$$-0{\cdot}6111 \pm 2{\cdot}145 \times 0{\cdot}3879$$
$$= -1{\cdot}4 \text{ and } 0{\cdot}2.$$

In any of these approaches, if the sample means are sufficiently close and the sample sizes are sufficiently large, the confidence interval for the difference in means may be narrow enough to allow one to conclude that the means are effectively equal for all practical purposes. The investigator must, however, be careful not to conclude that the two populations are identical unless there is good reason to believe that the variances are also equal.

4.4 **Inferences from proportions**

The sampling error of a proportion

This has already been fully discussed in §3.6. If individuals in an infinitely large population are classified into two types A and B, with probabilities π and $1 - \pi$, the number r of individuals of type A in a random sample of size n follows a binomial distribution. We shall now apply the results of §4.2 to prove the formulae previously given for the mean and variance of r.

Suppose we define a quantitative variable x, which takes the value 1 for each A individual and 0 for each B. We may think of x as a score attached to each member of the population. The point of doing this is that, in a sample of n consisting of r As and $n - r$ Bs,

$$\sum x = (r \times 1) + [(n - r) \times 0]$$
$$= r$$

and

$$\bar{x} = r/n, \ = p \text{ in the notation of §3.6.}$$

The sample proportion p may therefore be identified with the sample mean of x, and to study the sampling variation of p we can apply the general results established in §4.2. We shall need to know the population mean and standard deviation of x. From first principles these are

$$\mathrm{E}(x) = (\pi \times 1) + [(1 - \pi) \times 0] \qquad (4.12)$$
$$= \pi,$$

and

$$\mathrm{var}(x) = \mathrm{E}(x^2) - [\mathrm{E}(x)]^2$$
$$= (\pi \times 1^2) + [(1 - \pi) \times 0^2] - \pi^2$$
$$\pi(1 - \pi).$$

From (4.1),

$$\mathrm{var}(\bar{x}) = \frac{\pi(1-\pi)}{n}. \tag{4.13}$$

Writing (4.12) and (4.13) in terms of p rather than x and $\bar{x}$, we have

$$\mathrm{E}(p) = \pi$$

and

$$\mathrm{var}(p) = \frac{\pi(1-\pi)}{n},$$

as in (3.15) and (3.16). Since $r = np$,

$$\mathrm{E}(r) = n\pi$$

and

$$\mathrm{var}(r) = n\pi(1-\pi),$$

as in (3.13) and (3.14).

One more result may be taken from §4.2. As n approaches infinity, the distribution of $\bar{x}$ (that is, of p) approaches the normal distribution, with the corresponding mean and variance. The increasing symmetry has already been noted in §3.6.

Inferences from the proportion in a sample

Suppose that, in a large population, individuals drawn at random have an unknown probability π of being of type A. In a random sample of n individuals, a proportion p $(= r/n)$ are of type A. What can be said about π?

Suppose first that we wish to test a null hypothesis specifying that π is equal to some value π_0. On this hypothesis, the number of type A individuals, r, found in repeated random samples of size n would follow a binomial distribution. To express the departure of any observed value, r, from its expected value, $n\pi_0$, we could state the extent to which r falls into either of the tails of its sampling distribution. As in §4.1 this extent could be measured by calculating the probability in the tail area. The situation is a little different here because of the discreteness of the distribution of r. Do we calculate the probability of obtaining a larger deviation than that observed, $r - n\pi_0$, or the probability of a deviation at least as great? Since we are saying something about the degree of surprise elicited by a certain observed result, it seems reasonable to include the probability of this result in the summation. Thus, if $r > n\pi_0$ and the probabilities in the binomial distribution with parameters π_0 and n are $P_0, P_1, \ldots, P_n$, the P value for a one-sided test will be

$$P_+ = P_r + P_{r+1} + \ldots + P_n.$$

For a two-sided test we could add the probabilities of deviations at least as large as that observed, in the other direction. The P value for the other tail is

$$P_- = P_r' + P_{r'-1} + \ldots + P_0,$$

where r' is equal to $2n\pi_0 - r$ if this is an integer, and the highest integer less than this quantity otherwise. The P value for the two-sided test is then $P = P_- + P_+$.

For example, if $r = 8, n = 10$ and $\pi_0 = \frac{1}{2}$,

$$P_+ = P_8 + P_9 + P_{10}$$

and

$$P_- = P_2 + P_1 + P_0.$$

If $r = 17, n = 20$ and $\pi_0 = \frac{1}{3}$,

$$P_+ = P_{17} + P_{18} + P_{19} + P_{20}$$
$$P_- = 0.$$

If $r = 15, n = 20$ and $\pi_0 = 0.42$,

$$P_+ = P_{15} + P_{16} + P_{17} + P_{18} + P_{19} + P_{20}$$
$$P_- = P_1 + P_0.$$

An alternative, and perhaps preferable, approach is to regard a two-sided test at level α as being essentially the combination of two one-sided tests each at level $\frac{1}{2}\alpha$. The two-sided P value is then obtained simply by doubling the one-sided value.

Considerable simplification is achieved by approximating to the binomial distribution by the normal (§3.8). On the null hypothesis

$$\frac{r - n\pi_0}{\sqrt{[n\pi_0(1 - \pi_0)]}}$$

is approximately a standardized normal deviate. Using the continuity correction, (p.80), the tail area required in the significance test is approximated by the area beyond a standardized normal deviate

$$z = \frac{|r - n\pi_0| - \frac{1}{2}}{\sqrt{[n\pi_0(1 - \pi_0)]}}, \qquad (4.14)$$

and the result will be significant at, say, the 5% level if this probability is less than 0·05.

Example 4.6

In a clinical trial to compare the effectiveness of two analgesic drugs, X and Y, each of 100 patients receives X for a period of 1 week and Y for another week, the order of administration being determined randomly. Each patient then states a preference for one of the two drugs. Sixty-five patients prefer X and 35 prefer Y. Is this strong evidence for the view that, in the long run, more patients prefer X than Y?

Test the null hypothesis that the preferences form a random series in which the probability of an X preference is $\frac{1}{2}$. This would be true if X and Y were equally effective in all respects affecting the patients' judgements. The standard error of r is

$$\sqrt{(100 \times \tfrac{1}{2} \times \tfrac{1}{2})} = \sqrt{25} = 5.$$

The observed deviation, $r - n\pi_0$, is

$$65 - 50 = 15.$$

With continuity correction, the standardized normal deviate is $(15 - \frac{1}{2}/5 = 2{\cdot}90$. Without continuity correction, the value would have been $15/5 = 3{\cdot}00$, a rather trivial difference. In this case the continuity correction could have been ignored. The normal tail area for $z = 2{\cdot}90$ is 0·0037; the departure from the null hypothesis is highly significant, and the evidence in favour of X is strong. The exact value of P, from the binomial distribution, is 0·0035, very close to the normal approximation.

The 95% confidence limits for π are the two values, π_L and π_U, for which the observed value of r is just significant on a one-sided test at the $2\frac{1}{2}\%$ level (Fig. 4.8). These values may be obtained fairly readily from tables of the binomial distribution, and are tabulated in the Geigy Scientific Tables (1982, Vol. 2, pp. 89–102). They may be obtained also from Fisher and Yates (1963, Table VIII1).

The normal approximation may be used in a number of ways.

1 The tail areas could be estimated from (4.14). Thus, for 95% confidence limits, approximations to π_L and π_U are given by the formulae

$$\frac{r - n\pi_L - \frac{1}{2}}{\sqrt{[n\pi_L(1 - \pi_L)]}} = 1{\cdot}96$$

and

$$\frac{r - n\pi_U + \frac{1}{2}}{\sqrt{[n\pi_U(1 - \pi_U)]}} = -1{\cdot}96.$$

2 In method **1**, if n is large, the continuity correction of $\frac{1}{2}$ may be omitted.

Method **1** involves the solution of a quadratic equation for each of π_L and π_U; method **2** involves a single quadratic equation. A further simplification is as follows.

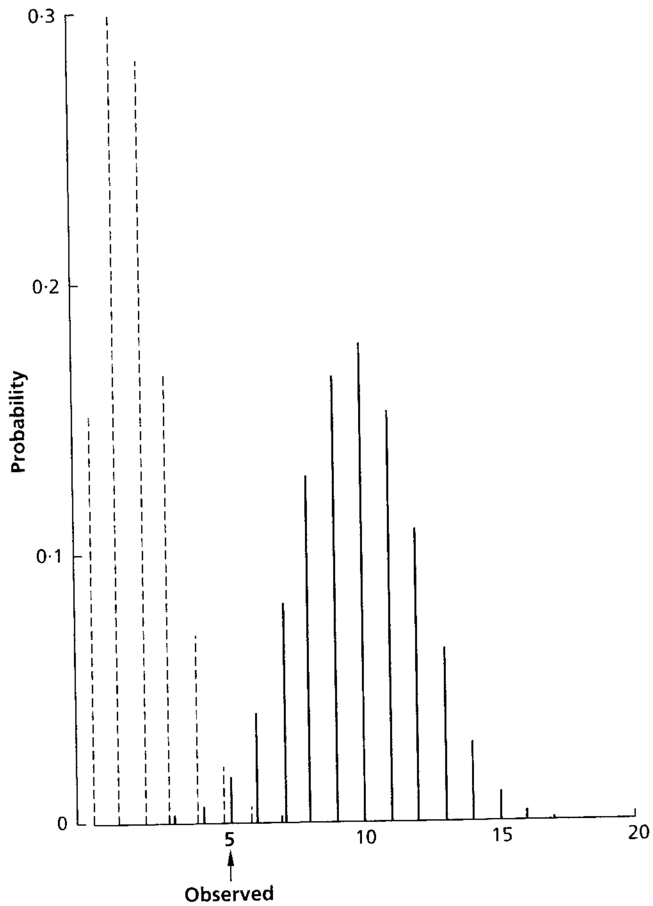

Fig. 4.8 Binomial distributions illustrating the 95% confidence limits for the parameter π based on a sample with five individuals of a certain type out of 20. For $\pi = \pi_L = 0{\cdot}09$ the probability of 5 or more is 0·025; for $\pi = \pi_U = 0{\cdot}49$ the probability of 5 or less is 0·025. (– – –) $\pi_L = 0{\cdot}09$; (———) $\pi_U = 0{\cdot}49$.

3 Replace $\pi_L(1 - \pi_L)$ and $\pi_U(1 - \pi_U)$ by $p(1 - p)$. This is not too drastic a step, as $p(1 - p)$ changes rather slowly with changes in p, particularly for values of p near $\frac{1}{2}$. Ignoring the continuity correction, as in **2**, we have the most frequently used approximation to the 95% confidence limits:

$$p \pm 1{\cdot}96\sqrt{(pq/n)},$$

where, as usual, $q = 1 - p$. The simplification here is due to the replacement of the standard error of p, which involves the unknown value π, by the approximate form $\sqrt{(pq/n)}$, which can be calculated entirely from known quantities.

Newcombe (1998a), in an extensive study of various methods, recommends method **2**. Method **3** has a coverage probability well below the nominal 95%, tending to shorten the interval unduly in the direction from p towards $\frac{1}{2}$.

Exact limits require the calculation of tail-area probabilities for the binomial distribution. The lower limit π_L, for instance, is the solution of

$$\sum_{j=r}^{n} \binom{n}{j} \pi_L^j (1-\pi_L)^{n-j} = 0{\cdot}025. \qquad (4.15)$$

The binomial tail areas can be obtained from many statistical packages. Alternatively, they can be obtained from tables of another important distribution: the F distribution. This is described in more detail in §5.1. We note here that the distribution is indexed by two parameters, ν_1 and ν_2, called *degrees of freedom* (DF). Table A4 shows the value of F exceeded with probability P for various combinations of ν_1, ν_2 and P, denoted by F_{P,ν_1,ν_2}.

It can be shown that the left-hand side of (4.15) equals the probability that a variable distributed as F with $2n-2r+2$ and $2r$ degrees of freedom exceeds $r(1-\pi_L)/(n-r+1)\pi_L$. Therefore

$$\frac{r(1-\pi_L)}{(n-r+1)\pi_L} = F_{0{\cdot}025,\,2n-2r+2,\,2r}.$$

That is,

$$\pi_L = \frac{r}{r+(n-r+1)F_{0{\cdot}025,\,2n-2r+2,\,2r}}$$

and similarly,

$$\pi_U = \frac{r+1}{r+1+(n-r)F^{-1}_{0{\cdot}025,\,2r+2,\,2n-2r}} \qquad (4.16)$$

(Miettinen, 1970).

Example 4.6, continued

With $n = 100$, $p = 0{\cdot}65$, the exact 95% confidence limits are found to be 0·548 and 0·743.

Method **1** will be found to give 0·548 and 0·741, method **2** gives 0·552 and 0·736, and method **3** gives

$$0{\cdot}65 \pm 1{\cdot}96\sqrt{\left[\frac{(0{\cdot}65)(0{\cdot}35)}{100}\right]}$$
$$= 0{\cdot}65 \pm (1{\cdot}96)(0{\cdot}0477)$$
$$= 0{\cdot}557 \text{ and } 0{\cdot}743$$

In this example method **3** is quite adequate.

Example 4.7

As a contrasting example with small numbers, suppose $n = 20$ and $p = 0{\cdot}25$. The exact 95% confidence limits are 0·087 and 0·491 (see Fig. 4.8).

Method **1** gives 0·096 and 0·494, method **2** gives 0·112 and 0·469, method **3** gives 0·060 and 0·440. Method **3** is clearly less appropriate here than in Example 4.6. In general, method **3** should be avoided if either np or $n(1-p)$ is small (say, less than 10). Methods **2** and **3** may also be unreliable when either $n\pi_L$ or $n(1-\pi_U)$ is less than 5. In this case, $n\pi_L$ is about 2 so it is not surprising that the lower confidence limit is not too well approximated by the normal approximation.

Example 4.8

As an example of finding the exact limits using the F distribution, consider the data of Fig. 4.8, where $r = 5$ and $n = 20$. Then,

$$\pi_L = 5/(5 + 16F_{0{\cdot}025,\,32,\,10})$$

and

$$\pi_U = 6/(6 + 15F^{-1}_{0{\cdot}025,\,12,\,30}).$$

The value of $F_{0{\cdot}025,\,32,\,10}$ can be obtained by interpolation in Table A4 where interpolation is linear in the reciprocal of the degrees of freedom. Thus,

$$F_{0{\cdot}025,\,32,\,10} = 3{\cdot}37 - (3{\cdot}37 - 3{\cdot}08) \times \left(\frac{1}{24} - \frac{1}{32}\right) \Big/ \frac{1}{24}$$
$$= 3{\cdot}30.$$

Therefore $\pi_L = 0{\cdot}0865$ and, since $F_{0{\cdot}025,\,12,\,30} = 2{\cdot}41$, $\pi_U = 0{\cdot}4908$. In this example, interpolation was required for one of the F values and then only in the first of the degrees of freedom, but in general interpolation would be required for both F values and in both the degrees of freedom. This is tedious and can be avoided by using a method based on the normal approximation except when this method gives values such that either $n\pi_L$ or $n(1-\pi_U)$ is small (say, less than 5).

It was remarked earlier that the discreteness of the distribution of r made inferences from proportions a little different from those based on a variable with a continuous distribution, and we now discuss these differences. For a continuous variable an exact significance test would give the result $P < 0{\cdot}05$ for exactly 5% of random samples drawn from a population in which the null hypothesis were true, and a 95% confidence interval would contain the population value of the estimated parameter for exactly 95% of random samples. Neither of these properties is generally true for a discrete variable. Consider a binomial variable from a distribution with $n = 10$ and $\pi = 0{\cdot}5$ (Table 3.4, p. 76). Using the exact test, for the hypothesis that $\pi = 0{\cdot}5$, significance at the 5% level is found only for $r = 0$, 1, 9 or 10 and the probability of one or other of these values is 0·022.

Therefore, a result significant at the 5% level would be found in only 2·2% of random samples if the null hypothesis were true. This causes no difficulty if the precise level of P is stated. Thus if $r = 1$ we have $P = 0{\cdot}022$, and a result significant at a level of 0·022 or less would occur in exactly 2·2% of random samples. The normal approximation with continuity correction is then the best approximate test, giving, in this case, $P = 0{\cdot}027$.

A similar situation arises with the confidence interval. The exact confidence limits for the binomial parameter are conservative in the sense that the probability of including the true value is *at least* as great as the nominal confidence coefficient. This fact arises from the debatable decision to include the observed value in the calculation of tail-area probabilities. The limits are termed 'exact' because they are obtained from exact calculations of the binomial distribution, rather than from an approximation, but not because the confidence coefficient is achieved exactly. This problem cannot be resolved, in the same way as for the significance test, by changing the confidence coefficient. First, this is difficult to do but, secondly and more importantly, whilst for a significance test it is desirable to estimate P as precisely as possible, in the confidence interval approach it is perfectly reasonable to specify the confidence coefficient in advance at some conventional value, such as 95%. The approximate limits using the continuity correction also tend to be conservative. The limits obtained by methods **2** and **3**, however, which ignore the continuity correction, will tend to have a probability of inclusion nearer to the nominal value. This suggests that the neglect of the continuity correction is not a serious matter, and may, indeed, be an advantage.

The problems discussed above, due to the discreteness of the distribution, have caused much controversy in the statistical literature, particularly with the analysis of data collected to compare two proportions, to be discussed in §4.5. One approach, suggested by Lancaster (1952, 1961), is to use *mid-P* values, and this approach has been advocated more widely recently (Williams, 1988; Barnard, 1989; Hirji, 1991; Upton, 1992; Berry & Armitage, 1995; Newcombe, 1998a). The mid-P value for a one-sided test is obtained by including in the tail only one-half of the probability of the observed sample. Thus, for a binomial sample with r observed out of n, where $r > n\pi_0$, the one-sided mid-P value for testing the hypothesis that $\pi = \pi_0$ will be

$$\text{mid-}P_{+} = \tfrac{1}{2}P_r + P_{r+1} + \ldots + P_n.$$

It has to be noted that the mid-P value is not the probability of obtaining a significant result by chance when the null hypothesis is true. Again, consider a binomial variable from a distribution with $n = 10$ and $\pi = 0{\cdot}5$ (Table 3.4, p. 76). For the hypothesis that $\pi = 0{\cdot}5$, a mid-P value less than 0·05 would be found only for $r = 0, 1, 9$ or 10, since the mid-P value for $r = 2$ is 2[0·0010

$+0{\cdot}0098 + \frac{1}{2}(0{\cdot}0439)] = 0{\cdot}0655$, and the probability of one or other of these values is 0·022.

Barnard (1989) has recommended quoting both the *P* and the mid-*P* values, on the basis that the former is a measure of the statistical significance when the data under analysis are judged alone, whereas the latter is the appropriate measure of the strength of evidence against the hypothesis under test to be used in combination with evidence from other studies. This arises because the mid-*P* value has the desirable feature that, when the null hypothesis is true, its average value is 0·5 and this property makes it particularly suitable as a measure to be used when combining results from several studies in making an overall assessment (meta-analysis; §18.10). Since it is rare that the results of a single study are used without support from other studies, our recommendation is also to give both the *P* and mid-*P* values, but to give more emphasis to the latter.

Corresponding to mid-*P* values are mid-*P* confidence limits, calculated as those values which, if taken as the null hypothesis value, give a corresponding mid-*P* value; that is, the 95% limits correspond to one-sided mid-*P* values of 0·025.

Where a normal approximation is adequate, *P* values and mid-*P* values correspond to test statistics calculated with and without the correction for continuity, respectively. Correspondingly, confidence intervals and mid-*P* confidence intervals can be based on normal approximations, using and ignoring the continuity correction, respectively. Thus, the mid-*P* confidence limits for a binomial probability would be obtained using method **2** rather than method **1** (p. 115).

Where normal approximations are inadequate, the mid-*P* values are calculated by summing the appropriate probabilities. The mid-*P* limits are more tedious to calculate, as they are not included in standard sets of tables and there is no direct formula corresponding to (4.16). The limits may be obtained fairly readily using a personal computer or programmable calculator by setting up the expression to be evaluated using a general argument, and then by trial and error finding the values that give tails of 0·025.

Example 4.7, continued

The mid-*P* limits are given by

$$P_0 + P_1 + P_2 + P_3 + P_4 + \tfrac{1}{2}P_5 = 0{\cdot}975 \text{ or } 0{\cdot}025,$$

where P_i is the binomial probability (as in (3.12)) for i events with $n = 20$ and $\pi = \pi_L$ or π_U. This expression was set up on a personal computer for general π, and starting with the knowledge that the confidence interval would be slightly narrower than the limits of 0·0865 and 0·4908 found earlier the exact 95% mid-*P* confidence limits were found as 0·098 and 0·470. Method **2** gives the best approximation to these limits but, as noted earlier, the lower confidence limit is less well approximated by the normal approximation, because $n\pi_L$ is only about 2.