Quick review

Last week, we thought about how we can improve our estimation of the *population* mean, μ. Recall that the *sample* mean, $\bar{x}$, is a *point estimate* of μ; when we calculate this, it gives us a single point on the number line.

1. It is very unlikely that this point estimate will "capture" μ;
2. It is also very unlikely that any two sample means will be the same as each other!

We thought about both of these issues last week through the *Vintage Clothing Co.* example.

The *Central Limit Theorem* tells us about the variability of the sample mean $\bar{x}$; specifically,

$$\bar{x} \sim N\left(\mu, \frac{\sigma^2}{n}\right), \qquad \text{approximately,}$$

where σ^2 is the population variance and n is the sample size. We can use this result to examine the variability of the sample mean, without having to repeatedly sample from the population (as we did last week!). We can also use this result to construct an *interval estimate*, or *confidence interval* for the population mean μ, within which we can have a certain level of "confidence" of capturing μ. The formula we derived was:

$$\bar{x} \pm z \times \sqrt{\frac{\sigma^2}{n}},$$

where z is a critical value from the *standard Normal distribution*; recall from last week that $z = 1.96$ for a 95% confidence interval, and $z = 1.65$ or 2.58 for a 90% or 99% confidence interval (respectively).

Example: The *Holiday Hypermarket*

The *Holiday Hypermarket* are a travel agency selling exotic holidays online and over the telephone. Their main call centre employs 500 salespeople. The management are wondering whether or not it is worthwhile having such a large telesales staff since online sales now account for most of their custom.

To investigate, they randomly sample 10% of their employees; for each, they look at the number of sales they made on Wednesday 1st February 2012. From this sample, the average (mean) number of sales made was $\bar{x} = 15.18$. Assuming $\sigma = 8.54$, obtain a 95% confidence interval for the population mean number of sales made.

...Solution...

Question: Where did $\sigma = 8.54$ come from? Recall from last week that σ^2 is the *population variance*, and so σ is the *population standard deviation*. Surely, it is unlikely that we know this value? Only if we have taken a census, or have some sort of "control" over the process variability, will we actually be able to work this out. Last week, we called this **Case 1: Known variance $\boldsymbol{\sigma^2}$**.

What if we don't know the value of σ (it's quite likely that we don't!)?

Recall that we can estimate σ^2 using the *sample variance* s^2.

1.3.2 Case 2: Unknown variance σ^2

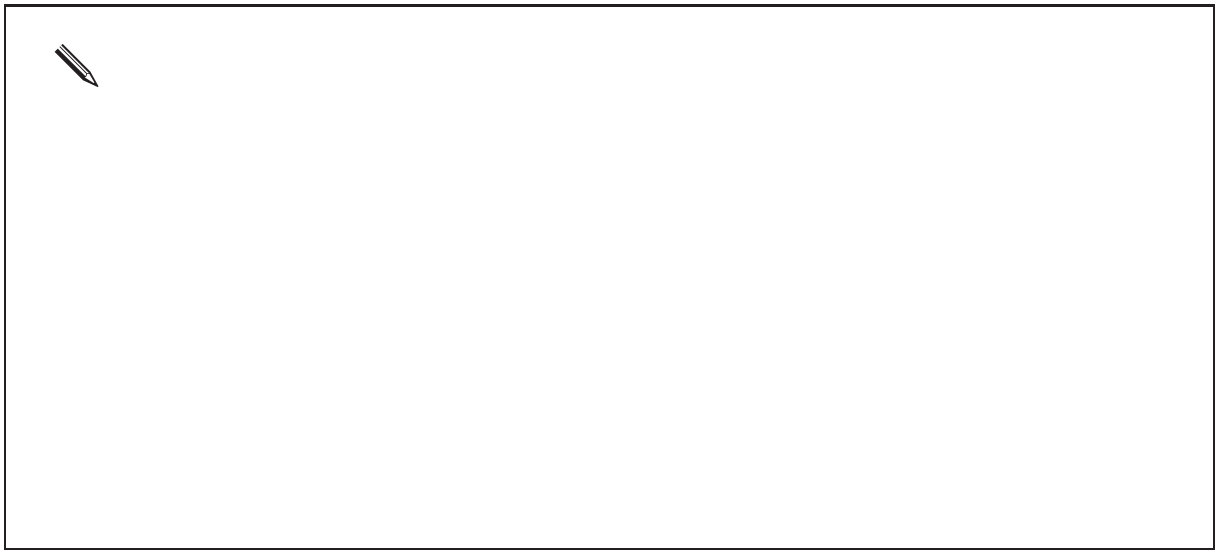
If the population variance σ^2 is unknown, we can no longer use the Normal distribution and instead have to use the t–distribution to calculate confidence intervals. We have seen that when our random sample follows a Normal distribution, or indeed any distribution (if the sample size is large), then the sample mean $\bar{x} \sim N(\mu, \sigma^2/n)$. From this, it follows that

$$Z = \frac{\bar{x} - \mu}{\sqrt{\sigma^2/n}},$$

where Z is the standard Normal distribution, i.e. $Z \sim N(0, 1)$. However, if the population variance is unknown, then the quantity

$$T = \frac{\bar{x} - \mu}{\sqrt{s^2/n}}$$

does *not* have a $N(0, 1)$ distribution (note that the *population* variance σ^2 in Z has been replaced with the *sample* variance s^2 in T). Instead it has a Student's t–distribution. This distribution is similar to the $N(0, 1)$ distribution in that it is symmetrical and bell–shaped, but it is more heavily tailed to allow for greater uncertainty in $\bar{x}$ since the true variability is now unknown. Its exact shape is determined by one parameter called the "degrees of freedom". Table 1.2 gives critical values of Student's t–distribution with various degrees of freedom. These numbers depend on two quantities: ν, the degrees of freedom, and p, a probability.



The expression for the confidence interval in this case is similar to the case where σ^2 is known:

$$\bar{x} \pm t_p \sqrt{s^2/n},$$

where σ^2 has been replaced with the sample variance s^2 and t_p is the appropriate value from the t–distribution tables. But how do we find this value?

First, we need to find p. If we are looking for the 95% confidence interval, we are looking for the value of p which satisfies the equation

$$\begin{aligned} 100(1-p)\% &= 95\%, \qquad \text{i.e.} \\ p &= 0.05. \end{aligned}$$

We would look up the value in t tables (Table 1.2) in the p column, or in this case the 5% column.

We also need to know which row to look in. The rows are given as the degrees of freedom, ν, where $\nu = n - 1$. Hence, if our sample was of size $n = 10$ and we were looking for the 95% confidence interval, we would look in the $\nu = 9$ row and the $p = 5\%$ column to give us a value of **2.262** to use in our calculation.

Example 1
A sample of size 15 is taken from a larger population; the sample mean is calculated as 12 and the sample variance as 25. What is the 95% confidence interval for the population mean μ?

We know that the confidence interval is given by

$$\bar{x} \pm t_p \sqrt{s^2/n},$$

where

$$\begin{aligned} n &= 15, \\ \nu &= n - 1 = 15 - 1 = 14, \\ p &= 5\%, \\ \bar{x} &= 12 \qquad \text{and} \\ s^2 &= 25. \end{aligned}$$

We can find our t value by looking in the $p = 5\%$ column and the $\nu = 14$ row, giving a value of 2.145. Putting what we know into our expression, we get

$$\begin{aligned} 12 &\pm t_{5\%} \sqrt{\frac{25}{15}} \\ 12 &\pm 2.145 \sqrt{\frac{25}{15}} \qquad \text{i.e.} \\ 12 &\pm 2.77. \end{aligned}$$

Hence, the confidence interval is $(9.23, 14.77)$.

Example 2
A credit card company wants to determine the mean income of its card holders. It also wants to find out if there are any differences in mean income between males and females. A random sample of 225 male card holders and 190 female card holders was drawn, and the following results obtained:

	Mean	Standard deviation
Males	£16 450	£3675
Females	£13 220	£3050

Calculate 95% confidence intervals for the mean income for males and females. Is there any evidence to suggest that, on average, males' and females' incomes differ? If so, describe this difference.

...95% confidence interval for male income...

95% confidence interval for female income
Again, the true population variance, σ^2, is unknown, so we can't use the approach of section 1.3.1, and so again we use the t–distribution as in section 1.3.2:

$$\bar{x} \pm t_p \times \sqrt{s^2/n}.$$

Now,

$$\begin{aligned} \bar{x} &= 13220, \\ s^2 &= 3050^2 \\ &= 9302500, \qquad \text{and} \\ n &= 190. \end{aligned}$$

Again, since the sample size is large, we use the ∞ row of table 1.1 to obtain the value of t_p, and so the 95% confidence interval for μ is found as

$$\begin{aligned} 13220 &\pm 1.96 \times \sqrt{9302500/190}, \qquad \text{i.e.} \\ 13220 &\pm 1.96 \times 221.27, \qquad \text{i.e.} \\ 13220 &\pm 433.69. \end{aligned}$$

So, the 95% confidence interval is (£12786.31, £13653.69).

Since the 95% confidence intervals for males and females *do not overlap*, there *is* evidence to suggest that males' and females' incomes, on average, are different. Further, it appears that male card holders earn more than women.

1.3.3 Confidence intervals: a general approach

In this section, we summarise the general procedure for calculating a confidence interval for the population mean μ.

Case 1: Known population variance σ^2

(i) Look out for "...the population variance/standard deviation is...", "the process variance/standard deviation is...", "$\sigma^2 = ...$"

(ii) Calculate the sample mean $\bar{x}$ from the data;

(iii) Calculate your interval! For example,

- for a 90% confidence interval, use the formula

$$\bar{x} \pm 1.64 \times \sqrt{\sigma^2/n};$$

- for a 95% confidence interval, use the formula

$$\bar{x} \pm 1.96 \times \sqrt{\sigma^2/n};$$

- for a 99% confidence interval, use the formula

$$\bar{x} \pm 2.58 \times \sqrt{\sigma^2/n}.$$

Case 2: Unknown population variance σ^2

(i) Look out for "...the sample variance/standard deviation is...", "$s^2 = ...$"

(ii) Calculate the sample mean $\bar{x}$ *and* the sample variance s^2 from the data;

(iii) For a $100(1-p)\%$ confidence interval, look up the value of t under column p, row ν of table 1.1, remembering that $\nu = n - 1$. Note that, for a 90% confidence interval, $p = 10\%$, for a 95% confidence interval, $p = 5\%$ and for a 99% confidence interval, $p = 1\%$;

(iv) Calculate your interval, using

$$\bar{x} \pm t_p \times \sqrt{s^2/n}.$$

Exercises

1. A company packs sacks of flour. The variance of the filling process is 100g. A sample of 50 bags is taken and weighed and the resulting sample mean is 750g. Compute a 95% and 99% confidence interval for the mean weight of a bag of flour.

2. A company manufactures bolts with a process variance of 50mm. A sample of 100 bolts is taken and measured and their average length is calculated as 98mm. What is the 95% confidence interval for the mean length of bolts? If the bolts are designed to be 100mm long, is the process satisfactory?

3. A class of students has sat an exam. A sample of 40 students is taken and their marks produced first. This sample has a mean of 55% and a sample variance of 100. Calculate the 95% confidence interval for the mean mark of the class as a whole.

4. A sample of 12 students is taken and their mean IQ calculated as 110 (with a sample variance of 220). What is the 95% and 99% confidence intervals for the population value based on this sample? What do you notice about the calculated interval as the confidence level increases? Do either of these two confidence intervals contain the known population mean IQ of 100?

5. The following are the number of cars caught speeding each day on one speed camera over a two week period.

10	12	15	9	8	12	11
6	15	17	12	10	9	7

   What is the 95% confidence interval for this sample? How does this compare with the whole Northumbria police average of 8 per day per camera?

		50%	20%	10%	5%	1%
	1	1.00	3.078	6.314	12.706	63.657
	2	0.816	1.886	2.920	4.303	9.925
	3	0.765	1.638	2.353	3.182	5.841
	4	0.741	1.533	2.132	2.776	4.604
	5	0.727	1.476	2.015	2.571	4.032
	6	0.718	1.440	1.943	2.447	3.707
	7	0.711	1.415	1.895	2.365	3.449
ν	8	0.706	1.397	1.860	2.306	3.355
	9	0.703	1.383	1.833	2.262	3.250
	10	0.700	1.372	1.812	2.228	3.169
	11	0.697	1.363	1.796	2.201	3.106
	12	0.695	1.356	1.782	2.179	3.055
	13	0.694	1.350	1.771	2.160	3.012
	14	0.692	1.345	1.761	2.145	2.977
	15	0.691	1.341	1.753	2.131	2.947
	16	0.690	1.337	1.746	2.120	2.921
	17	0.689	1.333	1.740	2.110	2.898
	18	0.688	1.330	1.734	2.101	2.878
	19	0.688	1.328	1.729	2.093	2.861
	20	0.687	1.325	1.725	2.086	2.845
	21	0.686	1.323	1.721	2.080	2.831
	22	0.686	1.321	1.717	2.074	2.819
	23	0.685	1.319	1.714	2.069	2.807
	24	0.685	1.318	1.711	2.064	2.797
	25	0.684	1.316	1.708	2.060	2.787
	26	0.684	1.315	1.706	2.056	2.779
	27	0.684	1.314	1.703	2.052	2.771
	28	0.683	1.313	1.701	2.048	2.763
	29	0.683	1.311	1.699	2.045	2.756
	⋮	⋮	⋮	⋮	⋮	⋮
	∞	0.674	1.282	1.645	1.960	2.576

Table 1.2: Tabulated values of t for which $\Pr(|T| > t) = p$, where T has a t–distribution with ν degrees of freedom