q=\frac{1}{2}$, is symmetrical about $x = 9$ (the average number of successes).

Of course, once we have the full set of probabilities like this, it is easy to determine the chances of the student passing the examination on guesswork alone. We need to find the probability of his getting 12 or more correct answers, and so we have

$$\begin{aligned}\text{Prob}(X\geqslant 12) &= \text{Prob}(X=12) + \text{Prob}(X=13) + \ldots + \text{Prob}(X=18)\\ &= 0.0708 + 0.0327 + 0.0117 + 0.0031 + 0.0006 + 0.0001\\ &\quad + 0.0000\\ &= 0.1190\end{aligned}$$

So the student has about a 1 in 8 chance of passing by guesswork alone. Although obviously the odds are against this happening, the probability is by no means negligible. To counteract this, in practice most multiple-choice papers have more than two suggested answers to each question. Suppose there were four options per question. Then with $p=\frac{1}{4}$ and $q=\frac{3}{4}$, the probability of passing on guesswork alone, i.e. $\text{Prob}(X\geqslant 12)$, is now only 0.0002.

Even for quite small values of n, calculating binomial probabilities can be very tedious, particularly probabilities of *ranges* of values like the $X\geqslant 12$ in the last example. However, for $n\leqslant 20$ and a wide selection of values of p, binomial probabilities can be read directly from pages 4 to 11 of *EST* – individual probabilities like $\text{Prob}(X=12)$ from the left-hand pages, and cumulative probabilities like $\text{Prob}(X\leqslant 12)$ or $\text{Prob}(X\geqslant 12)$ from the right-hand pages.

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 $\mu = np$.
The standard deviation of the number of successes is $\sigma = \sqrt{npq}$.

In the case of the 18-question examination paper with two options per question ($n=18$, $p=\frac{1}{2}$), the average number of correct answers obtained by guesswork is $\mu=np=18\times\frac{1}{2}=9$ (as you would surely 'expect'). The standard deviation of the number of correct answers is $\sigma=\sqrt{npq} = \sqrt{18\times\frac{1}{2}\times\frac{1}{2}} = \sqrt{4.5} = 2.12$. It is interesting to compare these figures with what we get when there are four options per question ($n=18$, $p=\frac{1}{4}$). Then, $\mu=np= 18\times\frac{1}{4} = 4.5$ – much lower, again as you would 'expect'; the standard deviation $\sigma=\sqrt{npq} = \sqrt{18\times\frac{1}{4}\times\frac{3}{4}} = 1.84$. Note that this is also lower than before, indicating a smaller variability around the mean value; whereas with $p=\frac{1}{2}$ there is a relatively large chance of the number of correct answers being substantially different from the average (either high or low), with $p=\frac{1}{4}$ it is more likely that we will definitely get a score close to the mean value.

The normal distribution

The normal distribution is the most important of all continuous probability distributions. One reason for this is that it very adequately describes the behaviour of results of many measuring processes. Typical examples are:

(a) the heights of females (or males) in a community;
(b) the net weights of cans of petfood, packets of breakfast cereals, etc.;
(c) the lengths of mass-produced components;
(d) the *actual* time of arrival of, say, the '8 o'clock bus'.

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 which behaves in a similar way and has a perfectly symmetrical shape. A typical normal curve is shown in Figure 1.16.

Such a curve is often described as being 'bell-shaped'. The problems

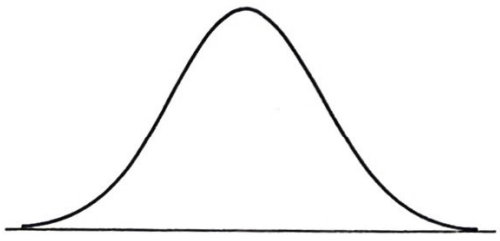


Figure 1.16

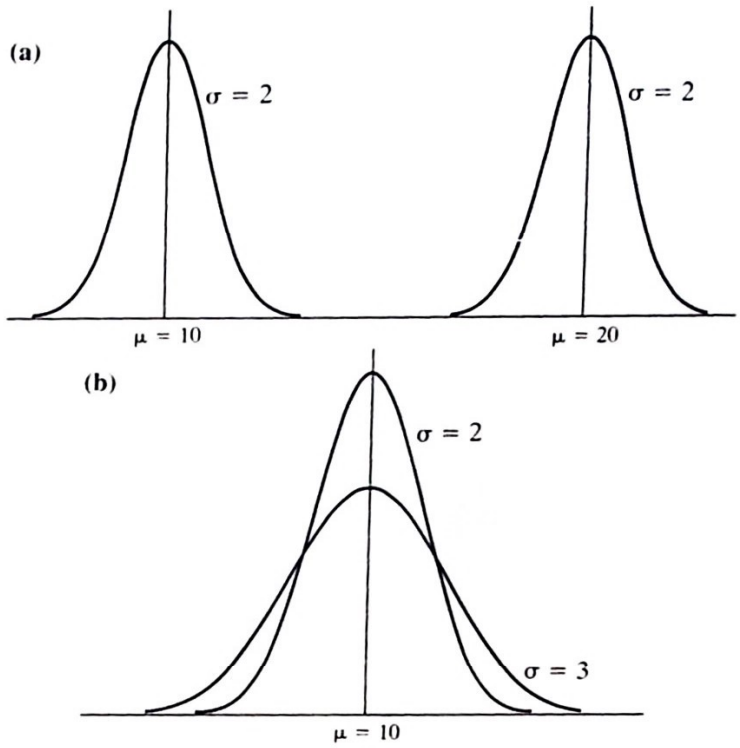


Figure 1.17

about where 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 1.17 illustrates normal distributions for varying means μ and standard deviations σ.

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 $N(\mu, \sigma^2)$. Thus the two distributions in Figure 1.17a could be denoted by N(10,4) and N(20,4) and those in Figure 1.17b 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 its position. Increasing the standard deviation flattens the shape (which 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 recorded the heights of 200 female students who attended our college. Figure 1.18 shows the histogram produced from this sample.

This histogram is reproduced in Figure 1.19, and superimposed on it is the normal curve whose mean and standard deviation are the same as those of the sample.

A vitally important feature of normal distributions is that, regardless of

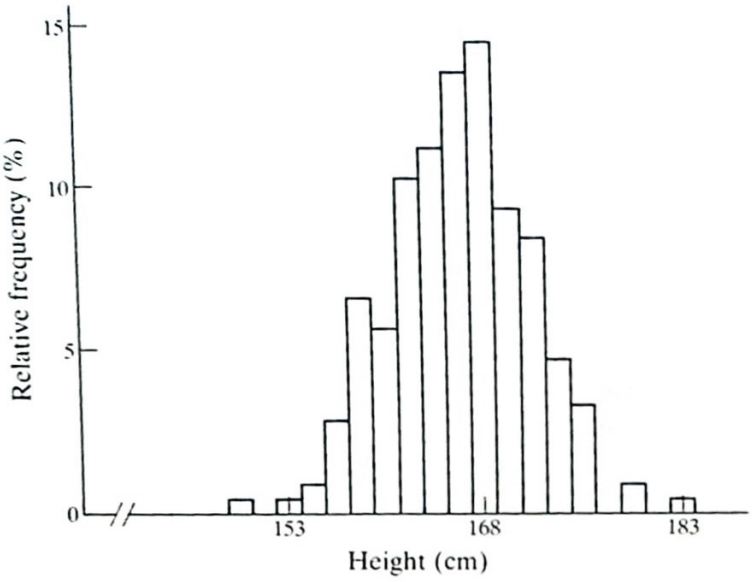


Figure 1.18

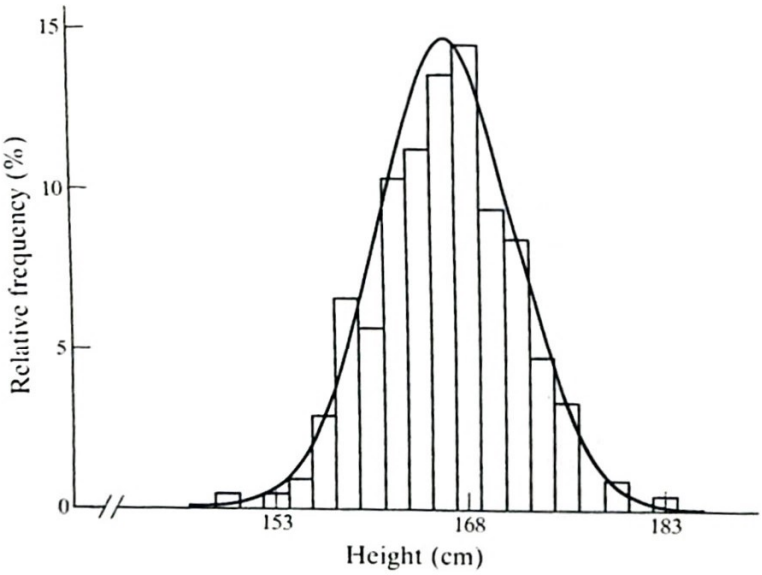


Figure 1.19

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 lies within 1.96 (i.e. roughly 2) standard deviations of μ;
(c) 99% of its area lies within 2.58 standard deviations of μ.

See Figure 1.20 for illustration of this point.

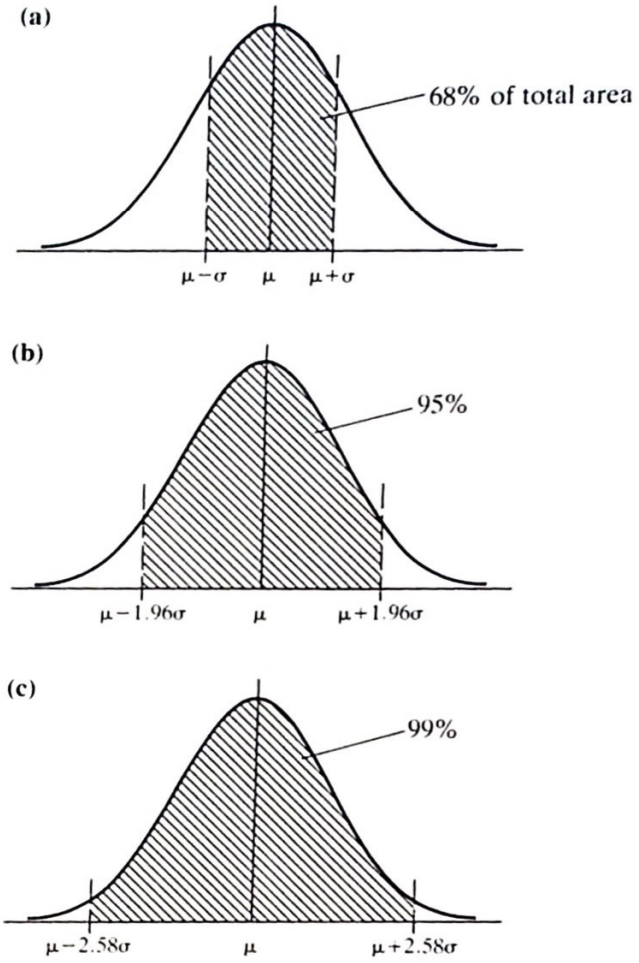


Figure 1.20

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%, and more than 2.5 standard deviations away is about 1%. Thus if we observed a value that is more than, say, 2 standard deviations away from the mean, then we might classify it as a somewhat unusual event which might cast doubt as to whether the observation even comes from this distribution – more of this idea later!

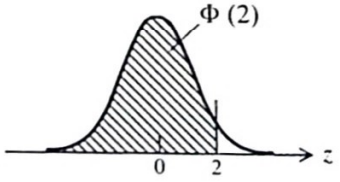
We recall that, for a continuous distribution, the probability of an observation lying between two specified values is given by the corresponding area under the graph of the pdf. Now, as we have said, normal distributions are encountered (at least approximately) in a wide variety of situations. Since different situations are likely to give rise to different values of the mean μ and/or standard deviation σ, we need some easy way of calculating areas under different normal curves. Clearly, it is impossible to provide a table for every possible normal distribution!

Fortunately, it is easy to relate *any* normal distribution to *one* particular normal distribution: this is the so-called *standard* normal distribution, which is simply the normal distribution whose mean is 0 and standard deviation is 1, i.e. $N(0,1^2)=N(0,1)$. So all we need in order to get probabilities for any normal distribution is a table of areas for the standard normal distribution. The basic idea is as follows. Suppose we have a variable X which is normally distributed with mean μ and standard deviation σ. Then if we define a new variable Z as

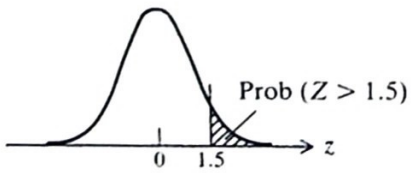
$$Z = \frac{X - \mu}{\sigma}$$

it turns out that the distribution of Z is always the *standard* normal distribution. Since any event involving X can be easily expressed in terms of an event involving Z (as we illustrate in the examples below), then a table of the *standard* normal distribution $N(0,1)$ is indeed all we need. Table A at the back of this book (or pages 18–19 of *EST*) gives the probabilities $\text{Prob}(Z \leqslant z)$, i.e. the cumulative distribution function of $N(0,1)$, which is often denoted by $\Phi(z)$. Here are some examples of the use of this table.

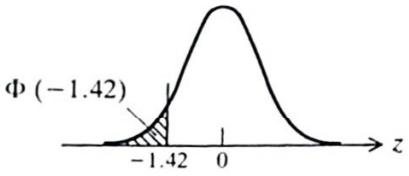
(a) $\Phi(2) = \text{Prob}(Z \leqslant 2)$
$= 0.9772$



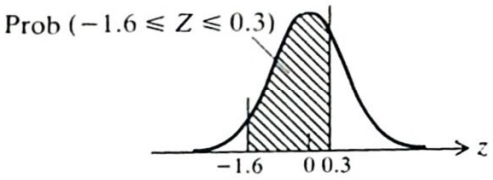
(b) $\text{Prob}(Z > 1.5) = (\text{area to the right of } z=1.5)$
$= 1 - \text{Prob}(Z \leqslant 1.5)$
$= 1 - (\text{area to the left of } 1.5)$
$= 1 - \Phi(1.5)$
$= 1 - 0.9332$
$= 0.0668$



(c) $\text{Prob}(Z<-1.42) = \Phi(-1.42)$
$= 0.0778$



(d) $\text{Prob}(-1.6 \leq Z \leq 0.3) =$ (area to the left of 0.3) − (area to the left of −1.6)
$= \Phi(0.3) - \Phi(-1.6)$
$= 0.6179 - 0.0548$
$= 0.5631$

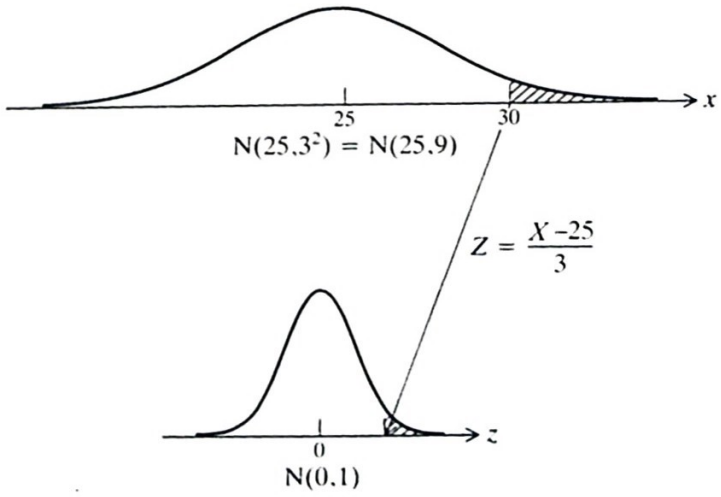


The uses of the relation $Z = (X-\mu)/\sigma$ on 'non-standard' normal distributions is shown in the following examples.

EXAMPLE 1

Suppose that turn-round time for servicing an aircraft between flights is normally distributed with mean 25 min and standard deviation 3 min. What is the probability that the turn-round will exceed 30 min?

Let X denote the servicing time – its normal curve is:



The shaded area corresponds to Prob($X>30$) and this is equal to the shaded area under the standard normal curve, if Z is defined according to the above rule, i.e.

$$Z=\frac{X-\mu}{\sigma}=\frac{X-25}{3}$$

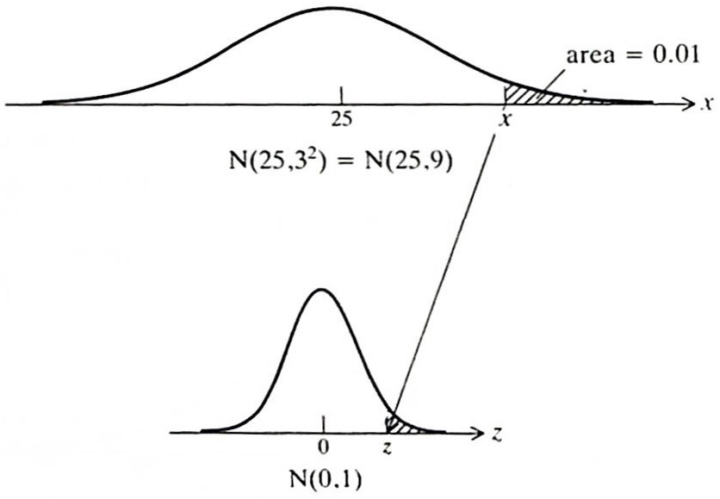
Using $Z = (X-\mu)/\sigma$, we find that $X=30$ corresponds to $Z = (30-25)/3 = 1.67$. So, we can now use Table A to obtain the required probability:

$$\begin{aligned}\text{Prob } (X>30) &= \text{Prob}(Z>1.67)\\ &= 1 - \text{Prob}(Z\leqslant 1.67)\\ &= 1 - \Phi(1.67)\\ &= 0.0475\end{aligned}$$

EXAMPLE 2

In the same situation, suppose we want to find the servicing time x, say, which is exceeded on 1% of occasions, i.e. Prob($X>x$) = 0.01.

In the diagrams below, both shaded areas correspond to what is sometimes called an *upper-tail* probability of 0.01:



The shaded area corresponds to Prob($X>30$) and this is equal to the shaded area under the standard normal curve if Z is defined according to the above rule, i.e.

$$Z = (X-\mu)/\sigma = (x-25)/3$$

First we use Table A to find z. As there has to be an area of 0.01 to the right of z (and thus an area of 0.99 to the left), we have

$$\text{Prob}(Z<z) = 0.99 \qquad \text{i.e.} \qquad \Phi(z) = 0.99$$

and so, reading from Table A, we obtain $z = 2.326$. But

$$z = \frac{x - \mu}{\sigma}$$

where $\mu=25$ and $\sigma=3$, and so

$$2.326 = \frac{x - 25}{3}$$

From this

$$x = 3\times2.326 + 25 = 31.98$$

What does this result imply? Simply that there is a chance of only about 1% that the servicing time will exceed 32 min.

As these two examples show, by using the relation $Z=(X-\mu)/\sigma$ in conjunction with the standard normal table (Table A), we are able to obtain probabilities in any normal distribution regardless of the value of its mean and standard deviation.

The normal approximation to the binomial distribution

Calculating binomial probabilities can get quite painful for large values of n (and remember that the tables in *EST* only go up to $n=20$). Fortunately, unless the probability p of success is close to 0 or 1, normal distributions can provide quite a good approximation to the binomial distributions.

To illustrate how this approximation is used (and how good it is), let us again consider the binomial distribution with $n=18$ and $p=0.5$. This particular distribution arose in the previous section as the distribution of the number of correct answers obtained by guessing between the two choices in each question. We found that $\mu=9$ and $\sigma=2.12$. Figure 1.21 shows the histogram of this distribution together with the normal distribution $N(9,2.12^2)$. Immediately we see just how well the normal distribution describes the behaviour of this binomial distribution. In fact, it can be shown that the closeness of these two distributions continually improves as n increases – in technical terms, we say that the binomial distribution is *asymptotically* normally distributed.

The calculation of approximate binomial probabilities from the normal distribution is really quite straightforward. For example, let us calculate

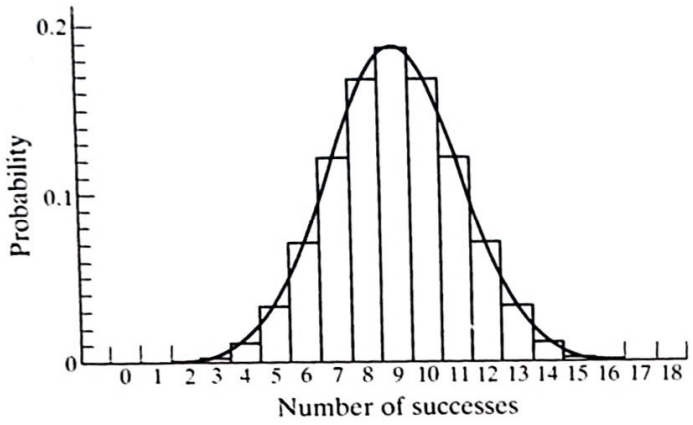


Figure 1.21

the probability of having 12 or more correct answers on the multiple-choice paper. Figure 1.22 indicates the area under the normal curve needed to approximate the required probability. Note that the shaded region on the normal scale starts at 11.5 rather than 12. This is to compensate for the fact that the binomial variable is discrete (integer-valued) while the normal variable is continuous – the 'correction' of 0.5 is called the *continuity correction*. The reason for its use is clear from the diagram: the area to the right of 12 would clearly be too small.

The value of Z corresponding to the 'corrected' binomial value 11.5 is

$$z = \frac{11.5 - 9}{2.12} = 1.18$$

(where, you will recall, $\mu = np = 18 \times \frac{1}{2} = 9$ and $\sigma = \sqrt{npq} = \sqrt{18 \times \frac{1}{2} \times \frac{1}{2}} = 2.12$). From the normal table (Table A)

$$\begin{aligned} \text{Prob}(Z \geqslant 1.18) &= 1 - \Phi(1.18) \\ &= 1 - 0.8810 \\ &= 0.1190 \end{aligned}$$

which agrees (to four decimal places) with the true probability of 0.1190 obtained earlier using exact binomial calculations.

As a general rule, if $n \geqslant 20$ and p is between $5/n$ and $1 - 5/n$, quite good results will be obtained using this normal approximation to the binomial distribution (and the larger n is, the better!). The continuity correction gives

$$\text{Prob}(X \leqslant x) = \Phi\left(\frac{x + 0.5 - \mu}{\sigma}\right)$$

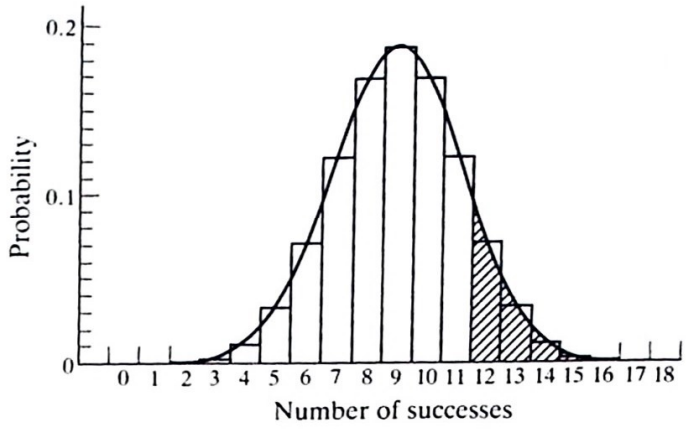


Figure 1.22

$$\text{Prob}(X=x) = \Phi\left(\frac{x + 0.5 - \mu}{\sigma}\right) - \Phi\left(\frac{x - 0.5 - \mu}{\sigma}\right)$$

and

$$\text{Prob}(X \geqslant x) = 1 - \Phi\left(\frac{x - 0.5 - \mu}{\sigma}\right)$$

where $\mu=np$ and $\sigma=\sqrt{npq}$ are the mean and standard deviation of the binomial distribution.

In later chapters you will find that the normal distribution is the asymptotic distribution of very many of the test statistics that we use.

The chi-squared distribtion

The final distribution we need to mention is the so-called *chi-squared* (χ^2) distribution (denoted by the Greek letter (chi) χ, pronounced 'ki'). It was discovered in 1876 by Friedrich Robert Helmert and rediscovered in 1900 by Karl Pearson. Since then it has been used in a wide variety of contexts but, although we too shall use it in later chapters, its complexity prevents us from examining its features in much detail. The shape of its probability density curve varies according to a quantity known as its number of degrees of freedom (denoted by the Greek letter (nu) ν in the tables). To give you some idea of how the shape depends on ν, we reproduce in Figure 1.23 the density curves for chi-squared distributions for $\nu=1$, 2, 4 and 10. As you see, it is difficult to speak of a typical shape for a chi-squared distribution! However, the sort of information about chi-squared distributions which we shall need later is summarized in Table B. This table gives the end-points for areas of certain sizes. For example, the entry for $\nu=2$ and $q=0.95$ is 5.991, and this tells us that the

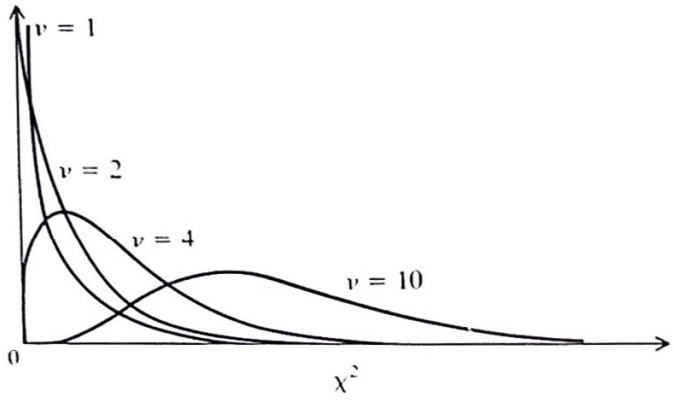


Figure 1.23

area to the left of 5.991 under the chi-squared pdf with $\nu=2$ degrees of freedom is 0.95, as shown in Figure 1.24. It follows that, in this case, with two degrees of freedom, the probability of a chi-squared value being *greater* than 5.991 is $1 - 0.95 = 0.05$, i.e. 5%.

The values tabulated in Table B are referred to as percentage points since they are the values for areas of 95% and 99% of the total (a more extensive range of percentage points of the chi-squared distribution is given on page 21 of *EST*).

A note on the use of the median instead of the mean

In this chapter we have referred to the median and the mean as measures of 'location' or 'average'. Classical statistical tests for location almost always use the mean; however, the mean is not used in distribution-free methods. There are a number of reasons for this. First, there are a few probability distributions that do not even have a mean! Thus, if our distribution-free techniques are to be valid for *all* (continuous) distributions, the mean must clearly be inappropriate. Further, even if a population's mean does exist, the distribution of the sample mean depends very much on that population's probability distribution (unless the sample size is large) – thus the sample mean is not useful for distribution-free purposes.

So, when a measure of location or average is needed in a distribution-free context, we shall invariably use the median. All continuous probability distributions have a median – it is defined as the point, ϕ, that divides the area under the probability density function into two equal parts. For symmetric distributions, ϕ is equal to the mean μ (if μ exists). For non-symmetric (or asymmetric) distributions, ϕ is different

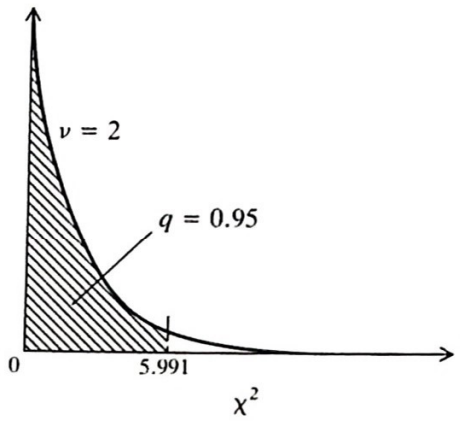


Figure 1.24

from μ; however, in some senses ϕ is then a *better* measure of 'average' or 'typical value' than μ since it is often the case in asymmetric distributions that the minority of extreme values have an undue influence on μ, making it 'untypical' of the distribution as a whole.

2

The underlying ideas of statistical tests

Introduction

Tests (usually called *hypothesis tests* or *significance tests*) often play a major part in statistical investigations. This book is almost entirely concerned with statistical tests, and it is important for us to be quite clear from the start on both the usefulness and the limitations of such procedures. To be honest, many people tend to look upon statisticians, and their techniques and conclusions, with suspicion or even downright distrust. 'You can prove *anything* with statistics' is a common gibe, and this is often supported by Benjamin Disraeli's famous definition of the three kinds of lies.* In this chapter, which in a sense is the most important (and perhaps the most difficult) one in the book, not only will we hope to dispel some of the myths about statistics (and statisticians!), but we shall lay the essential foundations for almost all subsequent chapters.

The basic idea of most statistical techniques is to increase our knowledge about populations using information in samples taken from them. 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 average yields from four different types of wheat.
(b) A new medical treatment is more effective than the existing treatment.
(c) Urban children are less healthy than rural children.
(d) There is no connection between sales of luxury goods and unemployment.

* That is, his 'lies, damned lies, and statistics'. Actually, although this famous saying is indeed usually ascribed to Disraeli, John Bibby writes in his amusing little book *Quotes, damned quotes, and . . .* (Demast Books, 1983) that: 'as far as I can tell, Winston Churchill knew that Mark Twain wrote that Benjamin Disraeli had said the oft-quoted phrase. However, it sounds most un-Disraelite to me. Lord Blake, biographer of Dizzy, has confirmed as much in a personal letter.

. . . and so on.

Whether or not particular hypotheses are eventually assessed to be reasonable will depend on the weight of evidence contained in data from the relevant population(s). Actually, as opposed to the gibe mentioned at the start of the chapter, we never claim to *prove* anything completely 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'.

The wary reader will note, both from the preamble and from the examples of hypotheses just given, that the use of statistical tests presupposes the meaningful existence of 'populations'. Traditionally, statisticians and users of statistical techniques seem to have paid scant attention to the validity of this presumption in their areas of application. However, as we have seen in Chapter 1, such validity is by no means a foregone conclusion. Partly because of this, other statisticians have more recently tended to suggest that the complete topic of statistical testing is inappropriate to the real world. The countless thousands of users of statistical tests over the last half-century might find it difficult to agree with such an extreme a view! We agree that more care and attention need to be paid to the questions of existence and identification of the populations under study than has normally been the case in traditional practice, but that, subject to the context being considered suitable as the result of such considerations, tests do have an important and useful role to play in helping to interpret data. As implied in Chapter 1, both the descriptions and the illustrations of the use of the techniques in this book should be read with the understanding that the circumstances of applications are deemed to be appropriate.

Anyone involved with statistical testing 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 terminology.

To these ends, in the next section we present five main steps which may 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'; however, in practice, our task will usually be easier than this, for the whole of step 2 and parts of steps 4

and 5 will follow automatically as soon as any specific test from the remaining chapters of this book has been chosen. However, this comprehensive discussion will show you more clearly what really happens in a complete sense in a statistical test, and all the examples given in later chapters effectively consist of these five steps plus some 'technical aids'.

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.

A car-accessory manufacturer has developed what it claims to be an 'improved' car wax polish. A motoring magazine decides to carry out an experiment to compare the new polish with the old so as to test the manufacturer's claim. Eighteen car owners were selected and asked to treat half the roof of their car with the new polish and half with the old. The owners were told to judge which polish gave the better finish and to send their findings to the magazine.

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

STEP 1 FORMULATION OF THE PRACTICAL PROBLEM IN TERMS OF HYPOTHESES

We shall begin by attempting to clarify exactly what is meant by the word 'hypothesis'. According to the authors' edition of the Concise Oxford Dictionary, a hypothesis is defined as '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 – 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 (denoted by H_1 or H_A).

It is perhaps unfortunate that the word 'alternative' can give the impression that H_1 is somehow inferior to H_0; for it is, in fact, the *more*

important hypothesis from the practical point of view. The 'A' might better stand for 'action' rather than 'alternative' since H_1 or H_A *indicates the range of situations that we want the test to be able to recognize or 'diagnose'* should they be true. A positive diagnosis will indicate that we should perhaps take action of some kind (for example, to adopt a new medical treatment if the test indicates it to be better than the existing one, or to recommend a new paint 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 the problem of the car wax polish can now be immediately identified as:

H_1: The new polish is an improvement on the old one.
H_0: The two polishes are equally good.

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

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 almost invariably do is to carry out some calculation on the data so as to summarize them in terms of a single representative figure. Any such quantity calculated from the data is called a *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:

(a) 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 tend to 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.
(b) 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 the guide as to what are or are not reasonable values of T under H_0.

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

> A car owner has to judge between the two polishes. Suppose we denote his judgment by '+' or '−' according to whether he judges the new or the old polish to be the better. Let us define T as the total number of '+'s, i.e. T is the number of car owners who judged the new polish as the better of the two.

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 the new polish is a real improvement (i.e. H_1 is true), then we can expect T to have a high value, whereas if the polishes were just as good as each other (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, the car owners' decisions as to which is the better polish will be rather like tossing a coin, i.e. they will each produce a '+' or a '−' with probability of $\frac{1}{2}$ each; consequently the null distributioin of T is binomial with $n=18$ and $p=\frac{1}{2}$ (cf. the example regarding the mutliple-choice exam paper in Chapter 1).

Once we have discovered the null distribution of T (i.e. its probability distribution when H_0 is true), we can discriminate between what are and are not reasonable values of T under H_0.

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 T will 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,

Exhibit 1006
Page 071 of 453

our conclusion is that we may 'reject H_0 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 H_0 in favour of H_1 and the value of T is then said to be *not significant*.

Unlike many books and teachers, 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: Average length = 60 cm.
H_1: Average 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. Furthermore, tests are always constructed so that lack of information, especially through having too little data, tends to produce *non-significant* values; in this case, it is obviously even more misleading to talk positively about 'accepting H_0'. So we prefer the less ambitious description of 'not rejecting H_0', which gives a more accurate impression of what the test has found (or rather, not found!).

In the car wax problem, remember that if H_1 is true then we expect more '+'s than '−'s (i.e. a high value of T); otherwise if H_0 is true we expect similar numbers of '+'s and '−'s (a moderate value of T) or, of course, fewer '+'s than '−'s (a low value of T) if the new polish is inferior to the old one. So we should reject H_0 in favour of H_1 (i.e. have positive evidence that the new polish is better) only if T is *large*, i.e. in symbols, if $T \geqslant c$, where c is some appropriately chosen 'critical value'. This critical value c is determined using our knowledge of the null distribution of T, which in this example is binomial with $n=18$ and $p=\frac{1}{2}$. Since the average value of T under this distribution is 9 (from $\mu=np=18\times\frac{1}{2}$), values of c that are possibly worth considering are, say, 10 or 11 or 12, . . . or 17, or even 18. The selection of a suitable critical value from these possibilities is the subject of step 4.

STEP 4 DECIDING ON A SIGNIFICANCE LEVEL, AND HENCE DETERMINING THE CRITICAL VALUE

We have now decided on the *type* of critical region appropriate for our test, namely $T \geqslant c$. But how do we decide on the details of precisely how large the critical region should be, i.e. how do we choose c? 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 H_0 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. Our standard practice throughout the book, for all tests and choices of hypotheses, is to provide details for significance levels of 5% and 1%, which are the ones most commonly used, and sometimes a couple of others.

Suppose that for the car wax problem we decide, at least in the first instance, on a significance level of about 5%. We would say 'about 5%' since it is usually impossible with a discrete null distribution to obtain critical regions that 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 $n=18$ and $p=\frac{1}{2}$, and we seek a value c such that $P(T \geqslant c) = 0.05$ when H_0 is true. Using the table of probabilities for the binomial distribution given in Chapter 1 (page 32) we see that $P(T \geqslant 13) = 0.0481$, which is the closest we can get to 0.05. Thus, we may state the test's decision rule as follows:

> If $T \geqslant 13$ we shall reject H_0 in favour of H_1 at a significance level of about 5%, otherwise we shall not reject H_0 (at this level).

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 $T \leqslant 12$, we cannot reject H_0 in favour of H_1, for then either the value of T is 'middling' and so entirely reasonable under H_0, or T is small – in which case, although it is unlikely to occur under H_0, it would be even more unlikely under H_1.

Now suppose that the results from the 18 car owners were

+ + + − + + + + + + + − + + + + + +

Counting the number of '+' signs gives $T=16$.

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 that the car waxes are equally good in favour of the alternative hypothesis that the new wax polish is indeed an improvement.

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 18 signs were '+'s, i.e. *every* car owner preferred the new wax. 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 17, 16 (as above), 15, 14 or even 13 '+'s! In practice, a much more informative conclusion may be 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 fall just 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:

(a) very little, if any, evidence (if the 'borderline' significance level exceeds 10%);
(b) some suspicion, but nowhere near conclusive evidence (between 5% and 10%);
(c) considerable evidence (between 1% and 5%);
(d) very strong evidence (between 0.1% and 1%);
(e) practically conclusive evidence (less than 0.1%).

How does this discussion affect our conclusion in the car wax problem? Previously we described the result $T=16$ as being significant at the 5% level. But the table on page 32 in Chapter 1 gives $P(T\geqslant 16)=0.0007$. Therefore the result is far stronger than merely 'significant at the 5% level'; it is in fact significant at the 0.07% level (or at any weaker level, i.e. at any level exceeding 0.07%). So, in the words of the preceding

paragraph, we have 'practically conclusive evidence' to support the manufacturer's claim that the new polish is better than the old.

SUMMARY

The above ideas on statistical testing are conveniently summarized in the flowchart shown in Figure 2.1. In this flowchart, notice that there is a choice to be made on whether or not to follow the *simple* route. The non-mature, relatively 'simple-minded', route (the left-hand side) follows our initial description of fixing a significance level and then essentially just answering yes or no to the question of whether the result is significant at this specific level. This rather unimaginative approach is still used in some textbooks but, as we have seen, it can be very wasteful of the information actually presented in the data. Nevertheless, if the ideas of statistical testing are new to you, you may find it easier to stay with this approach for the time being. The more sophisticated route (the right-hand side) summarizes the approach covered in all but the first paragraph in step 5 above.

Finally, we should remind the reader that, in practice, substantial parts of the above procedure will be simplified once we have decided which test to use. The test we have devised for the car wax problem is a standard procedure known as the 'sign test', to be covered in more detail in Chapter 3. The main practical simplifications are that, as soon as we have decided to use such a standard test,

(a) the statistic to be calculated is automatically decided upon, and
(b) details of critical values for particular significance levels are made directly available to us in tables, without us having to investigate the null distribution ourselves.

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 which, 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 $T \geqslant c$. We shall now see that H_1 is responsible for classifying a test as being either *one-sided* or *two-sided*.

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

H_1: The new polish is an improvement on the old one.

But suppose, more generally, that we were interested in finding out

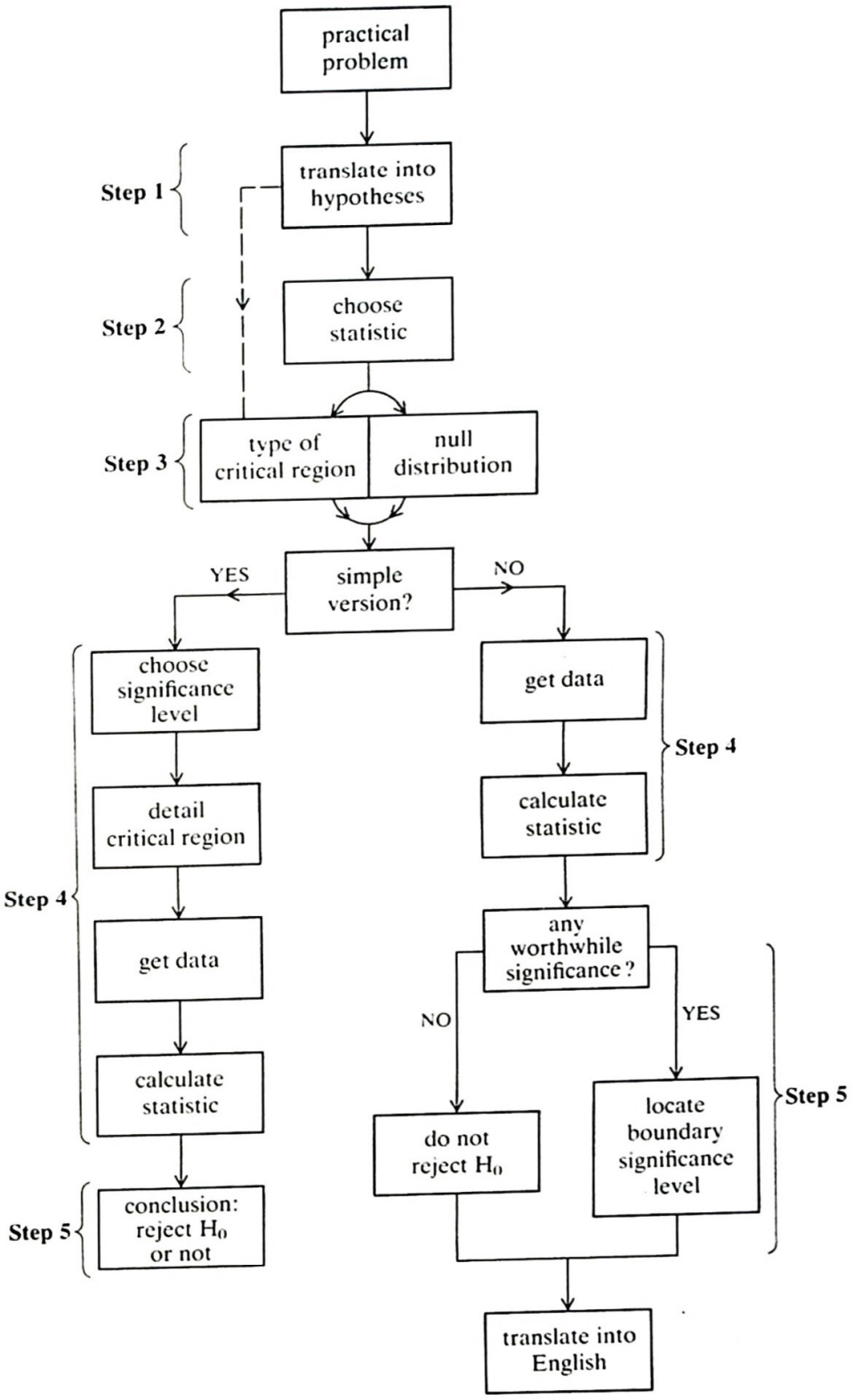


Figure 2.1 Flowchart for statistical testing.

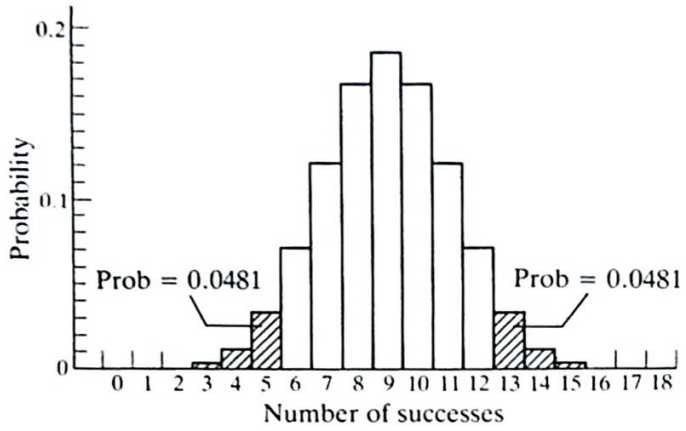


Figure 2.2 Critical region $T \leqslant 5$ and $T \geqslant 13$.

whether there is *any* kind of real difference between the two polishes, then the appropriate alternative hypothesis would be

H_1: There is *some* difference between the new and the old polishes.

Now, unlike the original alternative, this H_1 of some difference allows for the possibilities that the new polish could be either better or worse than the old polish. Because of the 'two-sidedness' of this H_1, the statistical test for this situation is referred to as a *two-sided test*, as opposed to the test based on the original 'one-sided' H_1 (which specifically required that the new polish be better than the old), which is called a *one-sided test*.

The critical region for such a two-sided test is quite different from that for the one-sided test. Under H_0, we know that the 'expected' or average value of T is 9 (i.e. $18\times\frac{1}{2}$). For the two-sided test, it is a difference in the actual value of T from its expected value of 9 in *either* direction that will constitute evidence for H_1. In symbols, therefore, we now need a critical region of the form 'distance between T and 9' $\geqslant c^*$ i.e. the difference between T and 9 is at least c. For example, if we choose $c=4$, the critical region would contain all possible values of T satisfying $T\leqslant 5$ and $T\geqslant 13$.

However, using the symmetry of the binomial distribution with $p=\frac{1}{2}$ (see Fig. 2.2) or directly from the table on page 32 in Chapter 1, it is easy to find that the probability of T attaining such a value is $2\times 0.00481 = 0.0962$ or 9.62%; this is the significance level corresponding to this critical region – and, of course, it shows the critical region to be unsatisfactory if we are interested in significance levels of around 5%. So let us make the critical region smaller by taking $c=5$ instead of $c=4$. This gives the critical region containing the values $T\leqslant 4$ and $T\geqslant 14$. The table

* This statement is sometimes written as $|T-9|\geqslant c$, where as we mentioned in Chapter 1 the vertical bars mean 'take the absolute value of $T-9$'.

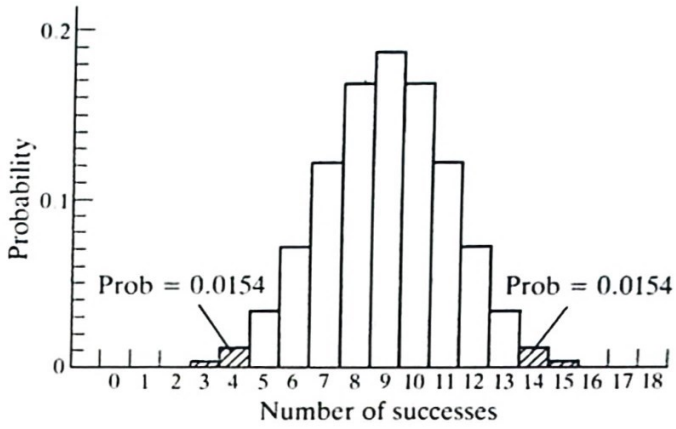


Figure 2.3 Critical region $T \leq 4$ and $T \geq 14$.

on page 32 shows us that the corresponding significance level is $2 \times 0.0154 = 0.0308$(See Fig. 2.3) – rather less than 0.05 but certainly more appropriate than the 0.0962 significance level for $c=4$.

So our decision rule for the two-sided test is to reject H_0 in favour of H_1 if $T \leqslant 4$ or $T \geqslant 14$, and otherwise not to reject H_0. This rule is the best we can find for the nominal significance level of 5% although, as we have just seen, the actual significance level is more like 3%.

The question of which form of alternative hypothesis is relevant, 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 recognizing differences in the opposite direction. This is well illustrated by the car wax problem. The value $T=13$, which is quite a likely result if the new polish is rather better than the old, is significant at the 5% level in the one-sided test, but not in the two-sided test. On the other hand, very small values are significant in the two-sided test (with its nominal 5% critical region $T \leqslant 4$ and $T \leqslant 14$) but not in the above-one-sided test.

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). However, we shall attempt in this book to give sufficient description of the illustrative problems to allow well informed choices to be made.

Of course, in our tables, etc., it will be necessary to distinguish between information relating to one-sided and two-sided tests. We shall use α_1 and α_2 to denote the significance levels corresponding respectively to one-sided and two-sided tests. Thus, for example, we would write the critical region for the above one-sided test as $T \geqslant 13$ for $\alpha_1 = 5\%$ (nominal) and the critical region for the two-sided test as $T \leqslant 4$ and $T \geqslant 14$ for $\alpha_2 = 5\%$ (nominal).

Best conservative critical values

We mentioned earlier that it is often impossible to obtain a critical region that exactly corresponds to a desired significance level (this was very marked in the above two-sided test!). So we follow what is in fact standard practice of indicating the largest critical region whose significance level is *no greater* than the nominal figure, i.e. we use the so-called '*best conservative*' critical regions. For example, suppose that, in the one-sided version of the car wax problem, we wanted to use a nominal significance level of $\alpha_1 = 1\%$ (so as to lessen the risk of rejecting H_0 when it is in fact true). From the table on page 32 we find that, if H_0 is true,

$$\text{Prob}(T \geqslant 14) = 0.0154$$
$$\text{Prob}(T \geqslant 15) = 0.0038$$

The 1% best conservative critical region is $T \geqslant 15$, even though $\text{Prob}(T \geqslant 14) = 0.0154$ is actually slightly closer to $0.01 = 1\%$ than 0.0038, the $\text{Prob}(T \geqslant 15)$; so we are erring 'on the safe side'! The practice of using best conservative critical values is adopted in *EST* and virtually all other books of statistics tables as well as in the tables in this book.

Power of tests

We shall often present more than one test for a given problem. How then do you, the reader, decide which to use in practice? We shall discuss the pros and cons of the tests as we go along. For example, one test may be particularly quick and simple to execute whereas a second, more complicated, test may be better able to diagnose when H_1 is true, i.e. it may be more *sensitive* or *powerful*. Whether one goes for the simple test or instead decides that the extra sensitivity or power justifies the harder work needed for the second test will be a matter of personal taste and circumstances.

Exact power comparisons are usually difficult or impossible, so we will usually make only qualitative comments. However, in some cases

approximate power calculations have been done (usually from computer simulations) and then we will be able to be more specific.

Finally, we shall go into a little more detail on some problems of interpretation and application of statistical tests. We have already mentioned, near the beginning of the chapter, that difficulties involved with clearly defining relevant populations rightly cause concern to some statisticians. Presuming that these difficulties can be overcome, an even nastier logical obstacle awaits us; namely, that the *null hypotheses are virtually never true*!

For illustration, let us consider again the example of the car polishes. We established earlier that the appropriate null hypothesis was

H_0: The two polishes are equally good.

'Equally good'? *Exactly* equally good? Even if the two polishes are made by the same manufacturer, there are bound to be some differences in chemical formulations and the conditions under which they were made, and consequently in their properties. The differences may be substantial, or they may be slight, but nonetheless differences there will be.

Further, again consider the four examples at the beginning of the chapter. The appropriate null hypotheses are as follows:

(a) There are no differences between the average yields from four different types of wheat. *No* differences? That implies they are all *precisely* the same. How likely is that?
(b) A new medical treatment is exactly as effective as the existing treatment. Perhaps the only way this can be envisaged as being possible is if both treatments were entirely useless, which is not a very interesting situation!
(c) Urban children are just as healthy (or unhealthy) as rural children. Apart from the supreme difficulties of uniquely defining what we mean by urban and rural and of measuring 'health', could anybody really even consider the possibility of exact equality of 'healthiness' in the two groups?
(d) There is no connection between sales of luxury goods and unemployment. No connection? Do we really mean none at all?

Put in more formal terms, null hypotheses are almost always expressed in terms of an exact identity of two or more distributions, or at least identity of various aspects of them; or, when just one population is the subject of study, H_0 usually proposes a specific value for some characteristic such as the mean or median, or an exact specification of that population's distribution. Consider the standard textbook example of testing a coin for unbiasedness or fairness, usually expressed in terms of H_0: $p=\frac{1}{2}$, where p is the probability of the coin turning up heads when tossed. There are, incidentally, plenty of problems here of the type discussed in Chapter 1 regarding definitions, stability, etc. As the coin is repeatedly tossed, its properties will change – it will become worn, quite possibly more on one side than the other, and it may pick up some dirt; thus p could keep changing. In any case, is it really true that the outcome is wholly independent of *how* the coin is tossed, and *who* tosses it? But, setting all that aside, and supposing that the event of

throwing a head really does have a meaningful probability p, is it really feasible that p could be equal to 0.5? That means *precisely* 0.5 and not 0.499 999 9999 or 0.500 000 0001 or anything else. Also, we have only used 10 decimal places here, whereas p (being a *continuous* quantity) should strictly be expressed to infinitely many decimal places for complete precision; and 0.500 000 00 . . . is only one possibility out of 'infinitely many'. Realizing this, would you bet on H_0 being true for any particular coin.

Having become aware of such uncomfortable facts of life, some statisticians have come to the rather extreme conclusion that statistical tests have no value. The authors of this book, along with the majority of researchers, do not concur with this extreme attitude. However, we do believe that users of tests should be a lot more aware of such conceptual problems, and their consequences, than is usually the case. What are the implications?

First, you must *never* make the common mistake of inferring from a non-significant result in a test that H_0 is somehow 'proved' or even supported in any positive way. For, as we have indicated, the chances are that you will never come across a true null hypothesis in your lifetime! This is why we were careful earlier in the chapter to talk of 'not rejecting' H_0 when we obtain a non-significant result, as opposed to 'accepting it'. It also follows that, when we are carrying out a two-sided test, the alternative hypothesis is essentially *always* true (we do not need any data at all to establish that!). Of course, in the one-sided case it is quite feasible for *both* H_0 and H_1 to be untrue – the true state can be the other side of H_0 from H_1; however, let us just consider two-sided cases for now.

The vital point to realize is that a two-sided H_1 contains absolutely every possibility other than the single case comprising H_0; in particular, it includes cases very close to H_0, and it includes everything in between. With this more refined way of looking at the situation, what is it that a statistical test does, and what do we want it to do? Bearing in mind what we have just been saying, what we want is for the test to tell us when H_0 is *substantially* in error. Very small differences between yields from various types of wheat, or between medical treatments, or between car polishes, are (usually) unimportant for us to know about. We certainly do want to know about larger differences; and the larger the differences, the more important it is that we know about them. This idea reflects very nicely what a test does. It is a good thing that generally we do not want to know about very small differences, because normally a test will not be able to tell us about them, for the distribution of the test statistic T will then be only minutely different from its null distribution. But if the differences are substantial, the distribution of T (and hence its behaviour) will be *considerably* different from the null distribution, and this will generally considerably increase the chance of T falling into the critical region, i.e. giving a significant result. The practical effect of all this is that, in some quite unquantifiable way, significant results from tests will tend to correspond to relatively substantial (and perhaps important) differences from H_0; i.e. tests will usually do just the kind of thing that we want them to do.

But beware! There is a way to upset this happy state of affairs; and that, believe it or not, is to collect a *lot* of data. Now usually, of course, it is imagined that the more information we collect, the better able we are to get the 'right' answer; and, strictly speaking, that is also true with statistical tests. The unusual aspect of tests is that, as we have already established, we know what the 'right' answer is even before we look at any data. The 'right' answer is that H_0 is false and H_1 (at least,

if it is two-sided) is true! So (paradoxically) a very large amount of data can make a test more sensitive (at least from a practical point of view) than does a moderate amount of data. A large amount of data makes it very likely that a test will give a significant result, irrespective of whether the difference from H_0 is large or small, whereas using the test with moderate amounts of data will tend to give significant results when the difference from H_0 is important and non-significant results when the difference is unimportant.

This is why, in spite of the basic logical difficulties, tests have been found so useful over the years by countless users of statistical techniques. For reasons of economy of time and effort, amounts of data used in tests have tended to be 'moderate' rather than 'very large'. So, maybe more by luck than by judgement, users of tests have found themselves getting what they perceive to be useful and 'correct' results from tests, i.e. correct from the point of view of practical importance, as opposed to correct from the point of view of the formal definitions of the hypotheses.

Nobody can define exactly what is meant by terms as 'moderate' or 'very large', so we will not try. The sort of sample sizes that we use in the illustrations in this book are typical of standard practice, which has stood up well to the test of time. A moral that results from this discussion is that you should not decry the use of small samples nor, following a related argument, the use of relatively unsophisticated tests. Remember that a significant result using a small sample and/or an unsophisticated test will, far more often than not, indicate a more important difference from H_0 than a result of the same significance using a larger sample and/or a more sophisticated test!

Conclusion

You may have found some of the discussion in this chapter quite difficult to take in. That would not be surprising, for several of the points we have made here are poorly understood by many users of statistical methods, with the result that such users tend to over- or under-estimate the extent of the evidence in their data. We would recommend that you now read the next one or two chapters, and then come back and reread this chapter a little later. You will then find these important ideas taking shape and meaning much more than they may have done at the first reading.

PART II

Tests involving one sample

The two chapters in this part of the book are both concerned with investigating a single population, as opposed to comparing two or more populations. In Chapter 3 we test hypotheses about the median of the population, and two tests are presented for this situation. First, we return to the sign test which, as we saw in Chapter 2, is extremely simple to apply and needs only minimal assumptions about the underlying population. The other test, Wilcoxon's signed-rank test, uses more detailed information from the sample and so naturally is more powerful than the sign test; however, the conditions for its validity are more restrictive than with the sign test.

Chapter 4 deals with a quite different one-sample problem, usually called 'goodness-of-fit'. In this case the hypotheses being tested concern the nature of the whole population distribution rather than just the median. Again, two tests are presented. One of these tests, the Kolmogorov–Smirnov test, is particularly appropriate for continuous populations, while the other one, the chi-squared test, is more convenient with discrete populations. However, as we shall see, both tests can be used for either type of population. More important differences between these two tests relate to the kind of null hypothesis being tested. If H_0 proposes a completely specified distribution (for example, that the population is normally distributed with mean 10 and standard deviation 2), then the Kolmogorov–Smirnov test is the more powerful of the two – though it is more time-consuming to apply. However, if H_0 just specifies that the population has one of a family of distributions (for example, that it is some unspecified binomial distribution) then, in general, only the chi-squared test can be used.

3

One-sample location tests

THE SIGN TEST

Introduction

The sign test is almost certainly the oldest of all formal statistical tests as there is published evidence of its use long ago by J. Arbuthnott (1710)! It exists in several forms, but the version presented here is the one already introduced in Chapter 2, i.e. a test for the median value of a population. However, in this section we examine the procedure in more detail then previously.

PRACTICAL EXAMPLE

The managing director of a mail-order company is concerned about the possible slow progress of orders through his administration department. He tells the administration manager to ensure that at least half of the orders received are processed within one day (eight working hours). Some weeks later he times 18 orders selected at random to check whether his demand has been met.

HYPOTHESES

Being somewhat distrustful of his administration manager's ability, the managing director's main aim in performing the test is to check for evidence that his instructions have *not* been met. This happens when *less* than half of the order are being processed within eight working hours, or in other words, that the median processing time ϕ exceeds 8 h. The hypotheses may be summarized as follows:

$$H_0: \phi = 8 \text{ h.}$$
$$H_1: \phi > 8 \text{ h.}$$

Clearly, H_1 indicates that the test is *one-sided*. Therefore, we shall initially present the details appropriate for such a one-sided test; the amendments needed for the two-sided version of the test will be given later.

DATA

The times spent by the 18 orders in the administration process were recorded as:

16 h 30 min	14 h 00 min	5 h 40 min	9 h 10 min	11 h 45 min
4 h 20 min	7 h 55 min	10 h 15 min	7 h 45 min	16 h 05 min
10 h 05 min	7 h 30 min	9 h 15 min	11 h 55 min	9 h 25 min
10 h 35 min	8 h 20 min	10 h 10 min		

STATISTIC

The test is called the *sign* test because the statistic used is computed from data that are in (or have been reduced to) the form of '+' and '−' signs. So here let us write a '+' for all those times greater than 8 h and a '−' for those times less than 8 h. The above data can thus be represented as:

+	+	−	+	+	−
−	+	−	+	+	−
+	+	+	+	+	+

It is worth mentioning here that, once he had definitely decided to use the sign test, the managing director could have saved himself some trouble in collecting these data. Rather than bothering to produce the detailed information above, all he really had to do was to check the progress of each order at precisely eight working hours (i.e. at the same time next day) after receipt at administration; those which had then been processed through administration would be recorded as '−'s, and the rest as '+'s.

The test statistic, which we will denote by S, may be either the number of '+' signs or the number of '−' signs, according to context. Clearly, if H_0 is true we would expect the number of '+'s and '−'s to be roughly equal, whereas if H_1: $\phi > 8$ h is true, we would expect a relatively large number of '+'s and correspondingly few '−'s. The table of critical values (Table C at the back of the book) is given in terms of *small* values of the statistic, and therefore here we need to define S as the number of '−' signs.

CRITICAL REGIONS

With S defined in this way, the smaller the value of S then the more evidence there is to support H_1. Thus critical regions are of the form:

$$S \leqslant \text{critical value}$$

Following the convention we introduced in Chapter 2, we denote significance levels relevant to one-sided tests by α_1 (and those relevant to

two-sided tests by α_2). Thus from Table C (or the more detailed table on page 29 of *EST*) we see that, with a sample size $n=18$, the critical regions are $S \leqslant 5$ and $S \leqslant 3$ for significance levels $\alpha_1=5\%$ and 1% respectively.

SOLUTION

The 18 signs include only five minuses and so $S=5$. This result is significant at the $\alpha_1=5\%$ level, and consequently the managing director has obtained some evidence in favour of H_1, i.e. it seems quite likely that his requirement is not being met by the administration department.

DISCUSSION

The sign test is clearly an extremely quick and simple one-sample test for the median value. The fact that the numerical data are dispensed with and only the limited information, as summarized by the '+' and '−' signs, is retained, does mean, however, that relatively large differences between the numbers of '+' and '−' signs are necessary for the test to produce significant results, particularly if the sample is small. In other words, the test tends to support H_1 only when it is intuitively obvious from the data that H_1 is probably true! However, this criticism becomes less true for larger sample sizes, as can be seen by inspection of the more extensive table of critical values in *EST*; for example, with $n=100$, a 41/59 split of signs is sufficient to reject H_0 at the $\alpha_1=5\%$ significance level.

It is worth mentioning at this point that, in order to produce tests which are truly distribution-free (i.e. make no detailed assumptions about the form of population distribution), it turns out to be *necessary* to use statistics that do not depend on the actual numerical values in the data, but instead rely on a cruder summary of them (such as '+' and '−' signs in the sign test). In two-sample and other situations to be covered in later chapters, the reduction in information needed is not so drastic as in the one-sample case, so that the resulting test procedures tend to have rather greater power and sensitivity than the sign test. However, it will still always be the case that actual numerical values from the data will *never* be used in the formation of the test statistics in order that the 'distribution-freeness' property be retained. A by-product of this policy is that the resulting test statistics will generally be rather quicker and easier to calculate than their equivalents in classical statistical methods.

One-sided and two-sided tests

As we have said, a choice has to be made between defining S as the number of '+' signs or as the number of '−' signs. For a one-sided test, the appropriate choice is the one *which would be expected to have the smaller value when H_1 is true*. For a two-sided test, we define S simply as whichever of the two choices actually gives the smaller number. Thus in the above example, had we been testing H_0: $\phi=8$ against the two-sided alternative H_1: $\phi \neq 8$, S would take the value 5 *either* if there were 13 '+'s and 5 '−'s (as in our problem) or if there were 5 '+'s and 13 '−'s. But in the one-sided case above, had there been 13 '−'s and 5 '+'s, the test

value would have been $S=13$ (since in that problem S is equal to the number of '−' signs) and, of course, this would have provided no evidence in favour of the stated alternative H_1: $\phi>8$.

So in the two-sided case, the critical regions are of the form

$$S \leqslant \text{critical value}$$

where S is *either* the number of '+' signs *or* the number of '−' signs, whichever is the smaller. Thus, for example, either 3 '−'s or 3 '+'s would lead to the rejection of H_0: $\phi=8$ in favour of H_1: $\phi\neq 8$ at the $\alpha_2=2\%$ level of significance.

What happens if any observations are equal to the hypothesized value of ϕ?

If we are sampling from a continuous distribution and recording observations to sufficient accuracy, it is extremely unlikely that any observations will be *exactly* equal to the hypothesized value (8 h 00 min in the above example). However, usually data cannot be measured to such extreme accuracy and, in any case, it is often convenient to round values to relatively large units even if greater accuracy is possible. (For example, the above data have clearly been rounded to the nearest 5 min, although it would surely have been possible to measure the times more accurately than that.) It is then quite likely that one or two observations might be recorded as equal to the hypothesized value. The usual procedure with the sign test is to discard such values from the sample and to reduce the sample size accordingly. So, for example, if the first observation in the above sample had been 8 h 00 min instead of 16 h 30 min, we would then have considered the sample to be of size 17, consisting of 5 '−'s and 12 '+'s (which would still have produced evidence significant at the $\alpha_1=5\%$ level in favour of H_1: $\phi>8$). This procedure is quite satisfactory unless there are *several* observations at the crucial value; then there is a case for retaining the original sample size and sharing out these awkward observations equally, i.e. regarding half of them as '−'s and half as '+'s.

Null distributions

The null distribution of S, which we recall is its distribution when H_0 is true, has already been discussed in Chapter 2. However, for completeness, here are the details again.

If H_0 is true, then every one of the n observations in the sample gives rise to a '+' or a '−' sign with probability $\frac{1}{2}$ each. Now as all observations are *independent* of each other (which follows from the concept of a random sample), we know that the total number of '−' signs (or of '+' signs) has the binomial distribution with parameters n (the sample size) and $p=\frac{1}{2}$ (see page 31).

Thus if S represents the statistic used in the above practical problem (i.e. the number of '−' signs in a sample of $n=18$ signs), then all probabilities to do with S may be read directly from the table given on page 32 (or from pages 10–11 in *EST*). In particular, recalling that our critical regions are of the form $S \leqslant$ critical value, we quote the following:

$$\text{Prob}(S\leqslant 3) = 0.0038$$

$$\text{Prob}(S\leqslant 4) = 0.0154$$

$$\text{Prob}(S \leqslant 5) = 0.0481$$

$$\text{Prob}(S \leqslant 6) = 0.1189$$

As we mentioned in Chapter 2, with a discrete-valued statistic it is not generally possible to find critical regions corresponding exactly to the nominal 5% or 1% figures traditionally used for significance levels. The convention that is commonly adopted, and which is used in our tables (and in *EST*), is to indicate the largest critical region whose significance level is no greater than the nominal figure (the so-called best conservative critical region). Thus, using the above probabilities, we see that the best conservative critical region corresponding to $\alpha_1=1\%$ is $S \leqslant 3$; in fact, it corresponds to a significance level of 0.38% whereas the next larger critical region, $S \leqslant 4$, corresponds to a significance level of 1.54%, which is *larger* than the nominal 1%. Similarly, the best conservative critical region for a nominal significance level of 5% is $S \leqslant 5$.

In the two-sided case, S is defined as the number of either '−' or '+' signs, whichever is the smaller: for example, that $S \leqslant 3$ corresponds to the number of '−' (or '+') signs being 0, 1, 2, 3, 15, 16, 17 or 18. Since binomial distributions with $p=\frac{1}{2}$ are symmetric, it is clear that $\text{Prob}(S \leqslant 3)$ is now $2\times 0.0038=0.0076$; similarly all other probabilities in the above list need doubling. You will then see the justification for using the same list of critical values in Table C for both cases, but with the α_2 significance levels being *double* those of α_1.

Large sample sizes

If n exceeds 50, critical values are not obtainable from Table C. However, it is easy to see how approximate critical values may be calculated. For large n, we have seen in Chapter 1 that binomial distributions are well approximated by normal distributions, particularly when $p=\frac{1}{2}$. In the one-sided case, we know that the null distribution of S is the binomial distribution with parameters n and $p=\frac{1}{2}$. This distribution has mean $\mu=np=\frac{1}{2}n$ and standard deviation $\sigma = \sqrt{np(1-p)} = \sqrt{n/4} = \frac{1}{2}\sqrt{n}$.

We can therefore obtain good approximate critical values using the integer part of $\mu - z\sigma - \frac{1}{2} = \frac{1}{2}n - \frac{1}{2}z\sqrt{n} - \frac{1}{2}$, where z is the required percentage point of the standard normal distribution (from Table A) and where we have used the continuity correction (see Ch. 1). For $\alpha_1=5\%$ and 1% the respective values of z are 1.6449 and 2.3263. So, for example, with $n=50$ and $\alpha_1=5\%$, the approximate critical value of S is the integer part of $\frac{1}{2}\times 50 - \frac{1}{2}\times 1.6449\sqrt{50} - \frac{1}{2} = 18.68$, i.e. 18, which is the same as the entry in Table C.

Critical values for the two-sided test are obtained similarly using the appropriate values of z; for example, with $\alpha_2=5\%$ and 1%, z is 1.96 and 2.5758 respectively.

THE WILCOXON SIGNED-RANK TEST

Introduction

We have already remarked on the fact that the representation of the data in the sign test by just '+'s and '−'s is a rather drastic reduction of the