9 Confidence Intervals

In Chapter 7 we discussed the problem of constructing a *point estimate* of an unknown parameter using a sample of data. If we have a "good" estimator and a big enough sample this gives us a single value that is likely to be "close" to the true value of the parameter. But how close is "close"? In order to address this question, it is useful to be able to specify a range of plausible values around our point estimate which indicate how close the estimate is likely to be to the true value and how "confident" we can be in this. Such a range of values is called a *confidence interval*. An important point to keep in mind when working with confidence intervals is that "confidence" is *not* the same as probability, as we shall see.

9.1 Introduction

Consider a situation in which we want to estimate some population parameter, say θ, the proportion of internet users who have more than one email account. We are going to collect a sample of data in order to do this. Suppose we can define two statistics, say L and U, which we will calculate from these data (in the same way we could calculate the statistics $\bar{X}$ or S^2). Suppose further that we have carefully chosen L and U in such a way that if we repeated the experiment many times (i.e. repeatedly took samples from the population and computed L and U) we know that on 95% of occasions θ would lie between L and U. In other words, before we collect our data we know that

$$\Pr(\theta \text{ lies between } L \text{ and } U) = \Pr(L < \theta < U) = 0.95.$$

Then, once we observe our data and calculate observed values $L = \ell$ and $U = u$ for the statistics, we can say we are 95% confident that

$$\ell < \theta < u.$$

This interval is called a 95% confidence interval for θ. We can define other confidence intervals, e.g. 90% or 99%, using the same ideas. Note that a 90% confidence interval would be narrower than a 95% confidence interval but a 99% confidence interval would be wider. Can you think why?

It is important to realise that once we have used our sample data to calculate a 95% confidence interval $\ell < \theta < u$, it would *not* be correct to say that θ lies between ℓ and u with probability 0.95. The value of θ is assumed to be fixed (albeit unknown) and therefore cannot have probabilities associated with it; θ either lies between the observed limits ℓ and u or it doesn't.

9.2 Confidence interval for the mean of a normal distribution when the variance is known

Suppose $X_1, X_2, \ldots, X_n$ are IID $N(\mu, \sigma^2)$ random variables. We want to construct a 95% confidence interval for μ assuming that σ^2 is known. (Of course it is very unlikely that we would know σ^2 if we didn't know μ but we will save the more complicated case where σ^2 is unknown until later and use this simpler set-up to illustrate the general ideas).

From Section 6.7.1 we know that the sample mean $\bar{X}$ is also normal with

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

Translating and scaling then gives

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1).$$

From Section 6.6.3 we know that $\Pr(-1.96 < Z < 1.96) = 0.95$ if $Z \sim N(0,1)$ and so

$$\Pr\left(-1.96 < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < 1.96\right) = 0.95.$$

Once the data have been observed, if the observed value of $\bar{X}$ is $\bar{x}$, then we can say we are 95% confident that

$$-1.96 < \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} < 1.96.$$

Rearranging this inequality gives a 95% confidence interval for μ of

$$\bar{x} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}.$$

Other confidence intervals, e.g. 99%, can be computed in exactly the same way, replacing 1.96 with the appropriate quantile from the standard normal distribution. In general a $100(1-\alpha)\%$ confidence interval for μ is given by

$$\bar{x} - z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$$

where in general z_β denotes the standard normal quantile such that $\Pr(Z < z_\beta) = \beta$, with $Z \sim N(0,1)$. For example, taking $\alpha = 0.05$ gives us the 95% confidence interval above with $z_{0.975} = 1.96$, whilst taking $\alpha = 0.01$ gives us a 99% confidence interval with $z_{0.995} = 2.58$. These quantiles can be calculated in R using the `qnorm` function, in these examples, `qnorm(0.975)` and `qnorm(0.995)`.

The confidence intervals presented above are two-sided but it is also possible to produce one sided confidence intervals. For example, $z_{0.95} = 1.64$ and so

$$\Pr\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < 1.64\right) = 0.95,$$

which leads to the one-sided 95% confidence interval

$$\mu > \bar{x} - 1.64\frac{\sigma}{\sqrt{n}}.$$

9.2.1 Example

In robotic technology, accuracy is of vital importance. For a particular robot used to apply adhesive to a specified location it is known that the errors in the placement of the adhesive are approximately normally distributed. The following data (in micrometres) are a random sample of 9 errors:

$$1.5 \quad -1.1 \quad 1.9 \quad 0.6 \quad 2.4 \quad 0.5 \quad 0.5 \quad 2.0 \quad -2.2$$

Obtain (a) 95% and (b) 90% confidence intervals for the population mean error μ assuming the population variance is known to be 4 square-micrometres.

Solution.

(a) We have $\sigma = \sqrt{4} = 2$ and calculate $n = 9$ and $\bar{x} = 0.6778$. Our 95% confidence interval for μ is therefore

$$0.6778 - 1.96\frac{2}{\sqrt{9}} < \mu < 0.6778 + 1.96\frac{2}{\sqrt{9}},$$

which simplifies to

$$-0.6289 < \mu < 1.9844.$$

Note: this confidence interval includes zero which suggests a mean error of 0 is compatible with these data. Given the need for accuracy in robotic technology, this is reassuring.

(b) 

9.3 Confidence interval for the mean of a normal distribution when the variance is unknown

When $X_1, X_2, \ldots, X_n$ are IID normal with mean μ and known variance σ^2, our derivation of confidence intervals for μ relied on the fact that we know

$$\frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}} \sim N(0, 1).$$

When the population variance σ^2 is unknown we replace σ^2/n (called the *standard error* of $\bar{X}$) with an estimator, namely S^2/n where S^2 is the sample variance. However, the distribution of $(\bar{X}-\mu)/\sqrt{S^2/n}$ is not normal. Although the underpinning theory is beyond the scope of this course, it can be shown that

$$\frac{\bar{X}-\mu}{\sqrt{S^2/n}} \sim t_{n-1},$$

where t_{n-1} denotes the Student's t-distribution on $n-1$ degrees of freedom. Figure 1 shows a plot of the pdf for the Student's t-distribution with various degrees of freedom ν, along with a plot of the standard normal $N(0,1)$ pdf. As you can see the Student's t-distribution is symmetrical and bell-shaped like the normal distribution but it has heavier tails, especially for small degrees of freedom ν. However, as the degree of freedom parameter ν gets larger, the t_ν distribution tends towards the $N(0,1)$ distribution. Intuitively, the heavier tails of the Student's t-distribution allow us to accomodate additional uncertainty in our confidence intervals now that σ^2 is not known.

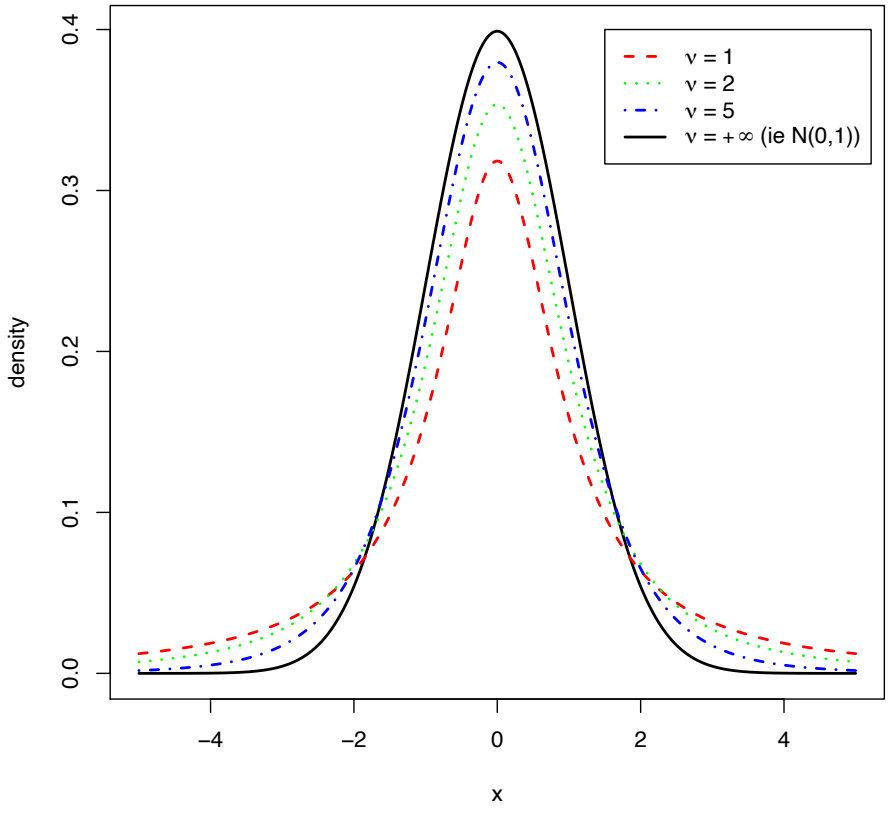

Figure 1: Probability density function for the t_ν distribution with various degrees of freedom ν. $\nu = +\infty$ gives the standard normal $N(0,1)$ pdf.

Using an analogous argument to that presented earlier in the case where σ^2 was known, once we have observed a sample with mean $\bar{x}$ and standard deviation s, a $100(1-\alpha)\%$ confidence

interval for the population mean μ is given by

$$\bar{x} - t_{n-1,1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{n-1,1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}},$$

where in general $t_{\nu,\beta}$ denotes the quantile of the t_ν distribution such that $\Pr(T < t_{\nu,\beta}) = \beta$, with $T \sim t_\nu$. For example, if $n = 10$ and $\alpha = 0.05$, $t_{n-1,1-\frac{\alpha}{2}} = t_{9,0.975} = 2.26$ which we can compute in R by typing `qt(0.975,9)`.

9.3.1 Example

Repeat Example 9.2.1, except this time assume the population variance is not known.

Solution.

(a) As before we calculate $n = 9$ and $\bar{x} = 0.6778$. Now we must also calculate $s = 1.5164$. To compute a 95% confidence interval we take $\alpha = 0.05$ and need to find $t_{9-1,1-0.05/2} = t_{8,0.975}$. Typing `qt(0.975,8)` in R gives $t_{8,0.975} = 2.31$ and so we have

$$0.6778 - 2.31\frac{1.5164}{\sqrt{9}} < \mu < 0.6778 + 2.31\frac{1.5164}{\sqrt{9}},$$

which simplifies to

$$-0.4898 < \mu < 1.8454.$$

(b)