Chapter 6
Evaluating the role of chance

6.1 Populations and samples

Suppose that as part of a general health cross-sectional survey, we wish to determine the proportion of men in a particular town who currently smoke cigarettes. For practical and financial reasons, it is impossible to interview every single man in this town, so we decide to select a random sample of 30 men. In this sample, the proportion of current smokers is 7/30 = 23%.

Usually it is impossible to examine every single individual of a population of interest, and we are limited to examining a sample (or subset) of individuals drawn from this population in the hope that it is representative of the whole population[a].

If we do not have information about the whole population, we cannot know the true proportion of the population. However, the proportion computed from a random sample can be used as an estimate of the proportion in the entire population from which the sample was drawn. In the above example, the sample proportion (23%) is our best guess of the true but unknown proportion of current smokers in the whole town.

6.1.1 How reliable is our sample proportion?

There is nothing special about the particular random sample we have used, and different random samples will yield slightly different estimates of the true population proportion. This implies that our sample estimate is subject to *sampling error*. The proportion of current smokers in the whole town is unlikely to be exactly the 23% found in our sample, due to sampling error. The question is, how far from 23% is it likely to be?

To try to answer this question, we first recognize that the sample we picked was only one of a very large number of possible samples of 30 individuals. Suppose we were able to look at 100 000 samples. For each sample, we interview 30 individuals and calculate the sample proportion of current smokers p. The value of p will vary from sample to sample. If we plot all values of p, we would see a distribution like the one shown in Figure 6.1.

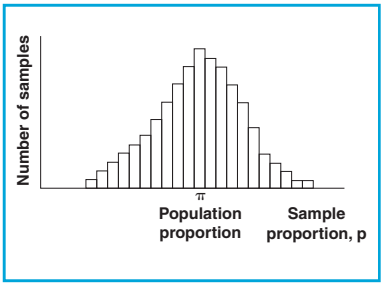

Figure 6.1.
Sampling distribution of p for 100 000 repeated samples of size 30.

This distribution is called the sampling distribution of p. It shows that although most sample estimates are likely to be concentrated around the true (population) proportion π, some will be a long way from this true value. The amount of spread tells us how precise our sample proportion p is likely to be, as an estimate of the true proportion π. If the distribution is wide, there is a lot of sampling error and our p may be a long way from

[a] To ensure representativeness, the sample of individuals should be *randomly selected* from the population of interest. That is, every individual in the population should have an equal chance of being included in the sample. The different ways in which a random sample can be drawn from a specific population are dealt with in Chapter 10.

the true value π. If it is narrow, there is little sampling error, and our p is likely to be very close to π.

We have already seen in Section 3.3.1 that the spread of a distribution can be measured by a quantity called the standard deviation. It can be shown that the standard deviation of a sampling distribution of p is given by

$$\mathrm{SE}(p)=\sqrt{\frac{\pi(1-\pi)}{n}}$$

where n represents the size of the sample. SE stands for *standard error*, which is the term we generally use for the standard deviation of a sampling distribution. The standard error is a measure of the precision with which our sample value p estimates the true value π. Notice that if we increase the sample size, n, we decrease the standard error and the sampling distribution will become narrower. This is just what we would expect, as larger samples should provide more reliable estimates of π.

When we actually do our survey, of course, we do not know the value of π (otherwise we would not need to do the survey!), and so we cannot actually use the above formula to calculate the SE(p). We can make a close estimation of it by replacing π in the formula with our sample estimate p, giving

$$\mathrm{SE}(p)=\sqrt{\frac{p(1-p)}{n}}$$

which can be rearranged as

$$\mathrm{SE}(p)=\sqrt{\frac{p^2(1-p)}{a}}$$

where a is the numerator of the proportion $p = a/n$ (in our sample a = 7, the observed number of current smokers). This last formula is particularly useful because it shows that the standard error is inversely related to the observed number of cases. It is the number of cases in the numerator of p that mainly determines the magnitude of the standard error, and not the sample size in itself.

It is possible to show mathematically that, in sufficiently large samples, approximately 95% of all the sample estimates will fall within 1.96 standard errors of the true value π; 2.5% of the sample estimates (one sample in 40) will be more than 1.96 SEs below the true value, and 2.5% (one in 40) will be more than 1.96 SEs above the true value (Figure 6.2).

Now what can we say about π from our single sample of 30 individuals? Our sample may have come from any part of the distribution shown in Figure 6.2. However, before drawing the sample, there was a 95% chance that the observed sample proportion p would lie within two standard errors (more precisely $1.96 \times \text{SE}$) of the true value π. As a logical consequence, intervals from samples of similar size but centred on each sample proportion p will include π if p is within two standard errors of π. Hence, an interval bounded by the following lower and upper limits

$$p - 1.96 \times \text{SE} \quad \text{and} \quad p + 1.96 \times \text{SE}$$

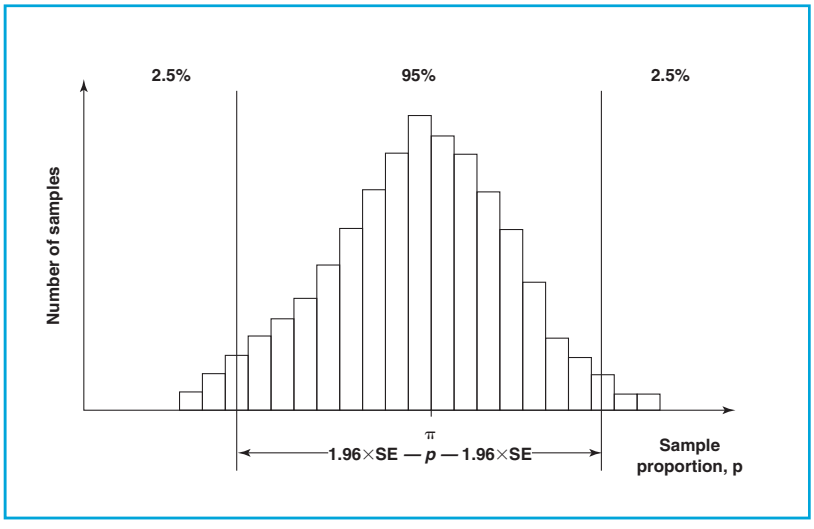

Figure 6.2.
Sampling distribution of p with 95% confidence limits.

(usually written $p \pm 1.96 \times \text{SE}(p)$) will include the true proportion π with probability 95%. These limits, calculated from the sample data, are called *lower* and *upper confidence limits*, respectively, and the interval between them is called a *95% confidence interval* of the unknown population proportion π.

In our example for estimating the proportion of men currently smoking,

$$n = 30,\ a = 7 \text{ and } p = 0.23$$

We estimate standard error of p to be

$$\text{SE}(p) = \sqrt{\frac{0.23^2(1-0.23)}{7}} = 0.076$$

A 95% confidence interval for the true proportion of men who currently smoke in the whole town is therefore given by

$$0.23 \pm 1.96 \times 0.076 = 0.081 \text{ to } 0.379$$

So our best estimate of the proportion of current smokers in the whole town is 23%, but the true value could easily be anywhere between 8% and 38%.

In strict terms, the confidence interval is a range of values that is likely to cover the true population value but we are still not certain that it will.

In reality, a confidence interval from a particular study may or may not include the actual population value. The confidence interval is based on the concept of repetition of the study under consideration. Thus if the study were to be repeated 100 times, we would expect 95 of the 100 resulting 95% confidence intervals to include the true population value.

If we want to be even more confident that our interval includes the true population value, we can calculate a *99% confidence interval*. This is done simply by replacing 1.96 with 2.58 in the above formula. That is, we use

$$p \pm 2.58 \times \sqrt{\frac{p^2(1-p)}{a}}$$

In our example, the corresponding 99% confidence interval for the proportion of current smokers in the whole town is 0.034 to 0.426, or roughly 3% to 43%.

Similarly, we can calculate a *90% confidence interval* by replacing 1.96 with 1.64 in the formula:

$$p \pm 1.64 \times \sqrt{\frac{p^2(1-p)}{a}}$$

In our example, the corresponding 90% confidence interval for the proportion of current smokers in the whole town is 0.105 to 0.355, or 11% to 36%.

6.1.2 How good are other sample estimates?

This useful way of describing sampling error is not limited to the sample proportion. We can obtain confidence intervals for any other sample estimates such as means, risks, rates, rate ratios, rate differences, etc. The underlying concept is similar to the one illustrated above for proportions. In all these cases, the confidence intervals provide an indication of how close our sample estimate is likely to be to the true population value.

Suppose that, as part of the same general health survey, we wished to determine the mean height of the men in the same town. We measured the 30 individuals from our sample and obtained a sample mean height of 165 cm. Again, the mean height of the male adult population in the whole town is unlikely to be exactly 165 cm. However, if a very large number of samples of 30 individuals were drawn from this population and for each one the mean height were calculated and plotted, a sampling distribution of the mean would be obtained. This sampling distribution will have a shape similar to that of the sampling distribution of a proportion (Figure 6.1), i.e., the distribution would be bell-shaped, with most sample estimates centred around the true population mean.

We can, therefore, obtain a 95% confidence interval in a similar way to that used for proportions:

$$\text{Sample mean} \pm 1.96 \times \text{SE of the mean}^{b}$$

The *standard error of a mean* can be estimated by

$$\text{SE (mean)} = \text{SD}/\sqrt{n}$$

where SD represents the standard deviation described in Section 3.3.1. Suppose that, in the above example, the standard deviation was found to be 7.1 cm. The standard error of the mean will be given by

$$\text{SE (mean)} = 7.1/\sqrt{30} = 1.3 \text{ cm}$$

A 95% confidence interval for this sample mean will be equal to

$$165 \text{ cm} \pm 1.96 \times 1.3 \text{ cm} = 162 \text{ cm to } 168 \text{ cm}$$

How do we interpret this confidence interval? If the study were done 100 times, of the 100 resulting 95% confidence intervals, we would expect 95 of them to include the true population value. Thus, the confidence interval from this particular sample of 30 men provides a range of values that is likely to include the true population mean, although we cannot be sure that it does.

As long as the sampling distribution of a particular estimate is approximately bell-shaped (i.e., it is what statisticians call a 'Normal distribution'), as it will always be if the sample size is sufficiently large, we can summarize the calculation of a 95% confidence interval as follows:

$$\text{Sample estimate} \pm 1.96 \times \text{SE(sample estimate)}$$

(To obtain a 90% or a 99% confidence interval, all we need to do is to replace 1.96 in the formula with, respectively, 1.64 or 2.58.)

In Example 6.1, men employed for 10 or more years were estimated to have an excess of 92 cancer deaths per 10 000 pyrs compared with those employed for less than 1 year, with a 95% confidence interval ranging from 61 to 122 deaths per 10 000 pyrs. This confidence interval was calculated using the above general formula as follows:

$$\text{Rate difference} \pm 1.96 \times \text{SE(rate difference)}$$

where the SE of the rate difference was about 15 deaths per 10 000 pyrs.

[b] The precise value to be used in this formula varies with the size of the sample and it is given in tables of the *t*-distribution. However, for large sample sizes (≥30) this factor is close to 1.96.

Example 6.1. *In a cohort study of 15 326 men employed in a particular factory, their cancer mortality was assessed in relation to duration of their employment (Table 6.1).*

Table 6.1.
Age-adjusted mortality rate ratios and rate differences of cancer (all sites combined) by duration of employment: hypothetical data.

Duration of employment (years)	No. of cases	Person -years	Rate[a]	Rate ratio (95% CI)[b]	Rate difference[a] (95% CI)[b]
<1[c]	44	40 056	11	1.0	0
1.0-1.9	67	21 165	32	2.9 (1.9–4.3)	21 (12–29)
2.0-4.9	19	3 105	61	5.6 (3.1–9.7)	50 (23–78)
5.0-9.9	48	5 067	95	8.6 (5.6–13.3)	84 (57–111)
≥10	43	4 192	103	9.3 (6.0–14.6)	92 (61–122)

[a] Rates per 10 000 person-years.
[b] CI = confidence interval.
[c] Baseline category.

The corresponding rate ratio was 9.3 with a 95% confidence interval ranging from 6.0 to 14.6. Thus, our best guess is that men employed for 10 or more years were nine times more likely to die from cancer than those employed for less than one year, but the true rate ratio might lie somewhere between 6.0 and 14.6 (and is unlikely to lie outside this range).

You might have noticed in this example that, whereas the confidence limits of a rate difference are equidistant from the sample estimate, this is not the case for the confidence limits of a rate ratio. This can be seen clearly in Figures 6.3 and 6.4(*a*). In contrast to the rate difference, the sampling distribution of a rate ratio is not symmetric, since the minimum possible value it can take is zero, whereas the maximum is infinity. To obtain a more symmetric distribution, a logarithmic transformation of the data was used. As a consequence of this transformation, the confidence limits are equidistant from the sample estimate of the rate ratio on the logarithmic scale (Figure 6.4(b)) but asymmetric when converted back to the original scale (Figure 6.4(a)) (see Appendix 6.1, at the end of this chapter, which provides formulae to calculate confidence intervals for difference and ratio measures).

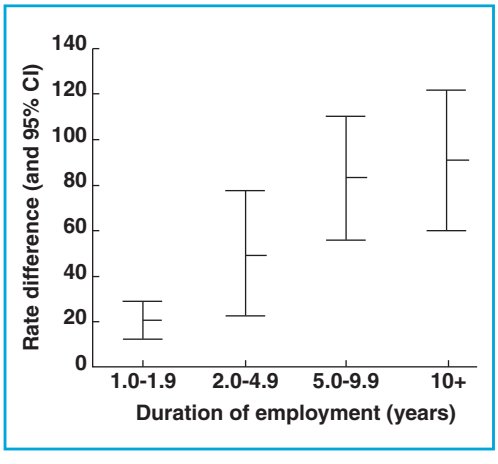

Figure 6.3.
Graphical display of rate differences (indicated by the middle horizontal lines) and their 95% confidence intervals (vertical lines) on an arithmetic scale (data from Table 6.1).

6.1.3 Display of confidence intervals

If we have two or more groups, we can display the sample estimates and their 95% confidence intervals in a graph. For instance, the rate ratios and rate differences from Table 6.1 and their respective confidence intervals are displayed in Figures 6.3 and 6.4.

The middle horizontal lines show the observed rate differences and rate ratios, while the vertical lines indicate the 95% confidence intervals. Note

how the confidence interval is much narrower when the number of cases is large (e.g., category 1–1.9 years (based on 67 cases)). It is the number of cases in the numerator of rates and proportions which determines the size of the standard error and, therefore, the width of the confidence interval.

6.1.4 Further comments

Statistical inference is a process by which we draw conclusions about a population from the results observed in a sample. The above statistical methods assume that the sample of individuals studied has been randomly selected from the population of interest, which was properly defined beforehand. That is, every individual in the population has an equal chance of being in the selected sample. Quite often in epidemiology, getting a truly random sample is impossible and thus we have to be concerned about *selection bias* (see Chapter 13). Confidence intervals convey only the effects of sampling variation on the precision of the sample estimates and cannot control for non-sampling errors such as bias in the selection of the study subjects or in the measurement of the variables of interest. For instance, if the smoking survey only included men who visited their doctors because of respiratory problems, the sample would be unrepresentative of the whole male population in the community. The statistical techniques described above assume that no such bias is present.

The other issue that needs to be kept in mind in epidemiological studies is that even when whole populations are studied, questions of random variability still need to be addressed. Death rates may be computed from national vital statistics, or incidence rates determined from cancer registries that cover whole populations, but there will still be random variation in the number of cases from one year to another.

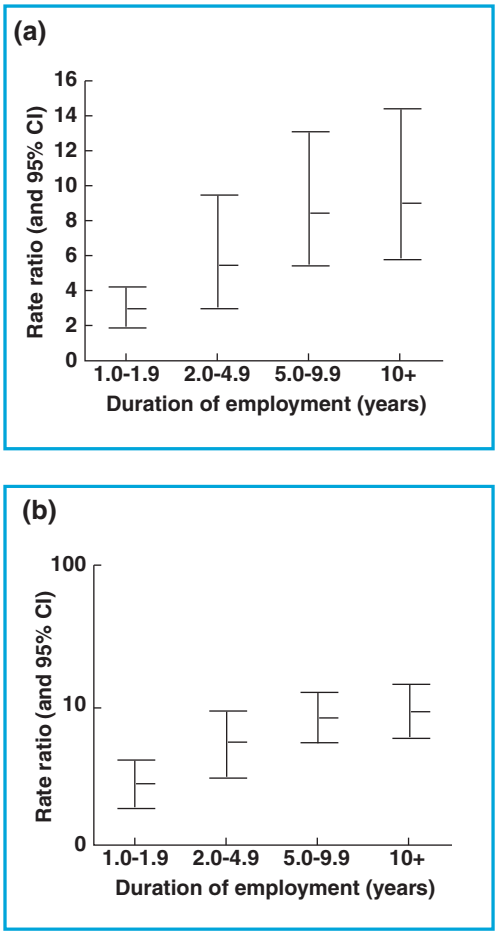

Figure 6.4.
Graphical display of rate ratios and their 95% confidence intervals: (*a*) on an arithmetic scale and (*b*) on a logarithmic scale (data from Table 6.1).

Example 6.2. *Table 6.2 shows that there is considerable random fluctuation in the number of female lip cancer cases registered from year to year in England and Wales.*

Table 6.2.
Number of incident cases of lip cancer by year of registration, females, England and Wales, 1979-89.[a]

Year of registration										
1979	**1980**	**1981**	**1982**	**1983**	**1984**	**1985**	**1986**	**1987**	**1988**	**1989**
58	57	60	46	57	64	53	47	51	58	62

[a] Data from OPCS (1983a to 1994b).

Even though the whole population of the country was studied to produce the data in Example 6.2, there was still random variability in the number of lip cancer cases from year to year, which cannot be explained in terms of changes in underlying risk. In these situations, we are sampling 'in time', that is the population in any particular year can be viewed as a 'sample'. The methods discussed above should still be used to assess the degree of precision of the observed rates. Of course, not all of the variation from year to year may be random and there may, for example, be an underlying trend, upwards or downwards, over a particular time period (see Section 4.3.2).

When the rates are based on small numbers of cases, so that the confidence intervals are very wide, it may be useful to reduce the random variation by pooling data over five or even ten adjacent calendar years and calculating the average annual incidence (or mortality) rates.

6.2 Testing statistical hypotheses

An investigator collecting data generally has two aims: (*a*) to draw conclusions about the true population value, on the basis of the information provided by the sample, and (*b*) to test hypotheses about the population value.

***Example 6.3**. In a clinical trial on metastatic cervical cancer, 218 patients were randomly assigned to a new treatment regimen or to the standard treatment. All patients were followed up for one year after their entry into the trial (or until death if it occurred earlier). The numbers of women still alive at the end of this follow-up period are shown in Table 6.3.*

Table 6.3.
Number of patients still alive one year after entry into the trial by type of treatment administered: hypothetical data.

		Type of treatment	
		New	**Standard**
Alive at the end of the first year	**Yes**	68	45
	No	40	65
	Total	108	110

The data in Example 6.3 show that 63.0% (68 out of 108) patients administered the new treatment were still alive compared with 40.9% (45 out of 110) of those given the standard treatment. From these results the new treatment appears superior but how strong is the evidence that this is the case?

In general, when comparing the effects of two levels of an exposure or of two treatments, two groups are sampled, the 'exposed' and the 'unexposed', and their respective summary statistics are calculated. We might wish to compare the two samples and ask: 'Could they both come from the same population?' That is, does the fact that some subjects were

exposed and others were not influence the outcome we are observing? If there is no strong evidence from the data that the exposure influences the outcome, we might assume that both samples came from the same population with respect to the particular outcome under consideration.

Statistical significance testing is a method for quantification of the chance of obtaining the observed results if there is no true difference between the groups being examined in the whole population.

6.2.1 The null hypothesis

An investigator usually has some general (or theoretical) hypothesis about the phenomenon of interest before embarking on a study. Thus, in Example 6.3, it is thought that the new treatment is likely to be better than the standard one in the management of metastatic cervical cancer patients. This hypothesis is known as the *study hypothesis*. However, it is impossible to prove most hypotheses conclusively. For instance, one might hold a theory that all Chinese children have black hair. Unfortunately, if one had observed one million Chinese children and found that they all had black hair, this would not have proved the hypothesis. On the other hand, if just one fair-haired Chinese child were seen, the theory would be disproved. Thus there is often a simpler logical setting for disproving hypotheses than for proving them. Obviously, the situation is much more complex in epidemiology. The observation that both the new and the standard treatments had similar effects in one single patient would not be enough to 'disprove' the study hypothesis that the two treatments were different in effect.

Consider again the cervical cancer clinical trial (Example 6.3). In addition to the study hypothesis that the new treatment is better than the standard one, we consider the *null hypothesis* that the two treatments are equally effective. If the null hypothesis were true, then for the population of all metastatic cervical cancer patients, the one-year survival experience would be similar for both groups of patients, regardless of the type of treatment they received. The formulation of such a null hypothesis, i.e., a statement that there is no true statistical association, is the first step in any statistical test of significance.

6.2.2 Significance test

After specifying the null hypothesis, the main question is:

If the null hypothesis were true, what are the chances of getting a difference at least as large as that observed?

For example, in the cervical cancer trial, what is the probability of getting a treatment difference at least as large as the observed 63.0% – 40.9% = 22.1%? This probability, commonly denoted by P (capital P rather than the small p we used earlier for the sample proportion), is determined by applying an appropriate statistical test.

A simple example

To understand the basis for a statistical significance test, let us look first at an example that uses numbers small enough to allow easy calculations. Suppose that an investigator has a theory that predicts there will be an excess of male births among babies resulting from *in-vitro* fertilization (IVF) and he therefore wants to study the question 'Are there more boys than girls among IVF babies?'

The investigator formulates a null hypothesis that there is an equal proportion (0.5 or 50%) of males and females in the population of IVF babies. Next, he samples five records from one IVF clinic and finds that they are all males in this sample of births. We can now calculate the probability of obtaining five males and no females if the null hypothesis of equal numbers of males and females were true.

Probability that the first sampled is a male	= 0.5	
Probability that the first and the second sampled are males	= 0.5 x 0.5	= 0.25
Probability that the first, second and third sampled are males	= 0.5 x 0.5 x 0.5	= 0.125
Probability that the first, second, third and fourth sampled are males	= 0.5 x 0.5 x 0.5 0.5	= 0.0625
Probability that all five sampled are males	= 0.5 x 0.5 x 0.5 x 0.5 x 0.5	= 0.03125

Thus, there is a 3.125% chance of obtaining five males in a row even if the true proportions of males and females born in the whole population were equal. We have just done a statistical significance test! It yields the probability (*P*) of producing a result as large as or larger than the observed one if no true difference actually existed between the proportion of males and females in the whole population of IVF babies.

What can we conclude from this probability? This *P*-value can be thought of as a measure of the consistency of the observed result with the null hypothesis. The smaller the *P*-value is, the stronger is the evidence provided by the data against the null hypothesis. In this example, the probability of obtaining five boys in a row if the null hypothesis were really true was fairly small. Hence, our data provide moderate evidence that the number of boys is greater than the number of girls among IVF babies.

However, in spite of this small probability, the null hypothesis may well be true. Our final interpretation and conclusions depend very much on our previous knowledge of the phenomenon we are examining. (The situation can be compared to tossing a coin. Even after getting five tails in a series of five tosses, most people would still believe in the null hypothesis of an unbiased coin. However, if the first 20 tosses were all tails the investigators would be very suspicious of bias, since the probability of this happening by chance is only $(0.5)^{20} = 0.00000095$.)

Comparing two proportions

In the cervical cancer treatment trial (Example 6.3), if the null hypothesis of no difference in survival between the two treatments is

true, what is the probability of finding a sample difference as large as or larger than the observed 22.1% (= 63.0% – 40.9%)? If the null hypothesis were true, the only reason for the observed difference to be greater than zero is sampling error. In other words, even if the true population difference in proportions were zero, we would not expect our particular sample difference to be exactly zero because of sampling error. In these circumstances how far, on average, can we reasonably expect the observed difference in the two proportions to differ from zero?

We have already seen in this chapter that the standard error of a proportion gives an indication of how precisely the population value can be estimated from a sample. We can define the standard error of the difference between two proportions in a similar fashion. Theoretically, if we were to repeat the above cervical cancer trial over and over again, each time using the same number of patients in each group, we would obtain a sampling distribution of differences in proportions of a shape similar to that shown in Figure 6.1. The spread of this sampling distribution could be summarized by using a special formula to calculate the standard error. Its application to the cervical cancer trial data yields a standard error equal to 6.6%. The essence of the statistical test we apply to this situation is to calculate how many standard errors away from zero the observed difference in proportions lies. This is obtained as follows:

$$\text{Value of the test statistic} = \frac{\text{observed difference in proportions} - 0}{\text{standard error of difference}}$$

In the cervical cancer trial, the test statistic has a value of

$$\frac{0.221 - 0}{0.066} = 3.35$$

The observed difference between the two treatments (22.1%) is 3.35 standard errors from the null hypothesis value of zero. A value as high as this is very unlikely to arise by chance, since we already know that 95% of observations sampled from a bell-shaped distribution (i.e., a Normal distribution) will be within 1.96 standard errors of its centre.

The larger the test value, the smaller the probability *P* of obtaining the observed difference if the null hypothesis is true. We can refer to tables which convert particular statistical test values into corresponding values for *P*. An extract from one such table, based on the Normal (bell-shaped) distribution, is shown on the next page.

In the cervical cancer example, the value of the test statistic is 3.35, even larger than the highest value shown in the extract on the next page (3.291), and so the probability *P* is less than 0.001. That is, if the new and the standard treatments were really equally effective, the

test statistic exceeds in absolute value[a]	0.674	with probability	0.5	(50%)
"	1.282	"	0.2	(20%)
"	1.645	"	0.1	(10%)
"	1.960	"	0.05	(5%)
"	2.576	"	0.01	(1%)
"	3.291	"	0.001	(0.1%)

[a] Absolute value means that the plus or minus signs should be ignored; for example, –1 and +1 have the same absolute value of 1.

chances of getting so great a difference in survival would be less than one in a thousand. According to conventional use of statistical terminology, we would say that the difference in percentages is *statistically significant at the 0.1% level*. Hence, there is strong evidence that the new treatment is associated with a better one-year survival than is the standard treatment.

6.2.3 Interpretation of *P*-values

Cox & Snell (1981) give the following rough guidelines for interpreting *P*-values:

If $P > 0.1$, the results are reasonably consistent with the null hypothesis.

If $P \approx 0.05$, there is moderate evidence against the null hypothesis.

If $P \leq 0.01$, there is strong evidence against the null hypothesis.

It is common to consider $P<0.05$ as indicating that there is substantial evidence that the null hypothesis is untrue. (The null hypothesis is rejected and the results of the study are declared to be statistically significant at the 5% level.) However, *this emphasis on $P<0.05$ should be discouraged*, since 0.05 is an arbitrary cut-off value for deciding whether the results from a study are statistically significant or non-significant. It is much better to report the actual *P*-value rather than whether *P* falls below or above certain arbitrary limiting values.

When *P* is large, investigators often report a 'non-statistically significant' result and proceed as if they had proved that there was no effect. All they really did was fail to demonstrate that there was a statistically significant one. The distinction between demonstrating that there is no effect and failing to demonstrate that there is an effect is subtle but very important, since the magnitude of the *P*-value depends on both the extent of the observed effect and the number of observations made. Therefore, a small number of observations may lead to a large *P*-value despite the fact that the real effect is large. Conversely, with a large number of observations, small effects, so small as to be clinically and epidemiologically irrelevant, may achieve statistical significance. These issues are of great importance to clinical and epidemiological researchers and are considered in detail in later chapters (13 and 15).

6.2.4 Comparing other sample estimates

Although we have introduced significance testing for one particular problem (comparing two proportions), the same procedure can be applied to all types of comparative study. For instance, it can be used to compare other sample estimates (e.g., means, rates) or to assess more generally associations between variables.

Example 6.4. *In a case–control study to investigate the association between past history of infertility and the risk of developing benign ovarian tumours, the data shown in Table 6.4 were obtained.*

Table 6.4. Distribution of benign ovarian tumour cases and controls according to past history of infertility: hypothetical data.

	Past history of infertility	
	Yes ('exposed')	**No ('unexposed')**
Women with benign ovarian tumours ('cases')	16	42
Healthy women ('controls')	9	120

$$\text{Odds ratio} = \frac{\text{odds of infertility among the cases}}{\text{odds of infertility among the controls}} = \frac{16/42}{9/120} = 5.08$$

(The calculation of odds ratios from case–control studies is discussed in detail in Chapter 9.)

In Example 6.4, the null hypothesis assumes that the two groups of women have a similar risk of getting benign ovarian tumours, i.e., the true odds ratio in the population is equal to one. After using an appropriate statistical test, as described in Appendix 6.1, the researchers obtained a *P*-value of 0.0003, i.e., if the null hypothesis were true, the probability of getting an odds ratio as large as, or larger than, 5.08 would be very small (less than 1 in 1000). Thus, the data from this study provide strong evidence against the null hypothesis.

6.3 Confidence intervals and hypothesis testing

Statistical significance testing is designed to help in deciding whether or not a set of observations is compatible with some hypothesis, but does not provide information on the size of the association (or difference of effects). It is more informative not only to think in terms of statistical significance testing, but also to estimate the size of the effect together with some measure of the uncertainty in that estimate.

This approach is not new; we used it in Section 6.1 when we introduced confidence intervals. We stated that a 95% confidence interval for a sample estimate could be calculated as

$$\text{Sample estimate} \pm 1.96 \times \text{standard error of a sample estimate}$$

A 95% confidence interval for the difference in two proportions can be calculated in a similar way:

$$\text{Observed difference} \pm 1.96 \times \text{standard error of the difference}$$

(The calculation of the standard error of a difference between proportions is illustrated in the Appendix, Section A6.1.2).

In the cervical cancer trial (Example 6.3), this 95% confidence interval is

$$0.221 \pm 1.96 \times 0.066 = 0.092 \text{ to } 0.350 = 9.2\% \text{ to } 35.0\%$$

Thus, it is plausible to consider that the real difference in one-year survival between the new treatment and the standard treatment lies somewhere between 9% and 35%.

This confidence interval is consistent with the result of the statistical test we performed earlier (Section 6.2.2). The value of the test for the null hypothesis of no difference between the two treatments was 3.35, which corresponded to $P<0.001$.

Note that if the 95% confidence interval for a difference *does not include the null hypothesis value of zero*, then P is lower than 0.05. Conversely, if this confidence interval includes the value 0, i.e. one limit is positive and the other is negative, then P is greater than 0.05.

This example shows that there is a close link between significance testing and confidence intervals. This is not surprising, since these two approaches make use of the same ingredient, the standard error. A statistical test is based on how many standard errors the observed sample estimate lies away from the value specified in the null hypothesis. Confidence intervals are calculated from the standard error and indicate a range of values that is likely to include the true but unknown population parameter; this range may or may not include the value specified in the null hypothesis and this is reflected by the value of the test statistic.

In Example 6.5, the P-values are large, indicating that the results are consistent with the null hypothesis of no difference in risk between relatives of mycosis fungoides patients and the general population of England and Wales. However, inspection of the confidence intervals reveals that the confidence interval for all malignancies is quite narrow, whereas the one for non-Hodgkin lymphoma is wide and consistent with an almost five-fold increase in risk as well as with a 50% reduction. This confidence interval is wide because it is based on only three cases.

To summarize, P-values should not be reported on their own. The confidence intervals are much more informative, since they provide an

Example 6.5. *Various studies have suggested that relatives of patients who develop mycosis fungoides (a particular form of non-Hodgkin lymphoma) are at increased risk of developing cancer, particularly non-Hodgkin lymphomas. To clarify this issue, data on the number of cancer cases that occurred among first-degree relatives of mycosis fungoides patients diagnosed in one London hospital were ascertained. The observed number of cancer cases was then compared with those that would have been expected on the basis of the national rates for England and Wales. The results from this study are shown in Table 6.5.*

Table 6.5.
Cancer incidence among first-degree relatives of mycosis fungoides patients (unpublished data).

Site	Number of cases observed (*O*)	Number of cases expected (*E*)[a]	Standardized incidence ratio (*O*/*E*)	95% confidence interval	*P*-value
All sites	34	36.8	0.9	0.6–1.3	0.719
Non-Hodgkin lymphomas	3	2.1	1.5	0.5–4.6	0.502

[a] Calculated using the national age-specific cancer incidence rates for England and Wales as the standard.

idea of the likely magnitude of the association (or difference of effects) and their width indicates the degree of uncertainty in the estimate of effect.

Appendix 6.1 gives formulae to calculate confidence intervals and statistical tests for the most commonly used epidemiological measures.

Further reading

* Gardner & Altman (1986) provide a simple overview of most of the concepts covered in this chapter and also give suggestions for presentation and graphical display of statistical results.

Box 6.1. Key issues

- Epidemiological studies are usually conducted in subsets or *samples* of individuals drawn from the population of interest. A sample estimate of a particular epidemiological measure is, however, unlikely to be equal to the true population value, due to *sampling error*.

- The *confidence interval* indicates how precise the sample estimate is likely to be in relation to the true population value. It provides a range of values that is likely to include the true population value (although we cannot be sure that a particular confidence interval will in fact do so).

- *Statistical significance* testing is used to test hypotheses about the population of interest. The *P*-value provides a measure of the extent to which the data from the study are consistent with the 'null hypothesis', i.e., the hypothesis that there is no true association between variables or difference between groups in the population. The smaller the *P*-value, the stronger is the evidence against the null hypothesis and, consequently, the stronger the evidence in favour of the study hypothesis.

- *P*-values should generally not be reported alone, since they do not provide any indication of the magnitude of the association (or difference of effects). For instance, small effects of no epidemiological relevance can become 'statistically significant' with large sample sizes, whereas important effects may be 'statistically non-significant' because the size of the sample studied was too small. In contrast, confidence intervals provide an idea of the range of values which might include the true population value.

- Confidence intervals and statistical significance testing deal only with sampling variation. It is assumed that non-sampling errors such as bias in the selection of the subjects in the sample and in the measurement of the variables of interest are absent.

Appendix 6.1
Confidence intervals and significance tests for epidemiological measures

This appendix provides formulae for the calculation of confidence intervals and statistical significance tests for the most commonly used epidemiological measures. The formulae presented here can only be applied to 'crude' measures (with the exception of the standardized mortality (or incidence) ratio). For measures that are adjusted for the effect of potential confounding factors, see Chapter 14. For measures not considered here, see Armitage and Berry (1994). Similar results may be easily obtained using computer packages such as EPI INFO, STATA or EGRET.

A6.1.1 Calculation of confidence intervals for measures of occurrence

Single proportion (prevalence or risk)

Prevalence is a *proportion* and therefore the standard error and the confidence interval can be calculated using the formula discussed in Section 6.1.1:

$$\mathrm{SE}(p) = \sqrt{\frac{p^2(1-p)}{a}}$$

where a is the number of cases and $p = a/n$ (n being the sample size).

A 95% confidence interval can be obtained as

$$p \pm 1.96 \times \mathrm{SE}(p)$$

For a 90% confidence interval, the value 1.96 should be replaced by 1.64 and for a 99% confidence interval by 2.58.

Risk is also a *proportion*. Thus the standard error and confidence interval can be obtained in exactly the same way, as long as all the subjects are followed up over the whole risk period of interest. If the follow-up times are unequal, life-table or survival analysis techniques must be used (see Chapter 12), including the appropriate standard error formulae.

The simple method for obtaining confidence intervals described above is based on approximating the sampling distribution to the Normal distribution. This 'approximate' method is accurate in sufficiently large samples (greater than 30).

An 'exact' method for calculating confidence intervals for proportions, based on the binomial distribution, is recommended for smaller samples. This method is, however, too complex for the calculations to be performed on a pocket calculator.

Single rate

If the number of cases that occur during the observation period is denoted by a and the quantity of person-time at risk by y, the estimated incidence rate (r) is

$$r = a/y$$

An 'approximate' standard error can be calculated as follows:

$$\mathrm{SE}(r) = r/\sqrt{a}$$

The 95% confidence interval for the observed rate (r) can then be obtained as

$$r \pm 1.96 \times \mathrm{SE}(r)$$

Table A6.1.1.
95% confidence limit factors for estimates of a Poisson-distributed variable.[a]

Observed number on which estimate is based	Lower limit factor	Upper limit factor	Observed number on which estimate is based	Lower limit factor	Upper limit factor	Observed number on which estimate is based	Lower limit factor	Upper limit factor
(a)	(L)	(U)	(a)	(L)	(U)	(a)	(L)	(U)
1	0.025	5.57	21	0.619	1.53	120	0.833	1.200
2	0.121	3.61	22	0.627	1.51	140	0.844	1.184
3	0.206	2.92	23	0.634	1.50	160	0.854	1.171
4	0.272	2.56	24	0.641	1.48	180	0.862	1.160
5	0.324	2.33	25	0.647	1.48	200	0.868	1.151
6	0.367	2.18	26	0.653	1.47	250	0.882	1.134
7	0.401	2.06	27	0.659	1.46	300	0.892	1.121
8	0.431	1.97	28	0.665	1.45	350	0.899	1.112
9	0.458	1.90	29	0.670	1.44	400	0.906	1.104
10	0.480	1.84	30	0.675	1.43	450	0.911	1.098
11	0.499	1.79	35	0.697	1.39	500	0.915	1.093
12	0.517	1.75	40	0.714	1.36	600	0.922	1.084
13	0.532	1.71	45	0.729	1.34	700	0.928	1.078
14	0.546	1.68	50	0.742	1.32	800	0.932	1.072
15	0.560	1.65	60	0.770	1.30	900	0.936	1.068
16	0.572	1.62	70	0.785	1.27	1000	0.939	1.064
17	0.583	1.60	80	0.798	1.25			
18	0.593	1.58	90	0.809	1.24			
19	0.602	1.56	100	0.818	1.22			
20	0.611	1.54						

[a] Data from Haenszel *et al.*, (1962)

These formulae are appropriate when the number of cases in the numerator of the rate, *a*, is greater than 30. If the number of cases is small, 'exact' confidence intervals, based on the Poisson distribution, can be obtained from Table A6.1.1. This table gives factors by which the observed rate is multiplied to obtain the lower and the upper limit of a 95% confidence interval:

Lower limit = $r \times$ lower limit factor (L)
Upper limit = $r \times$ upper limit factor (U)

Consider the following example. The total number of deaths from stomach cancer among males aged 45–54 years in Egypt during 1980 was 39 in 1 742 000 person-years (WHO, 1986). Thus, using the 'approximate' method,

r = 39/1 742 000 pyrs = 2.24 per 100 000 pyrs

SE(r) = 2.24 per 100 000 pyrs/$\sqrt{39}$ = 0.36 per 100 000 pyrs

95% CI(r) = 2.24 ± 1.96 × 0.36 = 1.53 to 2.95 per 100 000 pyrs

For the 'exact' method, the lower limit factor (L) and the upper limit factor (U) corresponding to 39 cases are obtained from the table by interpolation between the rows for 35 and 40 cases.

$$L = 0.697 + \left[(0.714 - 0.697) \times \frac{39 - 35}{40 - 35}\right] = 0.711$$

$$U = 1.39 - \left[(1.39 - 1.36) \times \frac{39 - 35}{40 - 35}\right] = 1.37$$

Thus, the limits of the 95% confidence interval are

Lower limit = 2.24 per 100 000 pyrs × 0.711 = 1.59 per 100 000 pyrs
Upper limit = 2.24 per 100 000 pyrs × 1.37 = 3.07 per 100 000 pyrs

In this example, the 'exact' and the 'approximate' confidence limits are relatively close to each other, because the rate was based on a sufficiently large number of cases. The larger the number of cases, the closer will be the confidence limits obtained by these two methods.

Let us now consider some data from Kuwait. The total number of deaths from stomach cancer among men aged 45–54 years in this country in

1980 was only 3 in 74 000 pyrs (WHO, 1983). The 'approximate' method gives

$$r = 3/74\,000 \text{ pyrs} = 4.05 \text{ per } 100\,000 \text{ pyrs}$$

$$\text{SE}(r) = 4.05 \text{ per } 100\,000 \text{ pyrs}/\sqrt{3} = 2.34 \text{ per } 100\,000 \text{ pyrs}$$

$$95\%\ \text{CI}(r) = 4.05 \pm 1.96 \times 2.34 = -0.54 \text{ to } 8.64 \text{ per } 100\,000 \text{ pyrs}$$

This method gives a negative value for the lower limit, which is meaningless, as incidence and mortality rates cannot be negative. By the 'exact' method, consulting again Table A6.1.1, the limits for the 95% confidence interval are:

$$\text{Lower limit} = 4.05 \text{ per } 100\,000 \text{ pyrs} \times 0.206 = 0.83 \text{ per } 100\,000 \text{ pyrs}$$
$$\text{Upper limit} = 4.05 \text{ per } 100\,000 \text{ pyrs} \times 2.92 = 11.83 \text{ per } 100\,000 \text{ pyrs}$$

In this example, the 'exact and' 'approximate' confidence intervals are clearly different. When the number of cases is less than about 30, it is desirable to use the 'exact' method.

A6.1.2 Calculation of confidence intervals for measures of effect

Ratio of proportions (prevalence ratio or risk ratio)

A formula for the confidence interval around a risk ratio estimate of effect must take into account the fact that the sampling distribution of possible values for the risk ratio is highly skewed to the right. The minimum possible value a risk ratio can take is zero and the maximum is infinity. To make the distribution more symmetric, it is necessary to first convert the estimated risk ratios into their natural logarithms (denoted ln). We can then use formulae analogous to those presented in Section A6.1.1 to calculate a confidence interval around the value of the logarithm of the risk ratio rather than the risk ratio itself.

Consider the following example, in which 1000 exposed subjects and 1500 unexposed subjects were followed up for one year. The follow-up was complete for each subject. At the end of this period, 60 subjects among the exposed and 45 among the unexposed had developed the outcome of interest (Table A6.1.2).

Table A6.1.2.
Results from a cohort study in which risks were calculated as measures of occurrence of the outcome of interest in each study group: hypothetical data.

		Exposure		Total
		Yes	No	
Outcome	**Yes**	60 (a)	45 (b)	105 (n_1)
	No	940 (c)	1455 (d)	2395 (n_0)
Total		1000 (m_1)	1500 (m_0)	2500 (N)

The risk ratio (R) and its natural logarithm can be calculated as

$$R = p_1/p_0$$

$$\ln R = \ln (p_1/p_0)$$

An 'approximate' standard error of the logarithm of R can be estimated by

$$\text{SE}(\ln R) = \sqrt{\frac{1}{a} + \frac{1}{b} - \frac{1}{m_1} - \frac{1}{m_0}}$$

An 'approximate' 95% confidence interval for ln R is then given by (ln R) ± 1.96 SE(ln R), and the 95% confidence interval for the risk ratio (R) obtained by taking antilogarithms.

Thus, in the example shown in Table A6.1.2,

Risk in the exposed (p_1) = 60/1000 = 0.06

Risk in the unexposed (p_0) = 45/1500 = 0.03

Risk ratio (R) = 0.06/0.03 = 2.0

ln R = ln 2.0 = 0.69

$$\text{SE}(\ln R) = \sqrt{\frac{1}{60} + \frac{1}{45} - \frac{1}{1000} - \frac{1}{1500}} = 0.19$$

95% CI (ln R) = 0.69 ± 1.96 × 0.19 = 0.318 to 1.062

The 'approximate' 95% confidence interval of the risk ratio (R) can then be obtained by taking antilogarithms:

95% CI (R) = $e^{0.318}$ to $e^{1.062}$ = 1.37 to 2.89

A similar approach can be applied when the measure of interest is a *prevalence ratio*.

'Exact' methods should be used when the risk ratio or the prevalence ratio is based on small numbers of cases, but the calculations are too complex to be shown here.

Difference of proportions (prevalence difference or risk difference)

The standard error of the difference between two proportions p_1 and p_0 can be estimated, approximately, as

$$\text{SE}\,(p_1 - p_0) = \sqrt{\frac{p_1^{\,2}(1-p_1)}{a} + \frac{p_0^{\,2}(1-p_0)}{b}}$$

where a and b are the numbers of cases in the two study groups.

In the example shown in Table A6.1.2,

$$\text{Risk difference } (p_1 - p_0) = 0.06 - 0.03 = 0.03$$

$$\text{SE}\,(p_1 - p_0) = \sqrt{\frac{0.06^2\,(1-0.06)}{60} + \frac{0.03^2\,(1-0.03)}{45}} = 0.0087$$

$$\begin{aligned}95\%\ \text{CI}\,(p_1 - p_0) &= (p_1 - p_0) \pm 1.96\ \text{SE}\,(p_1 - p_0)\\ &= 0.03 \pm 1.96 \times 0.0087\\ &= 0.013 \text{ to } 0.047 \text{ or } 1\% \text{ to } 5\%\end{aligned}$$

A confidence interval for a *difference in prevalences* will be calculated in the same way.

Rate ratio

Consider the results from another hypothetical cohort study, shown in Table A6.1.3.

Table A6.1.3. Results from a cohort study in which rates were calculated as measures of occurrence of the outcome of interest in each study group: hypothetical data.

	Exposure		Total
	Yes	**No**	
Cases	60 (a)	45 (b)	105 (n)
Person-years at risk (pyrs)	4150 (y_1)	6500 (y_0)	10 650 (y)
Rate per 1000 pyrs	14.5 (r_1)	6.9 (r_0)	9.9 (r)

As with a risk ratio, a rate ratio can only take values from zero to infinity. Thus to construct a confidence interval for an estimated rate ratio (RR), its natural logarithm needs to be calculated first:

$$\ln \text{RR} = \ln\,(r_1/r_0)$$

An 'approximate' standard error of the logarithm of a rate ratio (RR) can be obtained as follows:

$$\text{SE}\,(\ln \text{RR}) = \sqrt{(1/a + 1/b)}$$

where a and b are the numbers of cases in the exposed and unexposed groups, respectively.

In this example, the incidence rate in the exposed (r_1) is equal to 60/4150=14.5 per 1000 pyrs. The incidence rate in the unexposed group (r_0) is 45/6500=6.9 per 1000 pyrs. Thus the rate ratio and its logarithm are:

$$\text{RR} = 14.5 \text{ per 1000 pyrs}/6.9 \text{ per 1000 pyrs} = 2.1$$

$$\ln \text{RR} = \ln 2.1 = 0.74$$

An 'approximate' standard error for the logarithm of a rate ratio of 2.1 based on 60 cases in the exposed group and 45 cases in the unexposed group may be calculated as follows:

$$\text{SE}(\ln \text{RR}) = \sqrt{(1/60 + 1/45)} = 0.197$$

The 'approximate' 95% confidence interval of the logarithm of the rate ratio is given by

$$\begin{aligned} 95\%\ \text{CI}(\ln \text{RR}) &= \ln \text{RR} \pm 1.96\ \text{SE}(\ln \text{RR}) \\ &= 0.74 \pm 1.96 \times 0.197 \\ &= 0.35 \text{ to } 1.13 \end{aligned}$$

We can then obtain the 'approximate' 95% confidence interval for the rate ratio by taking the antilogarithms of these values:

$$95\%\ \text{CI}(\text{RR}) = e^{0.35} \text{ to } e^{1.13} = 1.42 \text{ to } 3.10$$

There is also an 'exact' method of calculating confidence intervals for rate ratios that are based on small numbers of cases, but its discussion is beyond the scope of this chapter (see Breslow & Day, 1987, pp. 93–95).

When the rate ratio is an SMR (or SIR) (see Section 4.3.3), it is possible to calculate an 'exact' 95% confidence interval by multiplying the observed SMR (or SIR) by the appropriate lower and upper limit factors, exactly as we did for a single rate.

For instance, if the number of observed (O) leukaemia deaths in a certain town were 20 and only 15 would have been expected (E) if the town had the same age-specific rates as the whole country, the SMR would be equal to 1.33. The lower and the upper limit factors when the observed number of cases is 20 (see Table A6.1.1) are 0.611 and 1.54, respectively. Thus,

$$\text{SMR} = O/E = 20/15 = 1.33$$

$$\begin{aligned} 95\%\ \text{CI}(\text{SMR}) &= 1.33 \times 0.611 \text{ to } 1.33 \times 1.54 \\ &= 0.81 \text{ to } 2.05 \end{aligned}$$

Rate difference

The standard error of the difference between two estimated rates (r_1 and r_0) is given by

$$\text{SE}\,(r_1 - r_0) = \sqrt{\frac{r_1^2}{a} + \frac{r_0^2}{b}}$$

where a and b refer to numbers of cases in the two groups.

The 95% confidence interval is given by

$$95\%\ \text{CI}\,(r_1 - r_0) = (r_1 - r_0) \pm 1.96\ \text{SE}(r_1 - r_0)$$

Thus in the example shown in Table A6.1.3,

$$r_1 - r_0 = 14.5 \text{ per 1000 pyrs} - 6.9 \text{ per 1000 pyrs} = 7.6 \text{ per 1000 pyrs}$$

$$\text{SE}\,(r_1 - r_0) = \sqrt{((0.0145)^2/60 + (0.0069)^2/45)} = 0.00214$$
$$= 2.14 \text{ per 1000 pyrs}$$

$$95\%\ \text{CI}\,(r_1 - r_0) = 7.6 \pm 1.96 \times 2.14 = 3.41 \text{ to } 11.79 \text{ per 1000 pyrs}$$

Odds ratio

Data from a case–control study can be presented in a 2 × 2 table, as shown below:

Table A6.1.4. Results from a case–control study: hypothetical data.

	Exposure		Total
	Yes	**No**	
Cases	457 (a)	26 (b)	483 (n_1)
Controls	362 (c)	85 (d)	447 (n_0)
Total	819 (m_1)	111 (m_0)	930 (N)

An 'approximate' standard error of the logarithm of an odds ratio (OR) can be calculated as

$$\text{SE}\,(\ln \text{OR}) = \sqrt{(1/a + 1/b + 1/c + 1/d)}$$

In the example shown in Table A6.1.4,

$$\text{OR} = \frac{\text{odds of exposure among the cases}}{\text{odds of exposure among the controls}} = \frac{457/26}{362/85} = 4.13$$

$$\ln \text{OR} = \ln 4.13 = 1.42$$

$$\text{SE}\,(\ln \text{OR}) = \sqrt{(1/457 + 1/26 + 1/362 + 1/85)} = 0.23$$

$$95\%\ \text{CI}\,(\ln \text{OR}) = 1.42 \pm 1.96 \times 0.23 = 0.97 \text{ to } 1.87$$

Thus an 'approximate' 95% confidence interval for the odds ratio can be obtained by taking antilogarithms:

$$95\% \text{ CI (OR)} = e^{0.97} \text{ to } e^{1.87} = 2.64 \text{ to } 6.49$$

It is also possible to calculate an 'exact' confidence interval for small samples, but the calculations are too complex to be carried out on a pocket calculator.

A.6.1.3 Statistical tests

Comparison of two proportions (prevalences or risks)

To test the null hypothesis that there is no true difference between two proportions (either *risks* or *prevalences*), the results from a study should first be arranged in a 2 × 2 table similar to Table A6.1.2. In this table, the *observed* (*O*) number of cases among exposed is *a*. We can calculate the *expected* (*E*) value in cell *a* and the *variance* (*V*), assuming that the null hypothesis of no difference between the two groups is true.

$$O = a$$

$$E = m_1 n_1 / N$$

$$V = \frac{n_1 n_0 m_1 m_0}{N^2 (N - 1)}$$

A special statistical test called the chi-squared (χ^2) test can then be applied to measure the extent to which the observed data differ from those expected if the two proportions were equal, that is, if the null hypothesis were true.

$$\chi^2 = (O - E)^2 / V$$

In epidemiology, this application of the χ^2 test takes the name of *Mantel–Haenszel test*.

In the example shown in Table A6.1.2,

$$O = 60$$

$$E = 1000 \times 105/2500 = 42$$

$$V = \frac{105 \times 2395 \times 1000 \times 1500}{2500^2 \times (2500 - 1)} = 24.15$$

$$\chi^2 = \frac{(60 - 42)^2}{24.15} = 13.42$$

Large values of χ^2 suggest that the data are inconsistent with the null hypothesis, and therefore that there is an association between exposure and outcome. The *P*-value is obtained by comparing the calculated value of χ^2 with tables of the chi-squared distribution.

In referring the calculated χ^2 test statistics to these tables, we need to know a quantity called the 'degrees of freedom' (d.f.), which takes into consideration the number of sub-groups or 'cells' in the table which contributed to the calculation. For 2×2 tables, the number of degrees of freedom (d.f.) is one.

If the null hypothesis is true,

χ^2 test statistic (with 1 d.f.) exceeds	0.45	with probability	0.5
"	1.32	"	0.25
"	2.71	"	0.1
"	3.84	"	0.05
"	5.02	"	0.025
"	6.63	"	0.01
"	7.88	"	0.005
"	10.83	"	0.001

Thus, if the χ^2 statistic with 1 d.f. exceeds 3.84, then $P<0.05$, indicating some evidence of a real difference in the proportions. If it exceeds 6.63, then $P < 0.01$, and there is strong evidence of a difference.

In the example shown in Table A6.1.2, the value of χ^2 was 13.42, which corresponds to $P < 0.001$. There is therefore strong evidence for an association. Thus we can conclude that the observed risk ratio of 2.0 is statistically significantly different from 1, and that there is very strong evidence that the risk is higher in those who were exposed than in those who were not.

In fact, we have already performed a test for difference in two proportions in Section 6.2.2. We then used a different test statistic which gave similar results to the more general Mantel–Haenszel type of test statistic used here.

Note that the statistical test is the same regardless of the measure of effect (*risk (or prevalence) ratio* or *risk (or prevalence) difference*) that we are interested in. However, the confidence intervals are calculated in a different way (see Section A6.1.2).

Comparison of two odds

The approach discussed above for comparison of proportions can also be used to test the null hypothesis that there is no difference between the odds of exposure among the cases and the odds of exposure among the controls, that is, the odds ratio is equal to one.
In the example shown in Table A6.1.4,

$$O = 457$$

$$E = \frac{819 \times 483}{930} = 425.35$$

$$V = \frac{483 \times 447 \times 819 \times 111}{930^2 \times (930 - 1)} = 24.43$$

$$\chi^2 = \frac{(457 - 425.35)^2}{24.43} = 41.0$$

The χ^2 gives a measure of the extent to which the observed data differ from those expected if the two odds of exposure were equal. This χ^2 value (with one degree of freedom) corresponds to $P<0.001$. Thus, there is strong evidence against the null hypothesis.

Comparison of two rates

In cohort studies, where rates rather than risks are used as the measure of disease frequency, consideration must be given to the distribution of person-years between exposed and unexposed groups.

Consider again the example shown in Table A6.1.3. The observed number of cases among those who were exposed is $a = 60$. The expected value in cell a and the variance assuming that the null hypothesis is true (i.e., that there is no true difference in rates between the exposed and the unexposed groups) can be calculated as follows:

$$E = ny_1/y \text{ and } V = ny_1y_0/y^2$$

Then

$$\chi^2 = (O - E)^2/V$$

In the example shown in Table A6.1.3,

$$O = 60$$

$$E = \frac{105 \times 4150}{10\,650} = 40.92$$

$$V = \frac{105 \times 4150 \times 6500}{(10\,650)^2} = 24.97$$

$$\chi^2 = \frac{(60 - 40.92)^2}{24.97} = 14.58$$

This χ^2 value with one degree of freedom corresponds to $P < 0.001$, pro-

viding strong evidence against the null hypothesis of no association between exposure and the incidence of the disease.

The same procedure applies when the rate ratio is an SMR. In this case, the variance is equal to the expected number of cases. Thus if the number of observed leukaemia deaths (O) in a particular town is 20 and the expected number (E) based on the national age-specific rates is 15,

$$\text{SMR} = O/E = 20/15 = 1.33$$

$$V = 15$$

$$\chi^2 = \frac{(20 - 15)^2}{15} = 1.67$$

This χ^2 value, with one degree of freedom, corresponds to $0.1 < P < 0.25$. Thus, these results are consistent with the null hypothesis of no difference in the age-specific mortality rates from leukaemia between the town and the whole country.

Note that the statistical test is the same regardless of the measure of effect (*rate ratio* or *rate difference*) we are interested in. However, the confidence intervals are calculated in a different way (see Section A6.1.2).

χ^2 test for a linear trend in proportions (prevalences or risks)

So far we have considered only situations where individuals were classified as either 'exposed' or 'unexposed'. In many circumstances, how-ever, individuals can also be classified according to levels of exposure. For instance, suppose that a survey was carried out to assess whether infection with human papillomavirus (HPV) was associated with number of sexual partners. The results from this hypothetical study are shown in Table A6.1.5.

Table A6.1.5. Distribution of women infected and not infected with human papillomavirus (HPV) by number of lifetime sexual partners: hypothetical data.

	Lifetime number of sexual partners				Total
	1	2–3	4–5	>5	
HPV-positive	19(a_0)	33(a_1)	51(a_2)	107(a_3)	210(n_1)
HPV-negative	71(b_0)	68(b_1)	42(b_2)	61(b_3)	242(n_0)
Total	90(m_0)	101(m_1)	93(m_2)	168(m_3)	452(N)
Percentage of HPV positive	21.1	32.7	54.8	63.7	46.5
Score	0(x_0)	1(x_1)	2(x_2)	3(x_3)	

The results seem to support the study hypothesis of a trend for an increase in the proportion of HPV-positive women with increasing number of sexual partners. Although there is an apparent linear trend in proportions in the above table, each proportion (or percentage) is subject to sampling variability. We can use a χ^2 test for a linear trend in proportions to assess whether this trend may be due to chance.

The first step is to assign a score x to each exposure category. For example, in Table A6.1.5 '0' was assigned to those women with 1 partner, '1' to those with 2–3 partners, and so on. The second step is to use the following formulae to obtain the values of T_1, T_2 and T_3. In these formulae, the symbol $\sum$ means sum and the subscript i stands for the subscripts 0, 1, 2, 3, etc.

$$T_1 = \sum a_i x_i = (19 \times 0) + (33 \times 1) + (51 \times 2) + (107 \times 3) = 456$$

$$T_2 = \sum m_i x_i = (90 \times 0) + (101 \times 1) + (93 \times 2) + (168 \times 3) = 791$$

$$T_3 = \sum m_i x_i^2 = (90 \times 0^2) + (101 \times 1^2) + (93 \times 2^2) + (168 \times 3^2) = 1985$$

The χ^2 test for trend has one degree of freedom and can be calculated as

$$\chi^2 = \frac{N\,(NT_1 - n_1 T_2)^2}{n_1 n_0\,(N\,T_3 - T_2^2)}$$

Thus, in our example

$$\chi^2 = \frac{452 \times (452 \times 456 - 210 \times 791)^2}{210 \times 242 \times (452 \times 1985 - 791^2)} = 52.41$$

A χ^2 of 52.41 with 1 d.f. corresponds to $P<0.0001$. We can therefore conclude that there is strong evidence of a linear trend for an increasing proportion of HPV-positive women as lifetime number of sexual partners increases.

χ^2 test for a linear trend in odds ratios

Consider data from a hypothetical case–control study carried out to assess whether there is a decreasing risk of developing epithelial benign ovarian tumours with increasing parity (Table A6.1.6).

The results from this study apparently support the study hypothesis. (The calculation of odds ratios from case–control studies is discussed in detail in Chapter 9.)

Table A6.1.6. Distribution of cases of benign tumours of the ovary and controls by parity: hypothetical data.

	Parity			Total
	0[a]	**1–2**	**≥3**	
Benign tumour cases	30 (a_0)	23 (a_1)	7 (a_2)	60 (n_1)
Controls	46 (b_0)	48 (b_1)	35 (b_2)	129 (n_0)
Total	76 (m_0)	71 (m_1)	42 (m_2)	189 (N)
Odds ratio	1	0.73	0.31	
Score	0 (x_0)	1 (x_1)	2 (x_2)	

[a] Taken as the baseline category in the calculation of odds ratios.

A χ^2 test for a linear trend can be used to test the hypothesis that there is a decreasing risk of ovarian benign tumours with increasing parity. The calculations are exactly the same as those used for the χ^2 test for a linear trend in proportions.

$$T_1 = \sum a_i x_i = (30 \times 0) + (23 \times 1) + (7 \times 2) = 37$$

$$T_2 = \sum m_i x_i = (76 \times 0) + (71 \times 1) + (42 \times 2) = 155$$

$$T_3 = \sum m_i x_i^2 = (76 \times 0^2) + (71 \times 1^2) + (42 \times 2^2) = 239$$

The χ^2 test for trend can then be calculated as:

$$\chi^2 = \frac{189 \times (189 \times 37 - 60 \times 155)^2}{60 \times 129 \times (189 \times 239 - 155^2)} = 6.15$$

This test result with 1 d.f. corresponds to $0.01 < P < 0.025$. Thus there is evidence that the risk of developing a benign ovarian tumour decreases with increasing parity.

χ^2 test for a linear trend in rates

A similar approach can be used to test for a trend in rates. Consider the following cohort study to test the hypothesis that the risk of breast cancer increases with increasing duration of oral contraceptive use.

Table A6.1.7. Distribution of breast cancer cases and person-years at risk by duration of oral contraceptive use: hypothetical data.

	Duration of oral contraceptive use (years)			Total
	0	1–2	≥3	
Breast cancer cases	62 (a_0)	42 (a_1)	22 (a_2)	126 (n)
Person-years at risk	31 200 (y_0)	25 100 (y_1)	11 600 (y_2)	67 900 (y)
Rate per 100 000 pyrs	198.7(r_0)	167.3(r_1)	189.7(r_2)	185.6(r)
Score	0 (x_0)	1 (x_1)	2 (x_2)	

The observed rates by exposure level suggest, if anything, a downward trend with increasing duration of oral contraceptive use. A χ^2-test for a linear trend in rates similar to the one described above for proportions and odds ratios can be calculated as follows:

$$T_1 = \sum a_i x_i = (62 \times 0) + (42 \times 1) + (22 \times 2) = 86$$

$$T_2 = \sum y_i x_i = (31\,200 \times 0) + (25\,100 \times 1) + (11\,600 \times 2) = 48\,300$$

$$T_3 = \sum y_i x_i^2 = (31\,200 \times 0^2) + (25\,100 \times 1^2) + (11\,600 \times 2^2) = 71\,500$$

The χ^2 test for a linear trend in rates, which has one degree of freedom, can be calculated as follows:

$$\chi^2 = \frac{[T_1 - (n/y)T_2)]^2}{(n/y^2)\ (yT_3 - T_2^2)}$$

Thus, in our example,

$$\chi^2 = \frac{[86 - (126/67\ 900 \times 48\ 300)]^2}{(126/67\ 900^2) \times (67\ 900 \times 71\ 500 - 48\ 300^2)} = 0.19$$

This test value with 1 d.f. corresponds to $P > 0.5$. Hence, the results of the study provide no support for an upward or downward trend in breast cancer rates with duration of oral contraceptive use.

Validity of χ^2 tests

If in a 2×2 table the total sample size (N) is less than 40 and the expected value in any of the cells is less than 5, the χ^2 test should not be used. In these circumstances, the Fisher's exact test will be the appropriate statistical test (see Kirkwood (1988)). For larger tables, the χ^2 test is valid if no more than 20% of the expected values are less than 5, and none is less than one.

Note that the expected value (E) for a particular cell is calculated as follows:

$$E = \frac{\text{Total of the relevant row} \times \text{total of the relevant column}}{N}$$

Thus in Table A6.1.4,

$$E\ (a) = n_1 m_1/N = (483 \times 819)/930 = 425.35$$

$$E\ (b) = n_1 m_0/N = (483 \times 111)/930 = 57.65$$

$$E\ (c) = n_0 m_1/N = (447 \times 819)/930 = 393.65$$

$$E\ (d) = n_0 m_0/N = (447 \times 111)/930 = 53.35$$

The χ^2 test is valid in this example since the total sample size (N) is greater than 40 and all of the expected cell values are well above 5.