4.5 **Comparison of two proportions**

As in the comparison of two means, considered in §4.3, we can distinguish between two situations according to whether individual members of the two samples are or are not paired.

Paired case

Suppose there are N observations in each sample, forming therefore N pairs of observations. Denoting the samples by 1 and 2, and describing each individual as A or not A, there are clearly four types of pairs:

	Sample		Number
Type	1	2	of pairs
1	A	A	k
2	A	Not A	r
3	Not A	A	s
4	Not A	Not A	m

If the number of pairs of the four types are as shown above, another way of exhibiting the same results is in the form of a two-way table:

		Sample 2		
		A	Not A	
Sample 1	A	k	r	$k+r$
	Not A	s	m	$s+m$
		$k+s$	$r+m$	N

The proportions of A individuals in the two samples are $(k+r)/N$ in sample 1 and $(k+s)/N$ in sample 2. We are interested in the difference between the two proportions, which is clearly $(r-s)/N$.

Consider first a significance test. The null hypothesis is that the expectation of $(r-s)/N$ is zero, or in other words that the expectations of r and s are equal. This can conveniently be tested by restricting our attention to the $r+s$ pairs in which the two members are of different types. Denote $r+s$ by n. On the null hypothesis, given n disparate or 'untied' pairs, the number of pairs of type 2 (or, indeed, of type 3) would follow a binomial distribution with a parameter equal to $\frac{1}{2}$. The test therefore follows precisely the methods of §4.4. A large-sample test is obtained by regarding

$$z = \frac{r - \frac{1}{2}n}{\frac{1}{2}\sqrt{n}} \tag{4.17}$$

as a standardized normal deviate. A continuity correction may be applied by reducing the absolute value of $r - \frac{1}{2}n$ by $\frac{1}{2}$. This test is sometimes known as McNemar's test. An alternative form of (4.17), with continuity correction included, is

$$z^2 = \frac{(|r - s| - 1)^2}{r + s}, \tag{4.18}$$

where z^2 may be regarded as a $\chi^2_{(1)}$ variate (§3.8). This is one of the few statistical calculations that really can be done in one's head; see also (5.8).

McNemar's test is based on the normal approximation to the discrete binomial distribution. The situation is similar to that considered in §4.4 except for the complication of the tied pairs, k and m; with the correction for continuity, (4.17) corresponds to (4.14) with $\pi_0 = \frac{1}{2}$. The significance test (4.18) will be satisfactory except for small values of $r + s$ (less than about 10), and in such cases an exact test may be carried out on the $r + s$ untied pairs. For this purpose the F distribution may be used, as in (19.28).

It should be noted that, although the significance test is based entirely on the two frequencies r and s, the estimated difference between the proportions of positives and therefore also its standard error depend also on N. That is, evidence as to the existence of a difference is provided solely by the united pairs; an assessment of the *magnitude* of that difference must allude to the remainder of the data. The distinction between statistical and clinical significance, referred to in §4.1, must be borne in mind.

The calculation of confidence limits for the difference between the two proportions involves accounting for the variation in the number of united pairs, $r + s$. This may be achieved by deriving the standard error from the properties of the multinomial distribution, which is an extension of the binomial distribution when there are more than two classes. Approximate confidence limits for the difference between the two proportions are then given by taking its standard error to be

$$\frac{1}{N}\sqrt{\left[r + s - \frac{(r - s)^2}{N}\right]} \tag{4.19}$$

(Fleiss, 1981), and using the usual normal theory. However, this method is unreliable for small frequencies, when the probability that the parameter is included within the interval tends to be considerably less than the nominal confidence coefficient (Newcombe, 1998c). In extreme cases the limits may fall outside the permissible range of [0, 1]. More satisfactory methods, such as Newcombe's recommended Method 10, require extensive computation, and are best implemented by a computer program such as that included in StatsDirect (1999).

There is a potential discrepancy between the test (4.17) and the confidence limits obtained using (4.19). If the test is just significant at the 5% level, with

$z = 1{\cdot}96$, the lower confidence limit is higher than zero. This arises because of the second term within the square root in (4.19). The discrepancy will be slight except for large differences between r and s.

When data of this type are obtained in a case–control study, emphasis is often directed to the odds ratio (see (4.22)), which can, for paired data, be estimated by the simple ratio r/s (see (19.26)). Tests and confidence limits can be obtained by the standard methods applied to the simple proportion $r/(r+s)$ (see (19.27)).

Example 4.9

Fifty specimens of sputum are each cultured on two different media, A and B, the object being to compare the ability of the two media to detect tubercle bacilli. The results are shown in Table 4.4. The null hypothesis that the media are equally effective is tested by the standardized normal deviate

$$z = \frac{12 - (\frac{1}{2})(14)}{\frac{1}{2}\sqrt{14}} = \frac{5}{1{\cdot}871} = 2{\cdot}67 \quad (P = 0{\cdot}008).$$

There is very little doubt that A is more effective than B. The continuity correction would reduce the normal deviate to $4{\cdot}5/1{\cdot}871 = 2{\cdot}41$, still a significant result ($P = 0{\cdot}016$).

The exact significance level is given by

$$P = 2 \times (\tfrac{1}{2})^{14}[1 + 14 + (14 \times 13/2)] = 0{\cdot}013.$$

Table 4.4 Distribution of 50 specimens of sputum according to results of culture on two media.

Type	Medium A	Medium B	Number of sputa
1	+	+	20
2	+	−	12
3	−	+	2
4	−	−	16
			50

Alternative layout

		Medium B		
		+	−	Total
Medium A	+	20	12	32
	−	2	16	18
Total		22	28	50

The mid-P value is obtained by taking only one-half of the last term in the above expression and is 0·007, close to the approximate value 0·008 obtained above.

The approximate 95% confidence limits for the difference between the proportions of positive sputum on the two media are, from (4.19),

$$\begin{aligned}\frac{(12-2)}{50} \pm \frac{1{\cdot}96\sqrt{(14-10^2/50)}}{50} \\ = 0{\cdot}20 \pm 0{\cdot}14 \\ = 0{\cdot}06 \text{ and } 0{\cdot}34.\end{aligned}$$

More exact limits (Newcombe, 1998c) are 0·06 and 0·33 (StatsDirect, 1999).

Unpaired case: two independent samples

Suppose these are two populations in which the probabilities that an individual shows characteristic A are π_1 and π_2. A random sample of size n_1 from the first population has r_1 members showing the characteristic (and a proportion $p_1 = r_1/n_1$), while the corresponding values for an independent sample from the second population are n_2, r_2, and $p_2 = r_2/n_2$. In the general formulae (4.8) and (4.9),

$$m_1 = \pi_1 \text{ and } v_1 = \pi_1(1-\pi_1)/n_1;$$

$$m_2 = \pi_2 \text{ and } v_2 = \pi_2(1-\pi_2)/n_2.$$

Hence,

$$\left.\begin{aligned}\mathrm{E}(p_1 - p_2) &= \pi_1 - \pi_2 \\ \text{and} \\ \mathrm{var}(p_1 - p_2) &= \frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}\end{aligned}\right\}. \tag{4.20}$$

For confidence limits, π_1 and π_2 are unknown and may be replaced by p_1 and p_2, respectively, to give

$$\mathrm{var}(p_1 - p_2) = \frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}, \tag{4.21}$$

where

$$q_1 = 1 - p_1$$

and

$$q_2 = 1 - p_2.$$

Approximate limits then follow by applying the usual normal theory. Newcombe (1998b) shows that this method tends to give confidence intervals with too low a coverage probability. His paper describes another method (Method 10) with

more acceptable properties. A more complex method of Miettenen and Nurminen (1985), implemented in StatsDirect (1999), also has satisfactory coverage properties.

Suppose we wish to test the null hypothesis that $\pi_1 = \pi_2$. Call the common value π. Then p_1 and p_2 are both estimates of π, and there is little point in estimating π (as in (4.21)) by two different quantities in two different places in the expression. If the null hypothesis is true, both samples are from effectively the same population, and the best estimate of π will be obtained by pooling the two samples, to give

$$p = \frac{r_1 + r_2}{n_1 + n_2}.$$

This pooled estimate is now substituted for both π_1 and π_2 to give

$$\text{var}(p_1 - p_2) = pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right),$$

writing as usual $q = 1 - p$. The null hypothesis is thus tested approximately by taking

$$z = \frac{p_1 - p_2}{\sqrt{\left[pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right]}}$$

as a standardized normal deviate.

Example 4.10

In a clinical trial to assess the value of a new method of treatment (A) in comparison with the old method (B), patients were divided at random into two groups. Of 257 patients treated by method A, 41 died; of 244 patients treated by method B, 64 died. Thus, $p_1 = 41/257 = 0{\cdot}1595$ and $p_2 = 64/244 = 0{\cdot}2623$.

The difference between the two fatality rates is estimated as $0{\cdot}1595 - 0{\cdot}2623 = -0{\cdot}1028$. For 95% confidence limits we take

$$\begin{aligned}\text{var}(p_1 - p_2) &= \frac{(0{\cdot}1595)(0{\cdot}8405)}{257} + \frac{(0{\cdot}2623)(0{\cdot}7377)}{244} \\ &= 0{\cdot}0005216 + 0{\cdot}0007930 \\ &= 0{\cdot}0013146\end{aligned}$$

and

$$\text{SE}(p_1 - p_2) = \sqrt{0{\cdot}0013146} = 0{\cdot}0363.$$

Thus, 95% confidence limits are

$$-0{\cdot}1028 \pm (1{\cdot}96)(0{\cdot}0363) = -0{\cdot}0317 \text{ and } -0{\cdot}1739,$$

the minus sign merely serving to indicate in which direction the difference lies.

For the significance test, we form the pooled proportion

$$p = 105/501 = 0{\cdot}2096$$

and estimate $\mathrm{SE}(p_1 - p_2)$ as

$$\sqrt{\left[(0{\cdot}2096)(0{\cdot}7904)\left(\frac{1}{257}+\frac{1}{244}\right)\right]}$$
$$= 0{\cdot}0364.$$

Thus, the normal deviate is

$$\frac{-0{\cdot}1028}{0{\cdot}0364} = -2{\cdot}82(P = 0{\cdot}005).$$

There is strong evidence of a difference in fatality rates, in favour of A.

In this example, the frequencies are sufficiently large to justify the approximate formula for confidence limits. Newcombe's (1998b) Method 10 gives 95% limits as (−0·0314, −0·1736); the method of Miettenen and Nurminen (1985) gives (−0·0316, −0·1743).

Note that, in Example 4.10, the use of p changed the standard error only marginally, from 0·0363 to 0·0364. In fact, there is likely to be an appreciable change only when n_1 and n_2 are very unequal and when p_1 and p_2 differ substantially. In other circumstances, either standard error formula may be regarded as a good approximation to the other, and used accordingly.

In epidemiological studies it is often appropriate to measure the difference in two proportions by their ratio, p_1/p_2, rather than their difference. This measure is referred to as the *risk ratio, rate ratio* or *relative risk*, depending on the type of study. In case–control studies the relative risk cannot be evaluated directly but, in many circumstances, the *odds ratio*, defined by

$$OR = \frac{p_1/(1-p_1)}{p_2/(1-p_2)} \tag{4.22}$$

is a good approximation to the relative risk (Chapter 19). From the point of view of significance testing it makes no difference which measure is used and the method above or the extensions given later in this section are appropriate.

The confidence limits for the risk ratio and odds ratio both involve the use of logarithms, and the *natural* or Napierian logarithm must be used. In natural logarithms the base is e (= 2·7183); $\log_e x$ is usually written as $\ln x$, and $\ln x = 2{\cdot}3026 \log_{10} x$. Most pocket calculators have a key for $\ln x$ and for the antilogarithm, the exponential of x, often written as e^x, so that the conversion formula is not normally required.

Writing R for p_1/p_2, we have

$$R = \frac{r_1/n_1}{r_2/n_2} \tag{4.23}$$

and approximately (using formula (5.19) given later)

$$\text{SE}(\ln R) = \sqrt{\left(\frac{1}{r_1} - \frac{1}{n_1} + \frac{1}{r_2} - \frac{1}{n_2}\right)} \tag{4.24}$$

and the 95% confidence interval for R is

$$\exp[\ln R \pm 1{\cdot}96\text{SE}(\ln R)].$$

For more exact limits, see Koopman (1984), Miettenen and Nurminen (1985) and Gart and Nam (1988).

For a case–control study (§19.4) we must change the notation since there are no longer samples from the two populations, but instead cases of disease and controls (non-cases) are sampled separately and their exposure to some factor established. Suppose the frequencies are as follows:

		Cases	Controls	
Factor	+	a	c	$a+c$
	−	b	d	$b+d$
		$a+b$	$c+d$	n

Then the observed odds ratio is given by

$$OR = \frac{ad}{bc} \tag{4.25}$$

and, approximately,

$$\text{SE}[\ln(OR)] = \sqrt{\left(\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}\right)}. \tag{4.26}$$

Confidence limits based on a normal approximation using (4.26) are sometimes known as *logit limits*; they are illustrated below in Example 4.11. If any of the cell frequencies are small, more complex methods should be used. Exact limits are mentioned later, on p. 137, but an adequate approximation is often given by the method of Cornfield (1956). The expected values of the frequencies used in (4.25) will give a 'true' odds ratio denoted by ψ. The 95% confidence limits are the two values of ψ for which, given the marginal totals, the observed value of a has a one-sided P value of 0·025. That is, the limits are the solutions of

$$\frac{a - A(\psi)}{\sqrt{\text{var}(a;\psi)}} = \pm 1{\cdot}96, \tag{4.27}$$

where $A(\psi)$ is the value of a which, with the observed marginal totals, would give an odds ratio of ψ, and var $(a; \psi)$ is the variance of a for that value of ψ. That is,

$$\frac{A(d-a+A)}{(a+b-A)(a+c-A)} = \psi \tag{4.28}$$

and

$$\text{var}(a;\psi) \simeq \left(\frac{1}{A} + \frac{1}{a+c-A} + \frac{1}{a+b-A} + \frac{1}{d-a+A}\right)^{-1}. \tag{4.29}$$

It is tedious to solve (4.27), but the calculation can readily be set up as a spreadsheet calculation and solved quickly by trial and error. Using a trial value for A, ψ is obtained from (4.28) and var$(a;\psi)$ from (4.29). These are then substituted in the left-hand side of (4.27). The aim is to choose values of A so that (4.27) gives $\pm 1{\cdot}96$. This is achieved by trying different values and iterating. See also Breslow and Day (1980, §4.3) or Fleiss (1981, §5.6).

Example 4.11

Liddell *et al.* (1984) reported on a case–control study investigating the association of bronchial carcinoma and asbestos exposure in the Canadian chrysolite mines and mills. The data were as follows:

Asbestos exposure	Lung cancer	Controls
Exposed	148 (a)	372 (c)
Not exposed	75 (b)	343 (d)

The calculations are:

$$OR = (148 \times 343)/(75 \times 372) = 1{\cdot}82$$

$$\ln OR = 0{\cdot}599$$

$$\text{SE}(\ln OR) = \sqrt{\left(\frac{1}{148} + \frac{1}{75} + \frac{1}{372} + \frac{1}{343}\right)} = \sqrt{0{\cdot}02569} = 0{\cdot}160.$$

95% confidence interval for $\ln OR = 0{\cdot}599 \pm 1{\cdot}96 \times 0{\cdot}160$
$= 0{\cdot}285$ to $0{\cdot}913$,

95% confidence interval for $OR = \exp(0{\cdot}285)$ to $\exp(0{\cdot}913)$
$= 1{\cdot}33$ to $2{\cdot}49$.

Cornfield's method, using (4.27)–(4.29), gives the same limits of 1·33 and 2·49.

2×2 tables and χ^2 tests

An alternative way of displaying the data of Example 4.10 is shown in Table 4.5. This is called a 2×2, or sometimes a *fourfold, contingency table*. The total

frequency, 501 in this example, is shown in the lower right corner of the table. This total frequency or *grand total* is split into two different dichotomies represented by the two horizontal rows of the table and the two vertical columns. In this example the rows represent the two treatments and the columns represent the two outcomes of treatment. There are thus $2 \times 2 = 4$ combinations of row and column categories, and the corresponding frequencies occupy the four *inner cells* in the body of the table. The total frequencies for the two row categories and those for the two columns are shown at the right and at the foot, and are called *marginal totals*.

We have already used a 2×2 table (Table 4.4) to display the results needed for a comparison of proportions in paired samples, but the purpose was a little different from the present approach, which is concerned solely with the *unpaired* case.

We are concerned, in Table 4.5, with possible differences between the fatality rates for the two treatments. Given the marginal totals in Table 4.5, we can easily calculate what numbers would have had to be observed in the body of the table to make the fatality rates for A and B exactly equal. In the top left cell, for example, this *expected* number is

$$\frac{105 \times 257}{501} = 53{\cdot}862,$$

since the overall fatality rate is 105/501 and there are 257 individuals treated with A. Similar expected numbers can be obtained for each of the four inner cells, and are shown in Table 4.6 (where the observed and expected numbers are distinguished by the letters O and E). The expected numbers are not integers and have been rounded off to 3 decimal places. Clearly one could not possibly observe 53·862 individuals in a particular cell. These expected numbers should be thought of as expectations, or mean values, over a large number of possible tables with the same marginal totals as those observed, when the null hypothesis is true.

Note that the values of E sum, over both rows and columns, to the observed marginal totals. It follows that the *discrepancies*, measured by the differences $O - E$, add to zero along rows and columns; in other words, the four discrepancies are numerically the same (12.862 in this example), two being positive and two negative.

In a rough sense, the greater the discrepancies, the more evidence we have against the null hypothesis. It would therefore seem reasonable to base a significance test somehow on these discrepancies. It also seems reasonable to take account of the absolute size of the frequencies: a discrepancy of 5 is much more important if $E = 5$ than if $E = 100$.

It turns out to be appropriate to calculate the following index:

$$X^2 = \sum \frac{(O - E)^2}{E}, \tag{4.30}$$

Table 4.5 2 × 2 table showing results of a clinical trial.

	Outcome		
Treatment	Death	Survival	Total
A	41	216	257
B	64	180	244
Total	105	396	501

Table 4.6 Expected frequencies and contributions to X^2 for data in Table 4.5.

		Outcome		
Treatment		Death	Survival	Total
A	O	41	216	257
	E	53·862	203·138	257
	$O - E$	−12.862	12·862	0
	$(O - E)^2$	165·431	165·431	
	$(O - E)^2/E$	3·071	0·814	
B	O	64	180	244
	E	51·138	192·862	244
	$O - E$	12·862	−12·862	0
	$(O - E)^2$	165·431	165·431	
	$(O - E)^2/E$	3·235	0·858	
Total	O	105	396	501
	E	105	396	
	$O - E$	0	0	

the summation being over the four inner cells of the table. The contributions to X^2 from the four cells are shown in Table 4.6. The total is

$$X^2 = 3·071 + 0·814 + 3·235 + 0·858$$
$$= 7·978.$$

On the null hypothesis, X^2 follows the $\chi^2_{(1)}$ distribution (see §3.8), the approximation improving as the expected numbers get larger. There is one degree of freedom because only one of the values of E is necessary to complete the whole table. Reference to Table A2 shows that the observed value of 7·978 is beyond

the 0·01 point of the $\chi^2_{(1)}$ distribution, and the difference between the two fatality rates is therefore significant at the 1% level. The precise significance level may be obtained by taking the square root of 7·978 (= 2·82) and referring to Table A1; this gives 0·005.

On p. 126, we derived a standardized normal deviate by calculating the standard error of the difference between the two proportions, obtaining the numerical value of 2·82. This agrees with the value obtained as the square root of X^2. In fact it can be shown algebraically that the X^2 index is always the same as the square of the normal deviate given by the first method. The probability levels given by the two tests are therefore always in agreement.

The X^2 index is often denoted by χ^2, although it seems slightly preferable to reserve the latter for the theoretical distribution, denoting the calculated value by X^2.

There are various alternative formulae for X^2, of which we may note one. Denote the entries in the table as follows:

		Column 1	Column 2	
Row 1		a	b	r_1
Row 2		c	d	r_2
		s_1	s_2	N

Then

$$X^2 = \frac{(ad - bc)^2 N}{r_1 r_2 s_1 s_2}. \tag{4.31}$$

This version is particularly suitable for use with a calculator.

We have, then, two entirely equivalent significance tests. Which the user chooses to use is to some extent a matter of taste and convenience. However, there are two points to be made. First, the standard error method, as we have seen, not only yields a significance test but also leads naturally into the calculation of confidence intervals. This, then, is a strong argument for calculating differences and standard errors, and basing the test on these values rather than on the X^2 index. The main counter-argument is that, as we shall see in §8.5 and §8.6 the X^2 method can be generalized to contingency tables with more than two rows and columns.

It is important to remember that the X^2 index can only be calculated from 2×2 tables in which the entries are frequencies. A common error is to use it for a

table in which the entries are mean values of a certain variable; this practice is completely erroneous.

A closely related method of deriving a significance test is to work with one of the frequencies in the 2×2 table. With the notation above, the frequency denoted by a could be regarded as a random variable and its significance assessed against its expectation and standard error calculated conditionally on the marginal totals. This method proceeds as follows, using O, E and V to represent the observed value, expected value and variance of a:

$$\left.\begin{aligned} E &= \frac{(a+b)(a+c)}{N} \\ V &= \frac{(a+b)(c+d)(a+c)(b+d)}{N^2(N-1)} \\ X^2 &= \frac{(O-E)^2}{V} \end{aligned}\right\}. \tag{4.32}$$

Apart from a factor of $(N-1)/N$, (4.32) is equivalent to (4.31). This form is particularly convenient when combining the results of several studies since the values of O, E and V may be summed over studies before calculating the test statistic. Yusuf *et al.* (1985) proposed an approximate method of estimating the odds ratio and its standard error by

$$\left.\begin{aligned} OR &= \exp\left(\frac{O-E}{V}\right) \\ \mathrm{SE}[\ln(OR)] &= \frac{1}{\sqrt{V}} \end{aligned}\right\}. \tag{4.33}$$

This method was introduced for the combination of studies where the effect was small, and is known to be biased when the odds ratio is not small (Greenland & Salvan, 1990).

More details of such methods are given in §15.6 and of their application in overviews or meta-analyses in §18.10.

Example 4.12

Consider the data of Example 4.11. We have

$$O = 148, \quad E = 123{\cdot}62, \quad V = 42{\cdot}038$$

and (4.33) gives

$$OR = \exp(24{\cdot}38/42{\cdot}038) = 1{\cdot}79$$
$$\mathrm{SE}[\ln(OR)] = 0{\cdot}154.$$

These values are close to those found in Example 4.11

Both the standard error test and the χ^2 test are based on approximations which are valid particularly when the frequencies are high. In general, two methods of improvement are widely used: the application of a continuity correction and the calculation of exact probabilities.

Continuity correction for 2 × 2 tables

This method was described by F. Yates and is often called *Yates's correction*. The $\chi^2_{(1)}$ distribution has been used as an approximation to the distribution of X^2 on the null hypothesis and subject to fixed marginal totals. Under the latter constraint only a finite number of tables are possible. For the marginal totals of Table 4.5, for example, all the possible tables can be generated by increasing or decreasing one of the entries by one unit at a time, until either that entry or some other reaches zero. (A fuller discussion follows later in this section.) The position, therefore, is rather like that discussed in §3.8 where a discrete distribution (the binomial) was approximated by a continuous distribution (the normal). In the present case one might base the significance test on the probability of the observed table or one showing a more extreme departure from the null hypothesis. An improvement in the estimation of this probability is achieved by reducing the absolute value of the discrepancy, $O - E$, by $\frac{1}{2}$ before calculating X^2. In Example 4.10, Table 4.6, this would mean taking $|O - E|$ to be 12·362 instead of 12·862, and the corrected value of X^2, denoted by X_c^2, is 7·369, somewhat less than the uncorrected value but still highly significant.

The continuity correction has a relatively greater effect when the expected frequencies are small than when they are large. The use of the continuity correction gives an approximation to the P value in the exact test described below. As in the analogous situations discussed earlier in this chapter (and see also p. 137), we prefer the mid-P value, which corresponds to the *uncorrected* χ^2 test. We therefore recommend that the continuity correction should not routinely be employed.

The continuity-corrected version of (4.31) is

$$X_c^2 = \frac{(|ad - bc| - \frac{1}{2}N)^2 N}{r_1 r_2 s_1 s_2}. \tag{4.34}$$

If the continuity correction is applied in the χ^2 test, it should logically be applied in the standard error test. The procedure there is to calculate $p_1 - p_2$ after the frequencies have been moved half a unit nearer their expected values, the standard error remaining unchanged. Thus, in Example 4.10, we should have $p_{1(c)} = 41{\cdot}5/257 = 0{\cdot}1615, p_{2(c)} = 63{\cdot}5/244 = 0{\cdot}2602$, giving $z_{(c)} = -0{\cdot}0987/0{\cdot}0364 = -2{\cdot}71$. Since $(-2{\cdot}71)^2 = 7{\cdot}34$, the result agrees with that for X_c^2 apart from rounding errors.

The exact test for 2 × 2 tables

Even with the continuity correction there will be some doubt about the adequacy of the χ^2 approximation when the frequencies are particularly small. An exact test was suggested almost simultaneously in the mid-1930s by R.A. Fisher, J.O. Irwin and F. Yates, and is often called 'Fisher's exact test'. It consists in calculating the exact probabilities of the possible tables described in the previous subsection. The probability of a table with frequencies

a	b	r_1
c	d	r_2
s_1	s_2	N

is given by the formula

$$\frac{r_1!\, r_2!\, s_1!\, s_2!}{N!\, a!\, b!\, c!\, d!}. \tag{4.35}$$

This is, in fact, the probability of the observed cell frequencies *conditional* on the observed marginal totals, under the null hypothesis of no association between the row and column classifications.

Given any observed table, the probabilities of all tables with the same marginal totals can be calculated, and the P value for the significance test calculated by summation. Example 4.13 illustrates the calculations and some of the difficulties of interpretation which may arise.

Example 4.13

The data in Table 4.7, due to M. Hellman, are discussed by Yates (1934).

There are six possible tables with the same marginal totals as those observed, since neither a nor c (in the notation given above) can fall below 0 or exceed 5, the smallest marginal total in the table. The cell frequencies in each of these tables are shown in Table 4.8.

The probability that $a = 0$ is, from (4.35),

$$P_0 = \frac{20!\, 22!\, 5!\, 37!}{42!\, 0!\, 20!\, 5!\, 17!} = 0{\cdot}03096.$$

Tables of log factorials (Fisher & Yates, 1963, Table XXX) are often useful for this calculation, and many scientific calculators have a factorial key (although it may only function correctly for integers less than 70). Alternatively the expression for P_0 can be calculated without factorials by repeated multiplication and division after cancelling common factors:

$$P_0 = \frac{22 \times 21 \times 20 \times 19 \times 18}{42 \times 41 \times 40 \times 39 \times 38} = 0{\cdot}03096.$$

Table 4.7 Data on malocclusion of teeth in infants (reproduced from Yates (1934) with permission from the author and publishers).

	Infants with		
	Normal teeth	Malocclusion	Total
Breast-fed	4	16	20
Bottle-fed	1	21	22
Total	5	37	42

Table 4.8 Cell frequencies in tables with the same marginal totals as those in Table 4.7.

0 20 20	1 19 20	2 18 20
5 17 22	4 18 22	3 19 22
5 37 42	5 37 42	5 37 42
3 17 20	4 16 20	5 15 20
2 20 22	1 21 22	0 22 22
5 37 42	5 37 42	5 37 42

The probabilities for $a = 1, 2, \ldots, 5$ can be obtained in succession. Thus,

$$P_1 = \frac{5 \times 20}{1 \times 18} \times P_0$$

$$P_2 = \frac{4 \times 19}{2 \times 19} \times P_1, \text{etc.}$$

The results are:

a	Probability
0	0·0310
1	0·1720
2	0·3440
3	0·3096
4	0·1253
5	0·0182
	1·0001

This is the complete *conditional distribution* for the observed marginal totals, and the probabilities sum to unity, as would be expected. Note the importance of carrying enough significant digits in the first probability to be calculated; the above calculations were carried out with more decimal places than recorded by retaining each probability in the calculator for the next stage.

The observed table has a probability of 0·1253. To assess its significance we could measure the extent to which it falls into the tail of the distribution by calculating the probability of that table or of one more extreme. For a one-sided test the procedure clearly gives

$$P = 0{\cdot}1253 + 0{\cdot}0182 = 0{\cdot}1435.$$

The result is not significant at even the 10% level.

For a two-sided test the other tail of the distribution must be taken into account, and here some ambiguity arises. Many authors advocate that the one-tailed P value should be doubled. In the present example, the one-tailed test gave $P = 0{\cdot}1435$ and the two-tailed test would give $P = 0{\cdot}2870$. An alternative approach is to calculate P as the total probability of tables, in either tail, which are at least as extreme as that observed in the sense of having a probability at least as small. In the present example we should have

$$P = 0{\cdot}1253 + 0{\cdot}0182 + 0{\cdot}0310 = 0{\cdot}1745.$$

The first procedure is probably to be preferred on the grounds that a significant result is interpreted as strong evidence for a difference in the *observed direction*, and there is some merit in controlling the chance probability of such a result to no more than half the two-sided significance level. The tables of Finney *et al.* (1963) enable one-sided tests at various significance levels to be made without computation provided the frequencies are not too great.

To calculate the mid-P value only half the probability of the observed table is included and we have

$$\text{mid-}P = \tfrac{1}{2}(0{\cdot}1253) + 0{\cdot}0182 = 0{\cdot}0808$$

as the one-sided value, and the two-sided value may be obtained by doubling this to give 0·1617.

The results of applying the exact test in this example may be compared with those obtained by the χ^2 test with Yates's correction. We find $X^2 = 2{\cdot}39 (P = 0{\cdot}12)$ without correction and $X_c^2 = 1{\cdot}14 (P = 0{\cdot}29)$ with correction. The probability level of 0·29 for X_c^2 agrees well with the two-sided value 0·29 from the exact test, and the probability level of 0·12 for X^2 is a fair approximation to the exact mid-P value of 0·16.

Cochran (1954) recommends the use of the exact test, in preference to the χ^2 test with continuity correction, (i) if $N < 20$, or (ii) if $20 < N < 40$ and the smallest expected value is less than 5. With modern scientific calculators and statistical software the exact test is much easier to calculate than previously and should be used for any table with an expected value less than 5.

The exact test and therefore the χ^2 test with Yates's correction for continuity have been criticized over the last 50 years on the grounds that they are conservative in the sense that a result significant at, say, the 5% level will be found in less than 5% of hypothetical repeated random samples from a population in which the null hypothesis is true. This feature was discussed in §4.4 and it was remarked that the problem was a consequence of the discrete nature of the data

and causes no difficulty if the precise level of P is stated. Another source of criticism has been that the tests are conditional on the observed margins, which frequently would not all be fixed. For example, in Example 4.13 one could imagine repetitions of sampling in which 20 breast-fed infants were compared with 22 bottle-fed infants but in many of these samples the number of infants with normal teeth would differ from 5. The conditional argument is that, whatever inference can be made about the association between breast-feeding and tooth decay, it has to be made within the context that exactly five children had normal teeth. If this number had been different then the inference would have been made in this different context, but that is irrelevant to inferences that can be made when there are five children with normal teeth. Therefore, we do not accept the various arguments that have been put forward for rejecting the exact test based on consideration of possible samples with different totals in one of the margins. The issues were discussed by Yates (1984) and in the ensuing discussion, and by Barnard (1989) and Upton (1992), and we shall not pursue this point further. Nevertheless, the exact test and the corrected χ^2 test have the undersirable feature that the average value of the significance level, when the null hypothesis is true, exceeds 0.5. The mid-P value avoids this problem, and so is more appropriate when combining results from several studies (see §4.4). As for a single proportion, the mid-P value corresponds to an uncorrected χ^2 test, whilst the exact P value corresponds to the corrected χ^2 test.

The probability distribution generated by (4.35), for different values of the cell frequencies, a, b, c and d, is called the *hypergeometric distribution*. When the null hypothesis is not true, the expected frequencies will have an odds ratio, ψ, different from 1. In that case, the probabilities for the various values of a are proportional to the expression (4.35) multiplied by ψ^a and are somewhat awkward to evaluate. Nevertheless, an exact test for the non-null hypothesis that $\psi = \psi_0$ can, in principle, be obtained in a manner similar to that used for the exact test of the null hypothesis. This enables exact confidence limits to be obtained by finding those values of ψ, denoted by ψ_L and ψ_U, that give the appropriate tail-area probabilities in the two directions. Mid-P significance levels will, in large samples, give confidence limits similar to the approximate limits given by (4.26)–(4.29). The inclusion of the observed table in the tail-area calculations will give rather wider limits. In Example 4.11, these wider 95% limits (obtained from an algorithm of Thomas (1971) and implemented in StatsDirect (1999)) are 1·32 and 2·53; Cornfield's (1956) method (4.27)–(4.29), with a continuity correction reducing by $\frac{1}{2}$ the absolute value of the numerator $a - A(\psi)$ in (4.27), gives 1·31 and 2·52.

4.6 **Sample-size determination**

One of the questions most commonly asked about the planning of a statistical study, and one of the most difficult to answer, is: how many observations should