8 Estimation

8.1 Sampling distributions

We saw in Chapter 3 how samples are drawn from much larger populations. Data are collected about the sample so that we can find out something about the population. We use samples to estimate quantities such as disease prevalence, mean blood pressure, mean exposure to a carcinogen, etc. We also want to know by how much these estimates might vary from sample to sample.

In Chapters 6 and 7 we saw how the theory of probability enables us to link random samples with the populations from which they are drawn. In this chapter we shall see how probability theory enables us to use samples to estimate quantities in populations, and to determine the precision of these estimates. First we shall consider what happens when we draw repeat samples from the same population. Table 8.1 shows a set of 100 random digits which we can use as the population for a sampling experiment. The distribution of the numbers in this population is shown in Figure 8.1. The population mean is 4.7 and the standard deviation is 2.9.

The sampling experiment is done using a suitable random sampling method to draw repeated samples from the population. In this case decimal dice were a convenient method. A sample of four observations was chosen: 6, 4, 6, and 1. The mean was calculated: 17/4 = 4.25. This was repeated to draw a second sample of four numbers: 7, 8, 1, 8. Their mean is 6.00. This sampling procedure was done 20 times altogether, to give the samples and their means shown in Table 8.2.

These sample means are not all the same. They show random variation. If we were able to draw all of the 3 921 225 possible samples of size 4 and calculate their means, these means themselves would form a distribution. Our 20 sample means are a sample from this distribution. The distribution of all possible sample means is called the **sampling distribution** of the mean.

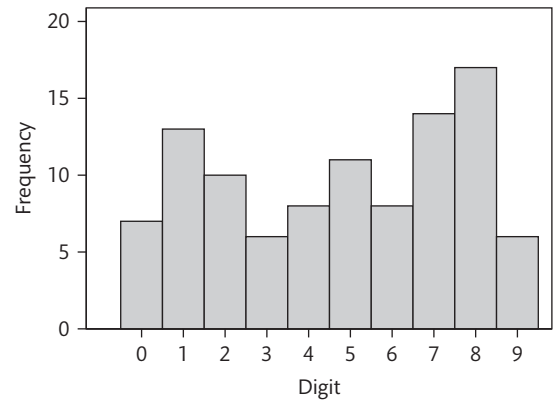

Figure 8.1 Distribution of the population of Table 8.1.

Table 8.1 Population of 100 random digits for a sampling experiment

9	1	0	7	5	6	9	5	8	8	1	0	5	7	6	5	0	2	1	2
1	8	8	8	5	2	4	8	3	1	6	5	5	7	4	1	7	3	3	3
2	8	1	8	5	8	4	0	1	9	2	1	6	9	4	4	7	6	1	7
1	9	7	9	7	2	7	7	0	8	1	6	3	8	0	5	7	4	8	6
7	0	2	8	8	7	2	5	4	1	8	6	8	3	5	8	2	7	2	4

Table 8.2 Random samples drawn in a sampling experiment

Sample	6	7	7	1	5	5	4	7	2	8
	4	8	9	8	2	5	2	4	8	1
	6	1	2	8	9	7	7	0	7	2
	1	8	7	4	5	8	6	1	7	0
Mean	4.25	6.00	6.25	5.25	5.25	6.25	4.75	3.00	6.00	2.75
Sample	7	7	2	8	3	4	5	4	4	7
	8	3	5	0	7	8	5	3	5	4
	7	8	0	7	4	7	8	1	8	6
	2	7	8	7	8	7	3	6	2	3
Mean	6.00	6.25	3.75	5.50	5.50	6.50	5.25	3.50	4.75	5.00

In general, the sampling distribution of any statistic is the distribution of the values of the statistic which would arise from all possible samples.

8.2 Standard error of a sample mean

For the moment we shall consider the sampling distribution of the mean only. As our sample of 20 means is a random sample from it, we can use this to estimate some of the parameters of the distribution. The 20 means have their own mean and standard deviation. The mean is 5.1 and the standard deviation is 1.1. Now the mean of the whole population is 4.7, which is close to the mean of the samples. But the standard deviation of the population is 2.9, which is considerably greater than that of the sample means. If we plot a histogram for the sample of means (Figure 8.2), we see that the centre of the sampling distribution and the parent population distribution are the same, but the scatter of the sampling distribution is much less.

Another sampling experiment, on a larger scale, will illustrate this further. This time our parent distribution will be the Normal distribution with mean 0 and standard deviation 1. Figure 8.3 shows the distribution of a random sample of 1000 observations from this distribution. Figure 8.3 also shows the distribution of means from 1000 random samples of size 4 from this population, the same sample size as in Figure 8.2. Figure 8.3

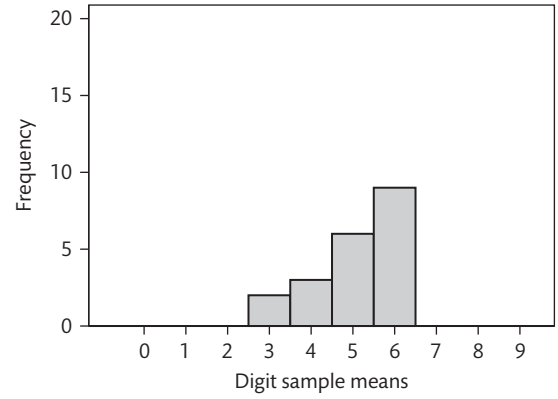

Figure 8.2 Distribution of the sample of the means of Table 8.2.

also shows the distributions of 1000 means of samples of size 9 and of size 16. In all four distributions the means are close to zero, the mean of the parent distribution. But the standard deviations are not the same. They are, in fact, approximately 1 (parent distribution); 1/2 (means of 4), 1/3 (means of 9), and 1/4 (means of 16). In fact, if the observations are independent of one another, the sampling distribution of the mean has standard deviation $\sigma/\sqrt{n}$ or $\sqrt{\sigma^2/n}$, where σ is the standard deviation of the parent distribution and n is the sample size (Appendix 8A). The mean of the sampling distribution is equal to the mean of the parent distribution. The actual, as opposed to simulated, distribution of the mean of four observations from a Normal distribution is shown in Figure 8.4.

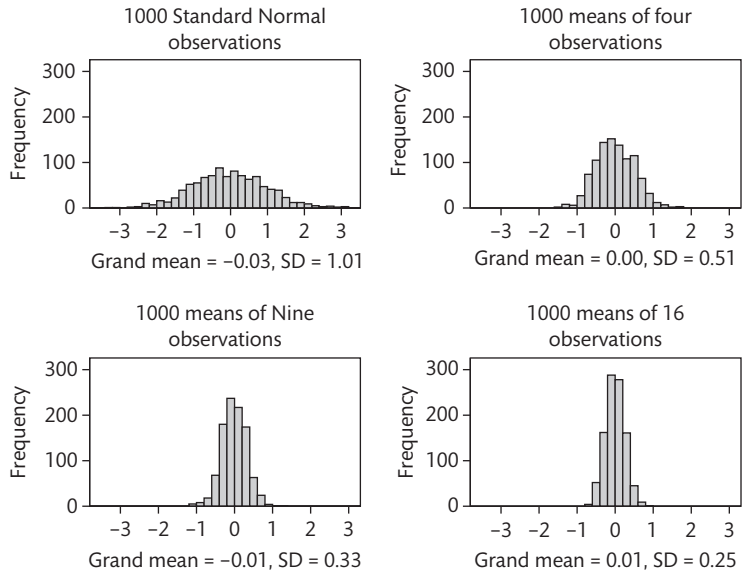

Figure 8.3 Samples of means from a Standard Normal variable.

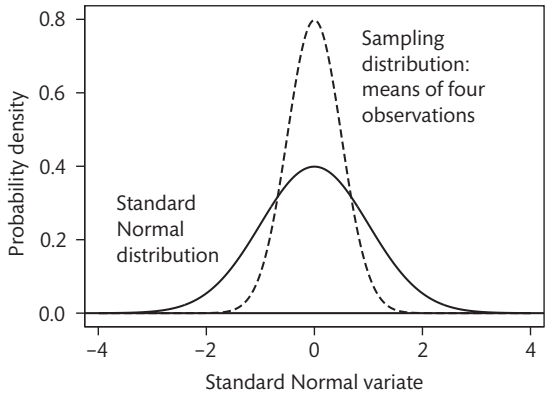

Figure 8.4 Sampling distribution of the mean of four observations from a Standard Normal distribution.

The sample mean is an estimate of the population mean. The standard deviation of its sampling distribution is called the **standard error** of the estimate. It provides a measure of how far from the true value the estimate is likely to be. In most estimation, the estimate is likely to be within one standard error of the true mean and unlikely to be more than two standard errors from it. We shall look at this more precisely in Section 8.3.

In almost all practical situations we do not know the true value of the population variance σ^2 but only its estimate s^2 (Section 4.7). We can use this to estimate the standard error by $s/\sqrt{n}$. This estimate is also referred to as the standard error of the mean. It is usually clear from the context whether the standard error is the true value or that estimated from the data.

When the sample size n is large, the sampling distribution of the sample mean, $\bar{x}$, tends to a Normal distribution (Section 7.3). Also, we can assume that s^2 is a good estimate of σ^2. So for large n, $\bar{x}$ is, in effect, an observation from a Normal distribution with mean μ and standard deviation estimated by $s/\sqrt{n}$. So with probability 0.95 or for 95% of possible samples, $\bar{x}$ is within 1.96 standard errors of μ. With small samples we cannot assume either a Normal distribution or, more importantly, that s^2 is a good estimate of σ^2. We shall discuss this in Chapter 10.

For an example, consider the 57 FEV1 measurements of Table 4.4. We have $\bar{x}$ = 4.062 litres, s^2 = 0.449 174, s = 0.67 litres. Then the standard error of $\bar{x}$ is $\sqrt{s^2/n}$ = $\sqrt{0.449\,174/57}$ = $\sqrt{0.007\,880}$ = 0.089. The best estimate of the mean FEV1 in the population is then 4.06 litres with standard error 0.089 litres.

The mean and standard error are often written as 4.062 ± 0.089. This is rather misleading, as the true value may be up to two standard errors from the mean with a reasonable probability. This practice is not recommended.

There is often confusion between the terms 'standard error' and 'standard deviation'. This is understandable, as

the standard error is a standard deviation (of the sampling distribution) and the terms are often interchanged in this context. The convention is this: we use the term 'standard error' when we measure the precision of estimates, and the term 'standard deviation' when we are concerned with the variability of samples, populations, or distributions. If we want to say how good our estimate of the mean FEV1 measurement is, we quote the standard error of the mean. If we want to say how widely scattered the FEV1 measurements are, we quote the standard deviation, *s*.

8.3 Confidence intervals

The estimate of mean FEV1 is a single value and so is called a **point estimate.** There is no reason to suppose that the population mean will be exactly equal to the point estimate, the sample mean. It is likely to be close to it, however, and the amount by which it is likely to differ from the estimate can be found from the standard error. What we do is find limits which are likely to include the population mean, and say that we estimate the population mean to lie somewhere in the interval (the set of all possible values) between these limits. This is called an **interval estimate**.

For instance, if we regard the 57 FEV measurements as being a large sample we can assume that the sampling distribution of the mean is Normal, and that the standard error is a good estimate of its standard deviation (see Section 10.6 for a discussion of how large is large). We therefore expect about 95% of such means to be within 1.96 standard errors of the population mean, μ. Hence, for about 95% of all possible samples, the population mean must be greater than the sample mean minus 1.96 standard errors and less than the sample mean plus 1.96 standard errors. If we calculated $\bar{x} - 1.96se$ and $\bar{x} + 1.96se$ for all possible samples, 95% of such intervals would contain the population mean. In this case these limits are $4.062 - 1.96 \times 0.089$ to $4.062 + 1.96 \times 0.089$ which gives 3.89 to 4.24, or 3.9 to 4.2 litres, rounding to two significant figures. 3.9 and 4.2 are called the **95% confidence limits** for the estimate, and the set of values between 3.9 and 4.2 is called the **95% confidence interval**. The confidence limits are the values at the ends of the confidence interval.

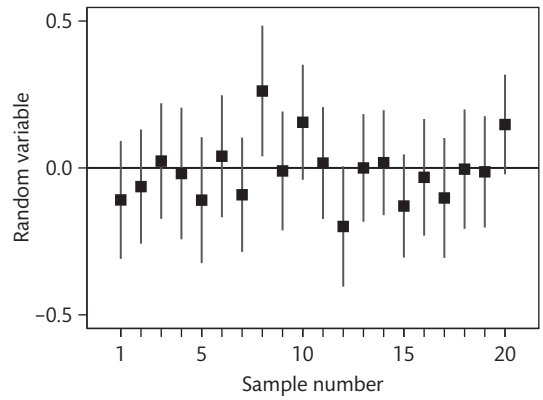

Figure 8.5 Mean and 95% confidence interval for 20 random samples of 100 observations from the Standard Normal distribution.

Strictly speaking, it is incorrect to say that there is a probability of 0.95 that the population mean lies between 3.9 and 4.2, though it is often put that way (even by me, though I try to avoid this). The population mean is a number, not a random variable, and has no probability. (This is the Frequency School view of probability, see Chapter 22 for a different, Bayesian view.) It is the probability that limits calculated from a random sample will include the population value which is 95%. Figure 8.5 shows confidence intervals for the mean for 20 random samples of 100 observations from the Standard Normal distribution. The population mean is, of course, 0.0, shown by the horizontal line. Some sample means are close to 0.0, some further away, some above, and some below. The population mean is contained by 19 of the 20 confidence intervals. In general, for 95% of confidence intervals it will be true to say that the population value lies within the interval. We just don't know which 95%. This is sometimes expressed by saying that we are 95% confident that the mean lies between these limits.

In the FEV1 example, the sampling distribution of the mean is Normal and its standard deviation is well estimated because the sample is large. This is not always true and although it is usually possible to calculate confidence intervals for an estimate, they are not all quite as simple as that for the mean estimated from a large sample. We shall look at the mean estimated from a small sample in Section 10.2.

There is no necessity for the confidence interval to have a probability of 95%. For example, we can also

calculate 99% confidence limits. The upper 0.5% point of the Standard Normal distribution is 2.58 (Table 7.2), so the probability of a Standard Normal deviate being above 2.58 or below −2.58 is 1% and the probability of being within these limits is 99%. The 99% confidence limits for the mean FEV1 are therefore, 4.062 − 2.58 × 0.089 and 4.062 + 2.58 × 0.089, i.e. 3.8 and 4.3 litres. These give a wider interval than the 95% limits, as we would expect as we are more confident that the population mean will be included. The probability we choose for a confidence interval is a compromise between the desire to include the estimated population value and the desire to avoid parts of scale where there is a low probability that the mean will be found. For most purposes, 95% confidence intervals have been found to be satisfactory.

Standard error is not the only way in which we can calculate confidence intervals, although at present it is the one used for most problems. Others are described in Section 8.9, Section 8.10, and Section 8.11. There are others, which I shall omit because they are rarely used.

8.4 Standard error and confidence interval for a proportion

The standard error of a proportion estimate can be calculated in the same way. Suppose the proportion of individuals who have a particular condition in a given population is p, and we take a random sample of size n, the number observed with the condition being r. Then the estimated proportion is r/n. We have seen (Section 6.4) that r comes from a Binomial distribution with mean np and variance $np(1-p)$. Provided n is large, this distribution is approximately Normal. So r/n, the estimated proportion, is from a Normal distribution with mean given by $np/n = p$, and variance given by

$$\begin{aligned}\text{VAR}\left(\frac{r}{n}\right) &= \frac{1}{n^2}\text{VAR}(r)\\ &= \frac{1}{n^2}np(1-p)\\ &= \frac{p(1-p)}{n}\end{aligned}$$

as n is constant, and the standard error is

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

We can estimate this by replacing p by r/n. As for the sample mean, this standard error is only valid if the observations are independent of one another. For example, in a random sample of first year secondary schoolchildren in Derbyshire (Banks *et al.* 1978), 118 out of 2 837 boys said that they usually coughed first thing in the morning. This gave a prevalence estimate of 118/2 837 = 0.041 6, with standard error $\sqrt{0.041\,6 \times (1 - 0.041\,6)/2\,837}$ = 0.003 7. The sample is large so we can assume that the estimate is from a Normal distribution and that the standard error is well estimated. The 95% confidence interval for the prevalence is thus 0.041 6 − 1.96 × 0.003 7 to 0.041 6 + 1.96 × 0.003 7 = 0.034 to 0.049. Even with this fairly large sample, the estimate is not very precise. This confidence interval, using the standard error, is called the **Wald interval**.

The standard error of the proportion is only of use if the sample is large enough for the Normal approximation to apply. A rough guide to this is that np and $n(1-p)$ should both exceed 5. This is usually the case when we are concerned with straightforward estimation. If we try to use the method for smaller samples, we may get absurd results. For example, in a study of the prevalence of HIV in ex-prisoners (Turnbull *et al.* 1992), of 29 women who did not inject drugs, one was HIV positive. The authors reported this to be 3.4%, with a 95% confidence interval −3.1% to 9.9%. The lower limit of −3.1%, obtained from the observed proportion minus 1.96 standard errors, is impossible. As Newcombe (1992) pointed out, the correct 95% confidence interval can be obtained from the exact probabilities of the Binomial distribution and is 0.1% to 17.8% (Section 8.9).

8.5 The difference between two means

In many studies we are more interested in the difference between two population parameters than in their absolute value. These could be means, proportions, the slopes of lines, and many other statistics. When samples are large we can assume that sample means and proportions

are observations from a Normal distribution, and that the calculated standard errors are good estimates of the standard deviations of these Normal distributions. We can use this to find confidence intervals.

For example, suppose we wish to compare the means, $\bar{x}_1$ and $\bar{x}_2$, of two independent large samples, sizes n_1 and n_2. The expected difference between the sample means is equal to the difference between the population means, i.e. $E(\bar{x}_1 - \bar{x}_2) = \mu_1 - \mu_2$. What is the standard error of the difference? The variance of the difference between two independent random variables is the sum of their variances (Section 6.6). Hence, the standard error of the difference between two independent estimates is the square root of the sum of the squares of their standard errors. The standard error of a mean is $\sqrt{s^2/n}$, so the standard error of the difference between two independent means is

$$\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

For an example, in a study of respiratory symptoms in schoolchildren (Bland *et al.* 1974), we wanted to know whether children reported by their parents to have respiratory symptoms had worse lung function than children who were not reported to have symptoms. Ninety-two children were reported to have cough during the day or at night, and their mean PEFR was 294.8 litre/min with standard deviation 57.1 litre/min, and 1643 children were not reported to have this symptom, their mean PEFR being 313.6 litre/min with standard deviation 55.2 litre/min. We thus have two large samples, and can apply the Normal distribution. We have

$$n_1 = 92, \quad \bar{x}_1 = 294.8, \quad s_1 = 57.1,$$
$$n_2 = 1643, \quad \bar{x}_2 = 313.6, \quad s_2 = 55.2$$

The difference between the two group means is $\bar{x}_1 - \bar{x}_2 = 294.8 - 313.6 = -18.8$. The standard error of the difference is

$$\sqrt{se_1^2 + se_2^2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$
$$= \sqrt{\frac{57.1^2}{92} + \frac{55.2^2}{1643}}$$
$$= 6.11$$

We shall treat the sample as being large, so the difference between the means can be assumed to come from a Normal distribution and the estimated standard error to be a good estimate of the standard deviation of this distribution. (For small samples see Section 10.3 and Section 10.6) The 95% confidence limits for the difference are thus $-18.8 - 1.96 \times 6.11$ and $-18.8 + 1.96 \times 6.11$, i.e. −6.8 and −30.8 l/min. The confidence interval does not include zero, so we have good evidence that, in this population, children reported to have day or night cough have lower mean PEFR than others. The difference is estimated to be between 7 and 31 litre/min lower in children with the symptom, so it may be quite small.

When we have paired data, such as a cross-over trial (Section 2.7) or a matched case–control study (Section 3.8), the two-sample method does not work. Instead, we calculate the differences between the paired observations for each subject, then find the mean difference, its standard error, and confidence interval as in Section 8.3.

8.6 Comparison of two proportions

We can apply the method of Section 8.5 to two proportions. The standard error of a proportion p is $\sqrt{p(1-p)/n}$. For two independent proportions, p_1 and p_2, the standard error of the difference between them is

$$\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

Provided the conditions of Normal approximation are met (see Section 8.4), we can find a confidence interval for the difference in the usual way.

For example, consider Table 8.3. The researchers wanted to know to what extent children with bronchitis in infancy get more respiratory symptoms in later life than others. We can estimate the difference between the proportions reported to cough during the day or at night among children with and children without a history of bronchitis before age 5 years. We have estimates of two proportions, $p_1 = 26/273 = 0.095\,24$ and $p_2 = 44/1\,046 = 0.042\,07$. The difference between them

Table 8.3 Cough during the day or at night at age 14 and bronchitis before age 5 (data from Holland *et al.* 1978)

Cough at 14	Bronchitis at 5		
	Yes	**No**	**Total**
Yes	26	44	70
No	247	1 002	1 249
Total	273	1 046	1 319

is $p_1 - p_2 = 0.095\,24 - 0.042\,07 = 0.053\,17$. The standard error of the difference is

$$\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

$$= \sqrt{\frac{0.095\,24 \times (1 - 0.095\,24)}{273} + \frac{0.042\,07 \times (1 - 0.042\,07)}{1\,046}}$$

$$= \sqrt{0.000\,315\,639 + 0.000\,038\,528}$$

$$= \sqrt{0.000\,354\,167}$$

$$= 0.018\,8$$

The 95% confidence interval for the difference is $0.053\,17 - 1.96 \times 0.018\,8$ to $0.053\,17 + 1.96 \times 0.018\,8 = 0.016$ to 0.090. Although the difference is not very precisely estimated, the confidence interval does not include zero and gives us clear evidence that children with bronchitis reported in infancy are more likely than others to be reported to have respiratory symptoms in later life. The data on lung function in Section 8.5 gives us some reason to suppose that this is not entirely a result of response bias (Section 3.9). As in Section 8.4, the confidence interval must be estimated differently for small samples.

This difference in proportions may not be very easy to interpret. The ratio of two proportions is often more useful. Another method, the odds ratio, is described in Section 13.7. The ratio of the proportion with cough at age 14 for bronchitis before 5 to the proportion with cough at age 14 for those without bronchitis before 5 is $p_1/p_2 = 0.09524/0.04207 = 2.26$. Children with bronchitis before 5 are more than twice as likely to cough during the day or at night at age 14 than children with no such history.

The standard error for this ratio is complex, and as it is a ratio rather than a difference it does not approximate well to a Normal distribution. If we take the logarithm of the ratio, however, we get the difference between two logarithms, because $\log(p_1/p_2) = \log(p_1) - \log(p_2)$ (Appendix 5A). We can find the standard error for the log ratio quite easily. We use the result that, for any random variable X with mean μ and variance σ^2, the approximate variance of $\log(X)$ is given by $\text{VAR}(\log_e(X)) = \sigma^2/\mu^2$ (see Kendall and Stuart 1969). Hence, the variance of $\log(p)$ is

$$\text{VAR}(\log(p)) = \frac{p(1-p)/n}{p^2} = \frac{1-p}{np}$$

For the difference between the two logarithms we get

$$\text{VAR}(\log_e(p_1/p_2)) = \text{VAR}(\log_e(p_1)) + \text{VAR}(\log_e(p_2))$$

$$= \frac{1-p_1}{n_1 p_1} + \frac{1-p_2}{n_2 p_2}$$

The standard error is the square root of this. (This formula is often written in terms of frequencies, but I think this version is clearer.) For the example the log ratio is $\log_e(2.263\,85) = 0.817\,07$ and the standard error is

$$\sqrt{\frac{1-p_1}{n_1 p_1} + \frac{1-p_2}{n_2 p_2}}$$

$$= \sqrt{\frac{1 - 0.095\,24}{273 \times 0.095\,24} + \frac{1 - 0.042\,07}{1\,046 \times 0.042\,07}}$$

$$= \sqrt{\frac{0.904\,76}{26} + \frac{0.957\,93}{44}}$$

$$= \sqrt{0.056\,57}$$

$$= 0.237\,84$$

The 95% confidence interval for the log ratio is therefore $0.817\,07 - 1.96 \times 0.237\,84$ to $0.817\,07 + 1.96 \times 0.237\,84 = 0.350\,89$ to 1.283 24. The 95% confidence interval for the ratio of proportions itself is the antilog of this: $e^{0.35089}$ to $e^{1.28324} = 1.42$ to 3.61. Thus we estimate that the proportion of children reported to cough during the day or at night among those with a history of bronchitis is between 1.4 and 3.6 times the proportion among those without a history of bronchitis.

The proportion of individuals in a population who develop a disease or symptom is equal to the probability that any given individual will develop the disease, called the **risk** of an individual developing a disease. Thus in Table 8.3 the risk that a child with bronchitis before age 5 will cough at age 14 is 26/273 = 0.09524, and the risk for a child without bronchitis before age 5 is 44/1046 = 0.04207. To compare risks for people with and without a particular risk factor, we look at the ratio of the risk with the factor to the risk without the factor, the **relative risk**. The relative risk of cough at age 14 for bronchitis before 5 is thus 2.26. To estimate the relative risk directly, we need a cohort study (Section 3.7) as in Table 8.3. We estimate relative risk for a case–control study in a different way (Section 13.7).

In the unusual situation when the samples are paired, either matched or two observations on the same subject, we use a different method (Section 13.9).

8.7 Number needed to treat

When a clinical trial has a dichotomous outcome measure, such as survival or death, there are several ways in which we can express the difference between the two treatments. These include the difference between proportions of successes, ratio of proportions (risk ratio or relative risk), and the odds ratio. The **number needed to treat** (**NNT**) is the number of patients we would need to treat with the new treatment to achieve one more success than we would on the old treatment (Laupacis *et al.* 1988). It is the reciprocal of the difference between the proportion of success on the new treatment and the proportion on the old treatment. For example, in the MRC streptomycin trial (Table 2.10) the survival rates after 6 months were 93% in the streptomycin group and 73% in the control group. The difference in proportions surviving was thus 0.93 – 0.73 = 0.20 and the number needed to treat to prevent one death over 6 months was 1/0.20 = 5. The smaller the NNT, the more effective the treatment will be.

The smallest possible value for NNT is 1.0, when the proportions successful are 1.0 and 0.0. This would mean that the new treatment was always effective and the old treatment was never effective. The NNT cannot be zero. If the treatment has no effect at all, the NNT will be infinite, because the difference in the proportion of successes will be zero. If the treatment is harmful, so that success rate is less than on the control treatment, the NNT will be negative. The number is then called the **number needed to harm** (**NNH**). The NNT idea caught on very quickly and has been widely used and developed, for example as the number needed to screen (Rembold 1998).

The NNT is an estimate and should have a confidence interval. This is apparently quite straightforward. We find the confidence interval for the difference in the proportions, then the reciprocals of these limits are the confidence limits for the NNT. For the MRC streptomycin trial the 95% confidence interval for the difference is 0.0578 to 0.3352, reciprocals 17.3 and 3.0. Thus the 95% confidence interval for the NNT is 3 to 17.

This is deceptively simple. As Altman (1998) pointed out, there are problems when the difference is not significant. The confidence interval for the difference between proportions includes zero, so infinity is a possible value for NNT, and negative values are also possible, i.e. the treatment may harm. The confidence interval must allow for this.

For example, Henzi *et al.* (2000) calculated NNT for several studies, including that of Lopez-Olaondo *et al.* (1996). This study compared dexamethasone against placebo to prevent postoperative nausea and vomiting. They observed nausea in 5/25 patients on dexamethasone and 10/25 on placebo. Thus the difference in proportions without nausea (success) is 0.80 – 0.60 = 0.20, 95% confidence interval –0.0479 to 0.4479 (Section 8.6). The number needed to treat is the reciprocal of this difference, 1/0.20 = 5.0. The reciprocals of the confidence limits are 1/(–0.0479) = –20.9 and 1/0.4479 = 2.2. But the confidence interval for the NNT is not –20.9 to 2.2. Zero, which this includes, is not a possible value for the NNT. As there may be no treatment difference at all, zero difference between proportions, the NNT may be infinite. In fact, the confidence interval for NNT is not the values between –20.9 and 2.2, but the values *outside* this interval, i.e. 2.2 to infinity (number needed to achieve an extra success, NNT) and minus infinity to –20.9 (number needed to achieve an extra failure, NNH). Thus the NNT is estimated to be anything

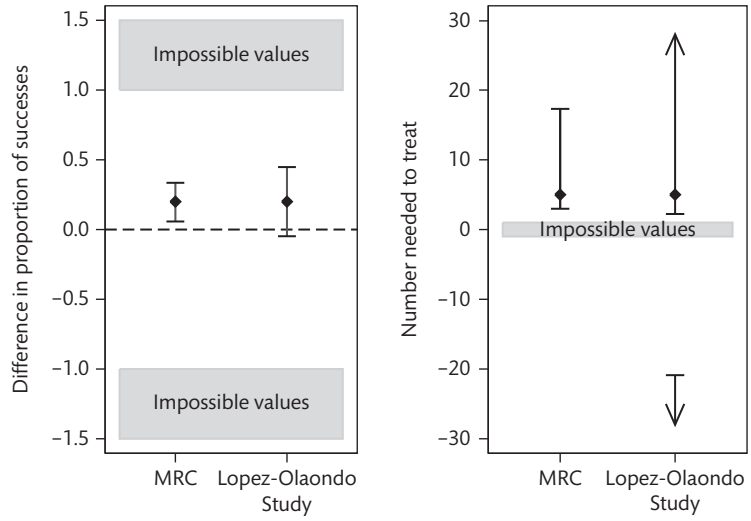

Figure 8.6 Confidence intervals for difference in proportion of successes and for number needed to treat for the data of MRC (1948) and Lopez-Olaondo *et al.* (1996).

greater than 2.2, and the NNH to be anything greater than 20.9. The confidence interval is in two parts, $-\infty$ to −20.9 and 2.2 to ∞. ('∞' is the symbol for infinity.) Henzi *et al.* (2000) quote this confidence interval as 2.2 to −21, which they say the reader should interpret as including infinity. Altman (1998) recommends 'NNTH = 21.9 to ∞ to NNTB 2.2', where NNTH means 'number needed to harm' and NNTB means 'number needed to benefit'. I prefer '$-\infty$ to −20.9, 2.2 to ∞'. Here $-\infty$ and ∞ each tell us that we do not have enough information to guide us as to which treatment should be used. The confidence intervals for the MRC and the Lopez-Olaondo trials are shown graphically in Figure 8.6.

Two-part confidence intervals are not exactly intuitive and I think that the problems of interpretation of NNT in trials which do not provide unequivocal evidence limit its value to being a supplementary description of trial results.

8.8 Standard error of a sample standard deviation

We can find a standard error and confidence interval for almost any estimate we make from a sample, but sometimes this depends on the distribution of the observations themselves. The sample standard deviation, s, is one such statistic. Provided the observations are independent and come from a Normal distribution, $(n-1)s^2/\sigma^2$ is from a Chi-squared distribution with $n-1$ degrees of freedom (Appendix 7A). The square root of this Chi-squared distribution is approximately Normal with variance 1/2 if n is large enough, so $\sqrt{(n-1)s^2/\sigma^2}$ has an approximately Normal distribution with variance 1/2. Hence s has an approximately Normal distribution with variance $\sigma^2/2(n-1)$. The standard error of s is thus $\sqrt{\sigma^2/2(n-1)}$, estimated by $s/\sqrt{2(n-1)}$. This is only true when the observations themselves are from a Normal distribution.

8.9 Confidence interval for a proportion when numbers are small

In Section 8.4 I mentioned that the standard error method for a proportion does not work when the sample is small. Instead, the confidence interval can be found using the exact probabilities of the Binomial distribution, the **Clopper Pearson method**. The method works like this. Given n, we find the value p_L for the parameter p of the Binomial distribution which gives a probability 0.025 of getting an observed number of successes, r, as big as or bigger than the value observed. We do this by calculating the probabilities from the formula in Section 6.4, iterating round different possible values of p until we get the right one. We also find the value p_U for the parameter p of the Binomial distribution which gives a probability 0.025 of getting an observed number

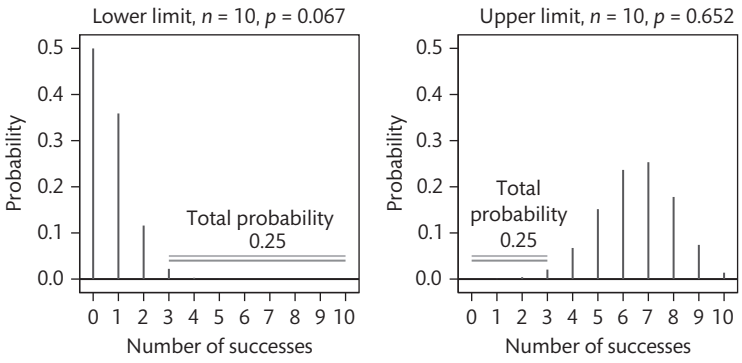

Figure 8.7 Distributions showing the calculation of the exact confidence interval for three successes out of ten trials.

of successes as small as or smaller than the value observed. The exact 95% confidence interval is p_L to p_U. For example, suppose we observe three successes out of 10 trials. The Binomial distribution with $n = 10$ which has the total probability for three or more successes equal to 0.025 has parameter $p = 0.067$. The distribution which has the total probability for three or fewer successes equal to 0.025 has $p = 0.652$. Hence the 95% confidence interval for the proportion in the population is 0.067 to 0.652. Figure 8.7 shows the two distributions. No large sample approximation is required and we can use this for any size of sample.

Unless the observed proportion is zero or one, these values are never included in the exact confidence interval. The population proportion of successes cannot be zero if we have observed a success in the sample. It cannot be one if we have observed a failure.

Although this interval is called 'exact', it can produce intervals which are too wide, in that more than 95% of possible samples give intervals which include the population proportion. Other methods have been developed, such as the Wilson interval, which give a proportion of intervals including the population proportion which is closer to 95% (see Brown *et al.* 2001, 2002).

8.10 Confidence interval for a median and other quantiles

In Section 4.5 we estimated medians and other quantiles directly from the frequency distribution. We can estimate confidence intervals for these using the Binomial distribution. This is a large sample method. The 95% confidence interval for the q quantile can be found by an application of the Binomial distribution (Section 6.4, Section 6.6) (see Conover 1980). The number of observations less than the q quantile will be an observation from a Binomial distribution with parameters n and q, and hence has mean nq and standard deviation $\sqrt{nq(1-q)}$. We calculate j and k such that:

$$j = nq - 1.96\sqrt{nq(1-q)}$$

$$k = nq + 1.96\sqrt{nq(1-q)}$$

We round j and k up to the next integer. Then the 95% confidence interval is between the jth and the kth observations in the ordered data. For the 57 FEV measurements of Table 4.4, the median was 4.1 litres (Section 4.5). For the 95% confidence interval for the median, $n = 57$ and $q = 0.5$, and

$$j = 57 \times 0.5 - 1.96\sqrt{57 \times 0.5 \times (1-0.5)} = 21.10$$

$$k = 57 \times 0.5 + 1.96\sqrt{57 \times 0.5 \times (1-0.5)} = 35.90$$

The 95% confidence interval is thus from the 22nd to the 36th observation, 3.75 to 4.30 litres from Table 4.4. Compare this to the 95% confidence interval for the mean, 3.9 to 4.2 litres, which is completely included in the interval for the median. This method of estimating percentiles is relatively imprecise. Another example is given in Section 20.7.

8.11 Bootstrap or resampling methods

Bootstrap or resampling methods (Efron and Tibshirani 1993) are an alternative way to find standard errors and confidence intervals. They takes their name, I think, from the expression 'raised up by your own bootstraps'. We use the sample itself, without any external structure of probability distributions. The idea is that the sample is a random sample of the population it represents, whatever that is, and that is the population about which we can draw conclusions. So if we draw an observation at random from our sample, it is also a random observation from the original population. All members of that population had an equal chance of being chosen. Now we record this observation, we put it back, and we draw another observation at random from our sample. It might be the same as the first one, but that doesn't matter. We record that, put it back, and draw another, and so on until we have a new sample of the same size as the first. It will contain some of the original observations, some of them repeated. That doesn't matter, the proportion of repetitions of each possible value is expected to be the same as its density in the original population. This procedure is called **resampling**.

For an example, consider Table 8.3. This comes from a sample of 1 319 children, for each of whom we have whether they had bronchitis before age 5 and cough during the day or at night reported at age 14. Using resampling, we can draw another sample of 1 319 observations. The corresponding table is shown in Table 8.4. For Table 8.3 we found the difference in the proportions with cough, the risk difference, to be 0.053 17 with standard error 0.018 82 and 95% confidence interval calculated from these 0.016 to 0.090. For Table 8.4, the risk difference is 0.073 38 with standard error = 0.019 49 and 95% confidence interval 0.035 to 0.112.

Now instead of resampling once, we can do it many times. I chose 1 000 resamplings. Each one produced a different estimate of the risk difference. These 1 000 estimates have a distribution, shown in Figure 8.8. The mean and standard deviation of this distribution are 0.053 17 and 0.019 38. This standard deviation provides an alternative estimate of the standard error of the difference between the risks, which does not make use of any theory about the Binomial distribution. These are the bootstrap estimates. We can use them to find a confidence interval, on the assumption that the resampling distribution is Normal, which appears to be a reasonable approximation from Figure 8.8. The 95% CI will be 0.053 17 – 1.96 × 0.019 38 to 0.053 17 + 1.96 × 0.019 38 = 0.015 to 0.091. This is very similar to the 0.016 to 0.090 found using the variance formula of the Binomial distribution.

We can also use the resampling distribution directly. The 95% CI will be from the 2.5th centile to the 97.5th centile. The 2.5th centile will be between the 25th and 26th of the 1 000 observations, which are 0.015 85 and 0.016 30, and the average of these is 0.016 08. The 97.5th centile will be between observations 975 and 976, which are 0.093 23 and 0.093 78, giving us 0.093 51. Hence the bootstrap confidence interval is 0.016 to 0.094. The corresponding point estimate is the median of this distribution, 0.053.

Table 8.4 Cough during the day or at night at age 14 by bronchitis before age 5, after resampling the data of Table 8.3 (data from Holland *et al.* 1978)

Cough at 14	Bronchitis at 5		
	Yes	**No**	**Total**
Yes	34	42	76
No	263	980	1 243
Total	297	1 022	1 319

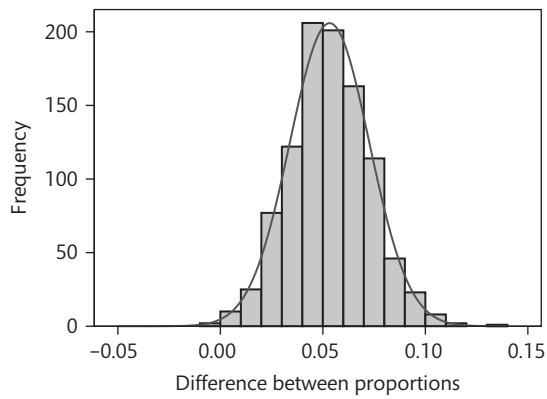

Figure 8.8 Histogram of 1 000 resampling estimates of the risk difference from Table 8.3 (data from Holland *et al.* 1978).

The two bootstrap estimates are very similar to one another and to the original and there seems little point in doing such a complicated procedure. Sometimes the bootstrap approach can be very useful, when the data do not meet the requirements of any straightforward conventional approach. We can get a bootstrap estimate for anything we can calculate from any sample. Bootstrap methods are particularly favoured by health economists, who often have cost data which have a few very extreme values, which conventional approaches might not accommodate well. This section gives the general idea of the bootstrap; there are many developments and variations. Methods which sample the possibilities in this way are (rather endearingly) called **Monte Carlo** methods.

8.12 What is the correct confidence interval?

A confidence interval only estimates errors caused by sampling. They do not allow for any bias in the sample and give us an estimate for the population of which our data can be considered a random sample. As discussed in Section 3.5, it is often not clear what this population is, and we rely far more on the estimation of differences than absolute values. This is particularly true in clinical trials. We start with patients in one locality, exclude some, allow refusals, and the patients cannot be regarded as a random sample of patients in general. However, we then randomize into two groups which are then two samples from the same population, and only the treatment differs between them. Thus the difference is the thing we want the confidence interval for, not for either group separately. Yet researchers often ignore the direct comparison in favour of estimation using each group separately.

For example, Salvesen *et al.* (1992) reported follow-up of two randomized controlled trials of routine ultrasonography screening during pregnancy. At ages 8 to 9 years, children of women who had taken part in these trials were followed up. A subgroup of children underwent specific tests for dyslexia. The test results classified 21 of the 309 screened children (7%, 95% confidence interval 3–10%) and 26 of the 294 controls (9%, 95% confidence interval 4–12%) as dyslexic. Much more useful would be a confidence interval for the difference between prevalences (–6.3 to 2.2 percentage points) or their ratio (0.44 to 1.34), because we could then compare the groups directly. See Bland and Altman (2011) for a fuller discussion.

8.13 Multiple choice questions: Confidence intervals

(Each branch is either true or false.)

8.1 The standard error of the mean of a sample:
- (a) measures the variability of the observations;
- (b) is the accuracy with which each observation is measured;
- (c) is a measure of how far a mean from a sample of this size is likely to be from the population mean;
- (d) is proportional to the number of observations;
- (e) is greater than the estimated standard deviation of the population.

8.2 95% confidence limits for the mean estimated from a set of observations:
- (a) are limits between which, in the long run, 95% of observations fall;
- (b) are a way of measuring the precision of the estimate of the mean;
- (c) are limits within which the sample mean falls with probability 0.95;
- (d) are limits calculated so as to include the population mean for 95% of possible samples;
- (e) are a way of measuring the variability of a set of observations.

8.3 If the size of a random sample were increased, we would expect:
- (a) the mean to decrease;
- (b) the standard error of the mean to decrease;
- (c) the standard deviation to decrease;
- (d) the sample variance to increase;
- (e) the degrees of freedom for the estimated variance to increase.

8.4 The prevalence of a condition in a population is 0.1. If the prevalence were estimated repeatedly from samples of size 100, these estimates would form a distribution which:
(a) is a sampling distribution;
(b) is approximately Normal;
(c) has mean = 0.1;
(d) has variance = 9;
(e) is Binomial.

8.5 It is necessary to estimate the mean FEV1 by drawing a sample from a large population. The accuracy of the estimate will depend on:
(a) the mean FEV1 in the population;
(b) the number in the population;
(c) the number in the sample;
(d) the way the sample is selected;
(e) the variance of FEV1 in the population.

8.6 In a study of 88 births to women with a history of thrombocytopenia (Samuels *et al.* 1990), the same condition was recorded in 20% of babies (95% confidence interval 13% to 30%, exact method):
(a) Another sample of the same size will show a rate of thrombocytopenia between 13% and 30%;
(b) 95% of such women have a probability of between 13% and 30% of having a baby with thrombocytopenia;
(c) It is estimated that between 13% and 30% of births to such women in the area would show thrombocytopenia;
(d) If the sample were increased to 880 births, the 95% confidence interval would be narrower;
(e) It would be impossible to get these data if the rate for all women was 10%.

8.14 Exercise: Confidence intervals in two acupuncture studies

Two short papers concerning adverse events associated with acupuncture appeared together in the *British Medical Journal*. They were very similar in the question they address and the methods used. Both papers referred to 'significant' events. The word is not used in its statistical sense.

White *et al.* (2001) recruited acupuncture practitioners through journals circulated to members of the British Medical Acupuncture Society and the Acupuncture Association of Chartered Physiotherapists. They asked acupuncturists to take part in a prospective survey, recording for each consultation adverse events, defined as 'any ill-effect, no matter how small, that is unintended and non-therapeutic, even if not unexpected'. Some events were considered to be 'significant', meaning 'unusual, novel, dangerous, significantly inconvenient, or requiring further information'.

White *et al.* reported that as the data were skewed, with extreme values present, confidence intervals were calculated using a bootstrapping procedure with 10 000 replications.

Data were collected from 78 acupuncturists, 31 822 (median 318, range 5 to 1911) consultations were included. Altogether, 43 'significant' events were reported, giving a rate of 14 per 10 000 (95% confidence interval 8 per 10 000 to 20 per 10 000). None of these events was a serious adverse event, a category which includes death, hospital admission or prolongation of existing hospital stay, persistent or significant disability or incapacity, or otherwise life-threatening. Hence the rate of serious events was estimated as 0 per 10 000 (95% confidence interval 0 per 10 000 to 1.2 per 10 000).

MacPherson *et al.* (2001) carried out a prospective audit of treatments undertaken during a 4-week period. They invited all 1 848 professional acupuncturists who were members of the British Acupuncture Council and were practising in the UK to record details of adverse events and mild transient reactions after treatment.

A total of 574 (31%) practitioners participated, reporting on 34 407 treatments. Practitioners were asked to give details of any adverse events they considered to be 'significant', using the same definition as White *et al.* (2001). There were no reports of serious adverse events, defined as described previously (95% confidence interval 0 to 1.1 per 10 000 treatments). Practitioners reported 43 minor adverse events, a rate of 1.3 (0.9 to 1.7) per 1 000 treatments.

MacPherson *et al.* concluded that 'In this prospective survey, no serious adverse events were reported after 34 407 acupuncture treatments. This is consistent, with 95% confidence, with an underlying serious adverse event rate of between 0 and 1.1 per 10 000 treatments.' They continue: 'Even given the potential bias of self reporting, this is important evidence on public health and safety as professional acupuncturists deliver approximately two million treatments per year in the United Kingdom. Comparison of this adverse event rate

for acupuncture with those of drugs routinely prescribed in primary care suggests that acupuncture is a relatively safe form of treatment'.

8.1 Are there any problems with the sampling methods used by White *et al.* and by MacPherson *et al.*? What alternative methods might have been used? Would they solve the problem?

8.2 Are there any problems with the data collection methods used in these studies? What alternatives could be used? Would they solve the problem?

8.3 White *et al.* reported the average age of their acupuncturists to be 47 (range 27 to 71) years. The median number of consultations for a practitioner was 318, range 5 to 1 911. What does this tell us about the shapes of the distributions of age and number of consultations?

8.4 Altogether, White *et al.* reported 43 'significant' events, giving a rate of 14 per 10 000 (95% confidence interval 8 per 10 000 to 20 per 10 000). What does this mean?

8.5 White *et al.* reported that none of the adverse events was serious (95% confidence interval 0 to 1.2 per 10 000 consultations). MacPherson *et al.* also reported that there were no records of serious adverse events (0 to 1.1 per 10 000 treatments). Can we conclude that there is no risk of serious events?

8.6 MacPherson *et al.* concluded that their data were consistent with an underlying serious adverse event rate of between 0 and 1.1 per 10 000 treatments. Is this a reasonable interpretation?

8.7 White *et al.* say '14 per 10 000 of these minor events were reported as significant. These event rates are per consultation, and they do not give the risk per individual patient'. Why do they not give the risk per individual patient?

8.8 MacPherson *et al.* said that further research measuring patients' experience of adverse events is merited. What would this tell us that these papers do not?

Appendix 8A: Standard error of a mean

When we calculate the mean of a sample of size n independent observations, we add n independent variables, each with variance σ^2. The variance of the sum is the sum of the variances (Section 6.6), $\sigma^2 + \sigma^2 + \cdots + \sigma^2 = n\sigma^2$. We divide this new variable by a constant, n, to get the mean. This has the effect of dividing its variance by the square of the constant, n^2. The variance on the mean is thus $n\sigma^2/n^2 = \sigma^2/n$. The standard error is the square root of this, $\sqrt{\sigma^2/n}$ or $\sigma/\sqrt{n}$.