Unit 8

Interval estimation

Introduction

In Unit 7, you have seen how a sample of data can be used to provide point estimates of a population parameter. For example, the sample mean provides an estimate of the population mean. However, different samples drawn from the same population will usually produce different estimates of the population parameter. To illustrate this point again, let us reconsider the example first considered in Example 1 of Unit 7.

Example 1 *Counts of the leech* Helobdella

An experiment was conducted to measure water contamination: 103 samples of water were collected, and the numbers of specimens of the leech *Helobdella* were counted in each sample. As in Unit 7, to avoid confusion, we will call these water samples 'volumes'. More than half of the volumes collected (58 of the 103) were free of this contamination, but all of the other volumes contained at least one leech: 25 contained exactly one leech, while the most contaminated volume contained eight leeches. The average number of leeches per volume was $\overline{x} = 0.816$. This provides an estimate of the Poisson population parameter λ, that is, of the underlying average number of leeches per volume.

Indeed, you saw in Activity 20(b) of Unit 7 that under a Poisson model, this is the maximum likelihood estimate of λ.

If, as in Activity 1 of Unit 7, a second experiment had been performed using the same water source, then the numbers of leeches collected would probably have been different, and the sample mean used to estimate the underlying population mean μ would also have been different. Quite conceivably, in this experiment, $\overline{x}$ might have been as low as 0.7 or even lower; it might have been as high as 1, or 2, or higher. In the initial experiment (the one actually performed) there were many 0s observed (volumes containing no leeches at all), quite a few 1s and 2s, and just a few 3s and 4s. As already noted above, none of the 103 volumes contained more than eight leeches. It would therefore be very surprising to find that the underlying mean number of leeches per volume was as high as 6, say, and it seems implausible (though, of course, it is not impossible) that the underlying mean could be as high as 8 or more.

This unit is about using the result of a statistical experiment to obtain some idea of a range of plausible values for some unknown population parameter (such as an average, or a rate, or a proportion). The minimum and maximum values of this range are called *confidence limits*, and the range of plausible values is called a *confidence interval*. In the same way that a single value obtained from the data is called a point estimate of a parameter, so a range of values obtained from the data is called an *interval estimate* of a parameter.

In Section 1, confidence intervals for a population mean are defined. To keep the details reasonably simple, attention is restricted to large samples of data so that the results of the Central Limit Theorem can be applied. In Section 2, the focus is on interpreting confidence intervals. It might be argued that knowing how to interpret confidence intervals is more important than knowing how to construct them, so this section is particularly important. In Section 3, the method of Section 1 is extended to a range of applications. You will learn how to construct confidence intervals for means, rates, proportions and differences of proportions. The methods of this section are also valid only for large samples and, because of their reliance on the Central Limit Theorem, produce only approximate confidence intervals.

Methods for calculating 'exact' confidence intervals for normal means and for differences between normal means are described in Section 4. An important feature of these methods is that they are applicable to small as well as large samples, and do not rely on any approximations. These confidence intervals require the use of quantiles of t-distributions, a family of probability distributions described in Subsection 4.2.

If you have studied confidence intervals before, you might have done so *after* studying 'hypothesis tests' – in which case confidence intervals might have been constructed *from* hypothesis tests. In this module, hypothesis tests are considered in Unit 9, and confidence intervals are considered first, in a direct fashion, in this unit. There is no cause for alarm: both the formulas for and interpretation of confidence intervals obtained from either route are precisely the same. The link between hypothesis tests and confidence intervals will be considered briefly in Subsection 3.3 of Unit 9.

Arguably, considering confidence intervals first is the more usual approach taken in textbooks.

1 Introducing confidence intervals

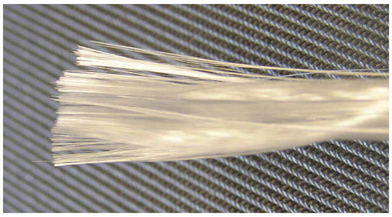
Confidence intervals for population means calculated from large samples of data are introduced in this section. When dealing with large samples, the Central Limit Theorem can be used. So, few assumptions are required about the population from which the sample is taken, and the calculations are relatively straightforward. This will allow you to concentrate on the definition of a confidence interval and then, especially in Section 2, on its interpretation. In Subsection 1.1, some settings in which confidence intervals might be needed are described. In Subsection 1.2, a confidence interval is defined and you will see how it is calculated.

1.1 Some examples

We begin with four examples, each of which is revisited later in the unit.

Example 2 *Strength of glass fibres*

Extremely thin fibres of glass are used in many materials, for insulation and for reinforcement. An experiment was conducted at the UK's National Physical Laboratory on the breaking strength of glass fibres. This was recorded (in unspecified units) for 63 glass fibres, each of length 1.5 cm. (Source: Smith, R.L. and Naylor, J.C. (1987) 'A comparison of maximum


A bundle of glass fibres

likelihood and Bayesian estimators for the three-parameter Weibull distribution', *Applied Statistics*, vol. 36, no. 3, pp. 358–69.)

The average strength in the 63 glass fibres in this sample is 1.51. This is useful information, which does indeed tell us something about the mean breaking strength of glass fibres of length 1.5 cm. But how confident can we be that this sample value is close to the population mean? If we were certain that breaking strength was the same in all glass fibres of this length, then we would need to measure breaking strength concentration in only one glass fibre to be sure to get the right answer. In fact, breaking strengths in the sample ranged from 0.55 to 2.24, and the sample standard deviation was 0.324. Thus if the experiment was repeated with a different sample of glass fibres, we would almost certainly obtain a different estimate of the mean. Intuitively, if the two estimates were close, then we might be reasonably confident that both values were close to the population mean. On the other hand, if the two estimates were very different, then we might have little confidence in either value.

A confidence interval provides a way of formalising this intuitive notion; it gives a range of plausible values for the population mean. Moreover, the confidence interval should be relatively narrow if there is little variability in sample means as experiments are repeated, and relatively wide if sample means differ considerably. An important point, however, is that a confidence interval can be calculated from a single sample, which is a great boon: the experiment does not actually have to be repeated in real life!

Example 3 *Accident counts*

In a 1960s investigation into accident-proneness in children, numbers of accidents were counted for 621 Californian boys over the eight-year period between the ages of 4 and 11. (Source: Mellinger, C.D. et al. (1965) 'A mathematical model with applications to a study of accident repeatedness among children', *Journal of the American Statistical Association*, vol. 60, no. 312, pp. 1046–59.) In this investigation an 'accident' was defined as one requiring professional medical attention.

BBC News Magazine, 22 April 2008, reported that in the UK children are nowadays more likely to injure themselves falling out of bed than climbing trees

The average number of accidents suffered by the children in the sample was 2.45, and the sample standard deviation was 2.03. This tells us something about the average frequency of accidents in 4- to 11-year-olds: accidents happen at a rate of about one every three years. There may, of course, be variation from child to child (some being more accident-prone than others) and at different ages. Indeed, it was an aim of the research exercise to study these complexities. (More recent research investigates, among others, effects of safety measures on accident rates.) However, we will look at the simple question: 'What is the average number of accidents we might expect a child to suffer between the ages of 4 and 11?'

In this case the data are counts, rather than measurements as in Example 2. However, as in that example, a different sample would almost certainly have produced a different estimate. This raises the question: 'Can we be confident that our sample estimate is close to the underlying population average number of accidents per child?'

Example 4 *Numbers of smokers*

Among many other things, the UK Office for National Statistics Opinions and Lifestyle Survey provides an estimate of the proportion, p, of 16- to 19-year-olds who smoke. In 2012, the survey took a random sample of 420 people in this age group. Let us suppose for a moment that the survey found that 63 of them were smokers. How can we use this information to calculate a confidence interval for p?

These data can be modelled by assuming that each person in the 16- to 19-year-old age group has the same probability p of being a smoker and treating each observation as an independent Bernoulli trial, with being a smoker being defined as a 'success'. Under this model, the total number of smokers in the 16- to 19-year-old age group in the sample follows a binomial $B(420, p)$ distribution. Our point estimate of p is the number of smokers divided by the number of people surveyed: this is $63/420 = 0.15$ (or 15%). As it happens, the Opinions and Lifestyle Survey just gives this percentage as 15 in a table of numbers rounded to the nearest whole percentage point. This means that the total number of smokers in the sample could actually have been any number from 61 to 65 inclusive to give the same result (and 63 was assumed just for purposes of illustration).

The remainder of the journey to calculating a confidence interval for p in this situation is undertaken later in this unit.

Example 5 *Sandflies*

Leishmaniasis is a disease giving rise to skin ulcers and sometimes further complications which occurs in many tropical and subtropical parts of the world. It is caused by parasites spread by the bite of infected female sandflies.

An experiment was performed in which sandflies were caught in two different light traps. (Source: Christensen, H.A., Herrer, A. and Telford, S.R. (1972) 'Enzootic cutaneous leishmaniasis in Eastern Panama. II: Entomological investigations', *Annals of Tropical Medicine and Parasitology*, vol. 66, no. 1, pp. 55–66.) The numbers of male and female flies were counted in each trap. The first trap was set 3 feet above the ground; the second trap was 35 feet above the ground. When the traps were inspected, the lower trap contained 173 male and 150 female sandflies; so the observed proportion of females was 150/323, or about 46%. The higher trap was found to contain 125 males and 73 females; in this case, the proportion of females was 73/198, or about 37%.

Thus there are two rather different estimates of the proportion of female sandflies in the population. The difference between the proportions at 3 feet and at 35 feet is about 9%. Whether or not the difference is a 'real' difference (fewer females venturing far above the ground) or due simply to random variation is the sort of question that is discussed in Unit 9. But can a confidence interval for this difference be obtained, and what might it tell us?

These four examples have several features in common. One is that they all involve data. Another is that in each case a descriptive population parameter has been identified and related to the data. In each case, the data can be used to obtain an estimate of the population parameter. And also, in each case, it would be useful to make that estimate an interval estimate: a range of plausible values for the parameter, not just a point estimate.

Activity 1 *Population parameters*

Identify the population parameters of interest in each of Examples 2–5.

As in Unit 7, in order to distinguish between an estimate and the population parameter being estimated, the 'hat' notation will be used. For example, if μ is a population mean, then $\widehat{\mu}$ is used to represent an estimate of μ; if p is a proportion or probability, then $\widehat{p}$ is used to denote an estimate of p.

1.2 A large-sample confidence interval

Having calculated a point estimate of the population parameter of interest, it would also be useful to know how close to the population value we might expect this estimate to lie. This has already been considered to some extent in Unit 7; there, properties of the *sampling distribution* of a point estimator, particularly the mean and variance of this distribution, were investigated. Making use of the sampling distribution of a point estimator is the idea behind the construction of a confidence interval; it is initially developed in Example 6 in the context of estimating the mean strength of glass fibres of length 1.5 cm.

For example, if the mean of a point estimator equals the parameter it is estimating, the point estimator is said to be unbiased.

Example 6 *Constructing a confidence interval*

The procedure described in Example 2 was to obtain $n = 63$ glass fibres of length 1.5 cm, measure the strength of each of these glass fibres, and calculate the sample mean. The sample mean $\overline{x}$ was 1.51 in the sample actually collected.

We want some way of expressing how confident we are that this estimate is close to the population mean μ. If we were to collect other samples of size $n = 63$ and find that the sample means calculated from these samples were clustered close to 1.51, then we would feel reassured that 1.51 was close to the population value μ. On the other hand, if the new sample means were very spread out, then we would have no reason to suppose that the value 1.51 was especially close to μ.

In practice, it is not usually feasible to repeat an experiment. However, thinking about the problem in this way suggests that the properties of the sampling distribution of the sample mean might be relevant because the sampling distribution tells us about the possible values we can expect to observe for the sample mean of different samples.

We know from Unit 7 that the sample mean is an unbiased estimator of the population mean, whatever the underlying distribution of the data, so all sampling distributions of the sample mean will have population mean μ. We can therefore concentrate on the variance of the sample mean.

Two sampling distributions of the sample mean are illustrated in Figure 1. For a sampling distribution with large variance, as in Figure 1(a), the sample means are very spread out; whereas for a sampling distribution with low variance, as in Figure 1(b), the sample means are closely clustered around the population mean. The case of Figure 1(b) should therefore result in a narrower range of plausible values for μ than that of Figure 1(a).

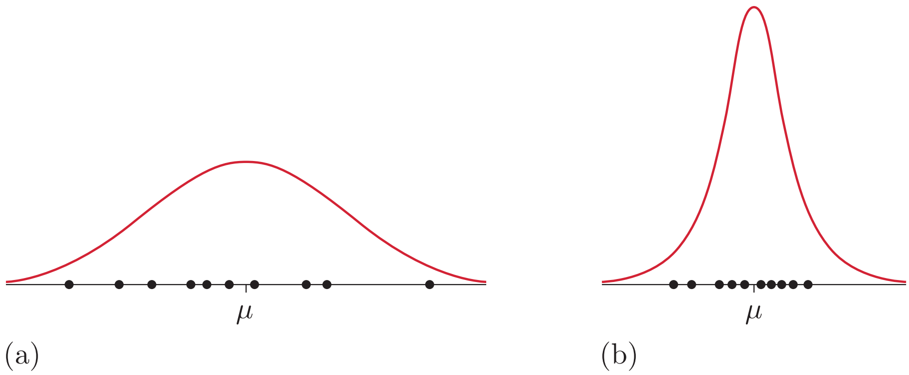


Figure 1 Sample means (the dots) and their sampling distributions

For a particular sample of size n from a population with mean μ and variance σ^2, the observed sample mean $\overline{x}$ is a single observation on the random variable $\overline{X}$ which represents the mean of a random sample of size n from the population. In Unit 6, you saw that, according to the Central Limit Theorem, provided that n is large enough, the distribution of $\overline{X}$ is approximately normal with mean μ and variance σ^2/n; that is,

See Subsection 6.2 of Unit 6.

$$\overline{X} \approx N\left(\mu, \frac{\sigma^2}{n}\right).$$

Assuming that a sample size of 63 is large enough for the Central Limit Theorem to be used, for samples of fibre strengths of size 63, we have

$$\overline{X} \approx N\left(\mu, \frac{\sigma^2}{63}\right),$$

the approximate sampling distribution of $\overline{X}$. Properties of the normal distribution can now be used to make probability statements about $\overline{X}$, the random variable representing means of samples of size 63 which, in turn, will shortly be used to construct a confidence interval for μ.

Activity 2 *Obtaining a normal probability*

Show that if $\overline{X} \approx N(\mu, \sigma^2/63)$, then

$$P\left(\mu - 1.96\,\frac{\sigma}{\sqrt{63}} \leq \overline{X} \leq \mu + 1.96\,\frac{\sigma}{\sqrt{63}}\right) = 0.95.$$

Hint: as in Subsection 4.4 of Unit 6, start by transforming from $\overline{X} \sim N(\mu, \sigma^2/63)$ to

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{63}}$$

so that $Z \sim N(0, 1)$.

Example 7 *Constructing a confidence interval, continued*

Let us return to the problem of constructing a confidence interval for the population mean μ based on the (approximate) sampling distribution of the sample mean for a sample of size $n = 63$ glass-fibre strengths, $\overline{X} \approx N(\mu, \sigma^2/63)$. In Activity 2, you showed that, with probability 0.95, $\overline{X}$ lies within 1.96 standard deviations of the mean μ; that is, within $1.96\,\sigma/\sqrt{63}$ units of μ. This is illustrated in Figure 2.

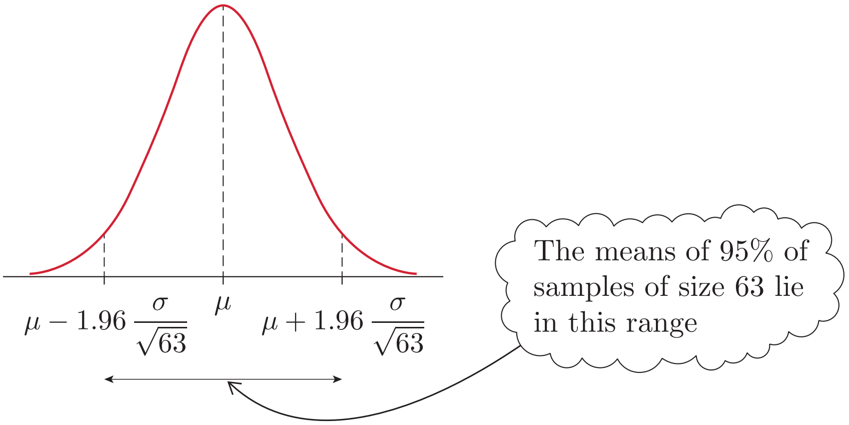


Figure 2 The sampling distribution of the mean for samples of size 63

In other words, the procedure 'draw a random sample of 63 glass fibres of length 1.5 cm and calculate the sample mean' will, with probability 0.95, produce a value within $1.96\,\sigma/\sqrt{63}$ of the population mean. Thus, if the population standard deviation σ is known – as we will assume for the moment – then the Central Limit Theorem can be used to determine how accurate our estimate of μ is likely to be, even though the true value of μ is not known.

Now, if $\overline{X}$ lies within $1.96\,\sigma/\sqrt{63}$ units of μ, then μ lies within $1.96\,\sigma/\sqrt{63}$ units of $\overline{X}$. To see this, think about $\overline{X}$ lying within c units of μ, as illustrated in Figure 3. Then the furthest μ can be from any value of $\overline{X}$ in the interval is either when $\overline{X}$ takes the value $\mu + c$, in which case $\mu = \overline{X} - c$, or when $\overline{X}$ takes the value $\mu - c$, in which case $\mu = \overline{X} + c$. Clearly, all values in between $\overline{X} - c$ and $\overline{X} + c$ are possible values of μ too, so we can say that if $\overline{X}$ lies within c units of μ, then μ lies within c units of $\overline{X}$.

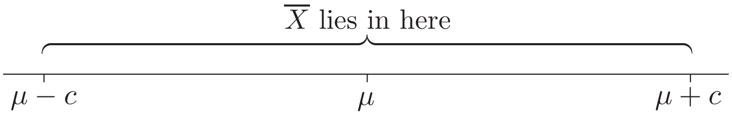


Figure 3 $\overline{X}$ lying within c units of μ

In particular, setting $c = 1.96\sigma/\sqrt{63}$, we can state that, with probability 0.95, the following interval, which has random variables as its limits,

$$\left(\overline{X} - 1.96\,\frac{\sigma}{\sqrt{63}}, \overline{X} + 1.96\,\frac{\sigma}{\sqrt{63}}\right), \tag{1}$$

contains the fixed, if unknown, value μ. In order to calculate a confidence interval for μ, this information must be related to the sample actually drawn in the experiment – that is, to the observation $\overline{x} = 1.51$ that was made on the random variable $\overline{X}$. This is done by replacing $\overline{X}$ in Interval (1) by the value $\overline{x}$ actually obtained, giving the interval

$$\left(\overline{x} - 1.96\,\frac{\sigma}{\sqrt{63}}, \overline{x} + 1.96\,\frac{\sigma}{\sqrt{63}}\right).$$

This is how we would like to define a 95% confidence interval for μ. There is just one difficulty: σ, the population standard deviation is not, in general, known. As a sample of 63 is reasonably large, then replacing σ in this expression by the sample standard deviation $s = 0.324$ should not introduce too much error. Since $\overline{x} = 1.51$, the lower bound of this interval, which is called the *lower 95% confidence limit* and denoted μ^-, is

$$\mu^- = \overline{x} - 1.96\,\frac{s}{\sqrt{63}} = 1.51 - 1.96\,\frac{0.324}{\sqrt{63}} \simeq 1.43.$$

The upper bound, or *upper 95% confidence limit*, denoted μ^+, is

$$\mu^+ = \overline{x} + 1.96\,\frac{s}{\sqrt{63}} = 1.51 + 1.96\,\frac{0.324}{\sqrt{63}} \simeq 1.59.$$

These confidence limits together define an approximate *95% confidence interval for* μ, which is written

$$(\mu^-, \mu^+) = (1.43, 1.59).$$

The result of the investigation may be summarised as follows. In a sample of 63 glass fibres of length 1.5 cm, the mean strength was 1.51, with approximate 95% confidence interval $(1.43, 1.59)$. We conclude that it is plausible that the population mean lies between about 1.43 and 1.59, while values outside the interval – such as 1.4 or 1.7, for instance – are not very plausible.

Confidence limits are generally reported to the same accuracy as the point estimate.

So far, the construction of a 95% confidence interval has been described for one particular dataset. The method used in Examples 6 and 7 can be applied to any random sample to obtain a 95% confidence interval for the population mean μ, as follows.

If we are given a random sample of size n from a population with mean μ and standard deviation σ, then provided that n is large enough, the Central Limit Theorem can be applied to approximate the distribution of the sample mean. This approximating distribution is normal with mean μ and variance σ^2/n:

$$\overline{X} \approx N\left(\mu, \frac{\sigma^2}{n}\right).$$

So, with probability approximately 0.95, the interval

$$\left(\overline{X} - 1.96\,\frac{\sigma}{\sqrt{n}}, \overline{X} + 1.96\,\frac{\sigma}{\sqrt{n}}\right)$$

This follows by replacing 63 by n in Activity 2 and Example 7.

contains μ. An approximate 95% confidence interval for μ is then obtained by replacing the random sample mean $\overline{X}$ by the value actually obtained, namely $\overline{x}$, and the population standard deviation σ by the sample standard deviation, s. This gives the approximate large-sample **95% confidence interval for $\boldsymbol{\mu}$**:

$$(\mu^-, \mu^+) = \left(\overline{x} - 1.96\,\frac{s}{\sqrt{n}}, \overline{x} + 1.96\,\frac{s}{\sqrt{n}}\right). \tag{2}$$

Here, μ^- and μ^+ are the **lower** and **upper 95% confidence limits for $\boldsymbol{\mu}$**.

Note that the confidence interval given in Interval (2) is approximate for two reasons:

- First, the Central Limit Theorem has been used to approximate the sampling distribution of the sample mean.
- Second, the population standard deviation σ has been replaced by s.

However, for large samples, the approximation will be good.

As was mentioned in Unit 6, the standard deviation of the sample mean $\overline{X}$, namely $\sigma/\sqrt{n}$, is often called the *standard error* of $\overline{X}$. Similarly, the estimated standard deviation of the sample mean $\overline{X}$, namely $s/\sqrt{n}$, is called the *estimated standard error* of $\overline{X}$.

Activity 3 *Accident counts*

In Example 3, data were described on the number of accidents suffered by children between the ages of 4 and 11 years. The sample size n is 621, the sample mean $\overline{x}$ is 2.45, and the sample standard deviation s is 2.03. Using these data, calculate an approximate 95% confidence interval for the mean number of accidents suffered by children between the ages of 4 and 11 years.

The confidence intervals described so far are 95% confidence intervals. The percentage '95%' is called the **confidence level** of the confidence interval. The confidence level 95% is determined by the numerical constant 1.96 used in Interval (2): 95% of values in a normal distribution lie within 1.96 standard deviations of the mean. Note that the constant 1.96 is $q_{0.975}$, the 0.975-quantile of the standard normal distribution. And the probability that a standard normal random variable Z lies between $-q_{0.975}$ and $q_{0.975}$ is 0.95:

In much of this unit, 1.96 is used as shorthand for its value correct to three decimal places, which is 1.960.

$$\begin{aligned} P\left(-q_{0.975} \leq Z \leq q_{0.975}\right) &= \Phi(q_{0.975}) - \Phi(-q_{0.975}) \\ &= \Phi(q_{0.975}) - (1 - \Phi(q_{0.975})) \\ &= 2\Phi(q_{0.975}) - 1 \\ &= 2 \times 0.975 - 1 = 0.95. \end{aligned}$$

Using a 95% confidence level is perhaps something of a default choice in many statistical investigations. However, other confidence levels may be obtained by using different quantiles. For example, for a normal distribution, 99% of values are within 2.576 standard deviations of the mean. The number 2.576 is the 0.995-quantile of the standard normal distribution. So, replacing 1.96 with $q_{0.995} = 2.576$ leads to an approximate 99% confidence interval for μ:

$$(\mu^-, \mu^+) = \left(\overline{x} - 2.576\,\frac{s}{\sqrt{n}}, \overline{x} + 2.576\,\frac{s}{\sqrt{n}}\right).$$

Similarly, 90% of values are within 1.645 standard deviations of the mean, where 1.645 is the 0.95-quantile of $N(0,1)$. So, replacing 1.96 with $q_{0.95} = 1.645$ leads to an approximate 90% confidence interval for μ:

$$(\mu^-, \mu^+) = \left(\overline{x} - 1.645\,\frac{s}{\sqrt{n}}, \overline{x} + 1.645\,\frac{s}{\sqrt{n}}\right).$$

More generally, $100(1-\alpha)\%$ of values lie between $q_{\alpha/2}$ and $q_{1-(\alpha/2)}$. Moreover, from Subsection 4.3 of Unit 6, for the standard normal distribution, $q_{\alpha/2} = -q_{1-(\alpha/2)}$. Therefore $100(1-\alpha)\%$ of values lie between $-q_{1-(\alpha/2)}$ and $q_{1-(\alpha/2)}$. This is illustrated in Figure 4 and confirmed in the following activity.

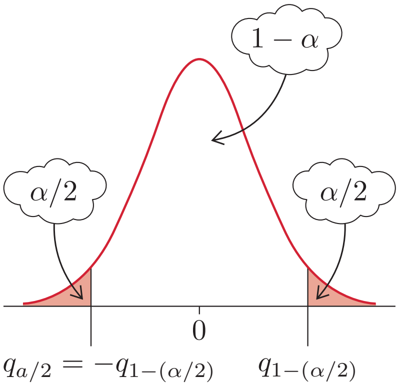


Figure 4 Quantiles of the standard normal distribution

Activity 4 *Proving that we're using the right quantile*

Show that for any α, $0 < \alpha < 1$, the probability that a standard normal random variable Z lies between $-q_{1-(\alpha/2)}$ and $q_{1-(\alpha/2)}$ is $1-\alpha$.

Notice that confidence levels are indexed by $100(1-\alpha)\%$ rather than, say, $100\,\gamma\%$ (and equivalently probabilities are indexed by $1-\alpha$ rather than γ). Moreover, this leads to use in the confidence interval of the $(1-(\alpha/2))$-quantile of the standard normal distribution. For example, a 99% confidence level corresponds to $\alpha = 0.01$ and to the 0.995-quantile of the standard normal distribution. This slightly awkward notation has its roots in the link between confidence intervals and hypothesis tests that we are leaving to one side for now.

γ is the Greek lower-case letter gamma.

Some of the quantiles of the standard normal distribution most widely used for $100(1-\alpha)\%$ confidence intervals are shown in Table 1; the quantiles given in this table are part of a larger table of quantiles that was first given in Unit 6 and is also given in the Handbook. In Table 1, normal quantiles are indexed by β, $0 < \beta < 1$.

Table 1 Quantiles of the standard normal distribution

β	0.9	0.95	0.975	0.99	0.995	0.999
q_β	1.282	1.645	1.960	2.326	2.576	3.090

Activity 5 *Matching confidence levels and quantiles*

Set $\beta = 1 - (\alpha/2)$ in Table 1. Complete the following expansion of Table 1 by inserting the confidence levels of the confidence intervals associated with each of the values of $1 - (\alpha/2)$ and $q_{1-(\alpha/2)}$.

Table 2

$1-(\alpha/2)$	0.9	0.95	0.975	0.99	0.995	0.999
Quantile, $q_{1-(\alpha/2)}$	1.282	1.645	1.960	2.326	2.576	3.090
Confidence level (%)						

The result of all this is that for any value of α between 0 and 1, a $100(1-\alpha)\%$ confidence interval can be obtained by replacing the number 1.96 in Interval (2) by $q_{1-(\alpha/2)}$, the $(1-(\alpha/2))$-quantile of the standard normal distribution. The $(1-(\alpha/2))$-quantile of the standard normal distribution is often denoted by z (just as a standard normal random variable is generally denoted by Z). This leads to the following definition of a large-sample confidence interval for a population mean, which is sometimes also called a z-interval.

A large-sample confidence interval for the population mean

Given a random sample $x_1, x_2, \dots, x_n$ of size n from a population with mean μ, an approximate $100(1-\alpha)\%$ confidence interval for μ, valid for large n, is given by

$$(\mu^-, \mu^+) = \left(\overline{x} - z\frac{s}{\sqrt{n}}, \overline{x} + z\frac{s}{\sqrt{n}}\right), \tag{3}$$

where $\overline{x}$ is the sample mean, s is the sample standard deviation, and z is $q_{1-(\alpha/2)}$, the $(1-(\alpha/2))$-quantile of the standard normal distribution; μ^- and μ^+ are, respectively, the lower and upper $100(1-\alpha)\%$ confidence limits for μ. This confidence interval is sometimes called a ***z*-interval**.

Remember that the approximate confidence interval for population mean μ given in the box is valid regardless of the underlying population distribution if n is large enough. The rule of thumb given in Unit 6 was that the Central Limit Theorem applies if $n \geq 25$. Here, we also need s to be a good estimate of σ which *could* be used as a reason to require n to be a little larger, but in M248 we continue to use the requirement that n be greater than or equal to 25 for a z-interval to be appropriate too.

Activity 6 *More confidence intervals for accident counts*

In Activity 3 you calculated an approximate 95% confidence interval for the mean number of accidents suffered by children between the ages of 4 and 11 years, using data first described in Example 3. Use these data to calculate approximate 90% and 99% confidence intervals for the mean number of accidents suffered by children between the ages of 4 and 11.

You may have noticed that of the three z-intervals you calculated in Activities 3 and 6, the 99% confidence interval is the widest, and the 90% confidence interval is the narrowest. This accords with the notion that the wider an interval of plausible values, the more confident you can be that the interval contains the population value. Mathematically, the width of the z-interval given by Interval (3) is given by

$$\mu^+ - \mu^- = \overline{x} + z\frac{s}{\sqrt{n}} - \left(\overline{x} - z\frac{s}{\sqrt{n}}\right) = 2z\frac{s}{\sqrt{n}}. \tag{4}$$

For given values of s and n, this width increases with increasing z. And $z = q_{1-(\alpha/2)}$, the point at which $P(Z \leq z) = 1 - (\alpha/2)$ when $Z \sim N(0,1)$, clearly increases with increasing values of the confidence level $100(1-\alpha)\%$. (Equivalently, it increases with decreasing values of α.) In short, the higher the confidence level, the wider the z-interval.

For examples of this, see the table in the solution to Activity 5.

So while, as we have already said, 95% is perhaps usually used as the choice of confidence level for confidence intervals, if you would like to be more sure that your confidence interval contains the population mean, then you should use a higher confidence level; 99% is then a popular choice. The trade-off, though, is a widening of the confidence interval, as discussed above. Similarly, if, for some reason, you are willing to be less certain that your confidence interval contains the population mean, then you can use a lower confidence level; 90% is then a popular choice. (And the corresponding confidence interval will be narrower.)

Activity 7 *Other factors affecting widths of confidence intervals*

By considering Equation (4), identify the two aspects of the data that affect the width of a z-interval. As these quantities increase, do they increase or decrease the width of the z-interval?

Exercise on Section 1

Exercise 1 *Fish traps*

One hundred fish traps were set, and later the number of fish caught in each trap was counted. The sample mean number of fish per trap was 4.04, and the sample standard deviation was 1.435. (Source: David, F.N. (1971) *A First Course in Statistics*, 2nd edn, London, Griffin.)

(a) Calculate an approximate 90% confidence interval for the mean catch per trap.

(b) State any assumptions you have made in calculating this confidence interval.

A 1920s Native American fish trap

2 Interpreting confidence intervals

While learning how to construct confidence intervals is important, it can be argued that knowing how to interpret confidence intervals is even more important. This is the subject of the current section.

In Section 1, you saw how to calculate an approximate 95% confidence interval for μ, the mean strength of glass fibres of length 1.5 cm. The confidence interval was $(1.43, 1.59)$. A key step in the argument leading to this result was the observation that, assuming σ to be known, with probability 0.95, the interval

$$\left(\overline{X} - 1.96\,\frac{\sigma}{\sqrt{63}}, \overline{X} + 1.96\,\frac{\sigma}{\sqrt{63}}\right)$$

This is a repeat of Interval (1).

contains μ.

This interval is a **random interval**: its endpoints are defined in terms of $\overline{X}$, which is itself a random variable, so the observed values of the random interval will vary according to the random sample selected. The key point about this statement is that it is a probability statement about a random interval, and not about the particular numerical interval $(1.43, 1.59)$ that was calculated from the sample actually obtained. To repeat, the probability associated with the confidence level, 0.95, is the probability that the random interval given in Interval (1) contains the true fixed, but unknown, value of μ (under, in this case, the sampling distribution of $\overline{X}$).

This distinction is important!

You might feel that this is rather unhelpful, since you are really interested in the sample actually obtained. However, another way to think of the probability statement is that it applies to the *statistical procedure* of drawing a random sample of size 63 from the population of glass fibres of length 1.5 cm, calculating the mean and standard deviation from this sample, and deriving the numerical values of the confidence limits. In other words, if this procedure were repeated many times, yielding confidence intervals (μ_1^-, μ_1^+), (μ_2^-, μ_2^+), (μ_3^-, μ_3^+), and so on, then about 95% of these intervals would contain the true (but unknown) value of μ; the other 5% would not contain μ. This is often known as the *repeated experiments* interpretation of confidence intervals; it is illustrated in Figure 5.

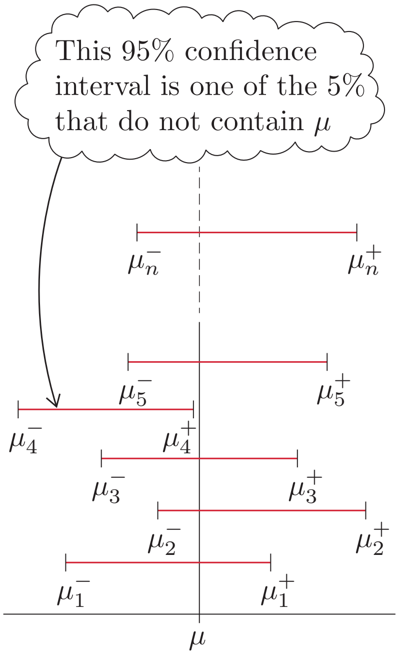


Figure 5 The repeated experiments interpretation of confidence intervals

The distinction is the same as the one between estimate (the confidence interval actually observed) and estimator (the procedure leading to the confidence interval).

The distinction between the numerical limits of the confidence interval calculated from the sample actually collected, and the statistical procedure involved in deriving it, is important. It is a common error to complete a statistical investigation with a 'confidence' statement such as 'with probability 0.95, the value of μ lies between 1.43 and 1.59', with the implication being that μ is a random variable about which a probability statement is being made. This is incorrect because μ is a fixed quantity, not a random variable to which we can attach meaningful probability statements. The random variables involved in the probability statement connected with a confidence interval are not μ but the estimators that give rise to the observed confidence limits.

To expand on the latter point, an observed confidence interval like $(1.43, 1.59)$ is sometimes referred to as a *realisation* of the random Interval (1) (in the same way that the value $\overline{x} = 1.51$ is a realisation of the random variable $\overline{X}$). The value of μ may or may not actually be between 1.43 and 1.59: the definition of a 95% confidence interval implies that if the sampling procedure were repeated a large number of times, then about 5% of the intervals generated in this way would miss the population mean completely. Thus there is no guarantee that the population mean glass-fibre strength lies in the interval calculated from a sample: quite possibly it might not. The 'confidence', implied in the statement 'The 95% confidence limits for the mean glass-fibre strength are 1.43 and 1.59', derives from the *statistical procedure* used to calculate the limits, rather than from the limits themselves.

In this sense, statistics is quite unlike cooking, the proof of the pudding lying not in the eating, but in the recipe!

The repeated experiments interpretation of confidence intervals is best illustrated by repeating experiments! This can readily be simulated on a computer. Therefore, in Computer Book B you have the opportunity to investigate the repeated experiments interpretation of confidence intervals using software designed for this purpose.

Refer to Chapter 4 of Computer Book B for the next part of the work in this subsection.

The interpretation of confidence intervals is summarised in the following box. It makes precise what we informally meant earlier by a confidence interval providing a range of plausible values for a parameter, given the data. It is phrased in terms of estimating the population mean, but the italicised passage applies to confidence intervals for other parameters also.

Interpreting confidence intervals

A $100(1-\alpha)\%$ confidence interval (μ^-, μ^+) for a population mean μ, calculated from a sample of size n with sample mean $\overline{x}$, may be interpreted as follows:

If a large number of samples of size n were drawn independently from the population, and a $100(1-\alpha)\%$ confidence interval calculated on each occasion, then approximately $100(1-\alpha)\%$ of these intervals would contain the population mean μ.

This is often known as the **repeated experiments** interpretation of confidence intervals.

Activity 8 *Interpreting confidence intervals*

In a study of anxiety prior to colonoscopy, the anxiety levels of 150 patients (77 male and 73 female) were assessed by means of a standardised questionnaire and summarised by means of an anxiety score: the higher the score, the higher the level of anxiety. The mean anxiety score for the female patients was 46.3, with 95% confidence interval $(44.9, 47.7)$. The mean score for the male patients was 36.9, with 95% confidence interval $(35.5, 38.3)$. (Source: Luck, A. et al. (1999) 'Effects of video information on precolonoscopy anxiety and knowledge: a randomised trial', *The Lancet*, vol. 354, no. 9195, pp. 2032–5.)

(a) Interpret the confidence interval for the women's mean anxiety score.

(b) Comment on the difference between men's and women's scores. Is the underlying mean anxiety score likely to be the same for women and men?

Activity 9 *Speed of light*

Suppose that 100 sets of 100 measurements are taken of the speed of light in a vacuum, denoted c, using a particular experimental technique. The true value of c is known, to a great degree of accuracy, by other methods; it is about 299 792 458 metres per second. For each set of 100 measurements, a 99% confidence interval for the mean is calculated. As it turns out, 60 of the confidence intervals do not include c. What might you conclude about the accuracy of the measurement technique?

Exercise on Section 2

Exercise 2 *Fish traps*

In Exercise 1, you found that the approximate 90% confidence interval for the mean catch per trap for 100 fish traps is $(3.80, 4.28)$. Interpret your confidence interval in terms of repeated experiments.

3 More large-sample confidence intervals

The method for calculating confidence intervals described in Subsection 1.2 applies to large samples for which the normal approximation given by the Central Limit Theorem may be used. The procedure may be summarised as follows.

- Use the normal approximation to the sampling distribution of the mean to write down a random interval with a specified probability of containing the population mean.
- Collect a random sample.
- Calculate numerical values for the limits of the random interval using the sample mean and sample standard deviation.
- Call your single observation on that random interval a confidence interval.

This method for calculating confidence intervals is general. It can be applied to all types of numerical data, not just continuous data. In fact, you have already seen an instance of this in Activities 3 and 6. The only requirements are that the sample is random, and that the sample size is large enough for the normal approximation given by the Central Limit Theorem to be used and for the sample standard deviation to lie close to the population standard deviation.

In this section, the method is extended to calculate approximate confidence intervals for parameters other than means. In Subsection 3.1, you will see how, in certain circumstances, a confidence interval for the mean can be used to determine a confidence interval for another parameter. In Subsection 3.2, the method is applied to calculate approximate confidence intervals for proportions, and in Subsection 3.3 for differences between proportions. In Subsection 3.4, you will learn how to use Minitab to calculate the large-sample confidence intervals introduced thus far. Finally, in Subsection 3.5, the general method is applied to calculate approximate confidence intervals for the Poisson parameter.

3.1 New confidence intervals from old

So far, the large-sample method has been applied to calculating confidence intervals for population means. In fact, the method can be extended to a range of other parameters using a simple trick. The idea is illustrated in Example 8.

Example 8 *Accident rates*

In Example 3, data were given on the number of accidents suffered by children between the ages of 4 and 11 years. Altogether 621 children were sampled. The average number of accidents suffered by each child was 2.45, and the sample standard deviation was 2.03. In Activity 3, you calculated a 95% confidence interval for the mean number of accidents per child; this was $(2.29, 2.61)$.

Hopefully few accidents here!

Other useful quantities may also be estimated from the data. For example, the average interval between accidents is $8/2.45 \simeq 3.27$ years; the number 8 comes from the fact that the period 4 to 11 years of age stretches from the fourth birthday to just before the twelfth birthday – an eight-year period. How should we construct a 95% confidence interval for the mean interval between accidents? Common sense might suggest that we apply the same operation to the limits of the confidence interval for the mean number of accidents $(2.29, 2.61)$ as to the sample mean. This gives the values $8/2.29 \simeq 3.49$ and $8/2.61 \simeq 3.07$, producing the interval $(3.07, 3.49)$.

Notice that the limits have been swapped around, so that the lower limit for the mean number of accidents corresponds to the upper limit for the mean interval between accidents, and the upper limit for the mean number of accidents corresponds to the lower limit for the mean interval between accidents. If this hadn’t been done, the lower bound of the interval would have been larger than the upper bound of the interval! A more formal justification that this is a reasonable thing to do will be given below; it will also be confirmed that the interval $(3.07, 3.49)$ is indeed a 95% confidence interval for the mean interval between accidents.

The method described in Example 8 may be justified as follows. For definiteness, the accident data will be used to illustrate the argument.

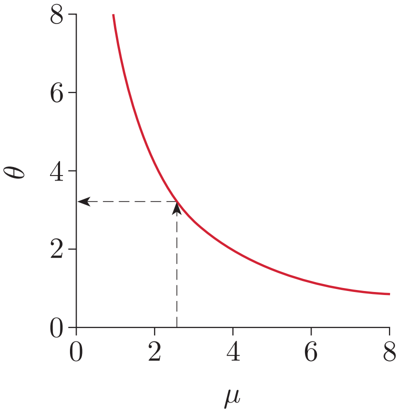


Figure 6 The transformation $\theta = 8/\mu$

The 95% confidence interval $(2.29, 2.61)$ for the mean number of accidents was based on a sample of 621 children. Suppose now that we wish to calculate a 95% confidence interval for the average interval between accidents. Let this mean interval be denoted by the symbol θ. Since accidents are counted over an eight-year period, θ is related to the mean number of accidents by the transformation

$$\theta = 8/\mu,$$

where $\mu > 0$. This transformation is represented in Figure 6. Using this transformation, θ is estimated by $\widehat{\theta} = 8/\widehat{\mu} = 8/2.45 \simeq 3.27$ years.

If $\overline{X}$ denotes the sample mean of a random sample of 621 children, then with approximate probability 0.95, the random interval

$$\left(\overline{X} - 1.96\,\frac{\sigma}{\sqrt{621}}, \overline{X} + 1.96\,\frac{\sigma}{\sqrt{621}}\right) \tag{5}$$

contains the population mean μ. But μ lies between the values $\overline{X} - 1.96\sigma/\sqrt{621}$ and $\overline{X} + 1.96\sigma/\sqrt{621}$ when $\theta = 8/\mu$ lies between the values

This argument assumes that the endpoints of the random interval are positive.

$$\frac{8}{\overline{X} - 1.96\sigma/\sqrt{621}} \quad \text{and} \quad \frac{8}{\overline{X} + 1.96\sigma/\sqrt{621}}. \tag{6}$$

These values are each of the form 8 divided by the corresponding value in Interval (5).

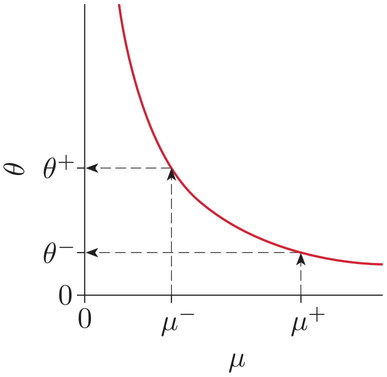


Figure 7 Transforming the limits

The first value in Expression (6) is larger than the second: this is because the function $\theta = 8/\mu$ in Figure 6 is decreasing, that is, its graph falls as you move to the right. Hence the lower limit for μ is transformed into the upper limit for θ, and vice versa; see Figure 7. It follows that, with probability approximately 0.95, the random interval

$$\left(\frac{8}{\overline{X} + 1.96\sigma/\sqrt{621}}, \frac{8}{\overline{X} - 1.96\sigma/\sqrt{621}}\right)$$

contains θ. An approximate confidence interval for θ is obtained by replacing $\overline{X}$ by the sample mean actually observed, namely $\overline{x} = 2.45$, and σ by the sample standard deviation $s = 2.03$. This is equivalent to applying the same transformation to the interval endpoints as to the estimate:

$$\theta^- = \frac{8}{\overline{x} + 1.96s/\sqrt{621}} = \frac{8}{\mu^+}, \qquad \theta^+ = \frac{8}{\overline{x} - 1.96s/\sqrt{621}} = \frac{8}{\mu^-}.$$

Thus an approximate 95% confidence interval for the mean interval between accidents is $(8/2.61, 8/2.29) = (3.07, 3.49)$ years.

So far, the procedure used in Example 8 has been justified for the particular transformation $\theta = 8/\mu$, where $\mu > 0$. Will this procedure work for any transformation? The answer is no. Consider, for example, the transformation $\theta = (\mu - 2.5)^2$ which is illustrated in Figure 8. This transforms the estimate $\widehat{\mu} = 2.45$ to $\widehat{\theta} = (2.45 - 2.5)^2 = 0.0025$, and the 95% confidence limits 2.29 and 2.61 to 0.0441 and 0.0121, respectively. Thus the estimate $\widehat{\theta}$ does not even lie between the transformed interval endpoints! The problem with the transformation in Figure 8 is that its graph falls, then rises: the function is neither increasing nor decreasing.

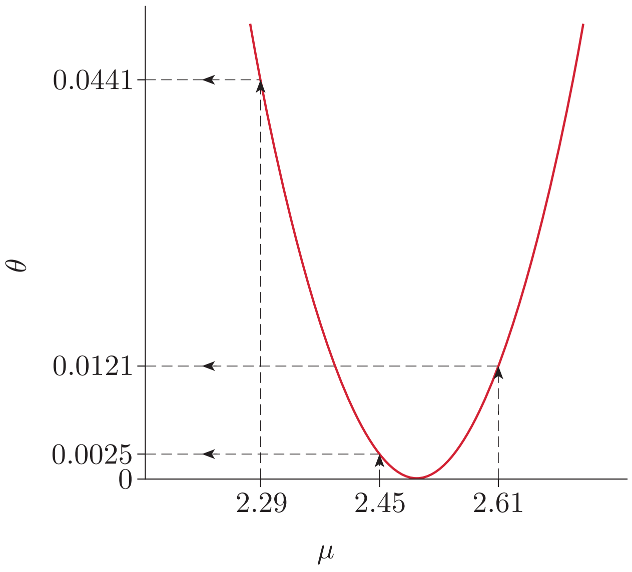


Figure 8 The transformation $\theta = (\mu - 2.5)^2$

However, the method does work whenever a transformation is either increasing or decreasing. This condition is sufficient to ensure that when the transformation is applied, values within the original interval are transformed to values within the new interval.

Instead of, or as well as, sketching a transformation, you can check whether it is increasing or decreasing or neither by considering its derivative. If the derivative of the transformation is positive, then the transformation is increasing; if the derivative of the transformation is negative, then the transformation is decreasing.

This mathematical check applies to all transformations that you will encounter in M248.

Example 9 *Derivatives of transformations*

The transformation $\theta = 8/\mu$ certainly appears to be decreasing in Figures 6 and 7. To check this mathematically, the derivative of θ with respect to μ is

$$\frac{d\theta}{d\mu} = -\frac{8}{\mu^2},$$

which is negative for any μ, and in particular for any $\mu > 0$. So this transformation is indeed decreasing.

On the other hand, the derivative of the transformation $\theta = (\mu - 2.5)^2$ is

$$\frac{d\theta}{d\mu} = 2(\mu - 2.5).$$

This is negative for $\mu < 2.5$, zero for $\mu = 2.5$, and positive for $\mu > 2.5$; see Figure 9. This confirms that this transformation is neither increasing, because its derivative is not positive for all $\mu > 0$, nor decreasing, because its derivative is not negative for all $\mu > 0$.

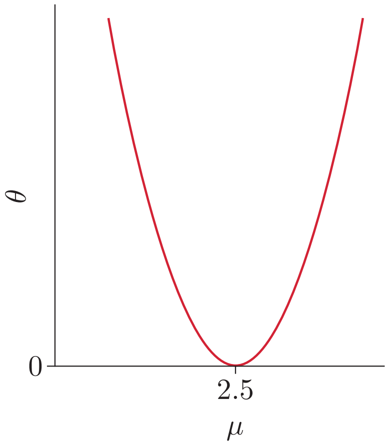


Figure 9 The transformation $\theta = (\mu - 2.5)^2$ with $\mu = 2.5$ marked

Derivatives of natural transformations

Activity 10 *More transformations*

(a) Sketch the graphs of the transformations $\theta = \mu/8$ and $\theta = e^{-\mu}$, for $\mu > 0$. Are they increasing, decreasing or neither?

(b) Confirm your answer to part (a) using differentiation.

The method described in this subsection may be summarised as follows.

Confidence intervals and transformations

Suppose that (μ^-, μ^+) is a $100(1-\alpha)\%$ confidence interval for μ, and $\theta = h(\mu)$. If the transformation h is either increasing or decreasing, then the limits $h(\mu^-)$ and $h(\mu^+)$ define a $100(1-\alpha)\%$ confidence interval for θ as follows.

- If h is increasing, then $(\theta^-, \theta^+) = (h(\mu^-), h(\mu^+))$. (7)
- If h is decreasing, then $(\theta^-, \theta^+) = (h(\mu^+), h(\mu^-))$. (8)

Strictly, you only need h to be increasing or decreasing for all $\mu^- \leq \mu \leq \mu^+$.

Example 10 *Annual accident rates*

A further parameter of interest relating to the accident data of Example 3 is the mean annual number of accidents suffered by children between the ages of 4 and 11 years. Let θ denote this parameter. Then $\theta = \mu/8$, where μ is the average number of accidents over the entire eight-year period. An estimate of θ is

$$\widehat{\theta} = \frac{\widehat{\mu}}{8} = \frac{2.45}{8} \simeq 0.306.$$

In Activity 10, you saw that the transformation $\theta = \mu/8$ is increasing. Hence a 95% confidence interval for θ may be derived from the 95% confidence interval for μ, $(2.29, 2.61)$, by applying this transformation to the confidence limits as in Interval (7):

$$(\theta^-, \theta^+) = \left(\frac{\mu^-}{8}, \frac{\mu^+}{8}\right) = \left(\frac{2.29}{8}, \frac{2.61}{8}\right) \simeq (0.286, 0.326).$$

The confidence limits for μ were obtained in Activity 3.

Activity 11 *Transformed anxiety scores*

Activity 8 concerned a study of anxiety prior to colonoscopy. One result of the study was that the mean anxiety score for male patients was 36.9, with 95% confidence interval $(35.5, 38.3)$. The anxiety score was derived from a questionnaire in such a way that its values range from a minimum of 20 to a maximum of 80. It could be argued that such a score would be more meaningful if transformed to a value between 0 and 100. Call the anxiety scores before and after transformation the ‘original’ and ‘transformed’ anxiety scores, respectively.

If x denotes a value of the original anxiety score and μ denotes its mean, then it turns out that $y = 5(x - 20)/3$, $20 \leq x \leq 80$, is the transformed anxiety score and $\theta = 5(\mu - 20)/3$, $20 \leq \mu \leq 80$, is its mean.

(a) Is the transformation $\theta = 5(\mu - 20)/3$ increasing, decreasing or neither?

(b) Give a point estimate and a 95% confidence interval for the mean transformed anxiety score.

3.2 Confidence intervals for proportions

In Subsection 3.1, you saw how, by transforming large-sample confidence intervals for the population mean, confidence intervals can be obtained for parameters that are certain types of functions of the mean. One advantage of the large-sample method for constructing confidence intervals is that it does not depend on modelling assumptions. In this subsection, the large-sample method is used in a slightly different way to calculate confidence intervals for proportions.

As you know, in some circumstances it is reasonable to assume that variation is described by a simple model. This is often true of data on the proportion of individuals with a particular attribute, for example: the proportion of people with a particular disease; the proportion of faulty components produced by a factory; the proportion of loan repayments made late; and so on. It is often reasonable to assume that such data arise from independent Bernoulli trials with common probability p, and hence that the total number of individuals with the attribute is binomial $B(n, p)$. These are the kinds of data that are considered in this subsection.

Suppose that x individuals out of n are observed to possess a particular attribute, and that x is an observation on the random variable $X \sim B(n, p)$. Interest centres on the population value of the parameter p. It turns out, however, that the kind of manipulation that we might like to do to obtain an exact confidence interval for the value of p is rather tricky for the binomial distribution. So we again turn to approximation by the normal distribution in order to obtain an approximate confidence interval for p.

Activity 12 *Normal approximation to the binomial*

For the binomial distribution, $B(n, p)$, we have $E(X) = np$ and $V(X) = np(1 - p)$; for the normal distribution, $N(\mu, \sigma^2)$, we have $E(X) = \mu$ and $V(X) = \sigma^2$.

(a) Use these facts to suggest the parameters for a normal distribution for X which approximates its true binomial distribution.

(b) Use the approximate normal distribution that you obtained in part (a) to suggest an approximate normal distribution for X/n.

It turns out that the approximate distribution that you obtained for X/n in Activity 12(b) is a good approximation to the distribution of X/n when both np and $n(1-p)$ are greater than about 5. In that case, we now have available an approximation to the distribution of X/n of the form

$$\frac{X}{n} \approx N\left(\mu, \frac{\sigma^2}{n}\right),$$

where $\mu = p$ and $\sigma^2 = p(1-p)$, that is,

$$\frac{X}{n} \approx N\left(p, \frac{p(1-p)}{n}\right). \tag{9}$$

It follows that the large-sample method for calculating confidence intervals that we used in Subsection 1.2, which was based on an approximate distribution of the form $N(\mu, \sigma^2/n)$, can be applied to the current problem, with mean $\mu = p$ and variance $\sigma^2 = p(1-p)$. The large-sample $100(1-\alpha)\%$ confidence interval for the mean μ of a random variable with known standard deviation σ may in general be written

$$(\mu^-, \mu^+) = \left(\widehat{\mu} - z\frac{\sigma}{\sqrt{n}}, \widehat{\mu} + z\frac{\sigma}{\sqrt{n}}\right), \tag{10}$$

where z is the $(1-(\alpha/2))$-quantile of the standard normal distribution. In the case of current interest, $\mu = p$, which is estimated by $\widehat{\mu} = \widehat{p} = x/n$, and $\sigma = \sqrt{p(1-p)}$, which is replaced by its estimate $\sqrt{\widehat{p}(1-\widehat{p})}$. This leads to the following result.

An approximate large-sample confidence interval for a proportion

An approximate $100(1-\alpha)\%$ confidence interval for a proportion p, obtained by observing x successes in a sequence of n independent Bernoulli trials each with probability of success p, is

$$(p^-, p^+) = \left(\widehat{p} - z\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}, \widehat{p} + z\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}\right), \tag{11}$$

where $\widehat{p} = x/n$ is the point estimate of p, and z is the $(1-(\alpha/2))$-quantile of the standard normal distribution.

Stating, as we did earlier, that the normal approximation underlying this confidence interval is valid when both np and $n(1-p)$ are at least 5 is not much good because we don't know p. But we can use its estimate $\widehat{p} = x/n$ once more to arrive at the following practical rule of thumb for employing Interval (11). As long as the number of 1s (successes) and the number of 0s (failures) in a sample are both more than about five (as in Example 11 below), approximate confidence intervals calculated using the approximate large-sample formula, Interval (11), should be reasonably accurate.

Example 11 *Smokers*

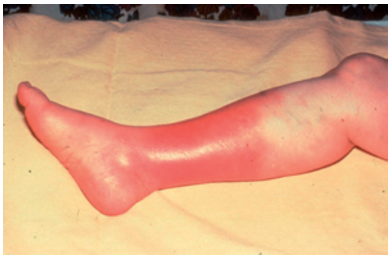
In Example 4, data were introduced on the number of smokers between the ages of 16 and 19 in the UK in 2012. From a random sample of 420 people in this age group, 15% were found to be smokers.

As explained in Example 4, the survey results do not give the actual number of smokers in the sample of 420. However, the number is not necessary to calculate the confidence interval. We know the percentage of the sample who smoke, so $\widehat{p} = 0.15$; this and the sample size are all we need to calculate an approximate confidence interval. In fact, an approximate 95% confidence interval for the unknown proportion p of smokers in the population is

$$\begin{aligned}(p^-, p^+) &= \left(\widehat{p} - z\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}, \widehat{p} + z\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}\right)\\ &= \left(0.15 - 1.96\sqrt{\frac{0.15 \times 0.85}{420}}, 0.15 + 1.96\sqrt{\frac{0.15 \times 0.85}{420}}\right)\\ &\simeq (0.15 - 0.03, 0.15 + 0.03)\\ &= (0.12, 0.18).\end{aligned}$$

The confidence interval is quite wide, the plausible range of values that it gives for the proportion of smokers in the population ranging from 12% to 18%.

Activity 13 *Cellulitis*



Cellulitis is an infection of the skin; its main symptom is the affected area of skin suddenly turning red, painful, swollen and hot.

A clinical trial was undertaken to examine the effect of taking a low dose of penicillin for twelve months on the recurrence of leg cellulitis in patients who had previously had two or more episodes of leg cellulitis. Of the 136 such patients randomly recruited to the trial from hospitals across the UK and Ireland, 30 patients had a recurrence of leg cellulitis during the twelve months of treatment. (Source: Thomas, K.S. et al. (2013) ‘Penicillin to prevent recurrent leg cellulitis’, *New England Journal of Medicine*, vol. 368, pp. 1695–703.)

Use these data to determine a 90% confidence interval for the proportion of patients with leg cellulitis whose cellulitis recurred during the time they were taking penicillin. State any modelling assumptions that you make.

3.3 Confidence intervals for differences between proportions

For the data described in Example 5, it is the difference between the proportions of female sandflies at 3 feet and 35 feet above ground level that is of interest. So we may wish to calculate a confidence interval for the difference between two proportions.

In general, suppose that x_1 individuals with a particular attribute are observed in a random sample of size n_1 from one population, and x_2 individuals with the same attribute in a random sample of size n_2 from a second population. Let p_1 and p_2 denote the proportions having the attribute in the two populations, and let $d = p_1 - p_2$ denote the difference between the proportions. An estimate of d is provided by the difference between the sample proportions:

$$\widehat{d} = \widehat{p}_1 - \widehat{p}_2 = \frac{x_1}{n_1} - \frac{x_2}{n_2}.$$

Suppose now that it is reasonable to assume that x_1 and x_2 are realisations of independent binomial random variables $X_1 \sim B(n_1, p_1)$ and $X_2 \sim B(n_2, p_2)$. When the numbers of successes and failures in both groups are sufficiently large, both X_1/n_1 and X_2/n_2 are approximately normal, by the argument made in Subsection 3.2 (see approximate Distributional Result (9)):

$$\frac{X_1}{n_1} \approx N\left(p_1, \frac{p_1(1-p_1)}{n_1}\right), \quad \frac{X_2}{n_2} \approx N\left(p_2, \frac{p_2(1-p_2)}{n_2}\right).$$

Let D denote the difference $(X_1/n_1) - (X_2/n_2)$. In Subsection 3.3 of Unit 6, you saw that: the mean of the difference between two independent random variables is equal to the difference between their means; the variance of their difference is equal to the *sum* of their variances; and if the random variables are normally distributed, then their difference is also normal. It follows that the distribution of D, which is a difference of two independent random variables with approximately normal distributions, is also approximately normally distributed:

$$D \approx N\left(p_1 - p_2, \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}\right).$$

Again using an argument similar to that used in Subsection 1.2 for constructing a confidence interval for a population mean, a random interval with probability $1 - \alpha$ of containing the difference $p_1 - p_2$ can be obtained. A confidence interval is then obtained by replacing the unknown parameters p_1 and p_2 in the expressions for the interval limits by their estimates $\widehat{p}_1 = x_1/n_1$ and $\widehat{p}_2 = x_2/n_2$, respectively.

An approximate large-sample confidence interval for the difference between two proportions

Suppose that $d = p_1 - p_2$ is the difference between two probabilities p_1 and p_2, and that x_1 and x_2 are observations on independent binomial random variables $X_1 \sim B(n_1, p_1)$ and $X_2 \sim B(n_2, p_2)$, respectively. Let $\widehat{p}_1 = x_1/n_1$, $\widehat{p}_2 = x_2/n_2$ and $\widehat{d} = \widehat{p}_1 - \widehat{p}_2$ denote the estimates of p_1, p_2 and d, respectively. Then an approximate $100(1-\alpha)\%$ confidence interval (d^-, d^+) for d is given by

$$\left(\widehat{d} - z\sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2}},\ \widehat{d} + z\sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2}}\right), \tag{12}$$

where z is the $(1-(\alpha/2))$-quantile of the standard normal distribution.

Notice that this does not reduce to taking the difference between confidence limits for single proportions.

Example 12 *Sandflies*

In Example 5, data were given on the proportions of female sand flies sampled at 3 feet and at 35 feet above ground level. The proportion of females observed at 3 feet was 150/323, and the proportion observed at 35 feet was 73/198. Suppose that we wish to calculate a 95% confidence interval for the population difference between proportions of females at these two heights. We will assume that sandflies at each height had the same probability of being caught, and that the true proportions of females are p_1 at 3 feet and p_2 at 35 feet. We will also assume that the sampling at the two levels was independent. With these assumptions, the data may be regarded as two independent random samples from binomial $B(n_1, p_1)$ and $B(n_2, p_2)$ distributions, with $n_1 = 323$ (at 3 feet) and $n_2 = 198$ (at 35 feet).

The sample proportions are

$$\widehat{p}_1 = \frac{150}{323} \simeq 0.464, \quad \widehat{p}_2 = \frac{73}{198} \simeq 0.369,$$

and the sample difference between proportions is

$$\widehat{d} \simeq 0.464 - 0.369 = 0.095.$$

Approximate 95% confidence limits for the population difference $d = p_1 - p_2$ are given by

$$\begin{aligned}
d^- &= \widehat{d} - z\sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2}} \\
&= 0.095 - 1.96\sqrt{\frac{0.464 \times 0.536}{323} + \frac{0.369 \times 0.631}{198}} \\
&\simeq 0.095 - 0.086 = 0.009, \\
d^+ &= \widehat{d} + z\sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2}} \\
&\simeq 0.095 + 0.086 = 0.181.
\end{aligned}$$

Hence an approximate 95% confidence interval for the difference between the proportions is $(0.009, 0.181)$. In percentage terms, a plausible range of values (at the 95% confidence level) stretches from a little below 1% to a little over 18%. This range does not include zero or any negative values: it seems that the proportion of females is likely to be higher at 3 feet than at 35 feet.

Activity 14 *Cellulitis: comparing penicillin and placebo*

As mentioned in Unit 3, a placebo is an inert substance. It is used in clinical trials to compare with active drugs, to control for any psychological benefits of simply taking medicine.

Activity 13 concerned a trial undertaken to examine the effect of penicillin on the recurrence of leg cellulitis in patients who had previously had two or more episodes of this condition. Of the 136 patients who received penicillin, 30 patients had a recurrence of leg cellulitis during the twelve months of treatment. In addition, a further 138 such patients received a placebo (no active treatment) instead of penicillin. Of these, 51 had a recurrence of leg cellulitis in a twelve-month period. Interest really centres on a comparison between the rates of recurrence of leg cellulitis in patients receiving penicillin and patients receiving the placebo.

(a) Calculate an approximate 95% confidence interval for the difference between the probabilities that patients receiving penicillin and patients receiving the placebo will suffer a recurrence of their leg cellulitis.

(b) Interpret this confidence interval.

3.4 Large-sample confidence intervals using Minitab

The large-sample confidence intervals described in Section 1 and in this section are straightforward to obtain using a calculator, once the sample mean and standard deviation, or the sample proportions, are known. However, calculating these summary statistics is often best done on a computer, especially for large datasets. In this subsection you will learn how to calculate large-sample confidence intervals using Minitab.

Refer to Chapter 5 of Computer Book B for the work in this subsection.

3.5 Confidence intervals for the Poisson parameter

In Subsection 3.2, it was argued that data on the proportion of individuals with a particular attribute often arise from a simple model, namely independent Bernoulli trials with common probability p, and hence that the total number of individuals with the attribute is binomial $B(n, p)$. Approximate large-sample confidence intervals for the proportion p were then derived using this modelling assumption.

Similarly, it is often assumed that data on counts of items or events follow a Poisson distribution with parameter $\lambda > 0$. Such data might include the number of catastrophic natural events over a decade, or the number of radioactive emissions in some time period, or the number of defective items leaving a factory, as just some examples.

Despite being counts, the numbers of accidents in Example 3 happen not to be especially well modelled by a Poisson distribution, so a less model-dependent large-sample approach was taken for those data in Subsection 1.2.

Suppose that a sample of n counts, $x_1, x_2, \ldots, x_n$, is observed and that each x_i is an observation on the random variable $X \sim \text{Poisson}(\lambda)$. From Unit 4, $E(X) = V(X) = \lambda$. Now let the sample mean be $\overline{x} = (x_1 + x_2 + \cdots + x_n)/n$. This is an observation on the random variable version of the sample mean, $\overline{X} = (X_1 + X_2 + \cdots + X_n)/n$, where $X_1, X_2, \ldots, X_n$ are independent random variables each following the Poisson(λ) distribution. The Central Limit Theorem can be invoked in order to derive large-sample confidence intervals for λ.

Activity 15 *Developing the large-sample confidence interval*

(a) Use the Central Limit Theorem to write down the approximate distribution of $\overline{X}$.

(b) Using Interval (10) and replacing any unknown parameters by their sample estimates, what is the form of the large-sample confidence interval for λ?

The result of Activity 15 is summarised in the following box.

An approximate large-sample confidence interval for the Poisson parameter

An approximate large-sample $100(1-\alpha)\%$ confidence interval for the Poisson parameter λ is

$$(\lambda^-, \lambda^+) = \left(\overline{x} - z\sqrt{\frac{\overline{x}}{n}}, \overline{x} + z\sqrt{\frac{\overline{x}}{n}}\right), \tag{13}$$

where $\overline{x}$ is the sample mean, and z is the $(1-(\alpha/2))$-quantile of the standard normal distribution.

It turns out that, as a rule of thumb, this confidence interval is valid when $n\lambda$ is at least 30.

Activity 16 *Confidence intervals for alpha particles*

It was argued in Example 13 of Unit 5 that counts of certain radioactive particle emissions are Poisson distributed. In particular, data from Rutherford and Geiger's classical 1910 experiment were discussed, in which $n = 2612$ counts of alpha particles were obtained with a sample mean number of emissions of $\overline{x} = 3.877$. Notice that $n\overline{x}$, which estimates $n\lambda$, is about 10 127, which is very much greater than 30.

(a) What is a 95% large-sample confidence interval for λ, the parameter of the Poisson distribution?

(b) The above relates to numbers of alpha particles emitted in periods of length $7\frac{1}{2}$ seconds. Suppose that, instead, interest centres on the rate of emissions per second. The mean such rate is therefore $\lambda/7.5$. What is a 95% large-sample confidence interval for $\lambda/7.5$?

Large-sample confidence intervals for Poisson means and rates can also be calculated using Minitab, in similar ways to those in Subsection 3.4. However, they are not included in this unit in an attempt to avoid it becoming too tedious.

Exercises on Section 3

Exercise 3 *Variance and standard deviation*

A source of environmental data as well as energy

Sometimes it is of interest to provide a confidence interval for the variance or for the standard deviation of a random variable, rather than for its mean. Suppose that, based on some environmental data, the 95% confidence interval $(1010.2, 1894.7)$ had been provided for a variance, σ^2. A critical reader of the researchers' report asked that a 95% confidence interval be provided for the standard deviation, σ, instead.

(a) For clarity, write $\theta = \sigma$ and $\beta = \sigma^2$. What transformation takes you from β to θ? By differentiation, ascertain whether, for $\beta > 0$, the transformation is increasing, decreasing or neither.

(b) Hence calculate an approximate 95% confidence interval for the standard deviation.

Exercise 4 *Curing infertility*

This question concerns women of child-bearing age who suffer from no-ovulation-cycle syndrome. It uses data from a 2009 Japanese study to compare follitropin alpha (hereafter, the 'study drug') with human menopausal gonadotropin (hereafter, the 'standard drug'). Out of 129 women who received the study drug, 102 were 'cured' (that is, successfully ovulated); out of 132 women who received the standard drug, 109 were cured. (Source: Kawasaki, Y. and Miyaoka, E. (2012) 'A Bayesian inference of $P(\pi_1 > \pi_2)$ for two proportions', *Journal of Biopharmaceutical Statistics*, vol. 22, no. 3, pp. 425–37.)

(a) Calculate an approximate 95% confidence interval for the probability that a woman receiving the study drug will be cured.

(b) Calculate an approximate 95% confidence interval for the difference between the proportions of women that are cured when given the study drug and when given the standard drug.

(c) Interpret the confidence interval that you obtained in part (b).

4 Exact confidence intervals for normal means

The big advantage of many of the methods for calculating confidence intervals discussed so far is that they are general, and require few assumptions about the underlying distribution of the data. For the remainder of this unit, however, we will switch attention away from confidence intervals based on arguments that assume large samples, and consider instead confidence intervals that are valid for any sample size but are based on specific assumptions about the distribution of the data. The main reason for doing so is to be able to provide better statistical inferences when the data consist of only (fairly) small samples.

In this section, confidence intervals for means of *normally distributed* random variables are discussed. The confidence intervals described in this section have the property that they are valid irrespective of the sample size; furthermore, they do not involve any approximations, and in this sense, they are *exact*. Thus when the underlying distribution is normal – and this assumption is the trade-off we are making to be able to obtain exactness – it is preferable to use the exact intervals and hence avoid any unnecessary approximations.

Exact confidence intervals for the mean of a normal population are partly derived in Subsection 4.1, and their derivation is completed in Subsection 4.3; confidence intervals for the difference between two normal means are discussed in Subsection 4.4. These intervals make use of a family of distributions not yet discussed in this module: the family of *t-distributions*. This family is described in Subsection 4.2. You will see how to obtain confidence intervals for normal means using Minitab in Subsection 4.5.

4.1 Revisiting the arguments leading to a confidence interval for a normal mean

Example 13 is typical of the sort of situation where an exact confidence interval for the mean of a normal distribution is appropriate.

Example 13 *A mechanical kitchen timer*

A kitchen timer is a small alarm clock that can be set to ring after any length of time between one minute and an hour. It is useful as a reminder to somebody working in a kitchen that some critical stage has been reached; the usefulness of such timers is not restricted to the kitchen, of course.

An enthusiastic cook was interested in the accuracy of his own kitchen timer, which was set by turning a dial, and on ten different occasions he set it to ring after a five-minute delay (300 seconds). The ten different time intervals recorded on a stopwatch are shown in Table 3 (overleaf).

Table 3 Time delay (seconds)

293.7	293.7	296.2	294.3	296.4	291.3	294.0	295.1	297.3	296.1

(Source: data provided by B.J.R. Bailey, University of Southampton)

Assuming that the stopwatch itself was an accurate measuring instrument, the only variability from the 300 seconds' delay intended in the times recorded would be due to difficulties in actually setting the time (that is, positioning the dial) and/or to malfunction in the operation of the timer.

The sample mean $\overline{x}$ is 294.81 seconds (that is, about 4 min 55 s or five seconds short of the five-minute interval set) and the sample standard deviation s is 1.77 seconds. The sample size, which is 10, is small, so the methods developed in Sections 1 and 3 may not be applicable. How can we calculate a confidence interval for the mean time recorded when a time of five minutes is set?

In Subsection 1.2, $n \geq 25$ was suggested as the rule of thumb to justify the use of large-sample methods with the sample mean.

Consider again the method introduced in Subsection 1.2. However, suppose now that the random variable X is normally distributed, $X \sim N(\mu, \sigma^2)$, and that $X_1, X_2, \ldots, X_n$ are independent observations on X. It follows that the sample mean is also normally distributed:

See Subsection 6.4 of Unit 6.

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

So, thanks to the normality (and independence) of $X_1, X_2, \ldots, X_n$, the *exact* normal distribution of the sample mean $\overline{X}$ may be used without calling on the Central Limit Theorem to provide an approximate normal distribution.

The properties of the normal distribution can now be used to make probability statements about $\overline{X}$. For example, with probability 0.95, $\overline{X}$ lies within $1.96\sigma/\sqrt{n}$ of the mean, μ. More generally, if z is the $(1-(\alpha/2))$-quantile of the standard normal distribution, then with probability $1-\alpha$, $\overline{X}$ lies within $z\sigma/\sqrt{n}$ units of μ. Equivalently, with probability $1-\alpha$, the random interval

$$\left(\overline{X} - z\frac{\sigma}{\sqrt{n}}, \overline{X} + z\frac{\sigma}{\sqrt{n}}\right)$$

contains μ. This statement is exact; no approximation is involved. If the true value of σ were known, then given a sample $x_1, x_2, \ldots, x_n$, an exact $100(1-\alpha)\%$ confidence interval for μ would be

$$(\mu^-, \mu^+) = \left(\overline{x} - z\frac{\sigma}{\sqrt{n}}, \overline{x} + z\frac{\sigma}{\sqrt{n}}\right). \tag{14}$$

One difficulty remains, however: this is that σ is not usually known. In Subsection 1.2, this difficulty was resolved by replacing σ with the sample standard deviation s. This is reasonable for large samples, since the calculation of s is then based on a large number of observations and is likely to produce a result close to the population standard deviation. However, when the sample size is small, as in Example 13, this may not be so. This is illustrated in Example 14.

Example 14 *Variability of the sample standard deviation*

For illustrative purposes, let us regard the first two values of Table 3 as a sample of size $n = 2$, the next two values as a separate (independent) sample of size $n = 2$, and so on, so that we have five independent samples of size $n = 2$. Then calculate the sample mean and sample standard deviation of each of these five samples. The results are shown in Table 4.

Table 4 Means and standard deviations

Sample	1	2	3	4	5
$\overline{x}$	293.7	295.3	293.9	294.6	296.7
s	0	1.34	3.61	0.78	0.85

The sample means from these five tiny samples are not too different from one another. The sample standard deviations from these five samples, on the other hand, vary widely, from 0 up to 3.61. This is a rather extreme example, in that most well-designed statistical experiments would use sample sizes rather larger than 2! However, it illustrates the point that for small samples, the sample standard deviation is very variable. In consequence, simply approximating σ by s in Interval (14) may produce very inaccurate confidence intervals.

The method used to derive the confidence interval for μ in Interval (14), assuming that σ is known, is based on the fact that

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

When $\overline{X}$ is standardised, this gives

See Subsection 4.4 of Unit 6.

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1). \tag{15}$$

When σ is not known, one solution is still to replace σ by S, the random variable representing the sample standard deviation for samples of size n. This defines a new random variable T:

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}}. \tag{16}$$

The expression for T in Equation (16) does not feature the unknown parameter σ. So it is possible to make probability statements about it such as

$$P\left(-t \leq \frac{\overline{X} - \mu}{S/\sqrt{n}} \leq t\right) = 1 - \alpha.$$

But, first, appropriate values of the quantity t have to be found. That is, before the random variable T can be used to make exact inferences about μ, we need to know its properties, and in particular its probability distribution. This is the subject of the next subsection; we will complete the argument that provides exact confidence intervals for a normal mean in Subsection 4.3 that follows it.

4.2 The family of t-distributions

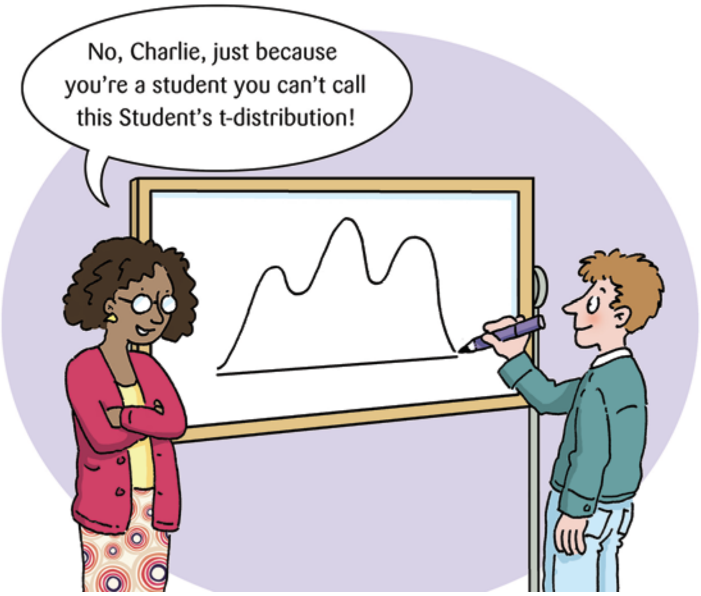
Since T in Equation (16) has been obtained from Z in Distributional Result (15) by replacing a constant, σ, by a random variable, S, you should not be surprised to learn that T is not normally distributed. For small samples, the distribution of T differs markedly from the standard normal distribution. For large samples, the two distributions are quite similar, thus providing further justification for the large-sample methods described in Sections 1 and 3.

ν is the Greek lower-case letter nu, pronounced 'new', to rhyme with 'mew'.

The distribution of T is known as *Student's t-distribution* or often just the *t-distribution* for short. In fact, there is not just one t-distribution, but a whole family of distributions indexed by a parameter ν called, unobviously, the *degrees of freedom*. If the random variable T has a t-distribution with ν degrees of freedom, then this is written

$$T \sim t(\nu).$$

Suppose now that $\overline{X}$ is the mean of a random sample of size n from a normal distribution with mean μ, and that S is the sample standard deviation. In this case, the random variable T defined in Equation (16) has Student's t-distribution with degrees of freedom $\nu = n - 1$. ('Student' was the pseudonym of W.S. Gosset (1876–1937).)


No, Charlie, just because you're a student you can't call this Student's t-distribution!

Student's t-distribution as a sampling distribution

In a random sample of size n with sample mean $\overline{X}$ and sample standard deviation S from a normal distribution with mean μ, the random variable

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

has **Student's *t*-distribution** (or simply a ***t*-distribution**) with $n-1$ degrees of freedom. This is written

$$T \sim t(n-1).$$

This result is stated without proof.

Note that the distribution of T does not depend on σ, the standard deviation of the normal random variable X. It is also the case that the degrees of freedom parameter, being the sample size minus one, will always take integer values: $\nu = 1, 2, 3, \ldots$. The value $\nu = 0$ is not included because the sample standard deviation is not defined when $n = 1$.

Like the standard normal random variable Z, the probability density function of each member of this family of distributions, for $\nu = 1, 2, 3, \ldots$, is symmetric about 0: the numerator of T, the difference $\overline{X} - \mu$, is as likely to be negative as it is to be positive. But in view of its dependence on two sample statistics (the sample standard deviation S and the sample mean $\overline{X}$), the random variable T is, in a sense, more variable than Z, and its probability density function has heavier tails than that of Z. This is clearly seen in Figure 10, which shows the probability density function of Z together with those of $t(1)$, $t(3)$ and $t(7)$ – that is, of the t-distributions with 1, 3 and 7 degrees of freedom. Indeed, as ν decreases, the $t(\nu)$ p.d.f. becomes less and less like the normal distribution, reflecting the very small sample size on which the sample standard deviation is then based.

The $t(1)$ distribution is also known as the Cauchy distribution.

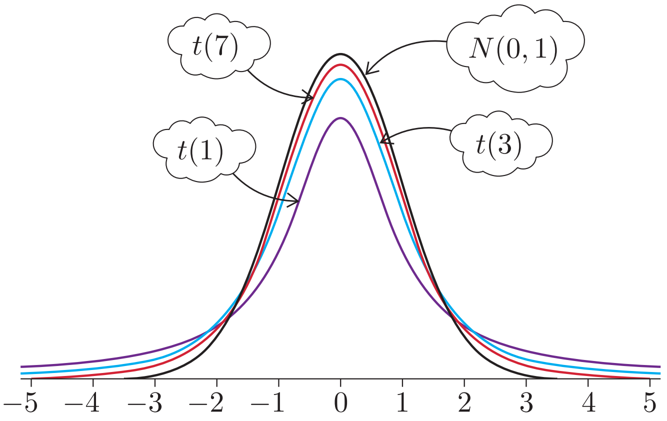
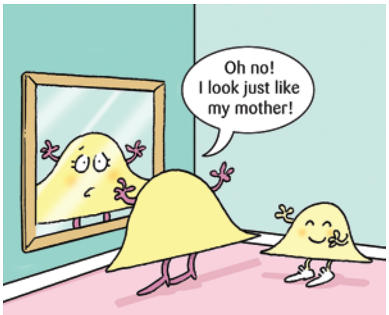


Figure 10 The p.d.f.s of $t(1)$, $t(3)$, $t(7)$ and $N(0,1)$

On the other hand, as the value of ν is increased, $t(\nu)$ becomes closer and closer to a standard normal distribution. This also makes sense: when the sample size n is large, S will generally produce a good estimate of σ, and the distribution of the random variable T will be close to the distribution of the standard normal random variable Z.



The t-distribution family at home

The general form of the p.d.f. of the $t(\nu)$ distribution is not as scary as some textbooks might lead you to believe. For those of you who are interested, the p.d.f., $f_\nu(t)$, is *proportional* to $g_\nu(t)$ where

You will not be expected to perform any calculations using this formula.

$$g_\nu(t) = \left(1 + \frac{t^2}{\nu}\right)^{-(\nu+1)/2}.$$

The symmetry of the p.d.f. is clear from the formula because it depends on t only through the value of t^2. In the following screencast, which is optional, it is shown how, mathematically, the p.d.f. proportional to $g_\nu(t)$ tends to the standard normal p.d.f. as ν becomes large.

Screencast 8.1 How the t-distribution tends to the normal distribution as ν becomes large (optional)

In order to calculate confidence intervals for the unknown parameter μ using Student's t-distribution, it will be necessary to employ quantiles of t-distributions. Computer software provides these, and later you will use Minitab to do so. For hand calculation, however, statistical tables remain useful. Quantiles of $t(\nu)$ for different values of ν are shown in Table 5; the values of ν are in the column headed 'df' (for degrees of freedom). This table is part of a larger table of quantiles that is given in the Handbook.

Table 5 Quantiles of t-distributions

df	0.90	0.95	0.975	0.99	0.995	0.999
1	3.078	6.314	12.71	31.82	63.66	318.3
2	1.886	2.920	4.303	6.965	9.925	22.33
3	1.638	2.353	3.182	4.541	5.841	10.21
4	1.533	2.132	2.776	3.747	4.604	7.173
5	1.476	2.015	2.571	3.365	4.032	5.893
6	1.440	1.943	2.447	3.143	3.707	5.208
7	1.415	1.895	2.365	2.998	3.499	4.785
8	1.397	1.860	2.306	2.896	3.355	4.501
9	1.383	1.833	2.262	2.821	3.250	4.297
10	1.372	1.812	2.228	2.764	3.169	4.144
11	1.363	1.796	2.201	2.718	3.106	4.025
12	1.356	1.782	2.179	2.681	3.055	3.930
13	1.350	1.771	2.160	2.650	3.012	3.852
14	1.345	1.761	2.145	2.624	2.977	3.787
15	1.341	1.753	2.131	2.602	2.947	3.733
16	1.337	1.746	2.120	2.583	2.921	3.686
17	1.333	1.740	2.110	2.567	2.898	3.646
18	1.330	1.734	2.101	2.552	2.878	3.610
19	1.328	1.729	2.093	2.539	2.861	3.579
20	1.325	1.725	2.086	2.528	2.845	3.552

Example 15 *Using the table of quantiles of t-distributions*

Four quantiles of t-distributions, for different degrees of freedom, are illustrated in Figure 11.

In Figure 11(a), for instance, the size of the right-hand shaded area is $\frac{1}{2}(1-0.95)=0.025$; so the corresponding quantile is $q_{0.975}$, the 97.5% point of $t(9)$. The value of $q_{0.975}$ is found by looking along the row of the table corresponding to $\nu=9$ and in the column headed 0.975: the entry gives the probability $P(T\leq t)=0.975$, so the value of t is 2.262.

By symmetry, the other quantile marked on Figure 11(a) corresponds to a probability of $1-0.975=0.025$ below it, and is minus the value of the quantile with a probability of 0.025 above it, that is, -2.262.

You should check that you can obtain the other quantiles shown in Figure 11, using the table of quantiles in the Handbook.

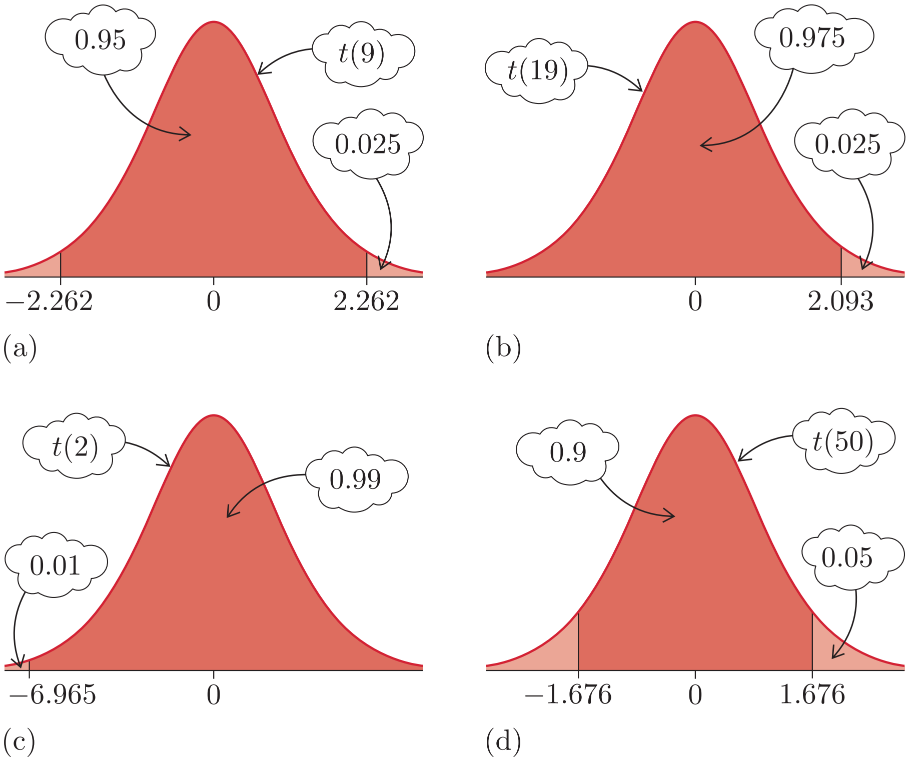


Figure 11 Quantiles of $t(\nu)$

Activity 17 *Quantiles of t-distributions*

Use the table of t-quantiles in the Handbook to find the following values of t.

(a) If $\nu=29$, determine t such that $P(T\leq t)=0.95$.

(b) If $\nu=30$, determine t such that $P(T\geq t)=0.05$.

(c) If $\nu = 5$, determine t such that $P(T \leq t) = 0.01$.

(d) If $\nu = 19$, determine t such that $P(|T| \leq t) = 0.99$.

You will not need to remember this formula.

A little-known fact is that – unlike normal quantiles and quantiles of other t-distributions – the α-quantile for the $t(2)$ distribution has an explicit algebraic formula. It is

$$\frac{2\alpha - 1}{\sqrt{2\alpha(1-\alpha)}}.$$

If you wish, you can check that this gives the values in the 'df = 2' row of Table 5; it does!

4.3 Confidence intervals for normal means

In this subsection, the problem of calculating confidence intervals for a normal mean μ is discussed further. In Subsection 4.2, you saw that if $\overline{X}$ is the sample mean and S is the sample standard deviation for a sample of size n from a normal distribution with mean μ, then

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t(n-1).$$

An important consequence of this result is that it is possible to make probability statements of the form

$$P\left(-t \leq \frac{\overline{X} - \mu}{S/\sqrt{n}} \leq t\right) = 1 - \alpha,$$

required at the end of Subsection 4.1, knowing now that in this statement t is the $(1 - (\alpha/2))$-quantile of the $t(n-1)$ distribution. This means that, with probability $1 - \alpha$, following a similar argument to that given in Subsection 1.2, $\overline{X}$ and μ lie no more than $tS/\sqrt{n}$ units apart. Equivalently, with probability $1 - \alpha$, the interval

$$\left(\overline{X} - t\frac{S}{\sqrt{n}}, \overline{X} + t\frac{S}{\sqrt{n}}\right)$$

contains μ. This is a random interval, centred on the sample mean $\overline{X}$, which with probability $1 - \alpha$ contains the unknown population mean μ. Note that the probability here is exact: no approximations have been used. Replacing the random variables $\overline{X}$ and S by the values obtained in the sample actually observed yields an exact confidence interval, sometimes called a t-interval because of its association with the t-distribution.

An exact confidence interval for a normal mean

Given a random sample of size n with sample mean $\overline{x}$ and sample standard deviation s from a normal distribution with mean μ, a $100(1-\alpha)\%$ confidence interval for μ based on this sample is

$$(\mu^-, \mu^+) = \left(\overline{x} - t\frac{s}{\sqrt{n}}, \overline{x} + t\frac{s}{\sqrt{n}}\right). \tag{17}$$

Here t is the $(1 - (\alpha/2))$-quantile of the t-distribution with $n - 1$ degrees of freedom. This confidence interval is exact and is sometimes referred to as a ***t*-interval**.

There is only one difference between a t-interval and a z-interval (Subsection 1.2): in a t-interval, you use t, the $(1 - (\alpha/2))$-quantile of the $t(n - 1)$ distribution, instead of z, the $(1 - (\alpha/2))$-quantile of the standard normal distribution, used in a z-interval.

Example 16 *Kitchen timer*

Following on from Example 13, if we assume that the time until the kitchen timer alarm bell sounds is normally distributed, but with unknown mean and variance, then Interval (17) may be used to construct a confidence interval for the unknown mean time delay μ. The sample statistics are $\overline{x} = 294.81$ seconds and $s = 1.77$ seconds. For a 90% confidence interval, for instance, since $\nu = n - 1 = 9$, the t-value required is the 0.95-quantile of $t(9)$, which is 1.833. Hence a 90% confidence interval for μ is

For a 90% confidence level, $\alpha = 0.1$, so $1 - (\alpha/2) = 0.95$.

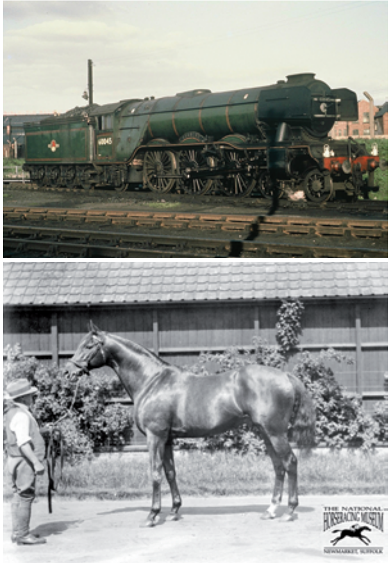
$$\begin{aligned}(\mu^-, \mu^+) &= \left(\overline{x} - t\frac{s}{\sqrt{n}}, \overline{x} + t\frac{s}{\sqrt{n}}\right) \\ &= \left(294.81 - 1.833\,\frac{1.77}{\sqrt{10}}, 294.81 + 1.833\,\frac{1.77}{\sqrt{10}}\right) \\ &\simeq (294.81 - 1.03, 294.81 + 1.03) \\ &= (293.78, 295.84).\end{aligned}$$

Notice that the confidence interval is usefully narrow; this is because the amount of variability in the data, as measured by the sample standard deviation, is small. Also, this interval of plausible values does not contain the number 300 (seconds, the five-minute delay that the timer was supposed to ring after). This suggests that the timer is consistently going off early.

Activity 18 *Coal consumption of the locomotive* Lemberg

In 1928 the London and North Eastern Railway ran the locomotive *Lemberg*, with an experimental boiler pressure of 220 pounds per square inch, in five trial runs. Several random variables were measured. One was the coal consumption, in pounds per drawbar horsepower hour. The five observations were as follows.

3.27 3.17 3.24 2.92 2.99



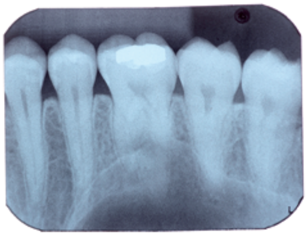
The locomotive *Lemberg* and the successful racehorse after which it was named

(a) Find the sample mean and sample standard deviation.

(b) Regarding these data as a random sample from a normal distribution, construct a 95% confidence interval for the mean coal consumption of the *Lemberg* at a boiler pressure of 220 pounds per square inch.

4.4 Confidence intervals for differences between normal means

In Subsection 3.3, methods were described for calculating confidence intervals for the difference between two proportions. Similarly, in applications involving continuous data, interest often focuses on estimating the difference between two means. In the rest of this section, the methods that have been described for calculating confidence intervals for single normal means are extended to differences between means arising from two independent samples of normally distributed data.



Example 17 *X-ray penetration*

The extent to which X-rays can penetrate tooth enamel has been investigated as the basis of a method for distinguishing between males and females in forensic medicine. The extent of penetration of the X-rays is called the 'spectropenetration gradient'. Spectropenetration gradients were measured for one tooth from each of eight females and eight males. (Source: Harraway, J.A. (1997) *Introducing Statistical Methods for Biological, Health and Social Sciences*, Dunedin, University of Otago Press.) Is there a difference between spectropenetration gradients for females and males of sufficient magnitude to be useful in distinguishing between the two? One way to investigate this question is to compare the mean spectropenetration gradients for females, μ_1 say, with the mean spectropenetration gradients for males, μ_2 say. This may be done by calculating a confidence interval for $\mu_1 - \mu_2$. The observations in the two groups can be assumed to be independent. Normal probability plots (not shown) suggest that it might be reasonable to assume a normal model for the variation observed in each group. How can we construct a confidence interval for the difference between the means μ_1 and μ_2?

Suppose that $X_1 \sim N(\mu_1, \sigma^2)$ and $X_2 \sim N(\mu_2, \sigma^2)$ are normal random variables with means μ_1 and μ_2, respectively, and variance σ^2 which is assumed to be the *same* in each distribution. (This is a standard initial assumption that will be discussed further later.) We wish to obtain a confidence interval for $\mu_1 - \mu_2$, the difference between the means.

Two random samples are drawn: a sample of size n_1 with sample mean $\overline{X}_1$ from the population with mean μ_1, and a sample of size n_2 with sample mean $\overline{X}_2$ from the population with mean μ_2. The sample means $\overline{X}_1$ and $\overline{X}_2$ may be used to estimate the population means μ_1 and μ_2, so it seems reasonable to base the estimate of the difference $d = \mu_1 - \mu_2$ on the difference between the sample means, $\overline{X}_1 - \overline{X}_2$. The variation in each population is normal, so the sample means are also normally distributed, as already used in Subsection 4.1:

$$\overline{X}_1 \sim N\left(\mu_1, \frac{\sigma^2}{n_1}\right), \quad \overline{X}_2 \sim N\left(\mu_2, \frac{\sigma^2}{n_2}\right).$$

Suppose also that the two samples are independent of one another. It follows that the difference $D = \overline{X}_1 - \overline{X}_2$ is also normally distributed, with mean the difference of the means and variance the sum of the variances (as already used in Subsection 3.3):

$$D = \overline{X}_1 - \overline{X}_2 \sim N\left(\mu_1 - \mu_2, \sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right). \tag{18}$$

As for the single sample case, the variance σ^2 is generally not known. When dealing with a single sample, the sample standard deviation S is used to estimate σ, and the t-distribution is used. In the present setting, however, there are two distinct (and independent) estimates of the common standard deviation: s_1, say, from the first sample, and s_2 from the second. It would make sense, intuitively, to combine these in some way. There are several ways of doing this, but the optimal combination turns out to be the following. It is the so-called 'pooled' estimate of the variance.

Pooled estimate of the variance

Given independent samples of size n_1 with sample variance s_1^2 and n_2 with sample variance s_2^2 from distributions with a common variance, the pooled estimate of the common variance is

$$s_P^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}. \tag{19}$$

The corresponding random variable is denoted S_P^2.

You will often use the corresponding estimate, s_P, of the common standard deviation, σ, which is obtained, of course, by taking the square root of s_P^2 as given in Equation (19).

The pooled variance estimate weights each sample variance in proportion to its sample size minus one. It therefore gives more weight to the estimate from the larger sample.

Using Distributional Result (18), it can be shown that

$$T = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{S_P\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} \sim t(n_1 + n_2 - 2). \tag{20}$$

You need not concern yourself with the details involved in obtaining this result.

That is, T has a t-distribution with $n_1 + n_2 - 2$ degrees of freedom. Given independent samples of sizes n_1 and n_2, the difference between the normal means is estimated by $\overline{x}_1 - \overline{x}_2$, and a $100(1 - \alpha)\%$ confidence interval for the unknown difference $d = \mu_1 - \mu_2$ is calculated as follows.

An exact confidence interval for the difference between normal means from two independent samples

Suppose that two independent samples have been collected from normal distributions with means μ_1 and μ_2, and common variance. If n_1 and n_2 are the sample sizes, $\overline{x}_1$ and $\overline{x}_2$ are the sample means, s_P is the pooled estimate of the common standard deviation, and t is the $(1-(\alpha/2))$-quantile of the t-distribution with n_1+n_2-2 degrees of freedom, then an exact $100(1-\alpha)\%$ confidence interval for the difference between the means $d=\mu_1-\mu_2$ is given by

$$(d^-,d^+)=\left(\overline{x}_1-\overline{x}_2-t\,s_P\sqrt{\frac{1}{n_1}+\frac{1}{n_2}},\overline{x}_1-\overline{x}_2+t\,s_P\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\right). \quad (21)$$

This confidence interval is sometimes referred to as a **two-sample *t*-interval**.

Interval (21) is exact, in the sense that no approximations are involved in its derivation. In particular, it is valid even for small samples. However, exactness comes at a price, in the form of assumptions that must be checked. The three assumptions required are as follows.

Assumptions for a two-sample t-interval

1 The variation in each population is adequately modelled by a normal distribution.

2 The samples drawn from each population are independent.

3 The variance is the same in the two populations.

Normal probability plots were discussed in Section 5 of Unit 6.

Assumption 1 may be checked informally by graphical methods, for example, by obtaining a histogram or a normal probability plot. The validity of assumption 2 depends on the design of the study and how the data were collected, and so can also be checked.

The length of the top section of an adult male thumb is approximately one inch. Such a ‘rule of thumb’ used to be used by carpenters.

But how should assumption 3 be checked? It will almost invariably be the case that the sample variances s_1^2 and s_2^2 for the two samples differ, and hence the question arises of how pronounced this difference might be before it suggests that the assumption of equal variances is faulty. In this module the following, standard, rule of thumb will be used: if the two sample variances differ by a factor of less than 3, it will be assumed that the assumption of equal variances is acceptable. This can be checked by dividing the larger of the two sample variances by the smaller, and observing whether or not this ratio is less than 3.

Activity 19 *X-ray penetration*

In Example 17, the problem set was to calculate a 95% confidence interval for the difference between the mean spectropenetration gradients of females and of males. The first sample consists of the spectropenetration gradients of 8 females; their sample mean is $\overline{x}_1 = 4.513$ and their sample variance is $s_1^2 = 0.5784$. The second sample consists of the spectropenetration gradients of 8 males; their sample mean is $\overline{x}_2 = 5.425$ and their sample variance is $s_2^2 = 0.5536$. Before using Interval (21), the assumptions must be checked. In Example 17, it was concluded that the normality assumption was reasonable; and the two samples may certainly be considered to be independent.

High values of the spectropenetration gradient imply less penetration by the X-rays.

(a) By comparing the sample variances, check that the assumption of a common variance is reasonable.

(b) Calculate the estimated difference between the mean spectropenetration gradient of females and the mean spectropenetration gradient of males, and the pooled estimate of the common standard deviation.

(c) Which quantile of which t-distribution is required in order to construct a 95% confidence interval in this case? Find the value of the appropriate quantile from the table in the Handbook.

(d) Hence calculate a 95% two-sample t-interval for the difference between the mean spectropenetration gradients of females and of males. On the basis of this confidence interval, might spectropenetration gradient be a useful tool in discriminating between males and females in forensic science?

4.5 Confidence intervals for normal means using Minitab

The calculation of sample means, sample standard deviations and pooled estimates can be very tedious, especially for larger samples. A computer is usually used to perform these arithmetic tasks. It can also be used to obtain appropriate t-values, and hence confidence intervals can be obtained at the touch of a key.

In this subsection, you will see how to use Minitab to perform these calculations. In addition, you will see how to calculate confidence intervals for differences between normal means when the assumption of a common variance does not hold.

If it is required to calculate a confidence interval for the difference between the means of two normal distributions, and the assumption that the two distributions share a common variance is untenable, then we must again resort to approximations, in place of the exact intervals of Subsection 4.4.

A Welsh tea interval. Even less fun than a Welch t-interval?

The method used by Minitab is called *Welch's t-interval* and generally produces confidence intervals that are slightly wider than those obtained using the equal variances assumption.

The explanation and formulas below are given for completeness and interest only; you do not need to engage with this material if you do not wish to. You certainly should, however, do the computer work which follows it (so you might wish to jump to that now). The method used by Minitab replaces the pooled variance used in the exact confidence interval, Interval (21), by the separate sample variances and uses an adjusted t-value that we will call t^*. So instead of estimating the variance in Distributional Result (18), namely

$$\sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right), \tag{22}$$

by

$$s_P^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right),$$

as is done in Interval (21), Expression (22) is estimated by

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

The problem is then that the exact Distributional Result (20) no longer holds. Instead, something called *Satterthwaite's approximation* is used, which replaces the degrees of freedom of the t-distribution, $n_1 + n_2 - 2$, by an appropriate approximate quantity. This quantity – which depends on the data through s_1^2 and s_2^2 – is given by

$$\nu_S = \left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)^2 \Bigg/ \left(\frac{s_1^4}{n_1^2(n_1 - 1)} + \frac{s_2^4}{n_2^2(n_2 - 1)} \right).$$

The resulting approximate $100(1 - \alpha)\%$ confidence interval for the difference between the means $d = \mu_1 - \mu_2$ when the assumption of common variances does not hold is given by

$$(d^-, d^+) = \left(\overline{x}_1 - \overline{x}_2 - t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}, \overline{x}_1 - \overline{x}_2 + t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \right),$$

where t^* is the $(1 - (\alpha/2))$-quantile of the t-distribution with ν_S degrees of freedom.

Refer to Chapter 6 of Computer Book B for the rest of the work in this subsection.

Exercises on Section 4

Exercise 5 *Silver content of Byzantine coins*

In Units 6 and 7, a small dataset on the percentages of silver in Byzantine coins was considered. The sample (from the first of four coinages) was of size $n = 9$, had sample mean 6.7444% and sample variance 0.2953%, and was argued to be plausibly modelled by the normal distribution. Construct a 99% confidence interval for the mean percentage of silver in such Byzantine coins.

Exercise 6 *Insect joint measurements*

Measurements were taken on the widths (in microns) of the first joint of the second tarsus (a segment of the leg) for two species of the insect *Chaetocnema* (flea beetles). Ten insects from each species were sampled, yielding the following sample means and variances:

$$\overline{x}_1 = 128.40, \quad s_1^2 = 48.93, \quad \overline{x}_2 = 122.80, \quad s_2^2 = 114.84.$$

(Source: Lindsey, J.C., Herzberg, A.M. and Watts, D.G. (1987) 'A method of cluster analysis based on projections and quantile–quantile plots', *Biometrics*, vol. 43, no. 2, pp. 327–41.)

Assume that the distribution of joint widths in each population is normal.

(a) Check that the assumption of common variance is reasonable, and calculate the pooled estimate of the common variance.

(b) Using a 95% confidence level, calculate a two-sample t-interval for the difference between the means.

Summary

In this unit, you have learned how to define, interpret and calculate confidence intervals for a population parameter and certain differences between population parameters using a random sample or samples from the population. A confidence interval is a realisation of a random interval with a specified probability of containing the population parameter value; this probability is called the confidence level. The definition leads to an interpretation of confidence intervals in terms of the sampling procedure used to generate them: if the entire procedure of collecting a random sample and calculating the confidence interval were repeated independently on a large number of occasions, then the proportion of confidence intervals containing the population parameter value would equal the confidence level.

For large samples, approximate confidence intervals are readily calculated with minimal assumptions by relying on the Central Limit Theorem. You have learned how to calculate large-sample confidence intervals for population means (z-intervals), for proportions and differences of proportions, for the Poisson parameter, and for transformations of these parameters.

When samples are drawn from normal populations, exact confidence intervals for population means may be calculated that do not rely on any approximations. These confidence intervals, which are sometimes called t-intervals, depend on a family of distributions called t-distributions. Quantiles of t-distributions are used in their calculation. You have seen how to calculate t-intervals for the mean of a normal population, and for the difference between two means using two independent normal samples, under appropriate assumptions.

You have used your computer to calculate confidence intervals using Minitab, and to explore the properties of confidence intervals using computer graphics.

Learning outcomes

After you have worked through this unit, you should be able to:

- understand how confidence intervals are defined in terms of random intervals
- interpret confidence intervals in terms of repeated experiments
- calculate a confidence interval for a population mean from a large sample, where 'large' is, in practice, at least 25
- obtain a confidence interval for $h(\mu)$, when h is either an increasing or a decreasing function, given a confidence interval for a parameter μ
- calculate an approximate confidence interval for a binomial parameter p from a sample of size n when np and $n(1-p)$ are both at least 5
- calculate an approximate confidence interval for the difference between binomial parameters from independent samples
- calculate an approximate confidence interval for the Poisson parameter λ from a sample of size n when $n\lambda$ is at least 30
- obtain quantiles of t-distributions from tables
- calculate a t-interval for the mean of a normal random variable
- calculate a two-sample t-interval for the difference between the means of two normal populations given independent samples from the populations
- state the assumptions required for calculating t-intervals
- use Minitab to calculate confidence intervals.

Solutions to activities

Solution to Activity 1

The population parameters are given in Table 6.

Table 6

Example 2	the mean breaking strength of glass fibres of length 1.5 cm
Example 3	the mean accident count for Californian boys between 4 and 11 years of age
Example 4	the probability that someone between 16 and 19 years of age smokes
Example 5	the difference between the proportions of female sandflies at 3 feet and at 35 feet above the ground

Solution to Activity 2

With $Z = (\overline{X} - \mu)/(\sigma/\sqrt{63})$ so that $Z \sim N(0,1)$,

$$\begin{aligned} &P\left(\mu - 1.96\,\frac{\sigma}{\sqrt{63}} \leq \overline{X} \leq \mu + 1.96\,\frac{\sigma}{\sqrt{63}}\right) \\ &\quad = P\left(-1.96 \leq Z \leq 1.96\right) = \Phi(1.96) - \Phi(-1.96) \\ &\quad = \Phi(1.96) - (1 - \Phi(1.96)) = 2\Phi(1.96) - 1 = 2 \times 0.975 - 1 = 0.95, \end{aligned}$$

using the table of the standard normal distribution (Table 5 of Unit 6 or in the Handbook).

Solution to Activity 3

Using the expression in Interval (2) with $\overline{x} = 2.45$ accidents and $s = 2.03$ accidents, the lower and upper 95% confidence limits are given by

$$\mu^- = 2.45 - 1.96\,\frac{2.03}{\sqrt{621}} \simeq 2.29,$$

$$\mu^+ = 2.45 + 1.96\,\frac{2.03}{\sqrt{621}} \simeq 2.61.$$

So an approximate 95% confidence interval for the mean number of accidents suffered by children between the ages of 4 and 11 is

$$(\mu^-, \mu^+) = (2.29, 2.61).$$

Solution to Activity 4

$$\begin{aligned} &P\left(-q_{1-(\alpha/2)} \leq Z \leq q_{1-(\alpha/2)}\right) = \Phi(q_{1-(\alpha/2)}) - \Phi(-q_{1-(\alpha/2)}) \\ &\quad = \Phi(q_{1-(\alpha/2)}) - (1 - \Phi(q_{1-(\alpha/2)})) = 2\Phi(q_{1-(\alpha/2)}) - 1 \\ &\quad = 2\{1 - (\alpha/2)\} - 1 = 1 - \alpha. \end{aligned}$$

Solution to Activity 5

The confidence levels of the confidence intervals associated with each of the values of $1-(\alpha/2)$ and $q_{1-(\alpha/2)}$ are $100(1-\alpha)\%$. So if $a = 1-(\alpha/2)$, then $\alpha = 2(1-a)$, so $100(1-\alpha) = 100(2a-1)$. For example, if $1-(\alpha/2) = 0.9$, then $\alpha = 2(1-0.9) = 0.2$, so $100(1-\alpha) = 100(1-0.2) = 80$. The completed table is given below.

Table 7

$1-(\alpha/2)$	0.9	0.95	0.975	0.99	0.995	0.999
Quantile, $q_{1-(\alpha/2)}$	1.282	1.645	1.960	2.326	2.576	3.090
Confidence level (%)	80	90	95	98	99	99.8

Solution to Activity 6

For an approximate 90% confidence interval, $z = q_{0.95} = 1.645$. The lower and upper 90% confidence limits are given by

$$\mu^- = 2.45 - 1.645\,\frac{2.03}{\sqrt{621}} \simeq 2.32,$$

$$\mu^+ = 2.45 + 1.645\,\frac{2.03}{\sqrt{621}} \simeq 2.58.$$

Thus an approximate 90% confidence interval for the mean number of accidents suffered by children between the ages of 4 and 11 is

$$(\mu^-, \mu^+) = (2.32, 2.58).$$

For an approximate 99% confidence interval, $z = q_{0.995} = 2.576$. The lower and upper 99% confidence limits are given by

$$\mu^- = 2.45 - 2.576\,\frac{2.03}{\sqrt{621}} \simeq 2.24,$$

$$\mu^+ = 2.45 + 2.576\,\frac{2.03}{\sqrt{621}} \simeq 2.66.$$

Hence an approximate 99% confidence interval is

$$(\mu^-, \mu^+) = (2.24, 2.66).$$

Solution to Activity 7

From Equation (4), the standard deviation of the sample, s, and the sample size, n, both affect the width of the z-interval. The larger the sample standard deviation, the wider the z-interval; the larger the sample size, the narrower the z-interval.

Solution to Activity 8

(a) If the entire experiment were repeated independently a large number of times, and on each occasion a random sample of 73 women were drawn and a 95% confidence interval were calculated, then about 95% of these intervals would contain the population value of the mean anxiety score. The 95% confidence interval actually observed, namely $(44.9, 47.7)$, is just one observation on a random interval, and may or may not contain the population mean.

(b) Suppose that the average anxiety score μ is indeed the same for women as it is for men. Then it cannot lie in both of the confidence intervals $(44.9, 47.7)$ for women and $(35.5, 38.3)$ for men, since these intervals do not overlap. In terms of repeated experiments, one, or perhaps even both, of the confidence intervals must have been among the unlucky 5% to miss the population value entirely. More reasonably, the very different confidence intervals might lead us to cast doubt on the notion that men and women share the same underlying anxiety score. We might perhaps conclude, on the contrary, that women tend to be more anxious than men prior to a colonoscopy. (Note, however, that to investigate whether the mean anxiety scores of men and women are equal, it is better to calculate a confidence interval for the difference between the two means. The formula required for this is not included in this unit.)

Solution to Activity 9

Each set of 100 measurements may be regarded as a random sample from the population of all measurements of the speed of light made with this technique. Let μ denote the mean of this population. Suppose that the instrument is not biased, so that $\mu = c$.

By its definition, a 99% confidence interval for μ is a realisation of a random interval with 0.99 probability of containing μ, and hence c (provided, of course, that $\mu = c$). Thus out of 100 such intervals, calculated from 100 independent samples of size 100, we would expect 99 intervals to contain c. There will, of course, be some random variation in this number: we would not be very surprised if, say, only 97 intervals contained c. However, we are told that only 40 intervals contain c. This is far fewer than expected, and hence might lead us to question the assumption that $\mu = c$. For example, we might suspect that our measuring instrument was biased after all, giving values systematically higher or lower than c.

Solution to Activity 10

(a) The graph of the transformation $\theta = \mu/8$ is increasing, as shown by Figure 12.

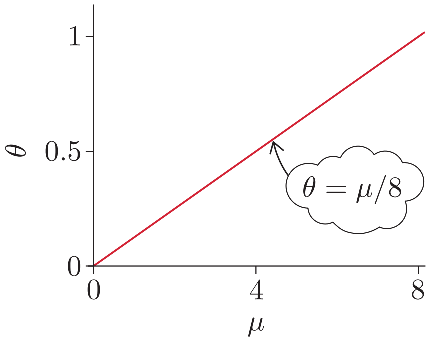


Figure 12 The transformation $\theta = \mu/8$

The graph of the transformation $\theta = e^{-\mu}$ is decreasing, as shown by Figure 13.

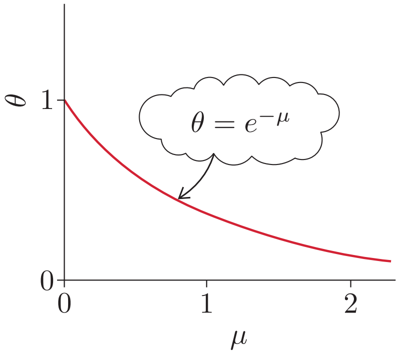


Figure 13 The transformation $\theta = e^{-\mu}$

(b) The derivative of the transformation $\theta = \mu/8$ is $d\theta/d\mu = 1/8$, which is positive for all $\mu > 0$, confirming that this transformation is increasing.

The derivative of the transformation $\theta = e^{-\mu}$ is $d\theta/d\mu = -e^{-\mu}$, which is negative for all $\mu > 0$, since $e^x > 0$ for any x, confirming that this transformation is decreasing.

Solution to Activity 11

(a) The graph of the transformation $\theta = 5(\mu - 20)/3$ is a line with intercept $-100/3$ and slope $5/3$, as shown in Figure 14.

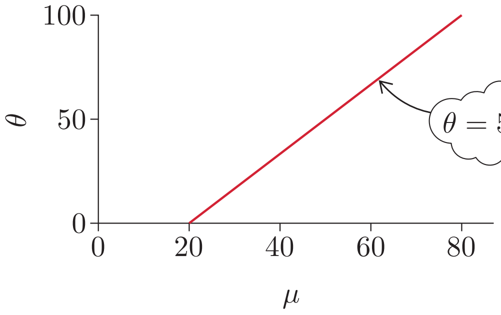


Figure 14 The transformation $\theta = 5(\mu - 20)/3$

Mathematically, $d\theta/d\mu = 5/3$, which is positive. Either way, the transformation is increasing.

(b) A point estimate for θ is $\widehat{\theta} = 5(\widehat{\mu} - 20)/3 = 5(36.9 - 20)/3 \simeq 28.2$.

Since the transformation is increasing, a 95% confidence interval for θ is, using Interval (7),

$$\begin{aligned}(\theta^-, \theta^+) &= \left(\frac{5(\mu^- - 20)}{3}, \frac{5(\mu^+ - 20)}{3}\right) \\ &= \left(\frac{5(35.5 - 20)}{3}, \frac{5(38.3 - 20)}{3}\right) \simeq (25.8, 30.5).\end{aligned}$$

Solution to Activity 12

(a) It seems reasonable to set $\mu = np$ and $\sigma^2 = np(1 - p)$ to obtain the approximating normal distribution for X as

$$X \approx N(np, np\,(1 - p)).$$

(b) From Subsection 3.1 of Unit 6, we know that if $X \sim N(\mu, \sigma^2)$, then $aX + b \sim N(a\mu + b, a^2\sigma^2)$. Applying this when $a = 1/n$, $b = 0$, yields

$$\frac{X}{n} \approx N\left(\frac{np}{n}, \frac{np\,(1 - p)}{n^2}\right) = N\left(p, \frac{p\,(1 - p)}{n}\right).$$

Solution to Activity 13

The estimate of p, the proportion of penicillin-taking leg cellulitis patients who suffer a recurrence of their cellulitis is

$$\widehat{p} = 30/136 \simeq 0.221.$$

Using Interval (11), the lower and upper confidence limits for p are given by

$$p^- = 0.221 - 1.645\sqrt{\frac{0.221 \times 0.779}{136}} \simeq 0.221 - 0.059 = 0.162,$$
$$p^+ = 0.221 + 1.645\sqrt{\frac{0.221 \times 0.779}{136}} \simeq 0.221 + 0.059 = 0.280.$$

So an approximate 90% confidence interval for the underlying proportion of leg cellulitis patients who suffer a recurrence of their cellulitis while taking penicillin is

$$(p^-, p^+) = (0.162, 0.280).$$

Here, it has been assumed that the recurrence or otherwise of leg cellulitis in different patients are independent events with the same probability of occurring, so that the data are binomially distributed. The assumption might, of course, not be valid.

Solution to Activity 14

(a) The estimated proportion of patients receiving penicillin who suffered a recurrence of their leg cellulitis was $\widehat{p}_1 = 30/136 \simeq 0.221$; the corresponding proportion for patients receiving the placebo was $\widehat{p}_2 = 51/138 \simeq 0.370$. The estimated difference between the proportions is

$$\widehat{d} = \widehat{p}_1 - \widehat{p}_2 = -0.149.$$

The lower and upper 95% confidence limits for d are obtained using Interval (12):

$$\begin{aligned} d^- &= -0.149 - 1.96\sqrt{\frac{0.221 \times 0.779}{136} + \frac{0.370 \times 0.630}{138}} \\ &\simeq -0.149 - 0.107 = -0.256 \end{aligned}$$

(which a fuller-accuracy calculation gives as -0.255, but the difference doesn't matter in the context of what is an approximate confidence interval anyway) and

$$\begin{aligned} d^+ &= -0.149 + 1.96\sqrt{\frac{0.221 \times 0.779}{136} + \frac{0.370 \times 0.630}{138}} \\ &\simeq -0.149 + 0.107 = -0.042. \end{aligned}$$

So an approximate 95% confidence interval for the underlying difference is

$$(d^-, d^+) = (-0.256, -0.042).$$

(b) If the entire trial were repeated independently a large number of times, and on each occasion a 95% confidence interval for the difference between proportions were calculated, then about 95% of these intervals would contain the population value of the difference between proportions of patients receiving penicillin and receiving the placebo suffering further leg cellulitis. In this instance, the 95% confidence interval does not contain any positive values, or the value zero which corresponds to no difference between the proportions of patients taking penicillin and taking the placebo suffering further leg cellulitis. This suggests that it is plausible that the proportion of patients receiving penicillin who had a recurrence of leg cellulitis is lower than the corresponding proportion of patients receiving the placebo; that is, the administration of penicillin seems to be having some effect on reducing leg cellulitis.

Solution to Activity 15

(a) The Central Limit Theorem says that $\overline{X}$ is approximately normally distributed with mean μ and variance σ^2/n, where $\mu = E(X)$ and $\sigma^2 = V(X)$. Using the formulas for the mean and variance of the Poisson(λ) distribution, this gives

$$\overline{X} \approx N\left(\lambda, \frac{\lambda}{n}\right).$$

(b) The large-sample confidence interval, Interval (10), involves σ, where the variance of the approximating normal distribution is σ^2/n. In this case, therefore, $\sigma = \sqrt{\lambda}$. Estimating λ by $\overline{x}$, Interval (10) becomes

$$(\lambda^-, \lambda^+) = \left(\overline{x} - z\sqrt{\frac{\overline{x}}{n}}, \overline{x} + z\sqrt{\frac{\overline{x}}{n}}\right).$$

Solution to Activity 16

(a) Using Interval (13),

$$\lambda^- = 3.877 - 1.96\sqrt{\frac{3.877}{2612}} \simeq 3.877 - 0.076 = 3.801,$$

$$\lambda^+ = 3.877 + 1.96\sqrt{\frac{3.877}{2612}} \simeq 3.877 + 0.076 = 3.953.$$

So an approximate 95% confidence interval for the underlying Poisson parameter is

$$(\lambda^-, \lambda^+) = (3.801, 3.953).$$

(b) An approximate 95% confidence interval for the per-second rate $\lambda/7.5$ is

$$\left(\frac{\lambda^-}{7.5}, \frac{\lambda^+}{7.5}\right) = (0.507, 0.527).$$

This is another application of the work of Subsection 3.1, the transformation $\theta = \lambda/7.5$ being an increasing one.

Solution to Activity 17

(a) Looking along the row of the table corresponding to $\nu = 29$ and in the column headed 0.95, you will find that $t = 1.699$.

(b) To find t such that $P(T \geq t) = 0.05$, you need to recognise that t satisfies $1 - P(T \leq t) = 0.05$ or $P(T \leq t) = 0.95$. Then the $\nu = 30$ row, 0.95 column gives $t = 1.697$.

(c) 0.01-quantiles are not provided in the table in the Handbook; they have to be found by a symmetry argument. The value of t such that $P(T \leq t) = 0.01$ is minus the value of t such that $P(T \leq t) = 1 - 0.01 = 0.99$. The $\nu = 5$ row, 0.99 column gives 3.365, so $t = -3.365$.

(d) The value of t such that $P(|T| \leq t) = 0.99$ is also the value of t such that $P(T \leq t) = 0.995$. You have used this kind of relationship for the normal distribution earlier and to make sense of Figures 11(a) and 11(d). As a reminder, from first principles, t is clearly positive, so

$$\begin{aligned} 0.99 = P(|T| \leq t) &= P(-t \leq T \leq t) = P(T \leq t) - P(T \leq -t) \\ &= P(T \leq t) - \{1 - P(T \leq t)\} = 2P(T \leq t) - 1. \end{aligned}$$

Thus we need $P(T \leq t) = (0.99 + 1)/2 = 0.995$, as claimed. The $\nu = 19$ row, 0.995 column gives $t = 2.861$.

Solution to Activity 18

(a) The sample mean and sample standard deviation are $\overline{x} = 3.118$ and $s \simeq 0.155$.

(b) For a 95% confidence interval, the required t-value is obtained from the t-distribution with $n - 1 = 5 - 1 = 4$ degrees of freedom: it is the 0.975-quantile of $t(4)$, $q_{0.975} = 2.776$.

The lower and upper 95% confidence limits for μ, the mean coal consumption in pounds per drawbar horsepower hour, are obtained using Interval (17):

$$\mu^- = 3.118 - 2.776 \times \frac{0.155}{\sqrt{5}} \simeq 3.118 - 0.192 = 2.926,$$

$$\mu^+ = 3.118 + 2.776 \times \frac{0.155}{\sqrt{5}} \simeq 3.118 + 0.192 = 3.310.$$

So a 95% confidence interval for μ is given by

$$(\mu^-, \mu^+) = (2.926, 3.310).$$

Solution to Activity 19

(a) The ratio of the larger of the two sample variances to the smaller is $0.5784/0.5536 \simeq 1.04$. Since this is less than 3, it is reasonable to proceed under the assumption of a common variance.

(b) The estimated difference between the mean spectropenetration gradient of females and the mean spectropenetration gradient of males is

$$\widehat{d} = \overline{x}_1 - \overline{x}_2 = 4.513 - 5.425 = -0.912.$$

The pooled estimate of the common variance is

$$\begin{aligned} s_P^2 &= \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2} \\ &= \frac{7 \times 0.5784 + 7 \times 0.5536}{14} = 0.5660. \end{aligned}$$

It is worth noting that in this case, the sample sizes in the two groups are equal, so the pooled estimate is in fact the average of the two sample variances. The pooled estimate of the common standard deviation is therefore

$$s_P = \sqrt{0.5660} \simeq 0.752.$$

(c) The required t-value is the 0.975-quantile of $t(14)$. The table in the Handbook lists the value as 2.145.

(d) Substituting all the necessary ingredients into Interval (21) gives the 95% confidence interval for the difference between mean spectropenetration gradients as (d^-, d^+), where

$$\begin{aligned} d^- &= \overline{x}_1 - \overline{x}_2 - t\, s_P \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \\ &= -0.912 - 2.145 \times 0.752 \times \sqrt{\frac{1}{8} + \frac{1}{8}} \\ &\simeq -0.912 - 0.807 = -1.719 \end{aligned}$$

and

$$d^+ \simeq -0.912 + 0.807 = -0.105.$$

Thus a 95% confidence interval for the difference between the mean spectropenetration gradients of females and of males is $(-1.719, -0.105)$.

This range of plausible values for the difference between means contains only negative values. This suggests that spectropenetration gradient might be a useful tool in discriminating between males and females in forensic science.

Solutions to exercises

Solution to Exercise 1

(a) We have $n = 100$, $\overline{x} = 4.04$, $s = 1.435$. For an approximate 90% confidence interval, z is the 0.95-quantile of the standard normal distribution, that is, $z = 1.645$. The lower and upper 90% confidence limits are

$$\mu^- = 4.04 - 1.645 \times \frac{1.435}{\sqrt{100}} \simeq 3.80,$$
$$\mu^+ = 4.04 + 1.645 \times \frac{1.435}{\sqrt{100}} \simeq 4.28.$$

So an approximate 90% confidence interval for the mean catch per trap is

$$(\mu^-, \mu^+) = (3.80, 4.28).$$

(b) In calculating this confidence interval, it has been assumed that the sample size is sufficiently large that the approximations involved in using the Central Limit Theorem, and in replacing the unknown population standard deviation by its sample value, do not introduce too much error.

Solution to Exercise 2

Suppose that the procedure of laying 100 fish traps, waiting the required time, lifting the traps, counting the fish in each trap, and calculating a 90% confidence interval for the mean catch per trap, was repeated on a large number of occasions in similar conditions. Then approximately 90% of the confidence intervals calculated in this way would contain the underlying population mean catch per trap. The 90% confidence interval actually observed, namely (3.80, 4.28), is just one observation on a random interval, and may or may not contain the population mean.

Solution to Exercise 3

(a) The transformation of interest is $\theta = \sqrt{\beta}$, both being equal to σ (this is equivalent to $\theta^2 = \beta = \sigma^2$).

The derivative of the transformation $\theta = \sqrt{\beta}$ is

$$\frac{d\theta}{d\beta} = \frac{1}{2\sqrt{\beta}},$$

which is positive for all $\beta > 0$. Hence this transformation is increasing.

(b) Since the transformation is increasing, a 95% confidence interval for $\theta = \sqrt{\beta} = \sigma$ is, using Interval (7),

$$(\theta^-, \theta^+) = \left(\sqrt{\beta^-}, \sqrt{\beta^+}\right) = \left(\sqrt{1010.2}, \sqrt{1894.7}\right) \simeq (31.8, 43.5).$$

Solution to Exercise 4

(a) The estimate of p, the probability of being cured after taking the study drug is $102/129 \simeq 0.791$. The number of successes (cures) is 102, and the number of failures is 27. Both are greater than 5, so the large-sample method can be used to calculate a 95% confidence interval for p. Using Interval (11), the lower and upper 95% confidence limits for p are

$$p^- = 0.791 - 1.96\sqrt{\frac{0.791 \times 0.209}{129}} \simeq 0.721,$$

$$p^+ = 0.791 + 1.96\sqrt{\frac{0.791 \times 0.209}{129}} \simeq 0.861.$$

So an approximate 95% confidence interval for p is

$$(p^-, p^+) = (0.721, 0.861).$$

(b) Let $n_1 = 129$ and $n_2 = 132$. The point estimate of the difference d between the cure probabilities in the two groups is

$$\widehat{d} = \widehat{p}_1 - \widehat{p}_2 = \frac{102}{129} - \frac{109}{132} \simeq 0.791 - 0.826 = -0.035.$$

Using Interval (12), the lower and upper 95% confidence limits for d are

$$d^- = -0.035 - 1.96\sqrt{\frac{0.791 \times 0.209}{129} + \frac{0.826 \times 0.174}{132}} \simeq -0.130$$

(which a fuller-accuracy calculation gives as -0.131) and

$$d^+ = -0.035 + 1.96\sqrt{\frac{0.791 \times 0.209}{129} + \frac{0.826 \times 0.174}{132}} \simeq 0.060.$$

So an approximate 95% confidence interval for d is

$$(d^-, d^+) = (-0.130, 0.060).$$

(c) If the entire trial were repeated independently a large number of times, and on each occasion a 95% confidence interval for the difference between proportions were calculated, then about 95% of these intervals would contain the population value of the difference between the proportions of women cured using study and standard drugs. In this instance, the estimated difference between cure probabilities (study drug minus standard drug) is negative, but with 95% confidence interval $(-0.13, 0.06)$. This interval of plausible values includes zero. Thus this trial suggests that there may be no difference between cure proportions of the study and standard drugs.

Solution to Exercise 5

For a 99% confidence interval, the required t-value is obtained from the t-distribution with $n-1=9-1=8$ degrees of freedom: it is the 0.995-quantile of $t(8)$, $q_{0.995}=3.355$.

The lower and upper 99% confidence limits for μ, the mean percentage of silver, are obtained using Interval (17):

$$\mu^- = 6.7444 - 3.355\sqrt{\frac{0.2953}{9}} \simeq 6.7444 - 0.6077 = 6.1367,$$

$$\mu^+ = 6.7444 + 3.355\sqrt{\frac{0.2953}{9}} \simeq 6.7444 + 0.6077 = 7.3521.$$

So, giving the result correct to two decimal places, a 99% confidence interval for μ is given by

$$(\mu^-, \mu^+) = (6.14, 7.35)\,\%.$$

Solution to Exercise 6

(a) The ratio of the larger to the smaller sample variance is $s_2^2/s_1^2 = 114.84/48.93 \simeq 2.35$. This is less than 3, so we conclude that the equal variance assumption is reasonable.

The pooled estimate of the variance is

$$\begin{aligned} s_P^2 &= \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2} \\ &= \frac{9\times 48.93 + 9\times 114.84}{18} = 81.885. \end{aligned}$$

(b) The sample difference is

$$\widehat{d} = \overline{x}_1 - \overline{x}_2 = 128.40 - 122.80 = 5.60.$$

There are $10+10-2=18$ degrees of freedom. Thus the required t-value is the 0.975-quantile of $t(18)$. From the table of t-quantiles, this is 2.101.

The lower and upper 95% confidence limits are obtained using Interval (21):

$$\begin{aligned} d^- &= 5.60 - 2.101\times\sqrt{81.885}\times\sqrt{\frac{1}{10}+\frac{1}{10}} \simeq 5.60 - 8.50 = -2.90, \\ d^+ &\simeq 5.60 + 8.50 = 14.10. \end{aligned}$$

Thus the two-sample t-interval for the difference between mean joint widths is

$$(d^-, d^+) = (-2.90, 14.10) \text{ microns}.$$

Acknowledgements

Grateful acknowledgement is made to the following sources:

Page 144: NoiseD. This file is licensed under the Creative Commons Attribution-Share Alike Licence
http://creativecommons.org/licenses/by-sa/3.0/

Page 145: Fox Photos/Getty Images

Page 146: CDC/Frank Collins Public Health Image Library

Page 152: iqoncept/www.123rf.com

Page 155: Mark R Harrington/Smithsonian National Museum of the American Indian

Page 156: © Lewis King/Dreamstime.com

Page 158: Colin Cramm/www.123rf.com

Page 159: Cathy Yeulet/www.123rf.com

Page 162: © Cathy Keifer

Page 165: Rightsholder unknown

Page 168: © U.S. National Library of Medicine

Page 170: © esuslo/Depositphotos.com

Page 171: © newlight/Depositphotos.com

Page 175: Adapted from a cartoon by John Landers. This file is licensed under the Creative Commons Attribution-Share Alike Licence
http://creativecommons.org/licenses/BY-NC-SA 4.0

Page 180 top: http://ttte.wikia.com/wiki/Lemberg This file is licensed under the Creative Commons Attribution-Share Alike Licence
http://creativecommons.org/licenses/by-sa/3.0/

Page 180 middle: © The National Horseracing Museum

Page 180 bottom: © kolaczan/Fotolia.com

Page 182: © iStockphoto.com/MizarLuca

Every effort has been made to contact copyright holders. If any have been inadvertently overlooked, the publishers will be pleased to make the necessary arrangements at the first opportunity.