Health Sciences M.Sc. Programme

Applied Biostatistics

Week 4: Standard Error and Confidence Intervals

Sampling

Most research data come from subjects we think of as samples drawn from a larger population. The sample tells us something about the population. The notion of sampling is a familiar one in health care. For example, if I want to measure a subject's blood glucose, I do not take all the blood. I draw a sample. One drop of blood is then used to represent all the blood in the body. I did this three times, from the same subject (myself) and got three measurements: 6.0, 5.9, and 5.8 mmol/L. Which of these was correct? The answer is that none of them were; they were all estimates of the same quantity. We do not know which of them was actually closest.

In research, we collect data on our research subjects so we can draw conclusions about some larger population. For example, in a randomised controlled trial comparing two obstetric regimes, the proportion of women in the active management of labour group who had a Caesarean section was 0.97 times the proportion of women in the routine management group who had sections (Sadler *et al*., 2000). (We call this ratio the relative risk.) This trial was carried out in one obstetric unit in New Zealand, but we are not specifically interested in this unit or in these patients. We are interested in what they can tell us about what would happen if we treated future patients with active management of labour rather than routine management. We want know, not the relative risk for these particular women, but the relative risk for all women.

The trial subjects form a sample, which we use to draw some conclusions about the population of such patients in other clinical centres, in New Zealand and other countries, now and in the future. The observed relative risk of Caesarean section, 0.97, provides an estimate of the relative risk we would expect to see in this wider population. If we were to repeat the trial, we would not get exactly the same point estimate. Other similar trials cited by Sadler *et al.* (2000) have reported different relative risks: 0.75, 1.01, and 0.64. Each of these trials represents a different sample of patients and clinicians and there is bound to be some variation between samples. Hence we cannot conclude that the relative risk in the population will be the same as that found in our particular trial sample. The relative risk which we get in any particular sample would be compatible with a range of possible differences in the population.

When we draw a sample from a population, it is just one of the many samples we could take. If we calculate a statistic from the sample, such as a mean or proportion, this will vary from sample to sample. The means or proportions from all the possible samples form the sampling distribution. To illustrate this with a simple example, we could put lots numbered 1 to 9 into a hat and sample by drawing one out, replacing it, drawing another out, and so on. Each number would have the same chance of being chosen each time and the sampling distribution would be as in Figure 1(a). Now we change the procedure, draw out two lots at a time and calculate the average. There are 36 possible pairs, and some pairs will have the same average (e.g. 1 and 9, 4 and 6 both have average 5.0). The sampling distribution of this average is shown in Figure 1(b).

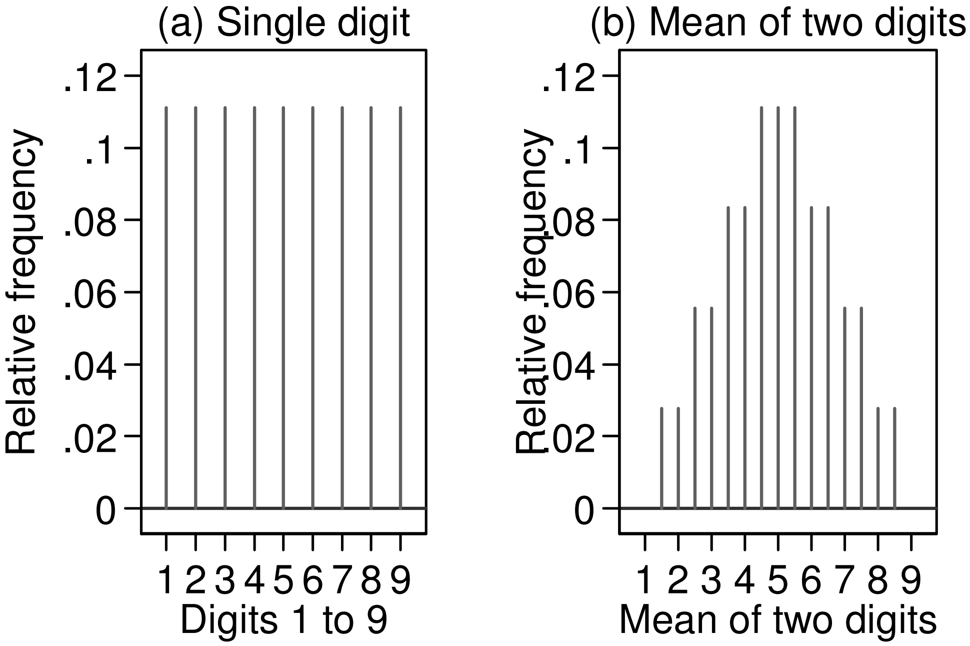


Figure 1. Sampling distribution for a single digit drawn at random and for the mean of two digits drawn together

There are three things which we should notice about Figure 1(b).

1. The mean of the distribution remains the same, 5.
2. The sampling distribution of the mean is not so widely spread as the parent distribution. It has a smaller variance and standard deviation.
3. It has a different shape to Figure 1(a). The sampling distribution of a statistic does not necessarily have the same shape as the distribution of the observations themselves, which we call the parent distribution. In this case, as so often, it looks closer to a Normal distribution than does the distribution of the observations themselves.

If we know the sampling distribution it can help us draw conclusions about the population from the sample, using confidence intervals and significance tests. We often use our sample statistic as an estimate of the corresponding value in population, for example using the sample mean to estimate the population mean. The sampling distribution tells us how far from the population value the sample statistic is likely to be. Any statistic which is calculated from a sample, such as a mean, proportion, median, or standard deviation, will have a sampling distribution.

Standard error

If the sample statistic is used as an estimate, we call the standard deviation of the sampling distribution the **standard error**. Rather confusingly, we use this term both for the unknown standard deviation of the sampling distribution and for the estimate of this standard deviation found from the data.

Table 1. FEV1 (litres) of 57 male medical students

2.85	3.19	3.50	3.69	3.90	4.14	4.32	4.50	4.80	5.20
2.85	3.20	3.54	3.70	3.96	4.16	4.44	4.56	4.80	5.30
2.98	3.30	3.54	3.70	4.05	4.20	4.47	4.68	4.90	5.43
3.04	3.39	3.57	3.75	4.08	4.20	4.47	4.70	5.00	
3.10	3.42	3.60	3.78	4.10	4.30	4.47	4.71	5.10	
3.10	3.48	3.60	3.83	4.14	4.30	4.50	4.78	5.10	

For an example, consider the 57 FEV1 measurements of Table 1. We have mean = 4.062 litres, standard deviation $s = 0.67$ litres. The standard error of the sample mean is found from the standard deviation divided by the square root of the sample size. I shall not go into why this is, but many statistics books give explanations, e.g. Bland (2000). For the FEV1 data, the standard error of the mean is $0.67/\sqrt{57} = 0.089$. The best estimate of the mean FEV1 in the population is then 4.062 litres with standard error 0.089 litres.

In general, standard errors are proportional to one over the square root of the sample size, approximately. To half the standard error we must quadruple the sample size.

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.

People find the terms ‘standard error’ and ‘standard deviation’ confusing. This is not surprising, as a standard error is a type of standard deviation. We use the term ‘standard deviation’ when we are talking about distributions, either of a sample or a population. We use the term ‘standard error’ when we are talking about an estimate found from a sample. 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*.

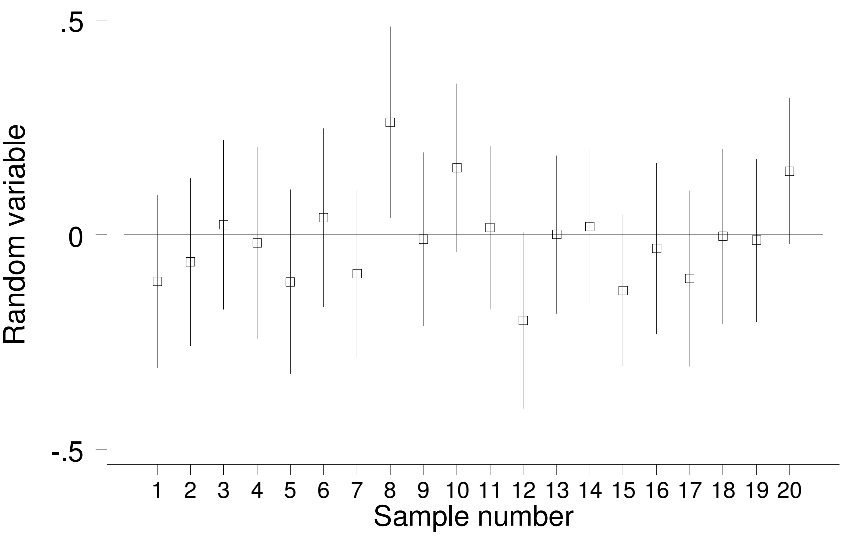
The standard error of an estimate tells us how variable estimates would be if obtained from other samples drawn in the same way as one being described. Even more often, research papers include confidence intervals and P values derived using them. Estimated standard errors can be found for many of the statistics we want to calculate from data and use to estimate things about the population from which the sample is drawn.

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**.

It is not possible to calculate useful interval estimates which always contain the unknown population value. There is always a very small probability that a sample will be very extreme and contain a lot of either very small or very large observations. We calculate our interval so that most of the intervals we calculate will contain the population value we want to estimate.

Figure 2. Simulation of 20 samples of 100 observations from a Standard Normal distribution, mean = 0 and SD =1.0, SE = 0.10.



In the FEV1 example, we have a fairly large sample, and so we can assume that the observed mean is from a Normal Distribution. For this illustration we shall also assume that the standard error is a good estimate of the standard deviation of this Normal distribution. (We shall return to this in Week 5.) We therefore expect about 95% of such means to be within 1.96 standard errors of the population mean. 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 mean minus 1.96 standard errors and mean plus 1.96 standard errors for all possible samples, 95% of such intervals would contain the population mean. In this case these limits are 4.062 – 1.96 × 0.089 to 4.062 + 1.96 × 0.089 which gives 3.89 to 4.24, or 3.9 to 4.2 litres, rounding to 2 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 ends of the confidence interval.

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 sloppy thinkers often put it that way. The population mean is a number, not a random variable, and has no probability. We sometimes say that we are 95% confident that the mean lies between these limits, but this doesn't help us understand what a confidence interval is. The important thing is: we use a sample to estimate something about a population. The 95% confidence interval is chosen so that 95% of such intervals will include the population value.

Confidence intervals do not always include the population value. If 95% of 95% confidence intervals include it, it follows that 5% must exclude it. In practice, we cannot tell whether our confidence interval is one of the 95% or the 5%.

Figure 2 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. Thus, 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%.

We expect to see 5% of the intervals having the population value outside the interval and 95% having the population value inside the interval. This is not the same as saying that 95% of further samples will have estimates within the interval. For example, if we look at the first interval in Figure 2, we can see that samples 8, 10, and 20 all have point estimates outside this interval. In fact, we expect about 83% of further samples to have their point estimates within a 95% confidence interval chosen at random.

The confidence interval need not 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, 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 since we are more confident that the mean will be included. We could also calculate a 90% confidence interval, which is 3.916 to 4.208, narrower than the 95% confidence interval. However, only 90% of such intervals will include the population value, 10% will not.

The probability we choose for a confidence interval is thus 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 and this is what is usually quoted in health research.

For the trial comparing active management of labour with routine management (Sadler *et al*., 2000), the relative risk for Caesarean section was 0.97. Sadler *et al.* quoted the 95% confidence interval for the relative risk as 0.60 to 1.56. Hence we estimate that in the population which these subjects represent, the proportion of women undergoing Caesarean section when undergoing active management of labour is between 0.60 and 1.56 times the proportion who would have Caesarean section with routine management.

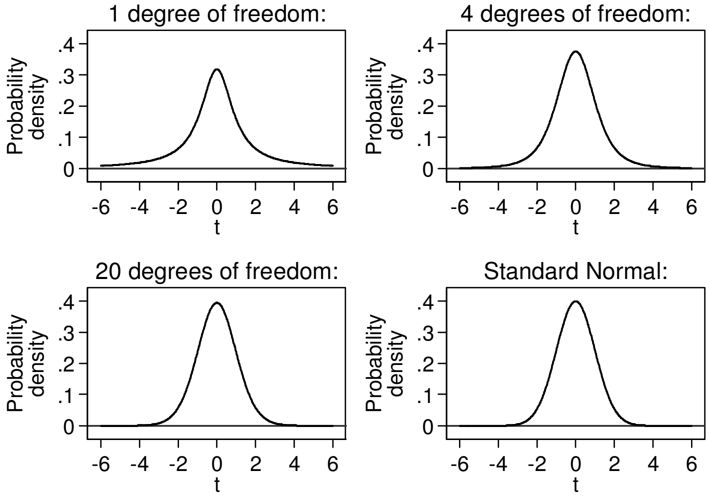
Confidence interval for the mean of a small sample

In the FEV1 example, we assumed that the observed mean is from a Normal Distribution and that the standard error is a good estimate of the standard deviation of this Normal distribution. This assumption may not be valid. We usually need 100 or more observations to make it. Although it was OK to illustrate the idea, we should not use the large sample Normal method for these data because the sample is too small. The standard error may not be sufficiently well estimated.

This problem was solved by W. G. Gossett, who published under the name “Student”. He found that the distribution of the standard error estimate depends on the distribution of the observations themselves.

There are several ways to estimate a confidence interval for the mean, but the most frequently used is Student’s t Distribution method. We must make the following assumption about the data: that the observations themselves come from a Normal distribution.

Figure 3. Some t Distribution curves.



In the large sample Normal method we calculate mean minus 1.96 standard errors and mean plus 1.96 standard errors. For the FEV1 sample, these limits are 4.062 – 1.96 × 0.089 to 4.062 + 1.96 × 0.089, which gives 3.89 to 4.24, or 3.9 to 4.2 litres. For a small sample from a Normal Distribution, we replace 1.96 by a number from the t distribution, also known as Student's t distribution.

The t Distribution is a family of distributions, like the Normal Distribution from which t is derived. The t family has one parameter, called the degrees of freedom. Figure 3 shows some members of the t family, with the Standard Normal Distribution for comparison. As the degrees of freedom increases the t distribution gets more and more like the Normal. We might expect this to happen, because the small sample method must turn into the large sample method as the sample gets bigger.

For a small sample from a Normal Distribution, we replace 1.96 in the confidence interval calculation by a number from the t distribution. Before the ubiquity of computers, we used tables of this distribution to look up the number we required. The value needed for the 95% confidence interval is usually tabulated as the value which would have 5% of the distribution further from the mean, which is zero, the itself. This is shown in Figure 4, for 4 degrees of freedom. Table 2 shows a simple table, shown these points of the t Distribution for several different values of the degrees of freedom parameter and several different outside probabilities. For a 90% confidence interval we would use the 10% point.

The degrees of freedom we use are the degrees of freedom for the standard deviation used to calculate the standard error: $n - 1 = 57 - 1 = 56$. From the table, 2.00 looks like the right value. We can calculate this more accurately using a computer, the way we now do this in practice: 2.0032407.

Figure 4. Two tailed probability point of the t Distribution

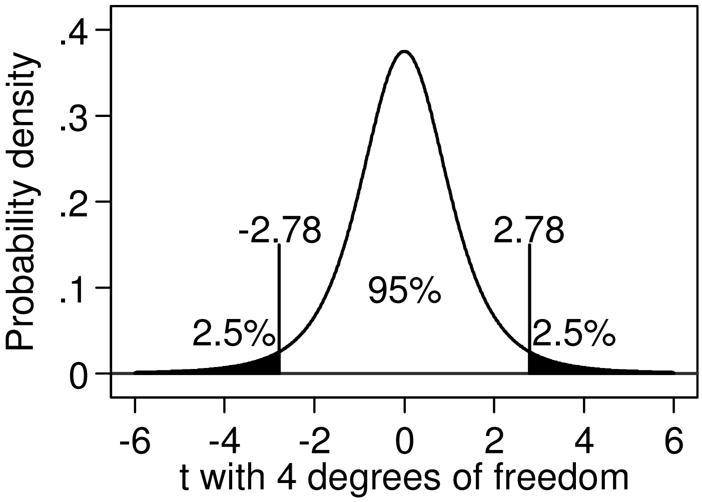


Table 2. Two tailed probability points of the t Distribution

D.f.	Probability				D.f.	Probability			
	0.10	0.05	0.01	0.001		0.10	0.05	0.01	0.001
	(10%)	(5%)	(1%)	(0.1%)		(10%)	(5%)	(1%)	(0.1%)
1	6.31	12.70	63.66	636.62	16	1.75	2.12	2.92	4.02
2	2.92	4.30	9.93	31.60	17	1.74	2.11	2.90	3.97
3	2.35	3.18	5.84	12.92	18	1.73	2.10	2.88	3.92
4	2.13	2.78	4.60	8.61	19	1.73	2.09	2.86	3.88
5	2.02	2.57	4.03	6.87	20	1.73	2.09	2.85	3.85
6	1.94	2.45	3.71	5.96	21	1.72	2.08	2.83	3.82
7	1.90	2.36	3.50	5.41	22	1.72	2.07	2.82	3.79
8	1.86	2.31	3.36	5.04	23	1.71	2.07	2.81	3.77
9	1.83	2.26	3.25	4.78	24	1.71	2.06	2.80	3.75
10	1.81	2.23	3.17	4.59	25	1.71	2.06	2.79	3.73
11	1.80	2.20	3.11	4.44	30	1.70	2.04	2.75	3.65
12	1.78	2.18	3.06	4.32	40	1.68	2.02	2.70	3.55
13	1.77	2.16	3.01	4.22	60	1.67	2.00	2.66	3.46
14	1.76	2.15	2.98	4.14	120	1.66	1.98	2.62	3.37
15	1.75	2.13	2.95	4.07	∞	1.65	1.96	2.58	3.29

D.f. = Degrees of freedom

∞ = infinity, same as the Standard Normal Distribution

The 95% confidence interval for the mean FEV1 is

$$4.062 - 2.003 \times 0.089 \text{ to } 4.062 + 2.003 \times 0.089$$

which gives 3.88 to 4.24, or 3.9 to 4.2 litres.

Using the large sample Normal method, we got:

$$4.062 - 1.96 \times 0.089 \text{ to } 4.062 + 1.96 \times 0.089$$

which gives 3.89 to 4.24, or 3.9 to 4.2 litres.

There is very little difference in this case, because the sample is not very small and because the data follow a Normal Distribution closely.

When the Normal assumption is not valid, there are several approaches which can be used, but they are beyond the scope of this module.

J. M. Bland
18 January 2012

References

Bland M. (2000) *An Introduction to Medical Statistics, 3rd. Edition* Oxford University Press.

Sadler LC, Davison T, McCowan LM. (2000) A randomised controlled trial and meta-analysis of active management of labour. *British Journal of Obstetrics and Gynaecology* **107**, 909-15.