Chapter 5

Confidence Intervals

Contents

5.1 Introduction

In the last chapter we learnt how to obtain *point* estimates population means, variances and proportions. The problem with point estimates is that they will almost certainly be *wrong*! Sample means will almost always differ from population means, for example.

Whenever you are given a point estimate, you should ask for some indication of how accurate the estimate is likely to be. If no-one can tell you, don't believe the estimate! An alternative to a point estimate is to provide an *interval* estimate: a range of values for the population parameter. Informally, we would say that the interval estimate is "correct" if it contains the true value of the population parameter. We will see in this chapter how we can use probability theory to construct interval estimates that have a high probability of being correct.

Informative interval estimates

We could always give a trivial interval estimate that was guaranteed to be correct. For example, if conducting an opinion poll for a UK general election, we could say, "The percentage of the Labour vote will be somewhere between 0% and 100%." Clearly, this gives no useful information. The narrower any interval, the more 'useful' it will be, but there will be a trade-off: narrower intervals are more likely to be 'wrong'.

5.2 $100(1-\alpha)\%$ Confidence Intervals

In the confidence interval method, we specify a probability $1-\alpha$ with which we would like the interval to contain the true value, *before we observe the data*. The resulting interval is known as a $100(1-\alpha)\%$ confidence interval. This probability of $1-\alpha$ will be large, but not too large: if we choose $\alpha = 0$, we'll see that the resulting 100% confidence interval will be uselessly wide. A common choice is $\alpha = 0.05$, resulting in a 95% confidence interval.

5.3 Confidence intervals for a population mean, with population variance known

Suppose we are going to obtain a random sample represented by $X_1, \ldots, X_n$, with $E(X_i) = \mu$ and $Var(X_i) = \sigma^2$. To simplify things, we'll start with the (unrealistic) assumption that σ^2 is known, and then relax this assumption later. The aim is to provide an interval estimate: a range of plausible values for μ, given $X_1, \ldots, X_n$.

1. **The general form of the confidence interval**
   We are going to construct an interval of the form

$$[\bar{X} - k, \bar{X} + k] \tag{5.1}$$

   Since $\bar{X}$ is a random variable, the endpoints of the interval are also random variables. Hence, *before we observe the data*, we think of a confidence interval as a *random* interval.

2. **Specifying the confidence level**
   We now specify the confidence level: the probability $1-\alpha$ with which we want the interval

to contain μ. Having specified $1-\alpha$, our objective is to determine k such that

$$P(\bar{X} - k \leq \mu \leq \bar{X} + k) = 1 - \alpha. \tag{5.2}$$

Note that the value of k is going to depend on $1-\alpha$: as $1-\alpha$ increases, k will increase (so we can't choose α to be *too* small, otherwise, we will get an unhelpfully wide interval). We can determine what k needs to be by considering the probability distribution of $\bar{X}$.

3. **The probability distribution of $\bar{X}$**
Recall that $\mathbb{E}(\bar{X}) = \mu$, and $Var(\bar{X}) = \sigma^2$. We assume that

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right). \tag{5.3}$$

Recall that this assumption is exactly true if the population distribution is normal, i.e. $X_i \sim N(\mu, \sigma^2)$. Otherwise, it is approximately true by the Central Limit Theorem for sufficiently large n.

We are nearly ready to determine the value of k such that equation (5.2) will hold. The remaining difficulty is that the distribution of $\bar{X}$ depends on μ, but the value of μ is unknown (its the very thing that we're trying to estimate). We deal with this problem by standardising $\bar{X}$.

4. **Standardising $\bar{X}$ to derive k**
We write

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1) \tag{5.4}$$

Now

$$P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = 1 - \alpha, \tag{5.5}$$

where $z_{\alpha/2}$ is the $(1-\alpha/2)$ quantile of the standard normal distribution. For example, if $1-\alpha = 0.95$, then $z_{\alpha/2} = 1.96$. Now,

$$P\left(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}\right) = 1 - \alpha, \tag{5.6}$$

and so

$$P\left(-z_{\alpha/2} \leq \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq z_{\alpha/2}\right) = 1 - \alpha, \tag{5.7}$$

for *any* value of μ. Rearranging the inequality we conclude that

$$P\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha. \tag{5.8}$$

Hence the required k in 5.2 is $z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$.

The confidence interval formula

Once we obtain the data and calculate the observed value of sample mean $\bar{x}$, we report the $100(1-\alpha)\%$ confidence interval to be

$$\left[\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]. \tag{5.9}$$

An example to help visualise the confidence interval method

Suppose we have $\sigma^2 = 225$ and $n = 25$. Before we obtain the data, we can say that the 95% confidence interval will be of the form

$$\left[\bar{X} - 1.96 \times \sqrt{\frac{225}{25}}, \bar{X} + 1.96 \times \sqrt{\frac{225}{25}}\right], \tag{5.10}$$

i.e $\bar{X} \pm 5.88$. Hence the random interval will contain μ as long as the random $\bar{X}$ is no more than a distance of 5.88 from μ. We have $\bar{X} \sim N(\mu, 9)$, and so this event will happen with probability 0.95, regardless of the value of μ, exactly as required. This is shown in Figure 5.1.

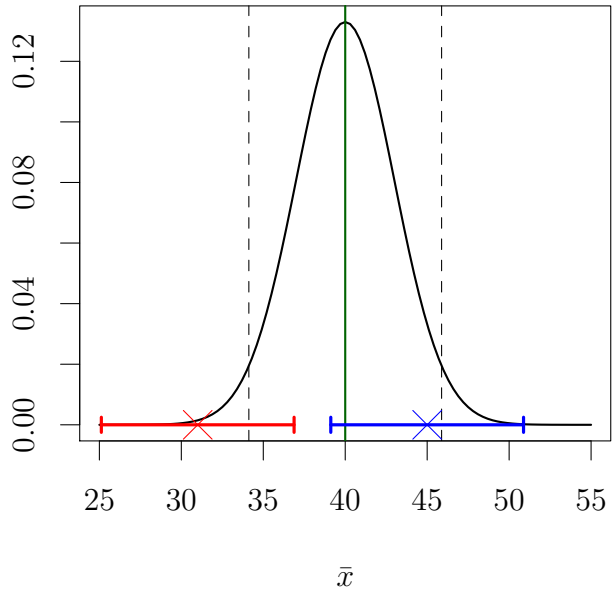

Figure 5.1: In this example, we suppose the unknown value of μ is 40, so that $\bar{X} \sim N(40, 9)$ (shown by the solid black line). The 95% confidence interval $\bar{X} \pm 5.88$ will contain μ as long as $\bar{X}$ falls no more than a distance of 5.88 from μ, with this distance indicated by the dashed lines. The red bar shows a CI that has 'missed': the observed $\bar{x}$ in this case was 31: too far from μ for the interval to include μ; the blue bar shows a CI that has 'hit': the observed $\bar{x}$ in this case was 45: close enough to μ for the interval to include μ. The probability that $\bar{X}$ will fall close enough to μ is exactly 0.95.

Exercise 5.1. *In the PISA example, suppose test scores are normally distributed with unknown population mean μ and known population variance 100^2. A sample of 1000 students is selected, and the observed sample mean is $\bar{x} = 498$. Calculate 95% and 99% confidence intervals for the population mean, using the following R output to help you.*

```
qnorm(c(0.95, 0.975, 0.99, 0.995))

## [1] 1.644854 1.959964 2.326348 2.575829
```

(Not all the R output is relevant: this is intentional.)

Comments

1. As α decreases, $z_{\alpha/2}$ increases, and so the confidence interval gets wider: the penalty for increasing the probability that the interval will contain μ is to have a wider, and hence less informative interval estimate.

2. For fixed α and σ, the width of the interval is proportional to $1/\sqrt{n}$: doubling a sample size from n to $2n$ will only shrink the width of a confidence interval by a factor of $\sqrt{2}$. To halve the width, we'd need to *quadruple* the sample size.

3. $\frac{\sigma}{\sqrt{n}}$ is the standard error of $\bar{X}$. Sometimes, instead of reporting a confidence interval, authors choose to present the estimate of a population parameter and the associated standard error. In the above exercise, you could have instead been asked to construct the confidence interval given a sample mean $\bar{x} = 498$ and a standard error of 3.16.

5.4 Interpretation of confidence intervals

Confidence intervals are notoriously easy to misinterpret! Without loss of generality, suppose we have chosen $\alpha = 0.05$. To interpret a 95% confidence interval, note the following.

1. If you do a large number of 'experiments' over your career, (approximately) 95% of the confidence intervals you produce will be 'correct' and contain the true value.

2. In a single experiment, *before* you collect your data, there is a 95% probability the 95% confidence interval *you are going to get* will contain the true value.

3. In a single experiment, *after* you have collected your data, and obtained, for example, a 95% confidence interval of $[77.2, 84.1]$ it is **wrong** to say, "The probability that μ lies in the interval [77.2, 84.1] is 0.95."

You may find it quite hard to reconcile statements (1) and (2) with statement (3). Technically, μ is constant, not a random variable, and so the probability of μ lying inside any *specified* interval can only be 1 (if it does), or 0 (if it doesn't). In any case, the following experiment may help you.

5.4.1 A simulation experiment to help interpret confidence intervals

We will conduct the following simulation experiment in R.

1. I will choose a value for the population mean μ, but will not tell you what it is.

2. 100 separate samples of size $n = 50$ will be generated from the $N(\mu, \sigma^2 = 25)$ distribution.

3. For each sample, we will calculate a 95% confidence interval for μ (so there will be 100 separate confidence intervals in total).

Before doing the experiment, we would expect about 95 of the 100 confidence intervals to contain μ, and any single sample would have the same probability of producing a confidence interval that contains μ. Now let's run the experiment and get the intervals. These are shown in Figure 5.2.

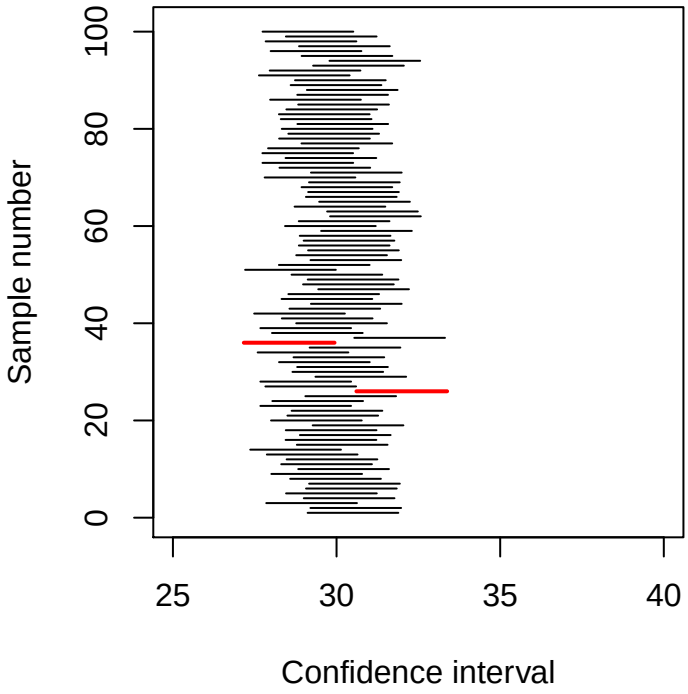

Figure 5.2: One hundred 95% confidence intervals for a population mean μ. Note the two red non-overlapping intervals: we couldn't say *both observed intervals* have a 95% chance of containing μ.

For example, in Sample number 1, the mean of the 100 observations was $\bar{x} = 30.5$, and so the calculation

$$\bar{x} \pm 1.96 \times \frac{5}{\sqrt{50}}$$

gives the 95% confidence interval [29.1, 31.9], shown by the short horizontal line at the bottom of Figure 5.2.

Sample number 26 produced an interval of $[30.6, 33.4]$, and sample number 36 produced an interval of $[27.2, 29.9]$. These two intervals are highlighted in red. If we tried to claim that

$$P(27.2 \leq \mu \leq 29.9) = 0.95, \tag{5.11}$$
$$P(30.6 \leq \mu \leq 33.4) = 0.95, \tag{5.12}$$

then, because the 'events' $\{27.2 \leq \mu \leq 29.9\}$ and $\{30.6 \leq \mu \leq 33.4\}$ are disjoint, we would have

$$\begin{aligned} P(\{27.2 \leq \mu \leq 29.9\} \cup \{30.6 \leq \mu \leq 33.4\}) &= P(27.2 \leq \mu \leq 29.9) + P(30.6 \leq \mu \leq 33.4) \\ &= 0.95 + 0.95 = 1.9, \end{aligned} \tag{5.13}$$

which is clearly nonsense, as probabilities can't be greater than 1.

To recap and summarise:

- The 95% probability refers to the probability *before* you get your data that your 95% confidence interval will contain the true value.

- *After* you get your data and calculate your interval, you cannot actually put any probability on it containing the true value[1]: it simply 'does or does not'.

- If you behave as if your 95% confidence interval *does* contain μ, you'll be 'doing the right thing' on 95% of occasions: you just won't know on *which* occasions you're 'doing the wrong thing'.

5.5 Some distribution theory: the χ^2 and Student t distributions

Before continuing with the study of confidence intervals, we introduce two more distributions that we will need later on.

5.5.1 The χ^2 distribution

If a random variable Y has the χ^2_ν distribution (the "chi-squared distribution with ν degrees of freedom"), we write $Y \sim \chi^2_\nu$ and the probability density function of Y is given by

$$f_Y(y) = \begin{cases} \frac{1}{2^{\nu/2}\Gamma(\nu/2)} y^{\nu/2-1} \exp\left(-\frac{y}{2}\right), & y \geq 0, \\ 0 & y < 0. \end{cases} \tag{5.14}$$

Note that we must have $\nu > 0$. Here Γ denotes the gamma function, defined by

$$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt. \tag{5.15}$$

It can be shown that

$$\mathbb{E}(Y) = \nu, \tag{5.16}$$
$$Var(Y) = 2\nu. \tag{5.17}$$

The χ^2_ν distribution is positively skewed, with the skew more apparent as the degrees of freedom ν decreases. Three χ^2 distributions are plotted in Figure 5.3

[1] Actually, you can, but only using *Bayesian Statistics*, which we teach in MAS364.

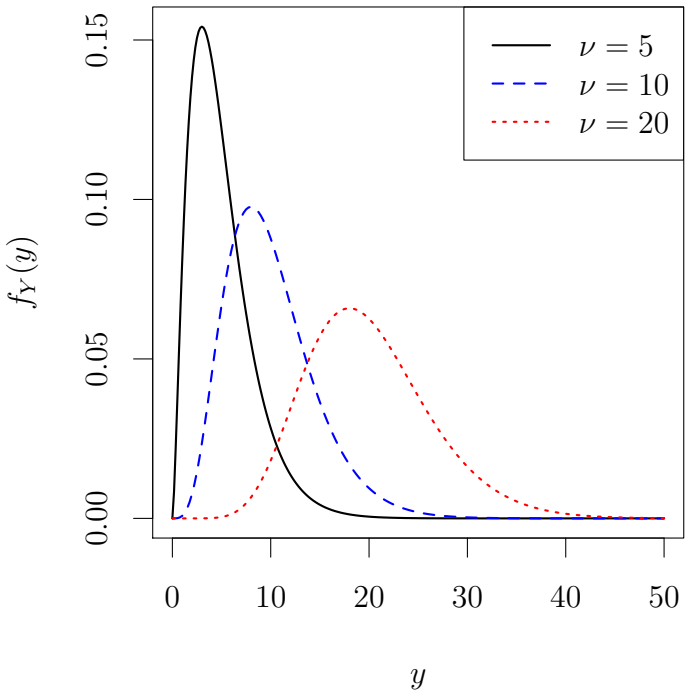

Figure 5.3: Three χ^2_ν distributions with $\nu = 5$, 10 and 20 degrees of freedom.

Notation: quantiles/percentiles of the χ^2 distribution

If Y has a chi-square distribution with ν degrees of freedom, we define the values $\chi^2_{\nu,\alpha}$ by

$$P(Y \leq \chi^2_{\nu,\alpha}) = 1 - \alpha, \tag{5.18}$$

so $\chi^2_{\nu,\alpha}$ is the $(1-\alpha)$ quantile or $100(1-\alpha)$ percentile of the χ^2_ν distribution. These values can be obtained in R: see Section 5.5.3. We illustrate this notation in Figure 5.4.

The relationship between the normal distribution and the χ^2 distribution

Theorem 5.1. *If $Z_1, \ldots, Z_n$ are n independent random variables each with the standard normal distribution then $\sum_{j=1}^n Z_j^2 \sim \chi^2_n$.*

Exercise 5.2. *Using the result above, show that if If $X_1, \ldots, X_n$ are n independent random variables each with the same normal distribution $N(\mu, \sigma^2)$ then*

$$\frac{\sum_{j=1}^n (X_j - \mu)^2}{\sigma^2} \sim \chi^2_n.$$

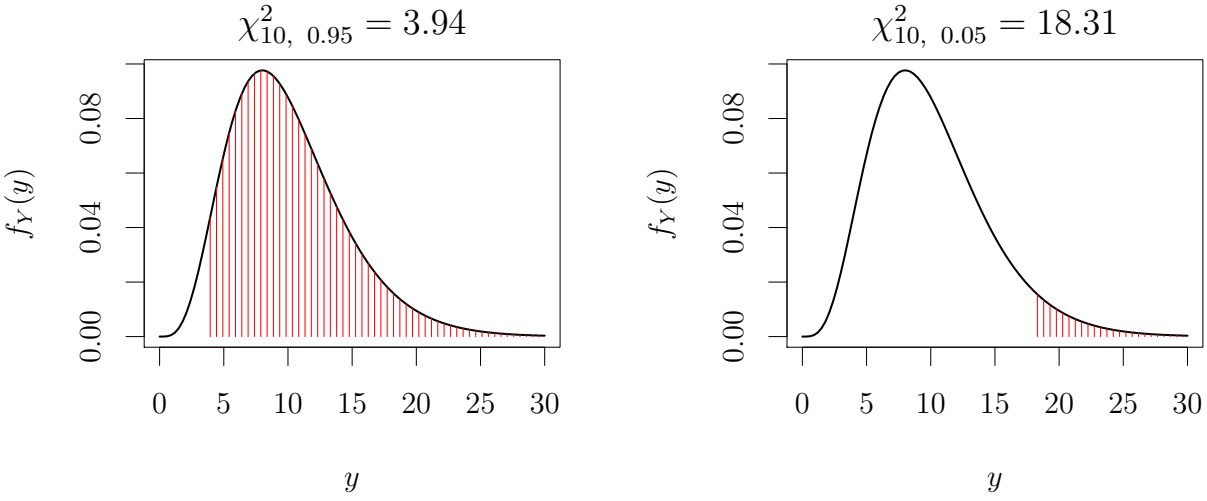

Figure 5.4: The 5th and 95th percentiles of the χ^2_{10} distribution, which are 3.94 and 18.31 respectively. Note the convention for the term α in $\chi^2_{\nu,\alpha}$ to refer the probability to the right (the shaded area in each plot), so that the 5th percentile is denoted by $\chi^2_{10,0.95}$.

5.5.2 The Student t distribution

If a random variable Y has a Student t distribution (or "Student's t distribution or just "t distribution" for short) with ν degrees of freedom, that is if

$$Y \sim t_\nu,$$

then Y has the density function

$$f_\nu(y) = \frac{1}{\sqrt{\pi\nu}} \frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)} (1 + y^2/\nu)^{-(\nu+1)/2}, \quad -\infty < y < \infty$$

If $\nu > 1$ then

$$\mathbb{E}(Y) = 0, \tag{5.19}$$

and if $\nu > 2$ then

$$Var(Y) = \frac{\nu}{\nu - 2} \tag{5.20}$$

The pdf f_ν is symmetric about zero. For large values of ν, it is very similar to the standard normal density $N(0, 1)$. Some t-distributions are plotted in Figure 5.5

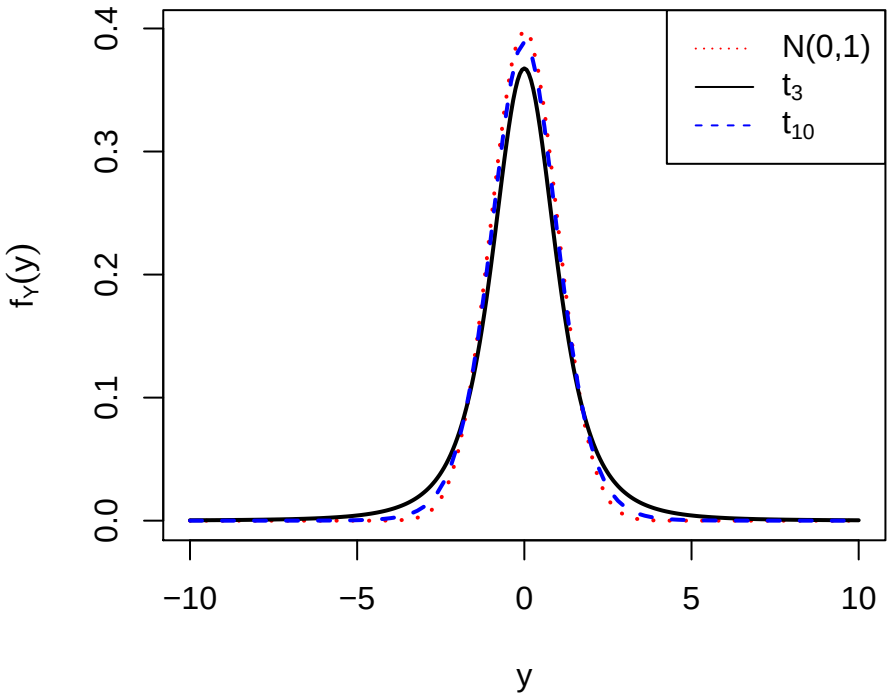

Figure 5.5: The t_3 and t_{10} distributions, together with the standard normal distribution. A t distribution with more than 30 degrees of freedom is hard to distinguish from a standard normal distribution. Note that the t distributions have extitheavier tails than the normal.

Notation: quantiles/percentiles of the t distribution

If T has a t distribution with ν degrees of freedom, we define the values $t_{\nu,\alpha}$ by

$$P(T \leq t_{\nu,\alpha}) = 1 - \alpha, \tag{5.21}$$

so $t_{\nu,\alpha}$ is the $(1-\alpha)$ quantile or $100(1-\alpha)$ percentile of the t_ν distribution.

The relationship between the normal distribution, the χ^2 distribution and the t distribution

All three distributions are related as follows.

Theorem 5.2. *If $Z \sim N(0,1)$ and $Y \sim \chi^2_\nu$, then*

$$T = \frac{Z}{\sqrt{Y/\nu}} \sim t_\nu, \tag{5.22}$$

so the ratio of a standard normal variable to the square root of a χ^2 variable has a t distribution.

We do not prove this result in this module, but a proof is given in MAS223.

5.5.3 The χ^2 and t distributions in R.

Cumulative probabilities and quantiles can be calculated in R in a similar way to using the `pnorm` and `qnorm` functions, but note that the value of the degrees of freedom parameter needs to be specified. For example,

```
pchisq(4.8, 7)

## [1] 0.3156451

qchisq(0.95, 10)

## [1] 18.30704

pt(-1, 3)

## [1] 0.1955011

qt(0.975, 10)

## [1] 2.228139
```

Hence (to three decimal places):

- for $Y \sim \chi^2_7$, we have $P(Y \leq 4.8) = 0.316$;
- for $Y \sim \chi^2_{10}$, we have $P(Y \leq 18.307) = 0.95$, and we write $\chi^2_{10,0.05} = 18.307$;
- for $T \sim t_3$, we have $P(T \leq -1) = 0.196$;
- for $T \sim t_{10}$, we have $P(T \leq 2.228) = 0.975$ and we write $t_{10,0.025} = 2.228$.

5.6 Confidence intervals for population variances

In the previous chapter, we saw that an unbiased estimator for a population variance σ^2 is

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2. \tag{5.23}$$

In the case where the population distribution is normal $N(\mu, \sigma^2)$, we can also construct a confidence interval for σ^2 and report an interval estimate. Recall that for a normal distribution, approximately 95% of the population will lie within 2σ of the population mean μ, so σ is a little more easy to interpret in this case, and a confidence interval will help describe our uncertainty about how much the population might vary.

In the previous section, to derive a $100(1-\alpha)\%$ confidence interval for the population μ, we considered an interval of the form

$$[\bar{X} - k, \bar{X} + k],$$

and derived k such that

$$P(\bar{X} - k \leq \mu \leq \bar{X} + k) = 1 - \alpha. \tag{5.24}$$

We'll do something similar here, but we'll instead consider an interval of the form

$$[aS^2, bS^2],$$

with $0 < a < 1$ and $b > 1$. One reason for doing this is that a variance can't be negative, and this will ensure the lower limit of the interval is non-negative.

Again, because S^2 is a random variable, the endpoints of the intervals are random variables, and so we will derive a and b such that

$$P(aS^2 \leq \sigma^2 \leq bS^2) = 1 - \alpha. \tag{5.25}$$

The next step is to consider the probability distribution of S^2.

5.6.1 The distribution of S^2

The following theorem gives us the distribution of S^2:

Theorem 5.3. *If $X_1, \ldots, X_n$ are n independent random variables each with the same normal distribution $N(\mu, \sigma^2)$ then*

$$\frac{(n-1)S^2}{\sigma^2} = \frac{\sum_{j=1}^{n}(X_j - \bar{X})^2}{\sigma^2} \sim \chi^2_{n-1}. \tag{5.26}$$

In other words, S^2 is a χ^2_{n-1} random variable, multiplied by $\sigma^2/(n-1)$. The proof of this result is outside the scope of this module, but note the similarity with the result in Exercise 5.2. The "$n-1$" relates to the fact that the terms in the sum cannot be regarded as n distinct sources of information; there is one constraint on them, namely that $\Sigma X_j = n\bar{X}$.

5.6.2 Deriving the confidence interval

Recall we are going to provide an interval of the form

$$[aS^2, bS^2],$$

so that our objective, is to find a and b for any choice of α such that

$$P(aS^2 \leq \sigma^2 \leq bS^2) = 1 - \alpha. \tag{5.27}$$

We have established

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}, \tag{5.28}$$

and so

$$P\left(\chi^2_{n-1,1-\alpha/2} \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi^2_{n-1,\alpha/2}\right) = 1 - \alpha. \tag{5.29}$$

Rearranging the inequalities, we get

$$P\left(\frac{(n-1)S^2}{\chi^2_{n-1,\alpha/2}} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha/2}}\right) = 1 - \alpha, \tag{5.30}$$

hence we have identified

$$a = \frac{(n-1)}{\chi^2_{n-1,\alpha/2}}, \qquad b = \frac{(n-1)}{\chi^2_{n-1,1-\alpha/2}}. \tag{5.31}$$

Once we have obtained and calculated the sample variance

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2, \tag{5.32}$$

we report the $100(1-\alpha)$ confidence interval as

$$\left[\frac{(n-1)s^2}{\chi^2_{n-1,\alpha/2}}, \frac{(n-1)s^2}{\chi^2_{n-1,1-\alpha/2}}\right]. \tag{5.33}$$

Note that if we have

$$P(aS^2 \leq \sigma^2 \leq bS^2) = 1 - \alpha. \tag{5.34}$$

then we also have

$$P(\sqrt{a}S \leq \sigma \leq \sqrt{b}S) = 1 - \alpha. \tag{5.35}$$

So confidence intervals for σ can be calculated in the obvious way from corresponding intervals for σ^2.

5.7 Confidence intervals for a population mean, with population variance unknown

In many situations, it is unlikely that we would know σ^2 but not μ, so we need to consider how to get a confidence interval for μ when σ^2 is unknown. Throughout this section we will assume that the population is normally distributed, so our sample of random variables $X_1, X_2, \ldots, X_n$ all have the $N(\mu, \sigma^2)$ distribution.

Previously, the confidence interval took the form

$$\left[\bar{X} - c_1\frac{\sigma}{\sqrt{n}}, \bar{X} + c_1\frac{\sigma}{\sqrt{n}}\right], \tag{5.36}$$

for some suitable constant c_1 (in fact, we have $c_1 = z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$ for a $100(1-\alpha)\%$ confidence interval). If σ^2 is unknown, perhaps we can replace it with its estimator S^2, so that the interval is of then of the form

$$\left[\bar{X} - c_2\frac{S}{\sqrt{n}}, \bar{X} + c_2\frac{S}{\sqrt{n}}\right], \tag{5.37}$$

for some other constant c_2? It turns out that we *can* do this, but we'll need a little more theory. The endpoints of the above interval are again random variables, but now we have to deal with the fact that *both* $\bar{X}$ and S^2 are random variables.

Confidence intervals for the population mean using the t distribution

We will consider the following function of $\bar{X}$ and S^2:

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}. \tag{5.38}$$

We now have the following result:

Theorem 5.4. *If $X_1, \ldots, X_n$ are independent $N(\mu, \sigma^2)$ random variables, then T has the Student t distribution with $n-1$ degrees of freedom:*

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}.$$

Note that the distribution of T depends on n, but not on μ or σ. This gives us a way to construct the confidence interval. We have,

$$P(-t_{n-1;\alpha/2} \leq T \leq t_{n-1;\alpha/2}) = 1 - \alpha. \tag{5.39}$$

Now substitute for T from (5.38) and rearrange to obtain

$$P\left(\bar{X} - \frac{S}{\sqrt{n}} t_{n-1;\alpha/2} \leq \mu \leq \bar{X} + \frac{S}{\sqrt{n}} t_{n-1;\alpha/2}\right) = 1 - \alpha. \tag{5.40}$$

Replacing the random variables $\bar{X}$ and S by their observed values, $\bar{x}$ and s, we find that a $100(1-\alpha)\%$ confidence interval for μ is given by:

$$\left[\bar{x} - \frac{s}{\sqrt{n}} t_{n-1;\alpha/2}, \quad \bar{x} + \frac{s}{\sqrt{n}} t_{n-1;\alpha/2}.\right] \tag{5.41}$$

Simulation experiment

Unlike the known population variance case, the width of the confidence interval cannot be known in advance: it will depend on the estimated population variance s^2. We illustrate this with a simulation experiment.

1. We suppose the true population distribution is $N(30, 5^2)$ (but that both mean and variance are unknown to us).

2. Twenty separate samples of size 10 are obtained from this distribution.

3. For each sample, a 95% confidence interval is calculated:

$$\left[\bar{x} - \frac{s}{\sqrt{10}} t_{9;0.025}, \quad \bar{x} + \frac{s}{\sqrt{10}} t_{9;0.025}.\right] \tag{5.42}$$

4. The twenty separate confidence intervals are plotted in Figure 5.6.

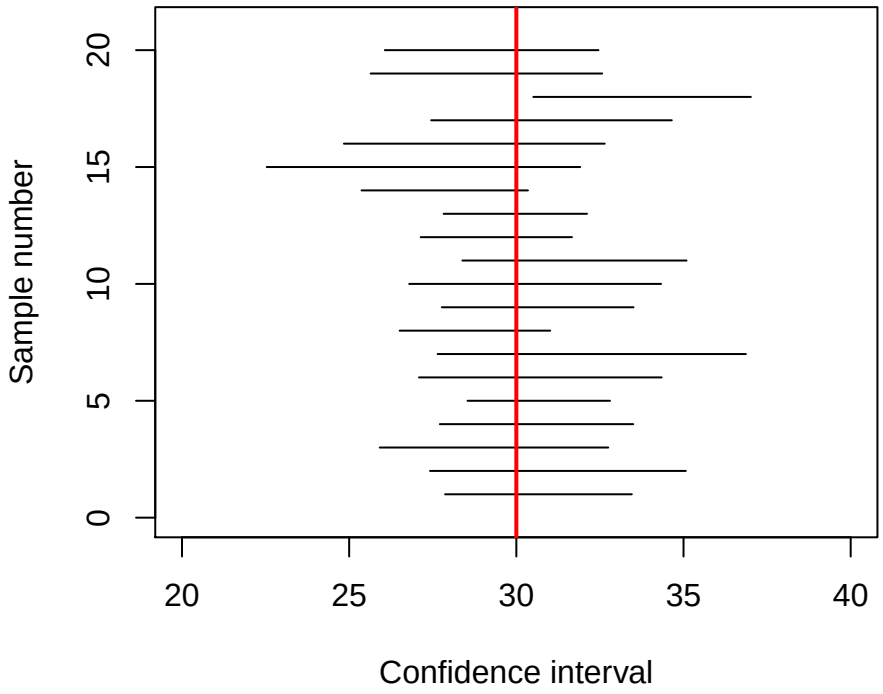

Figure 5.6: 95 % confidence intervals from twenty separate samples of data. Note how some intervals are wider than others, resulting from larger estimates of the population variance. Note also that one out of the twenty intervals has failed to contain the true population mean (30), as expected.

Exercise 5.3.

Patients with high level of low-density lipoprotein (LDL) cholesterol in their blood are at risk of developing cardiovascular disease, potentially resulting in heart attacks and/or strokes. Drugs known as "Statins" can be used to lower LDL cholesterol levels[2]*.*

Suppose a new cholesterol lowering drug has been developed. It is first tested in a small study: 10 patients are given the drug (with a daily dose over twelve weeks), and 10 patients are given a placebo (with patients unaware of whether they are receiving the active drug or the placebo or not). The outcome measure for each patient is the percentage reduction in LDL cholesterol, from the beginning to the end of the twelve week period. The data are given below.

```
drug
```

```
##  [1] -25.7 -11.6   6.6 -28.7 -15.0 -12.3  -4.8 -17.1  11.8 -15.8
```

```
sum(drug)
```

```
## [1] -112.6
```

```
sum(drug^2)
```

```
## [1] 2742.92
```

```
placebo
```

```
##  [1]   5.2  13.1  -6.1 -15.2  24.4 -33.0  11.7  -0.1  13.6   5.5
```

```
sum(placebo)
```

```
## [1] 19.1
```

```
sum(placebo^2)
```

```
## [1] 2503.37
```

Calculate

- *95% confidence intervals for the population mean percentage reductions, for both the drug and the placebo.*
- *95% confidence intervals for the population standard deviation percentage reductions, for both the drug and the placebo.*

Use (some of) the following R output to help you.

[2] see the NHS advice at http://www.nhs.uk/Conditions/Cholesterol-lowering-medicines-statins/Pages/Introduction.aspx, and for an extensive statistical analysis of the effectiveness of statins, see http://www.cochrane.org/CD008226/HTN_effect-atorvastatin-cholesterol

```r
qt(c(0.95, 0.975), 9)
```

```
## [1] 1.833113 2.262157
```

```r
qchisq(c(0.025, 0.975), 9)
```

```
## [1]  2.700389 19.022768
```

What do the confidence intervals suggest?

5.8 Confidence intervals for a population proportion

We may want to estimate the proportion of individuals in a population that have a certain property, for example

- the proportion of all graduates who are currently employed;
- the proportion of mobile phones (of a particular make and model) that will develop a fault in their first year of use;
- the proportion of UK voters who will vote for a particular party at the next general election.

Representing the population with binary variables

The main idea is to represent each randomly sampled observation as a binary variable (a variable that can only take the values 1 or 0); we can then use previous results from this chapter. For example, if we are interested in the proportion θ of all graduates who are currently employed, and we are going to randomly sample n graduates and ask them their employment status, then before we take the sample, define the random variable X_i as follows.

$$X_i = \begin{cases} 1 & \text{if the } i\text{-th graduate that we are going to sample is employed,} \\ 0 & \text{otherwise.} \end{cases} \tag{5.43}$$

The population proportion is the population mean

We have $X_i \sim Bernoulli(\theta)$ and the population mean μ is given by

$$\mu := \mathbb{E}(X_i) = 1 \times \theta + 0 \times (1 - \theta) = \theta. \tag{5.44}$$

Hence estimating a population proportion is equivalent to estimating a population mean, where each member of the population takes a value 1 or 0. Note that the population variance is given by

$$\sigma^2 := Var(X_i) = \mathbb{E}(X_i^2) - \mathbb{E}(X_i)^2 = \theta(1 - \theta). \tag{5.45}$$

so, in effect, there is only a single uncertain population parameter here: θ; if we know θ, we know both the population mean and population variance.

Estimating the population proportion with the sample proportion

Once we've observed the data, define p to be the proportion of sample members we observed to have the property of interest (e.g., the proportion of sampled graduates we observed to be employed). Just as the population proportion and the population mean (using binary variables) are the same thing, so are the sample proportion and sample mean. We define the constant x_i as follows.

$$x_i = \begin{cases} 1 & \text{if the } i\text{-th graduate that we did sample was employed,} \\ 0 & \text{otherwise.} \end{cases} \tag{5.46}$$

hence we have

$$p = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x} \tag{5.47}$$

As we usually estimate a population mean μ with a sample mean $\bar{x}$, here we can say that we estimate a population proportion θ with a sample proportion p.

Estimating the population variance

In this special case of binary variables, we have $\sigma^2 = \theta(1-\theta)$. Given that we have estimated θ using p, we can estimate σ^2 using $p(1-p)$.

Constructing the confidence interval

To get the confidence interval, we start with the usual confidence interval for a population mean μ, with known variance σ^2:

$$\left[\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]. \tag{5.48}$$

We now write the sample proportion p instead of $\bar{x}$, and substitute the estimated population variance $p(1-p)$ for σ^2. We then have a $(1-\alpha)100\%$ confidence interval of

$$\left[p - z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}, p + z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}\right]. \tag{5.49}$$

Notes:

1. This is only an *approximate* interval: each X_i has a Bernoulli distribution, not a normal distribution, and so $\bar{X}$ (interpreted as the random sample proportion) can only, at best, be approximately normally distributed, appealing to the Central Limit Theorem.

2. Even though we have replaced a population variance by an estimate, we do not make the switch from normal distribution to t-distribution. Informally, there is only one uncertain population parameter θ here (θ determines both the population mean and variance), and the confidence interval we use is only approximate in any case.

Exercise 5.4. *A survey has been conducted to estimate support for an independent Scotland*[3]*. 1067 voters in Scotland were asked: "Should Scotland be an independent country?". The responses were as follows: Yes: 43%, No: 45%, Don't know: 10%, Refused: 3%. Assuming each*

[3] See, for example, http://whatscotlandthinks.org/

respondent was selected at random from the population of eligible voters, calculate an approximate 95% confidence interval for the proportion of "Yes" voters in Scotland, ignoring the "Don't know" and "Refused" responses. What would the CI have been, assuming the same observed proportions, but with a sample size of 100 voters?