Lecture 2

CONFIDENCE INTERVALS II

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.

Quick review

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.

Quick review

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 $z = \mathbf{1.96}$ for a 95% confidence interval, and $z = \mathbf{1.65}$ or $\mathbf{2.58}$ 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.

Example: The *Holiday Hypermarket*

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.

Example: The *Holiday Hypermarket*

We are told that 10% of employees are randomly sampled; there are 500 employees in total, giving $n = 50$.

We are also given that

$$\bar{x} = 15.18 \qquad \text{and} \qquad \sigma = 8.54.$$

Thus, the 95% confidence interval for μ is given by

$$15.18 \pm 1.96 \times \sqrt{\frac{8.54^2}{50}} \qquad \text{i.e.}$$

$$15.18 \pm 2.367,$$

giving (12.81 sales, 17.55 sales).

Example: The *Holiday Hypermarket*

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.

Case 1: Known variance σ^2

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.

Case 2: Unknown variance σ^2

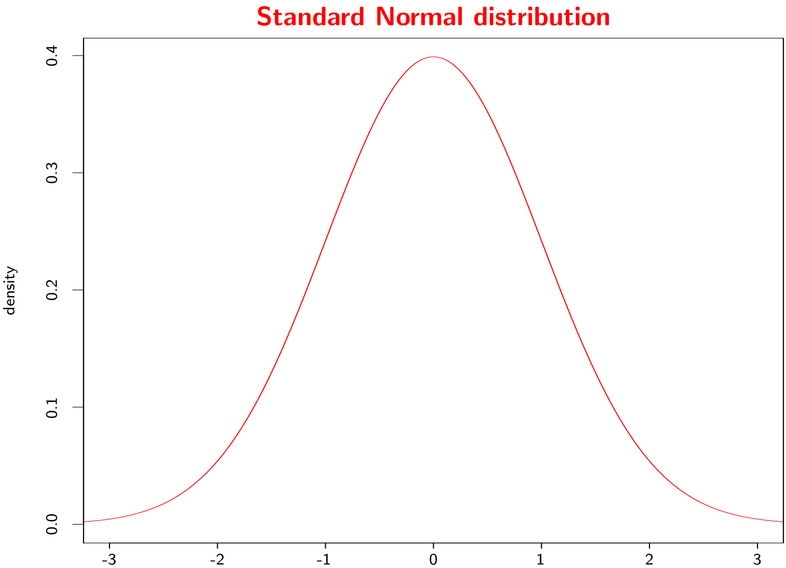
If the population variance is unknown (which is usually the case), the quantity

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

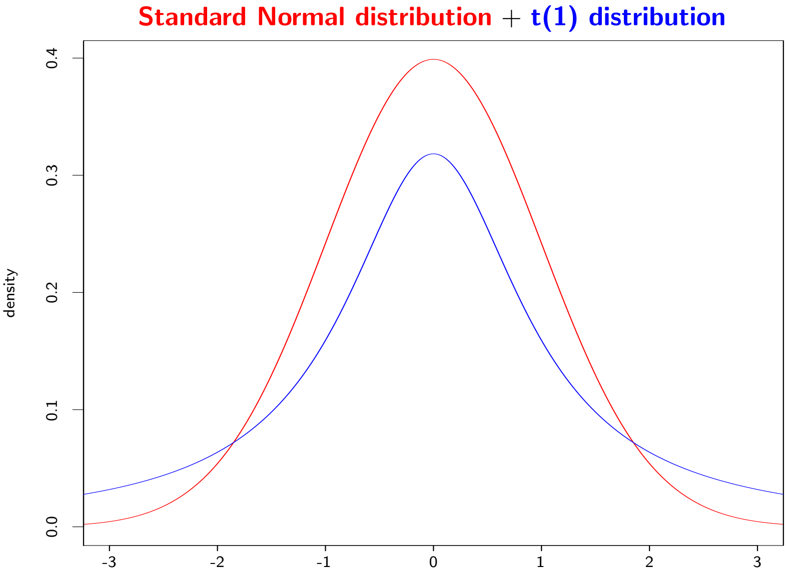
does **not** have a $N(0,1)$ distribution, but a **Student's t–distribution**.

- This is similar to the normal distribution (i.e. symmetric and bell–shaped), but is more 'heavily tailed';
- The *t*–distribution has one parameter, called the "degrees of freedom" ($\nu = n - 1$).

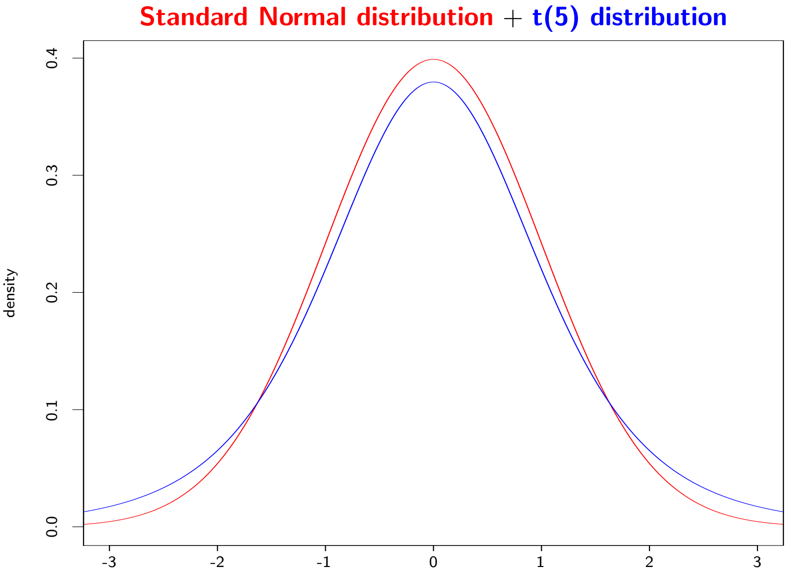
Case 2: Unknown variance σ^2



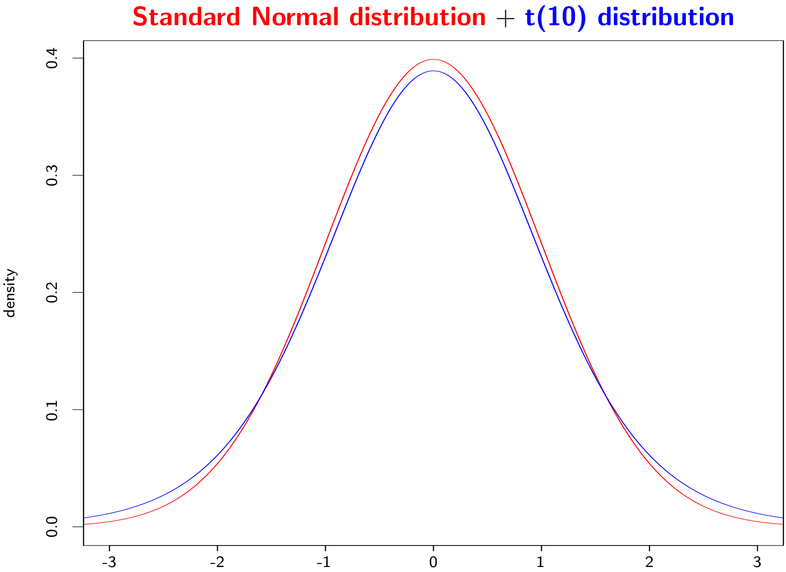
Case 2: Unknown variance σ^2



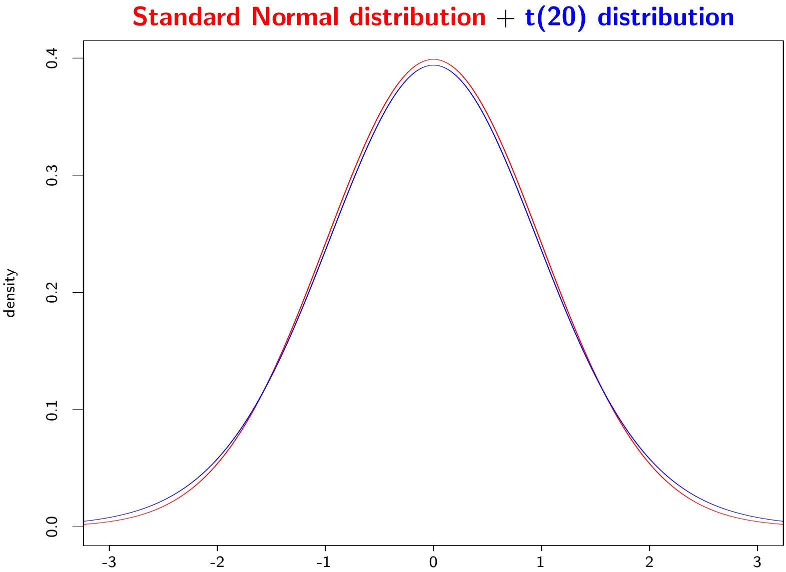
Case 2: Unknown variance σ^2



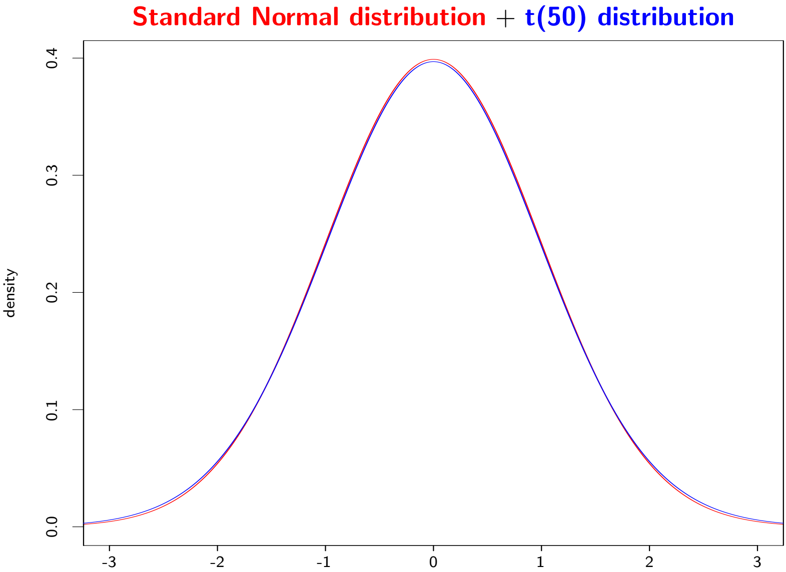
Case 2: Unknown variance σ^2



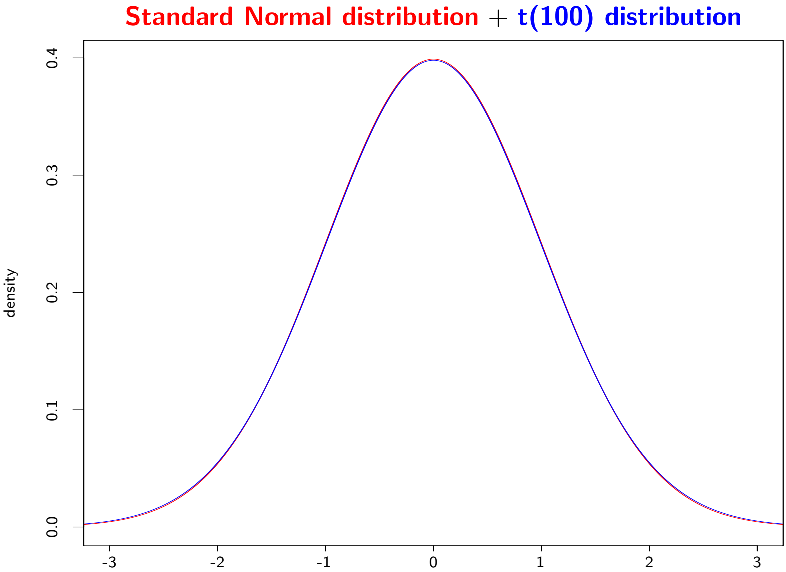
Case 2: Unknown variance σ^2



Case 2: Unknown variance σ^2



Case 2: Unknown variance σ^2



Case 2: Unknown variance σ^2

So if we don't know σ^2, the formula for the confidence interval becomes:

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

We find t_p from statistical tables (Table 1.2 in the notes). We read along the p column and down the ν row.

- For a 90% confidence interval, $p = 10\%$;
- For a 95% confidence interval, $p = 5\%$;
- For a 99% confidence interval, $p = 1\%$.
- The degrees of freedom, $\nu = n - 1$.

Example 1 (page 16)

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?

Example 1 (page 16)

We know that the confidence interval is given by

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

where

$$\begin{aligned} n &= 15, \\ \bar{x} &= 12 \qquad \text{and} \\ s^2 &= 25. \end{aligned}$$

Also, to find t, we know that

$$\begin{aligned} \nu &= n - 1 = 15 - 1 = 14 \qquad \text{and} \\ p &= 5\%. \end{aligned}$$

Example 1 (page 16)

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

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

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

Example 2 (page 17)

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

The true population variance, σ^2, is unknown (we have the **sample** variance), and so we have **case 2** and need to use the t distribution. Thus,

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

From the question,

$$\bar{x} = 16450, \qquad s^2 = 3675^2 = 13505625, \qquad \text{and} \qquad n = 225.$$

But what about t_p?

95% confidence interval for male income

The value t_p must be found from Table 1.2.

- Recall that the degrees of freedom, $\nu = n - 1$, and so here we have $\nu = 225 - 1 = 224$
- But Table 1.2 only gives value of ν up to 29 – for higher values, we use the ∞ row
- Since we require a 95% confidence interval, we read down the $p = 5\%$ column, giving a t value of 1.96

95% confidence interval for male income

Thus, the 95% confidence interval for μ is found as

$$\begin{aligned} 16450 &\pm 1.96 \times \sqrt{13505625/225}, \qquad \text{i.e.} \\ 16450 &\pm 480.2. \end{aligned}$$

So, the 95% confidence interval for **male income** is

$$(£15969.80,\ £16930.20).$$

95% confidence interval for female income

Again, the true population variance, σ^2, is unknown, and so we have case 2. Thus,

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

Now,

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

95% confidence interval for female income

As before, since the sample size is large we use the ∞ row of Table 1.2 to obtain the value of t_p, giving:

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

So, the 95% confidence interval for **female income** is

$$(£12786.31,\ £13653.69).$$

95% confidence interval for female income

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.

Confidence intervals: a general approach

We now summarise the general procedure for calculating a confidence interval for the population mean μ.

Case 1: Known population variance σ^2

(i) Key words: "**population variance**", "**population standard deviation**", "$\sigma = ...$"

(ii) Calculate the sample mean $\bar{x}$ from the data (if not given)

(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

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

Confidence intervals: a general approach

Case 2: Unknown population variance σ^2

(i) Key words: "**sample variance**", "**sample standard deviation**", "$s = \ldots$"

(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

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