MAS1403/ACE2013

Quantitative Methods for Business Management

Statistics for Marketing and Management

Semester 2, 2008–09

Lecturer: Dr. Lee Fawcett

New arrangements for 2009

Lecture times have now changed! The Monday 10 o'clock group has now moved to **Tuesday at 4**; the Monday 12 o'clock group has moved to **Monday at 10** (but in the same place).

Tutorial arrangements have also changed:

- **Group A**: Wednesday 9, Fine Art LT
- **Group B**: Wednesday 10, Herschel LT3
- **Group C**: Wednesday 11, Herschel LT3
- **Group D**: Thursday 9, Agriculture Clement Stephenson LT
- **Group E**: Thursday 12, Herschel LT1
- **Group F**: Thursday 3, Herschel LT1
- **Group G**: Friday 9, Bedson LG.38
- **Group H**: Friday 11, Herschel LT3

A **Minitab practical** will take the place of the tutorial in week 7

New arrangements for 2009

There will be three **CBA**s

There will be one **written assignment** over the Easter holidays

There will be an **exam** at the end of Semester 2 covering material from the *entire year*!

You should refer to the **week–by–week schedule** for this course for CBA/assignment deadlines, computer practicals etc. etc.

Lecture 1

ESTIMATION

Recap and introduction

Recall that data can be summarised in **two** ways:

1. **Graphical summaries**
   - Stem–and-leaf plots;
   - Bar charts;
   - Histograms;
   - Relative frequency histograms;
   - Frequency polygons.

2. Numerical summaries

- **Measures of location**
  - (i) Sample mean;
  - (ii) Sample median;
  - (iii) Sample mode.
- **Measures of spread**
  - (i) Range;
  - (ii) Variance (and standard deviation);
  - (iii) Interquartile range.

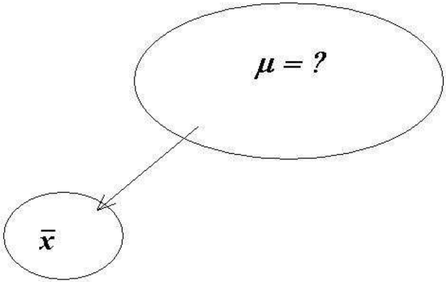
What does our sample tell us about the population?

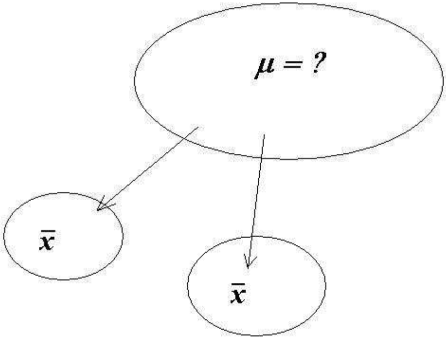
- We can rarely observe the entire population, so the **population mean** and **population variance** are hardly ever known *exactly*;
- These unknown quantities are called **parameters**;
- We use Greek letters to denote them – μ for the mean, and σ^2 for the variance (and so σ for the standard deviation);
- We hope that the **sample mean** ($\bar{x}$) will be quite close to the true mean (μ);
- **But how do we know if it is?**

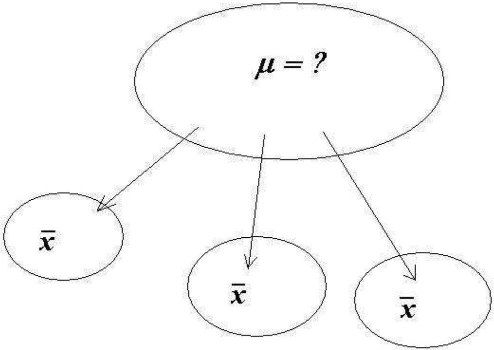
The distribution of the sample mean

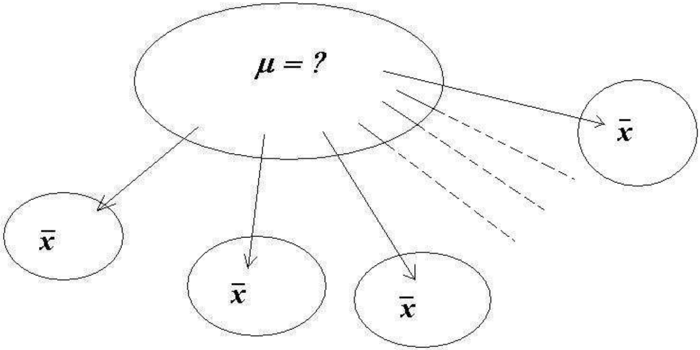
- Let $x_1, x_2, \ldots, x_n$ be a random sample from a $N(\mu, \sigma^2)$ distribution. We can calculate the mean from this sample – call this $\bar{x}_1$;
- Let $x_1, x_2, \ldots, x_n$ be a random sample from another $N(\mu, \sigma^2)$ distribution. We can calculate the mean from this sample too – call this $\bar{x}_2$;
- We can calculate the means from many samples, and look at the distribution of the $\bar{x}$'s!

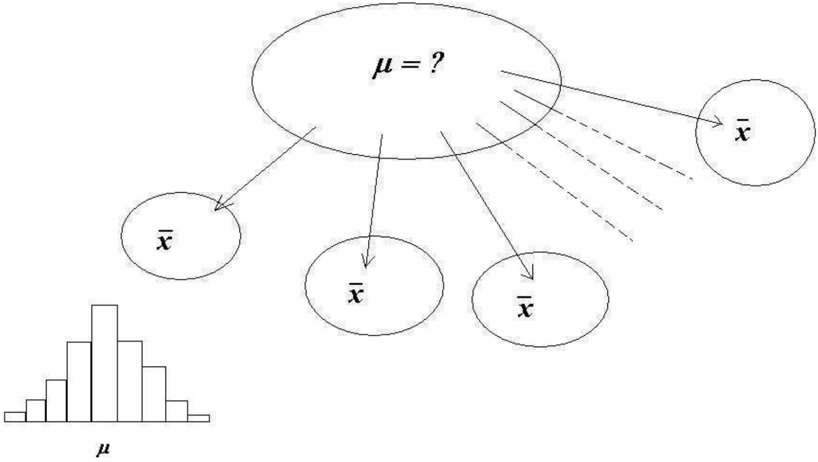
$\mu = ?$


$\mu = ?$
$\bar{x}$


μ = ?
x̄
x̄


$\mu = ?$
$\bar{x}$
$\bar{x}$
$\bar{x}$


$\mu = ?$
$\bar{x}$
$\bar{x}$
$\bar{x}$
$\bar{x}$



$\mu = ?$
$\bar{x}$
$\bar{x}$
$\bar{x}$
$\bar{x}$
μ

- It turns out that, if the populations from which the samples were drawn follow normal distributions, then the $\bar{x}$'s will also follow a normal distribution; in fact,

$$\bar{x} \sim N(\mu, \sigma^2/n).$$

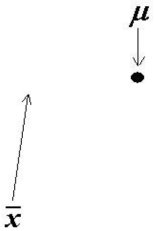
- The **Central Limit Theorem** goes one step further and says that, if n is large, then this result will (approximately) hold **no matter what the 'parent' population distribution**!

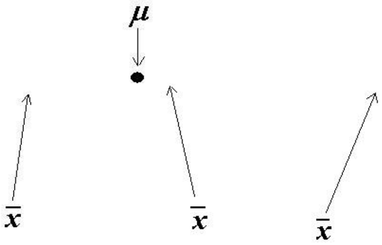
Interval estimation

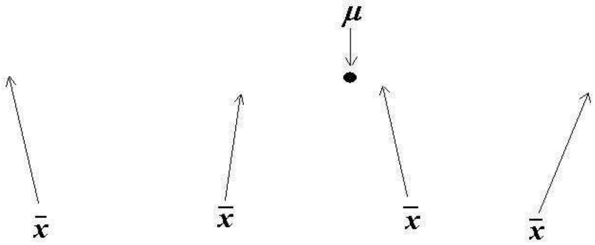
$\bar{x}$ is a **point estimate** of the population mean μ. We can improve estimation by constructing an **interval estimate**.

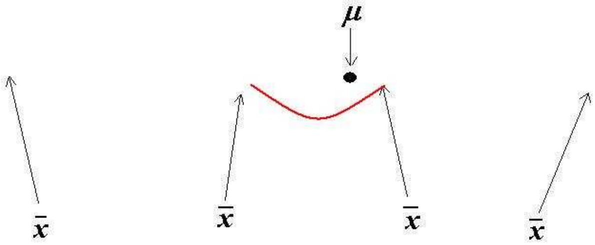
- To construct such an interval, we first calculate the sample mean $\bar{x}$;
- We then go a little bit to the left of $\bar{x}$ and a little bit to the right of $\bar{x}$ to create an interval to (hopefully!) 'capture' μ;
- It's more likely that μ will fall within this interval than exactly 'on top of' the point estimate.

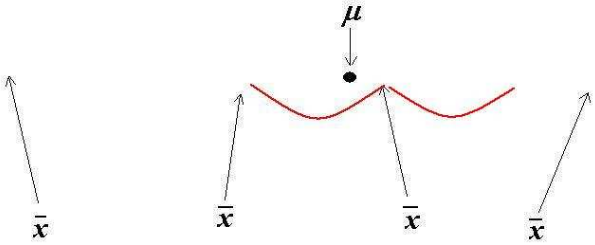

μ


μ
$\bar{x}$

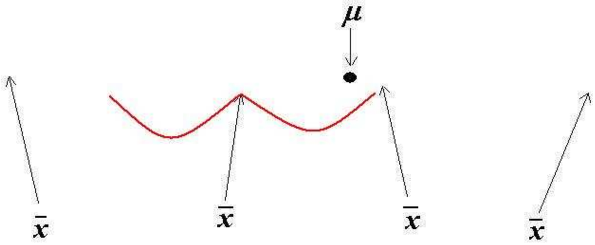

μ
$\bar{x}$
$\bar{x}$
$\bar{x}$

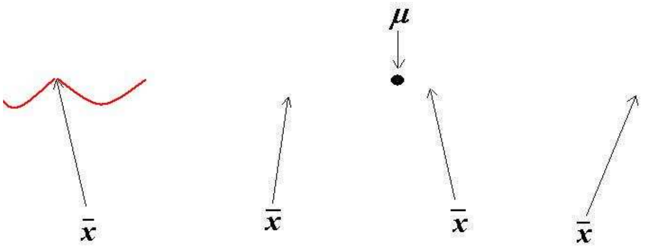

μ
$\bar{x}$
$\bar{x}$
$\bar{x}$
$\bar{x}$

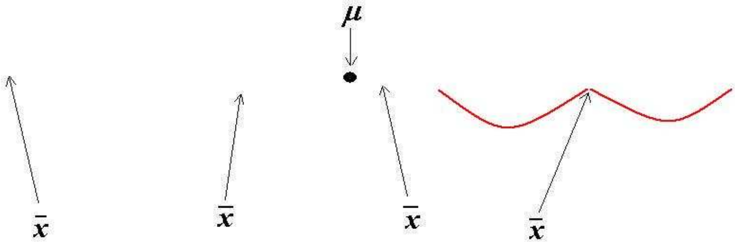

μ
$\bar{x}$
$\bar{x}$
$\bar{x}$
$\bar{x}$



μ
$\bar{x}$
$\bar{x}$
$\bar{x}$
$\bar{x}$


μ
$\bar{x}$
$\bar{x}$
$\bar{x}$
$\bar{x}$


μ
x̄
x̄
x̄
x̄


μ
$\bar{x}$
$\bar{x}$
$\bar{x}$
$\bar{x}$

But how much do we go to the left and right? This depends on:

(i) The size of our sample;

(ii) How 'confident' we want to be that our interval captures μ, and

(iii) What (if anything) we know about the population.

Construction of a confidence interval

- We know from the Central Limit Theorem that

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

- We can 'standardise' $\bar{x}$, using "slide–squash", i.e.

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

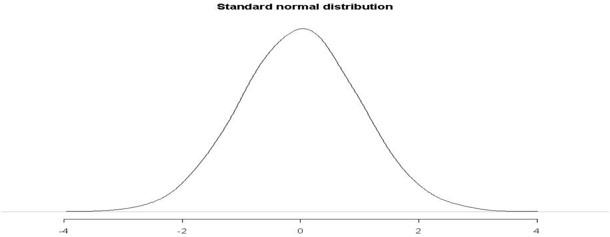
where $Z \sim N(0,1)$.

Construction of a confidence interval

- We know that (from tables)

$$\Pr(-1.96 < Z < 1.96) \quad = \quad 0.95;$$

We can think about this graphically:

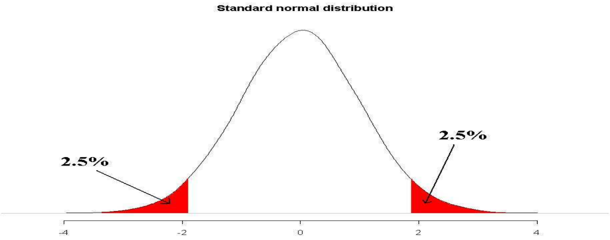


Construction of a confidence interval

- We know that (from tables)

$$\Pr(-1.96 < Z < 1.96) = 0.95;$$

We can think about this graphically:

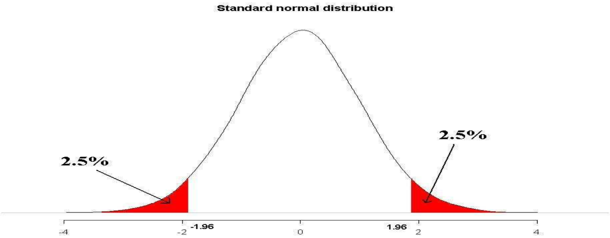


Construction of a confidence interval

- We know that (from tables)

$$\Pr(-1.96 < Z < 1.96) = 0.95;$$

We can think about this graphically:

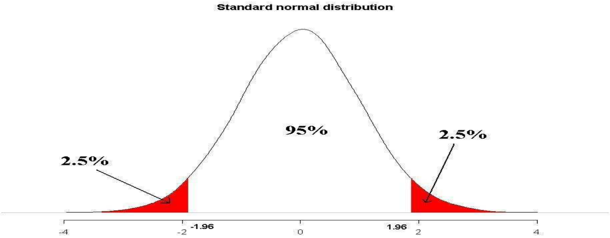


Construction of a confidence interval

- We know that (from tables)

$$\Pr(-1.96 < Z < 1.96) = 0.95;$$

We can think about this graphically:



- Thus,

$$\Pr\left(-1.96 < \frac{\bar{x} - \mu}{\sqrt{\sigma^2/n}} < 1.96\right) = 0.95;$$

Construction of a confidence interval

- Rearranging the LHS, we get

$$\Pr\left(\bar{x} - 1.96 \times \sqrt{\sigma^2/n} < \mu < \bar{x} + 1.96 \times \sqrt{\sigma^2/n}\right).$$

If we want a 99% confidence interval, the only thing that will change is the value 1.96.

Case 1: Known variance σ^2

If we know the population variance σ^2, we can just bung our numbers into the formula on the previous slide! Remember, the (95%) confidence interval is

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

where

- $\bar{x}$ is the **sample mean**;
- σ^2 is the **population variance**, and
- n is the **sample size**.

Example 1 (page 5)

A coffee machine fills cups with hot water; the variance of the filling process is known to be $\sigma^2 = 10$ml.

A sample of 100 filled cups gives a sample mean and we have calculated a sample mean of $\bar{x} = 40$ml.

What is the 95% confidence interval of the population mean μ?

Example 1 (page 5)

We already have a formula for the 95% confidence interval:

$$\bar{x} \quad \pm \quad 1.96\sqrt{\sigma^2/n}.$$

So, inputting our values, we get

$$\begin{aligned} 40 \quad &\pm \quad 1.96\sqrt{10/100}, \qquad \text{i.e.} \\ 40 \quad &\pm \quad 0.61. \end{aligned}$$

Hence, the 95% confidence interval for the population mean μ is $(39.39, 40.61)$.

Example 1 (page 5)

What would happen if the sample size increased to 200 and everything else remained the same? We'd get

$$\begin{aligned} 40 &\pm 1.96\sqrt{10/200}, \qquad \text{i.e.} \\ 40 &\pm 0.44. \end{aligned}$$

Hence, the 95% confidence interval for the population mean μ is $(39.56, 40.44)$.

This should be intuitive, since as the sample size increases we are becoming more sure of our estimate for the population value.

Example 1 (page 5)

What would be the 99% confidence interval in this case? From tables for the standard normal distribution, we can find that

$$\Pr(-2.58 < Z < 2.58) = 0.99;$$

hence, the 99% confidence interval is given by

$$\bar{x} \quad \pm \quad 2.58\sqrt{\sigma^2/n},$$

in this case giving

$$\begin{aligned} 40 &\quad \pm \quad 2.58\sqrt{10/200}, \qquad \text{i.e.} \\ 40 &\quad \pm \quad 0.58. \end{aligned}$$

Hence, the 99% confidence interval for the population mean μ is $(39.42, 40.58)$.

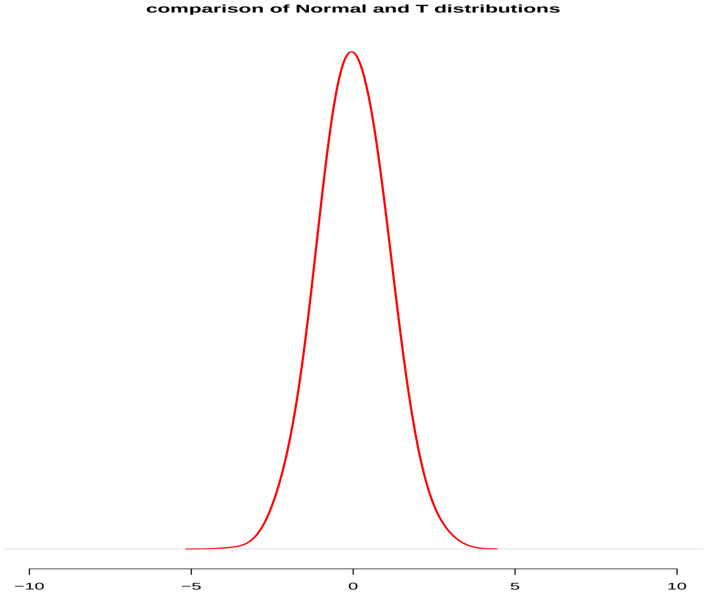
Case 2: Unknown variance σ^2

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

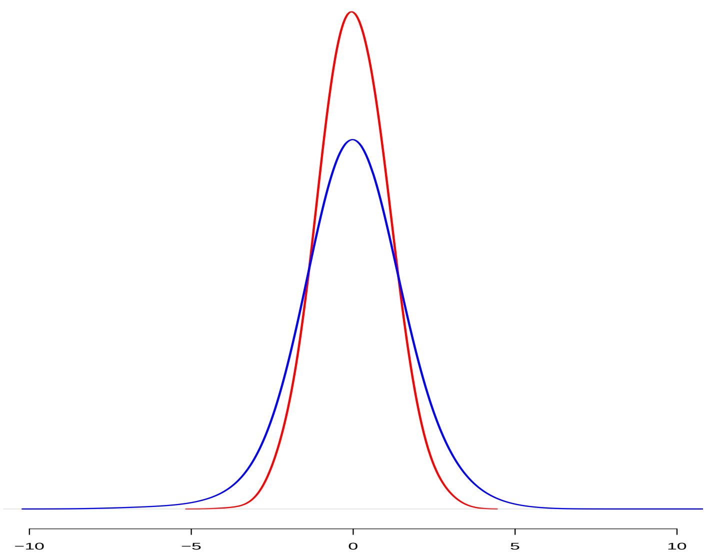
$$T = \frac{\bar{x} - \mu}{\sqrt{s^2/n}}$$

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" ($\nu = n - 1$).


comparison of Normal and T distributions
−10
−5
0
5
10

comparison of Normal and T distributions



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 1.1 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 2 (page 7)

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 2 (page 7)

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 2 (page 7)

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 $(9.23, 14.77)$.

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

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

(ii) 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, $p = 1\%$;

(iii) Calculate your interval, using

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

Application of Confidence Intervals

You might be asking: "why do we bother calculating confidence intervals?".

- By calculating a confidence interval for the population mean, it allows us to see how confident we are of the point estimate we have calculated. The wider the range, the less precise we can be about the population value.
- If we have a known (or target) value for a population and this does not fall within the confidence interval of our sample, this could suggest that there is something different about this sample.
- It allows us to start looking at differences between groups. If the confidence intervals for two samples do not overlap, this could suggest that they are from separate populations.

Example 1.4.1 (page 9)

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.

Example 1.4.1 (page 9)

95% confidence interval for male income

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

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

Here,

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

The value t_p must be found from table 1.1.

- Recall that the degrees of freedom, $\nu = n - 1$, and so here we have $\nu = 225 - 1 = 224$;
- But table 1.1 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 5% column, giving a t value of 1.96.

Example 1.4.1 (page 9)

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 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,

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

Example 1.4.1 (page 10)

Again, since the sample size is large, we use the ∞ row of table 1.1 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 is (£12786.31,£13653.69).

Example 1.4.1 (page 10)

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.