

# Investigating associations between two variables

## Chapter questions

- ▶ What are bivariate data?
- ▶ What are explanatory and response variables?
- ▶ What are two-way frequency tables and how do we interpret them?
- ▶ How do we construct and interpret segmented bar charts from two-way frequency tables?
- ▶ How do we construct and interpret parallel dot plots?
- ▶ How do we construct and interpret back-to-back stem plots?
- ▶ How do we construct and interpret parallel boxplots?
- ▶ What is a scatterplot, how is it constructed and what does it tell us?
- ▶ What do we mean when we describe the association between two numerical variables in terms of direction, form and strength?
- ▶ What is the difference between observation and experimentation?
- ▶ What is the difference between association and causation?

In this chapter we begin our study of **bivariate data**, data which is recorded on two variables from the same subject.

## 2A Bivariate data – Classifying the variables

### Learning intentions

- ▶ To introduce bivariate data.
- ▶ To be able to classify data as categorical or numerical.
- ▶ To be able to identify explanatory and response variables.

So far you have learned how to display, describe and compare the distributions of single variables. In the process you learned how to use data to answer questions like ‘What is the favourite colour of prep-grade students?’ or ‘How do the weights of tuna fish vary?’ In each case we concentrated on investigating the statistical variables individually.

However, questions like ‘Does the new treatment for headache work more quickly than the old treatment?’, ‘Are city voters more likely to vote for the Greens party than country voters?’ or ‘Can we predict a student’s test score from the time (in hours) they spent studying for the test?’ cannot be answered by considering variables separately. All of these questions relate to situations where the two variables are linked in some way (associated) so that they vary together. The data generated in these circumstances is called **bivariate data**.

Analysing bivariate data requires a new set of statistical tools. Developing and applying these tools is the subject of the next four chapters.

### Categorical and numerical variables

You will recall from Chapter 1 we defined two classifications of variables, categorical and numerical variables:

- **Categorical variables** generate data values that are names or labels, such as *favourite pet* (dog, cat, rabbit, bird) or *coffee size* (small, medium, large).
- **Numerical variables** generate data values that are numbers, usually resulting from counting or measuring, such as *number of brothers* (0, 1, 2, ...) or *hand span* (cm).

The first step in investigating the association between two variables is to classify each variable as either categorical or numerical. Consider again the previous questions:

- Does the new treatment for headache work more quickly than the old treatment?

The two variables in this question are *type of treatment*, a categorical variable taking the values ‘new’ and ‘old’, and *time taken for the headache to be relieved*, a numerical variable, measured in minutes. Thus, investigation of a question like this can be classified as **investigating the association between a categorical variable and a numerical variable**.

- Are city voters more likely to vote for the Greens party than country voters?

This question involves two variables, *place of residence*, which is a categorical variable taking the values ‘city’ and ‘country’, and *vote for the Greens*, which also is a categorical variable taking the values ‘yes’ and ‘no’. Investigation of a question like this can be classified as **investigating the association between two categorical variables**.

- Can we predict a student’s test score (%) from time (in hours) spent studying for the test?

The variable *test score* is a numerical variable, as is *time spent studying for the test*. Investigation of a question like this can be classified as **investigating the association between two numerical variables**.

As discussed in Chapter 1, categorical variables can be further classified as nominal or ordinal, and numerical variables can be further classified as discrete or continuous.



### Example 1 Identifying associations as categorical or numerical

For each of the following questions, determine if they involve investigating associations between

- one numerical variable and one categorical variable or
  - two categorical variables or
  - two numerical variables.
- Are younger people (age measured in years) more likely to believe in astrology (measured as ‘yes’ or ‘no’) than older people?
  - Do people who weigh more (weight measured in kg) tend to have higher blood pressure (blood pressure measured in mmHg)?
  - Are people who have a driver’s licence (measured as ‘yes’ or ‘no’) more likely to be in favour of lowering the driving age (measured as ‘yes’ or ‘no’)?

#### Solution

- One numerical variable (*age*) and one categorical variable (*belief in astrology*)
- Two numerical variables (*weight* and *blood pressure*)
- Two categorical variables (*have a driver’s licence* and *support for lowering the driving age*)

## Identifying response and explanatory variables

When investigating associations between variables, it is helpful to think of one of the variables as the **explanatory variable**. The other variable is then called the **response variable**. We use the explanatory variable to explain changes that might be observed in the response variable.

For example, the question, ‘Are city voters more likely to vote for the Greens party than country voters?’, suggests that knowing a person’s place of residence might be useful in

explaining voting preference. In this situation *place of residence* is the explanatory variable and *vote for Greens* is the response variable.

It is important to be able to identify the explanatory and response variables before you explore the association between them. Consider the following examples.



### Example 2 Identifying the response and explanatory variables

We wish to investigate the question, ‘Does the time it takes a student to get to school depend on their mode of transport?’ The variables here are *time* and *mode of transport*. Which is the response variable (RV) and which is the explanatory variable (EV)?

#### Explanation

In asking the question in this way we are suggesting that a student’s *mode of transport* might explain the differences we observe in the time it takes students to get to school.

#### Solution

EV: *mode of transport*  
RV: *time*



### Example 3 Identifying the response and explanatory variables

Can we predict people’s height (in cm) from their wrist measurement? The variables in this investigation are *height* and *wrist measurement*. Which is the response variable (RV) and which is the explanatory variable (EV)?

#### Explanation

Since we wish to predict height from wrist circumference, we are using *wrist measurement* as the predictor or explanatory variable. *Height* is then the response variable.

#### Solution

EV: *wrist measurement*  
RV: *height*

It is important to note that, in Example 3, we could have asked the question the other way around; that is, ‘Can we predict people’s wrist measurement from their height?’ In that case *height* would be the explanatory variable, and *wrist measurement* would be the response variable. The way we ask our statistical question is an important factor when there is no obvious explanatory variable.

### Response and explanatory variables

When investigating the association between two variables the *explanatory variable (EV)* is the variable we expect to explain or predict the value of the *response variable (RV)*.

**Note:** The explanatory variable is sometimes called the independent variable (IV) and the response variable the dependent variable (DV).

## Exercise 2A

### Identifying variables as categorical or numerical

#### Example 1

- 1 For each of the following questions, determine if they involve investigating associations between:
  - one numerical and one categorical variable or
  - two categorical variables or
  - two numerical variables.
  - a Are full-time and part-time students equally likely travel to university by car?
  - b Do Year 11 students watch more hours of television each week than Year 12 students?
  - c Do countries with higher household incomes (\$) tend to have lower infant mortality rates (deaths/1000 births)?
  - d Is there a relationship between attitude to gun control and country of birth?

### Identifying explanatory and response variables

#### Example 2

- 2 For each of the following situations identify the explanatory variable (EV) and response variable (RV). In each situation the variable names are *italicised*.

#### Example 3

- a We wish to investigate whether a fish's *toxicity* can be predicted from its *colour*. We want to be able to predict *toxicity* from *colour*.
  - b The relationship between *weight loss* and *type of diet* is to be investigated.
  - c We wish to investigate the relationship between a used car's *age* and its *price*.
  - d It is suggested that the *cost* of heating in a house depends on the type of *fuel* used.
  - e The relationship between the *price* of a house and its *location* is to be investigated.
- 3 The following pairs of variables are related. In each case identify which is likely to be the explanatory variable and which is the response variable, and the level of measurement of each variable (categorical or numerical). The variable names are italicised.
    - a *exercise level* (1 = light, 2 = moderate, 3 = a lot) and *age* (years).
    - b *years of education* (years) and *salary level* (\$ per annum).
    - c *comfort level* (0 = uncomfortable, 1 = comfortable) and *temperature* (°C).
    - d *time of year* (summer, autumn, winter, spring) and *incidence of hay fever* (1 = never, 2 = sometimes, 3 = regularly).
    - e *age group* (less than 25, 25 - 40, more than 40) and *musical taste* (classical, rock, rap, country, indie, dance, jazz).
    - f *AFL team supported* and *state of residence*.

## Exam 1 style questions

- 4 Respondents to a survey question "How concerned are you about climate change?" were asked to select from the following responses:  
1 = not at all, 2 = a little, 3 = moderately, 4 = extremely  
The data which was collected in response to this question is:
- A** nominal                      **B** ordinal                      **C** discrete  
**D** continuous                  **E** numerical
- 5 The variables *weight* (light, medium, heavy) and *height* (less than 160cm, 160-175cm, over 175cm) are:
- A** both nominal variables  
**B** both ordinal variables  
**C** a nominal and an ordinal variable respectively  
**D** an ordinal and a nominal variable respectively  
**E** both continuous variables
- 6 Researchers believe that reaction time might be lower in cold temperatures. They devise an experiment where *reaction time* in seconds is measured at three different *temperature* levels (1 = less than 8°C, 2 = from 8°C to 18 °C, 3 = more than 18°C). The response variable, and its classification are:
- A** *reaction time*, categorical                      **B** *temperature*, categorical  
**C** *reaction time*, numerical                      **D** *temperature*, numerical  
**E** *temperature*, ordinal

## 2B Investigating associations between categorical variables

### Learning intentions

- ▶ To be able to summarise data from two categorical variables using two-way frequency tables.
- ▶ To be able to appropriately percentage two-way frequency tables.
- ▶ To be able to use a percentaged two-way frequency table to identify and describe an association between two categorical variables.

If two variables are related or linked in some way, we say they are associated. To begin the investigation of an association between two categorical variables we create a **contingency table** or a **two-way frequency table**. It is called a two-way frequency table because it is summarising data from two variables.

## Constructing a two-way frequency table

It has been suggested that city and country people have differing attitudes to gun control; that is, that support for gun control depends on where a person lives. How might we investigate this relationship? Suppose we ask a sample of three people about their *attitude to gun control*, and we also record their *residence*. The resulting data for the three people might look like this:

Subject no.	<i>Residence</i>	<i>Attitude to gun control</i>
1	City	For
2	Country	For
3	City	Against

The first thing to note is that these two variables, *attitude to gun control* (for or against) and *residence* (city or country), are both categorical variables. Categorical data are usually presented in the form of a frequency table.

Suppose we continue until we have interviewed a sample of 100 people, and we find that there are 58 who live in the country and 42 who live in the city. We can present this result in a frequency table as shown to the right.

<i>Residence</i>	Frequency
Country	58
City	42
Total	100

From this table, we can see that there were more country than city people in our sample.

Suppose also when we record the attitude to gun control, we might have 62 'for' and 38 'against' gun control. Again, we could present these results in a frequency table as shown to the right.

<i>Attitude to gun control</i>	Frequency
For	62
Against	38
Total	100

From this table, we can see that more people in the sample were for gun control than against gun control. However, we cannot tell from the information contained in the tables whether *attitude to gun control* depends on the *residence* of the person. To do this we need to construct a **two-way frequency table**, which gives both the *attitude to gun control* and the *residence* for each person in the sample.

We begin by counting the number of people in the sample who are:

- from the country and for gun control
- from the city and for gun control
- from the country and against gun control
- from the city and against gun control.



Suppose again from our sample of 100 people we find the following frequencies:

- 32 country people are for gun control
- 30 city people are for gun control
- 26 country people are against gun control
- 12 city people are against gun control.

### Explanatory and response variables in two-way frequency tables

Before we set up the two-way frequency table, we need to decide which is the explanatory variable and which is the response variable of the two variables. Since we think that a person's attitude to gun control might depend on their place of residence, but not the other way around, then:

- *residence* is the explanatory variable (EV)
- *attitude to gun control* is the response variable (RV).

In two-way frequency tables, it is conventional to let the categories of the *response variable* label the *rows* of the table and the categories of the *explanatory variable* label the *columns* of the table. Following this convention, we can create the following two-way frequency table.

	<i>Residence</i>	
<i>Attitude to gun control</i>	Country	City
For	32	30
Against	26	12

To complete the table, it is usual to calculate the row and column sums, as shown below.

	<i>Residence</i>			
<i>Attitude to gun control</i>	Country	City	Total	
For	32	30	62	Row sum
Against	26	12	38	Row sum
Total	58	42	100	
	Column sum	Column sum		

The shaded regions in the table are called the **cells** of the table. It is the numbers in these cells that we look at when investigating the relationship between the two variables.



#### Example 4 Constructing a two-way frequency table

The following data were obtained when a sample of ten Year 9 students were asked if they intended to go to university (*university*). The gender of the student was also recorded.



Student	Gender	University	Student	Gender	University
1	Female	Yes	6	Male	Yes
2	Male	Yes	7	Female	Yes
3	Female	No	8	Male	No
4	Female	Yes	9	Female	No
5	Male	No	10	Female	Yes

Create a two-way frequency table from these data.

### Explanation

- We first need to identify the explanatory variable and the response variable.
- Create the table showing the values of *Gender* labelling the columns, and *University* labelling the rows.
- Consider Student 1, who is female and indicated yes to go to university. Place a mark in the corresponding cell of the table.
- Go through the data set one person at a time, placing a mark in the appropriate cell for each person.
- Finally, tally the marks in each cell, and then calculate the row and column sums. Make sure the total adds to the number of students in the sample.

### Solution

It is possible that a student's intention to go to university may depend on their gender, but not the other way around. Thus, *gender* is the explanatory variable and *university* is the response variable.

	Gender	
University	Male	Female
Yes		
No		

	Gender	
University	Male	Female
Yes		
No		

	Gender	
University	Male	Female
Yes		
No		

	Gender		
University	Male	Female	Total
Yes	2	4	6
No	2	2	4
Total	4	6	10

Consider again the two-way frequency table created to investigate the association between place of residence and attitude to gun control. This table tells us that more country people are in favour of gun control than city people. But is this just due to the fact that there were more country people in the sample? To help us answer this question we need to express the frequencies in each cell as **percentage frequencies**.

### Percentaged two-way frequency table

When the two-way frequency table has been constructed so that the values of the explanatory variable label the rows, then we calculate **column percentages** to help us investigate the association. This will give us the percentage of country and the percentage of city people for and against gun control, which can then be compared.

Column percentages are determined by dividing each of the cell frequencies by the relevant column sums. Thus, the percentage of:

- country people who are for gun control is:  $\frac{32}{58} \times 100 = 55.2\%$
- country people who are against gun control is:  $\frac{26}{58} \times 100 = 44.8\%$
- city people who are for gun control is:  $\frac{30}{42} \times 100 = 71.4\%$
- city people who are against gun control is:  $\frac{12}{42} \times 100 = 28.6\%$

**Note:** Unless small percentages are involved, it is usual to round percentages to one decimal place in tables.

	<i>Residence</i>	
<i>Attitude to gun control</i>	Country	City
For	55.2%	71.4%
Against	44.8%	28.6%
Total	100.0%	100.0%

### Using percentages to identify relationships between variables

Calculating the values in the table as percentages enables us to compare the attitudes of city and country people on an equal footing. From the table, we see that 55.2% of country people in the sample were for gun control compared to 71.4% of the city people. This means that the city people in the sample were more supportive of gun control than the country people. This reverses what the frequencies showed.

The fact that the percentage of ‘country people for gun control’ differs from the percentage of ‘city people for gun control’ indicates that a person’s attitude to gun control depends on their residence. Thus, we can say that the variables *attitude to gun control* and *residence* are **associated**.

If the variables *attitude to gun control* and *residence* were not associated, we would expect approximately equal percentages of country people and city people to be ‘for’ gun control. Finding a single row in the two-way frequency distribution in which percentages are clearly different is sufficient to identify a relationship between the variables.

We could have also arrived at this conclusion by focusing our attention on the percentages ‘against’ gun control. We might report our findings as follows.

### Report

In this sample of 100 people, a higher percentage of city people were for gun control than country people: 71.4% to 55.2%. This indicates that a person’s attitude to gun control is associated with their place of residence.

We will now consider a two-way percentage frequency table that shows no evidence of a relationship. Consider the following table that summarises responses to the question ‘Should mobile phones be banned in cinemas?’ These responses were obtained from 100 students in Year 10 and Year 12 – we are interested in investigating whether there is an association between these variables.

	<i>Year level</i>	
<i>Should mobile phones be banned in cinemas?</i>	Year 10	Year 12
Yes	87.9%	86.8%
No	12.1%	13.2%
Total	100.0%	100.0%

When we look across the first row of the table, we see that the percentages in favour are very similar. In this case, we might report our findings as follows.

### Report

In this sample of 100 Year 10 and Year 12 students, we see that the percentage of Year 10 and Year 12 students in support of banning mobile phones in cinemas is similar: 87.9% to 86.8%. This indicates that a person’s support for banning mobile phones in cinemas is not associated with their year level.


**Example 5** Identifying and describing an association from a percentaged two-way table ( $2 \times 2$ )

Are males and females in Year 9 equally likely to indicate an intention to go to university? Data from interviews with 200 Year 9 students are summarised in the following table. Write a brief report addressing this question and quoting appropriate percentages.

	Gender		
University	Male	Female	Total
Yes	50	54	104
No	55	41	96
Total	105	95	200

**Explanation**

- Determine the column percentages and complete the table as shown.
- Select an appropriate row to compare the male and female percentages.
- Construct a report.

**Solution**

	Gender	
University	Male	Female
Yes	47.6%	56.8%
No	52.4%	43.2%
Total	100.0%	100.0%

We can see from the top row that a greater proportion of females than males (56.8% compared with 47.6%) were intending to go to university. Report: In this sample of 200 Year 9 students, a greater proportion of females than males (56.8% compared with 47.6%) were intending to go to university. There is an association between gender and intention to go to university.

**Two-way frequency tables for categorical variables taking more than two values**

The table below displays the *smoking status* for a group of adults (smoker, past smoker, never smoked) by *educational level* (Year 9 or less, Year 10 or 11, Year 12, university). This is still a two-way frequency table (because involves two variables), each of these variables can take three values, and so we call this a  $3 \times 3$  table.

Smoking status	Education level (%)			
	Year 9 or less	Year 10 or 11	Year 12	University
Smoker	34.0	31.7	26.5	18.4
Past smoker	36.0	33.8	30.9	28.0
Never smoked	30.0	34.5	42.6	53.6
Total	100.0	100.0	100.0	100.0

Again, we look for an association between variables by comparing the percentages across one of the rows. The following report has been prepared using the percentages in the ‘Smoker’ row.

### Report

From Table 3.3 we see that the percentage of smokers steadily decreases with education level, from 33.9% for Year 9 or below to 18.4% for university. This indicates that smoking is associated with level of education.



### Example 6 Identifying and describing associations from a percentaged two-way table (3 × 3)

A survey was conducted with 1000 males under 50 years old. As part of this survey, they were asked to rate their interest in sport as ‘high’, ‘medium’, or ‘low’. Their age group was also recorded as ‘under 18’, ‘19–25’, ‘26–35’ and ‘36–50’. The results are displayed in the table.

Interest in sport	Age group (%)			
	Under 18 years	19–25 years	26–35 years	36–50 years
High	56.5	50.2	40.7	35.0
Medium	30.1	34.4	36.8	44.7
Low	13.4	15.4	22.5	20.3
Total	100.0	100.0	100.0	100.0

- Which is the explanatory variable, *interest in sport* or *age group*?
- Is there an association between *interest in sport* and *age group*? Write a brief response quoting appropriate percentages.

#### Explanation

- Age is a possible explanation for the level of interest in sport, but interest in sport cannot explain age.

#### Solution

*Age group* is the EV.

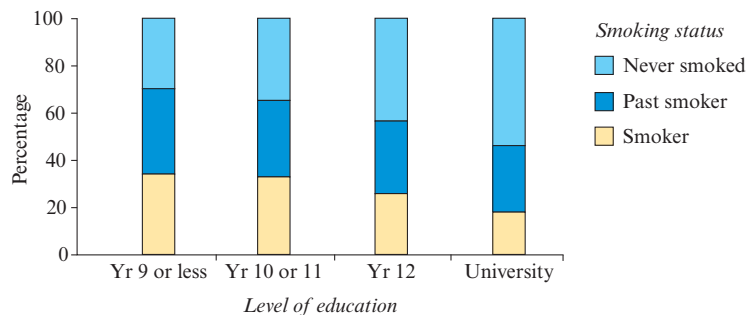
**b** If we look across all rows, we can see that the percentages are different for each age group. Select one row to compare and discuss – here we have chosen ‘high’.

There is an association between the level of interest in sport and age. A high level of interest in sport is seen to decrease steadily across the age categories from 56.5% for under 18 years, 50.2% for 19–25 years, 40.7% for 26–35 years to, at its lowest, 35% for 36–50 years.

## The segmented bar chart

A visual display which can be used to display the information in a two-way frequency table is a **segmented bar chart**. A segmented bar chart consists of separate bars for each value of the explanatory variables, with each bar separated into parts (segments) that show the percentage for each value of the response variable.

The following segmented bar chart below that has been constructed from the table displaying the smoking status of adults (smoker, past smoker, never smoked) by level of education (Year 9 or less, Year 10 or 11, Year 12, university).

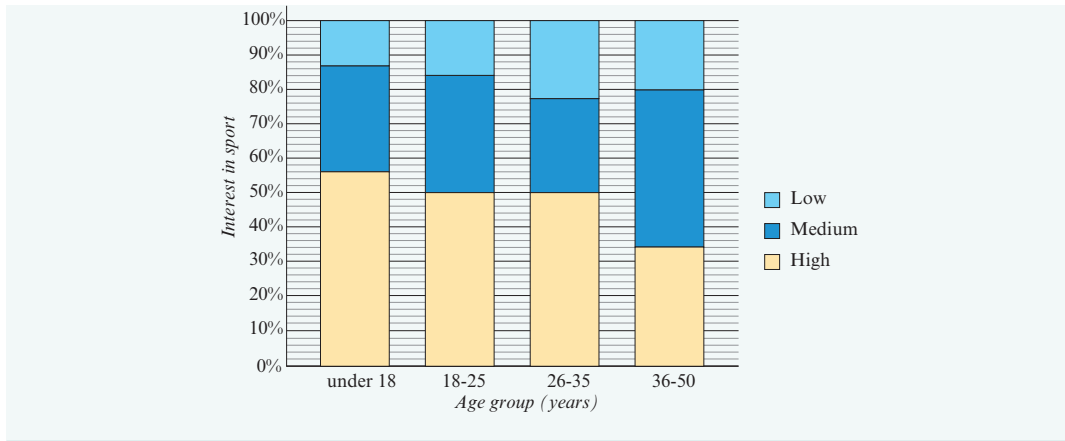


### Example 7 Constructing a segmented bar chart

Construct a segmented bar chart to display the association *interest in sport* and *age group* displayed in the table in Example 6.

#### Solution

- 1 Since *age group* is the EV, this variable will label the horizontal axis.
- 2 The vertical axis should be scaled from 0% to 100%, in intervals of 10%.
- 3 There will be a bar for each value of *age group*, that is, a bar for each column of the table.
- 4 Mark off, and colour, with each value of *interest in sport* assigned in the same colour.
- 5 Add a Key showing which colour has been assigned to each value of *interest in sport*.

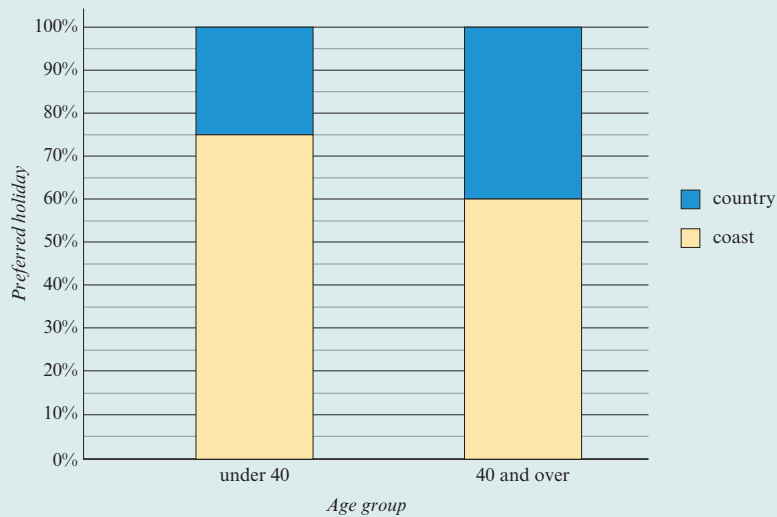


The percentage segmented bar chart allows an easier visual comparison of the percentages than does the percentage two-way table, and can be used to investigate the association between two categorical variables, as shown in the following example.



### Example 8 Identifying and describing associations from a segmented bar chart

The percentage segmented bar chart below shows the association between *preferred holiday* (country or coast) and *age group* (under 40, 40 or over) for a sample of 800 visitors to a travel website.



Does the percentage segmented bar chart support the contention there is an association between *preferred holiday* and *age group*?

#### Explanation

If we look across all rows, we can see that the heights of the segments, and thus the percentages, are different for each age group. Select one row to compare and discuss – here we have chosen ‘coast’.

#### Solution

There is an association between the holiday preference and age. Those aged under forty are more likely to choose a coastal holiday (75%) than those aged forty or over (60%).





## Exercise 2B

### Constructing a two-way frequency table

Example 4

- 1 The following data were obtained when a sample of 20 Year 12 students were asked if they intended to go to university (*university*). The *gender* of the student was also recorded.

Student No.	Gender	Intends to go to university	Student No.	Gender	Intends to go to university
1	F	Yes	11	F	Yes
2	M	Yes	13	M	Yes
3	F	No	13	F	No
4	F	Yes	14	F	Yes
5	M	No	15	M	No
6	M	Yes	16	M	Yes
7	F	Yes	17	F	Yes
8	M	No	18	M	No
9	F	No	19	F	No
10	F	Yes	20	F	Yes

- a Identify which is the explanatory and which is the response variable.  
 b Create a two-way frequency table from the data, with the values of the explanatory variable labelling the columns.

Example 5

- 2 The following data were obtained when a sample of 30 adults were asked if they supported *reducing university fees*. They were also classified by their *age group*: 17–18 years, 19–25 years, or 26 years or more. The results are given in the table below.

Age group	Reduce fees	Age group	Reduce fees	Age group	Reduce fees
17–18	Yes	26 or more	Yes	26 or more	No
19–25	Yes	17–18	Yes	19–25	Yes
26 or more	No	19–25	Yes	17–18	No
17–18	Yes	17–18	Yes	26 or more	Yes
19–25	Yes	17–18	Yes	17–18	No
26 or more	Yes	26 or more	No	26 or more	Yes
17–18	Yes	19–25	Yes	19–25	Yes
19–25	No	26 or more	Yes	17–18	Yes
26 or more	No	17–18	No	19–25	No
19–25	No	17–18	Yes	26 or more	Yes

- a Identify which variable is the explanatory variable and which is the response variable.
- b Create a two-way frequency table from these data, with the values of the explanatory variable labelling the columns.
- c Calculate the column percentages for the table.

Using two-way tables to identify associations between two categorical variables

Example 6

3 A survey was conducted with 242 university students. For this survey, data were collected on the students' *enrolment status* (full-time, part-time) and whether or not each *drinks alcohol* ('Yes' or 'No'). Their responses are summarised in the table opposite.

<i>Drinks alcohol</i>	<i>Enrolment status (%)</i>	
	Full-time	Part-time
Yes	80.5	81.8
No	19.5	18.2
Total	100.0	100.0

- a Which variable is the explanatory variable?
- b Is there an association between drinking alcohol and enrolment status? Write a brief report quoting appropriate percentages.

4 The table opposite was constructed from data collected to see if *handedness* (left, right) was associated with *gender* (male, female).

<i>Handedness</i>	<i>Gender</i>	
	Male	Female
Left	22	16
Right	222	147

- a Which variable is the response variable?
- b Convert the table to percentages by calculating column percentages.
- c Is *handedness* associated with *gender*? Write a brief explanation using appropriate percentages.

5 A survey was conducted with 59 students studying Business and 51 students studying Arts at university to determine whether they exercise, 'regularly', 'sometimes' or 'rarely'. Their responses are summarised in the percentaged two-way frequency table.

<i>Exercise</i>	<i>Course (%)</i>	
	Business	Arts
Rarely	28.8	39.2
Sometimes	52.5	54.9
Regularly	18.6	5.9
Total	99.9	100.0

- a Which is the explanatory variable?
- b Is the variable *exercise* nominal or ordinal?
- c What percentage of Arts students exercised sometimes?
- d Is there an association between how regularly these students exercise and their course? Write a brief response quoting appropriate percentages.

**Example 7**

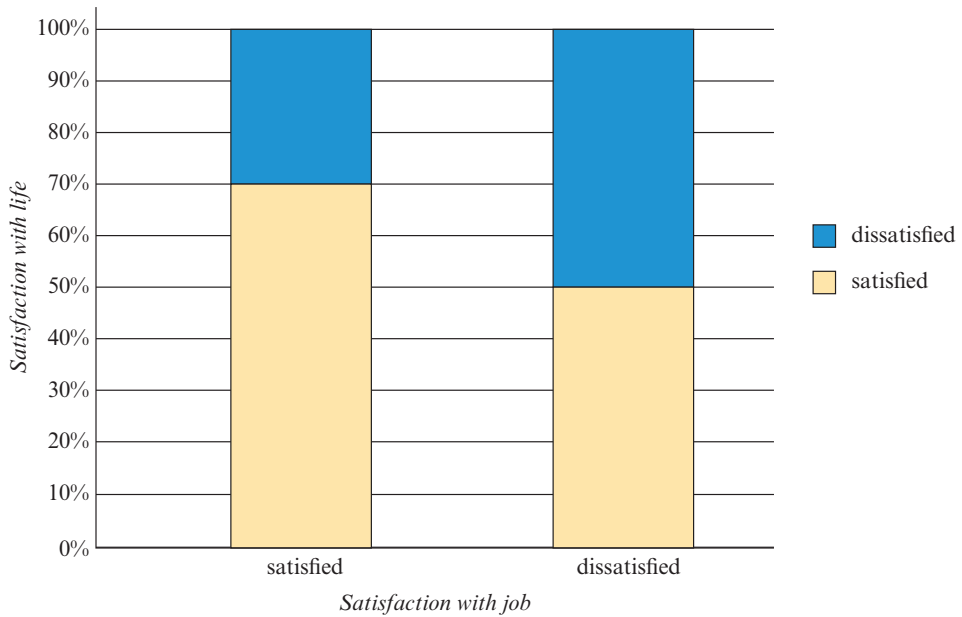
- 6 It was suggested that students in Dr Evans' mathematics class would achieve higher grades than students in Dr Smith's mathematics class. The following table shows the results for each class that year.

<i>Exam grade</i>	<i>Class</i>		<i>Total</i>
	<i>Dr Evans</i>	<i>Dr Smith</i>	
Fail	2	3	5
Pass	11	20	31
Credit or above	5	9	14
Total	18	32	50

- a Construct a percentaged two-way frequency table.
- b Construct a percentaged segmented bar chart.
- c Write a brief report on the association between teacher and grade.

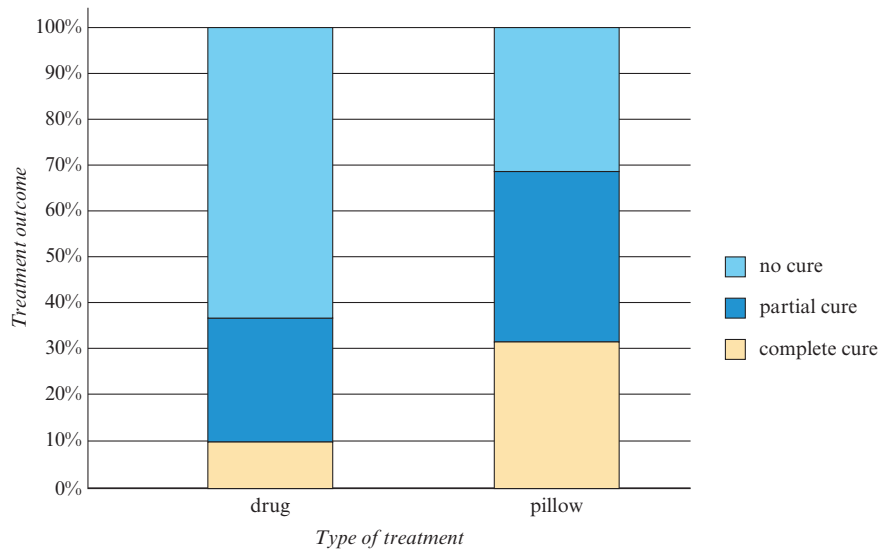
**Example 8**

- 7 Are those people who are satisfied with their job more likely to be satisfied with their life? Data collected from a survey of 200 adults are summarised in the following percentaged segmented bar chart.



Does the data support the contention that people who are satisfied with their job are more likely to be satisfied with their life? Write a brief report quoting appropriate percentages.

- 8 Researchers predicted that using a special pillow would be more effective in curing snoring than treatment with drugs. The association between the outcome of treatment and type of treatment is shown in the following percentaged segmented bar chart.



- a** Identify which variable is the explanatory variable and which is the response variable.
- b** Does the data support the contention the special pillow is more effective at treating snoring than the drug treatment? Write a brief report quoting appropriate percentages.
- 9** As part of the General Social Survey conducted in the US, respondents were asked to say whether they found life *exciting*, *pretty routine* or *dull*. Their marital status was also recorded as married, widowed, divorced, separated or never married. The results are organised into a table as shown.

Attitude to life	Marital status (%)				
	Married	Widowed	Divorced	Separated	Never
Exciting	47.6	33.8	46.7	45.9	52.3
Pretty routine	48.7	54.3	47.6	44.6	44.4
Dull	3.7	11.9	6.7	9.5	3.2
Total	100.1	100.0	100.0	100.0	99.9

- a** What percentage of widowed people found life ‘dull’?
- b** What percentage of people who were never married found life ‘exciting’?
- c** What is the likely explanatory variable in this investigation?
- d** Is the variable *attitude to life* nominal or ordinal?
- e** Does the information you have been given support the contention that a person’s attitude to life is related to their marital status? Justify your argument by quoting appropriate percentages.

## Exam 1 style questions

Use the following information to answer Questions 10-12

The data in the following table was collected to investigate the association between tertiary qualifications and happiness.

	Tertiary qualification		
Happy with life	Yes	No	Total
Yes	116	138	254
No	12	34	46
Total	128	172	300

- 10** The percentage of participants in the study who do not have a tertiary education is closest to:
- A** 57.3%      **B** 80.2%      **C** 54.3%      **D** 11.3%      **E** 19.8%
- 11** Of those people in the study who did not have a tertiary education, the percentage who are happy with their lives is closest to:
- A** 57.3%      **B** 80.2%      **C** 54.3%      **D** 11.3%      **E** 19.8%
- 12** The data in the table supports the contention that there is an association between *tertiary qualifications* and *happiness* because:
- A** 84.7% of people are happy.
- B** more people without a tertiary qualification are happy than people with a tertiary qualification.
- C** 90.6% of people with a tertiary qualification are happy, compared to 80.2% of those without a tertiary qualification.
- D** 54.3% of happy people do not have a tertiary qualification.
- E** 57.3% of people do not have a tertiary qualification, compared to 42.7% who do.

## 2C Investigating the association between a numerical and a categorical variable

### Learning intentions

- ▶ To be able to use parallel dot plots to identify and describe the association between a numerical variable and a categorical variable for small data sets.
- ▶ To be able to use back-to-back stem plots to display and describe the association between a numerical variable and a categorical variable for small data sets.
- ▶ To be able to use parallel boxplots to display the association between a numerical variable and a categorical variable which can take two or more values.

In the previous section, we learned how to identify and describe associations between two categorical variables. In this section, we will learn to identify and describe associations between a numerical variable and a categorical variable. Suppose, for example, we wish to investigate the association between attendance at a revision class, and test score. Here we can actually identify two variables. One is the variable *test score*, a numerical variable, and the other is the variable *attended revision class*, which is a categorical variable taking the values ‘yes’ or ‘no’.

The outcome of such an investigation will be a brief written report that compares the distribution of the numerical variable across two or more groups, the number of groups equal to the number of values which the categorical variable can take. The starting point for these investigations will be, as always, a graphical display of the data. Here our options are **parallel dot plots**, **back-to-back stem plots** or the **parallel boxplots**.

Using a graphical display of the data, as well as the values of the relevant summary statistics, we can compare the distributions of the numerical variable for each value of the categorical variable according to:

- shape
- centre
- spread

If any of these are noticeably different for differing values of the categorical variable we will conclude that the two variables are associated. Because it is often difficult to clearly identify the shape of a distribution with a small amount of data, we usually confine ourselves to comparing centre and spread, using the medians and *IQRs*, when using dot plots and back-to-back stem plots.

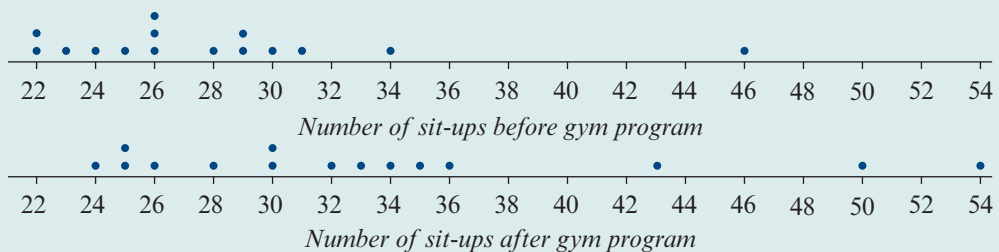
## Using parallel dot plots and back-to-back stem plots to identify and describe associations

For small data sets, parallel dot plots and back-to-back stem plots are ideal displays for identifying and describing associations between a numerical and a categorical variable.



### Example 9 Using a parallel dot plot to identify and describe associations

The parallel dot plots below display the distribution of the number of sit-ups performed by 15 people before and after they had completed a gym program.



Do the parallel dot plots support the contention that the number of sit-ups performed is associated with completing the gym program? Write a brief explanation that compares the distributions.

**Solution**

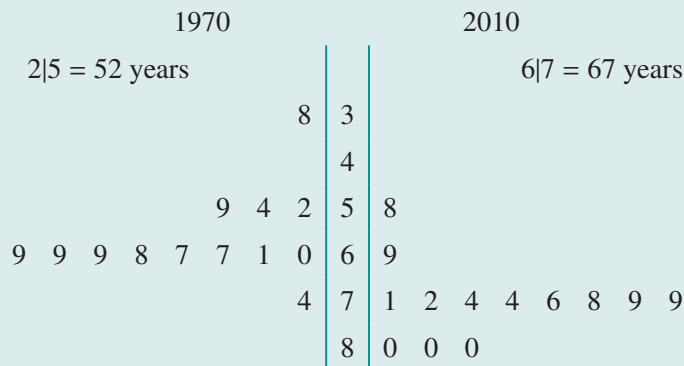
Here the numerical variable *sit-ups* is the response variable and the categorical variable *gym program*, taking the values 'before' and 'after', is the explanatory variable.

- 1 Locate the median number of sit-ups performed before and after the gym program,  $M = 26$  and  $M = 32$  sit-ups respectively.
- 2 Determine the *IQR* of sit-ups performed before and after the gym program,  $IQR = 6$  and  $IQR = 10$  sit-ups respectively.
- 3 There are reasonable difference in both the median and *IQR* of the number of sit-ups performed before and after the gym program, evidence that the number of sit-ups performed is associated with completing the gym program. Report your conclusion, backed up by a brief explanation.

The median number of sit-ups performed after attending the gym program ( $M = 32$ ) is higher than the median number of sit-ups performed before attending the gym program ( $M = 26$ ). The variability in the number of sit-ups has also increased from  $IQR = 6$  to  $IQR = 10$ . Thus we can conclude that the number of sit-ups performed is associated with completing the gym program.

**Example 10** Using a back-to-back stem plot to identify and describe associations

The back-to-back stem plot below displays the distribution of life expectancy (in years) for the same 13 countries in 1970 and 2010.



Do the back-to-back stem plots support the contention that life expectancy has changed between these two time periods?

**Solution**

Here the numerical variable *life expectancy* is the response variable and the variable *year* is the explanatory variable. While *year* can be considered a numerical variable, because it is only taking two values (1970 and 2010) we are treating it as a categorical variable in this example.

- 1 Determine the median life expectancies for 1970 and 2010. You should find them to be 67 and 76 years, respectively.



- 2 Determine the quartiles, and hence the values of the *IQR* for 1970 and 2010. You should find them to be 12.5 years and 8 years respectively.
- 3 These differences in median and *IQR* between 1970 and 2010 are sufficient to conclude that the distribution of life expectancy had changed over this time period. Report your conclusion, supported by a brief explanation.

Report: There is an association between year and life expectancy. The median life expectancy has increased between 1970 and 2010, from  $M = 67$  years to  $M = 76$  years. Over the same period life expectancy has also become less variable (*IQR* in 1970 = 12.5 years; *IQR* in 2010 = 8 years).

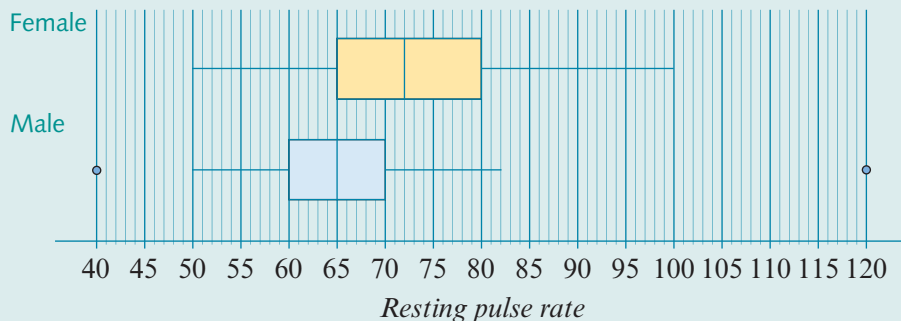
## Using parallel boxplots to identify and describe associations

The statistical tool most commonly used for investigating associations between a numerical and a categorical variable is the **parallel boxplot**. In a parallel boxplot, there is one boxplot for each category of the categorical variable. Associations can then be identified by comparing the way in which the distribution of the numerical variable changes between categories in terms of shape, centre and spread. We should also mention outliers when describing the distributions.



### Example 11 Comparing distributions across two groups using parallel boxplots

Use the following parallel boxplots to compare the pulse rates (in beats/minute) for a group of 70 male students and 90 female students.



#### Solution 0.5

Here the numerical variable *resting pulse rate* is the response variable and the categorical variable *gender* is the explanatory variable.

- 1 Compare the medians: The median for females is about 72, which is higher than that for males, which is about 65.
- 2 Compare the spread: The *IQR* for females is 15, which is more than the *IQR* for males, which is 10.

- 3 Compare the shape: Both distributions are approximately symmetric.
- 4 Locate any outliers. There are two outliers for the males, one at 40 and one at 120.
- 5 Write the report comparing the distributions.

### Report

There is an association between resting pulse rate and gender. On average, the resting pulse rate for males is lower (median: male = 65, female = 72) and less variable than that for females (*IQR*: male = 10, female = 15). The distributions of resting pulse rates for both male and female students were approximately symmetric. One male was found to have an extremely low pulse rate of 40, while another had an extremely high pulse rate of 120.



### Example 12 Comparing distributions across more than two groups using parallel boxplots

Use the parallel boxplots below to compare the salary distribution for workers in a certain industry across four different age groups: 20–29 years, 30–39 years, 40–49 years and 50–65 years.



### Solution

Here the numerical variable *salary* is the *response* variable and the categorical variable *age group* is the *explanatory* variable.

- 1 Compare the medians: The median salary increased from \$64 000 for 20–29 year-olds to \$72 000 for 50–65 year-olds.
- 2 Compare the *IQR*s: The *IQR* increased from around \$12 000 for 20–29-year-olds to around \$20 000 for 50–65-year-olds.
- 3 Comparing the shapes: The shape of the distribution of salaries changes with the age group, from symmetric to positively skewed.
- 4 Locate the outliers: There are no outliers in the 20–29 and 30–39 age group. Outliers also begin to appear at \$110 000 for the 40–49 age group, and at \$119 000, \$126 000 and \$140 000 for the 50–65 age group.
- 5 Write the report comparing the distributions.

## Report

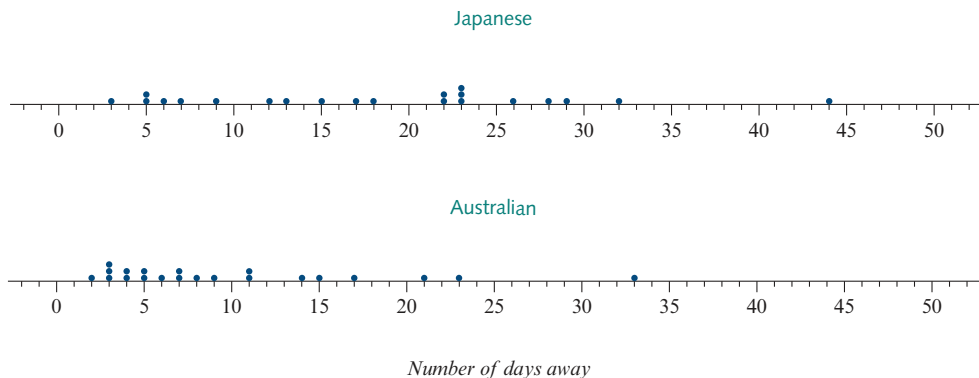
In this industry there is an association between salary and age group. The median salaries increase across the age groups, from \$64 000 for 20–29 year-olds to \$72 000 for 50–65 year-olds. The salaries also became more variable, with the *IQR* increasing from around \$12 000 for 20–29-year-olds to around \$20 000 for 50–65-year-olds. The shape of the distribution of salaries changes with age group, from symmetric for 20–29-year-olds, to progressively more positively skewed as age increases. There are no outliers in the 20-29 and 30-39 age group. Outliers also begin to appear at \$110 000 for the 40-49 age group, and at \$119 000, \$126 000 and \$140 000 for the 50-65 age group.



## Exercise 2C

## Example 9

- 1 Data was collected to compare the the number of days spent away from home (*number of days away*) by 21 tourists from each of Japan and Australia (*country of origin*). The data collected is displayed in the parallel dot plots below.



- a Identify each of the variables and classify each as categorical or numerical.
- b Use the parallel dot plots to compare the distributions of *number of days away* for the two different nationalities.

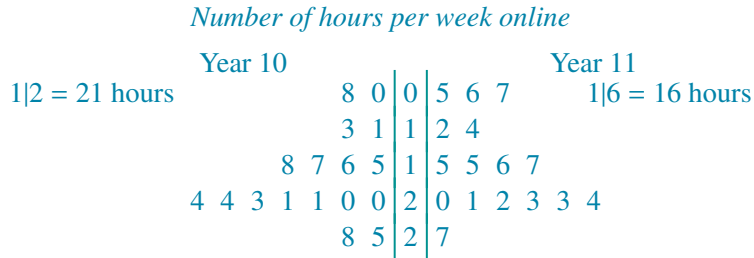
## Example 10

- 2 The back-to-back stem plot shown compares the distribution of the *age* patients (in years) admitted to a small hospital during one week, and their *gender*.

- a Classify each variable as categorical or numerical.
- b Do the back-to-back plots support the contention that the age of the patients is associated with their gender? Write a brief explanation that compares these distributions in terms of centre and spread.

Females		Males
9	0	
5 0	1	3 6
7	2	1 4 5 6 7
7 1	3	4
3 0	4	0 7
0	5	
	6	
9	7	
0 4 = 40 years		4 0 = 40 years

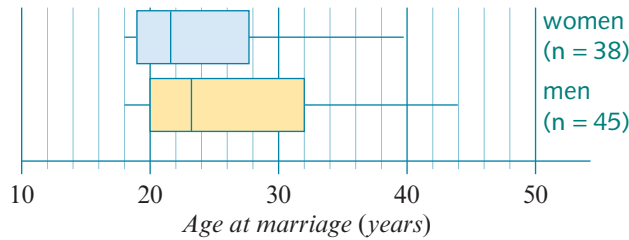
- 3 The following back-to-back stem plot displays the distributions of *the number of hours per week spent online* by a group of students, and their *year level*.



- a Classify each of the variables as categorical or numerical.
- b Use the stem plots to compare these distributions in terms of centre and spread. Draw an appropriate conclusion about the association between year level and the number of hours students spend online each week.

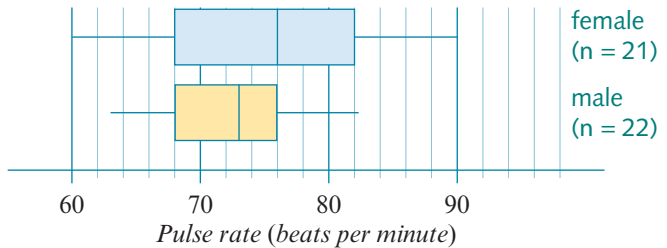
**Example 11**

- 4 The parallel boxplots show the distribution of ages of 45 men and 38 women when first married.



- a Identify each of the variables and classify as categorical or numerical.
- b Use the boxplots to compare these distributions in terms of shape, centre and spread and draw an appropriate conclusion about the association between gender and the age when first married.

- 5 The parallel boxplots show the distribution of pulse rates of 21 females and 22 males.

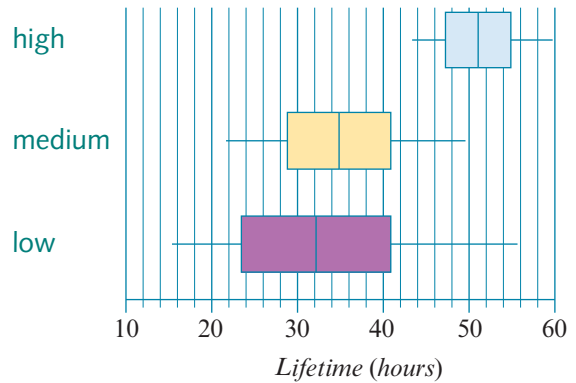


- a Identify each of the variables and classify as categorical or numerical.
- b Use the boxplots to compare these distributions, and draw an appropriate conclusion about the association between gender and pulse rate.

## Example 12

**6** The parallel boxplots show the distribution of the lifetime (in hours) of three differently priced batteries (low, medium, high).

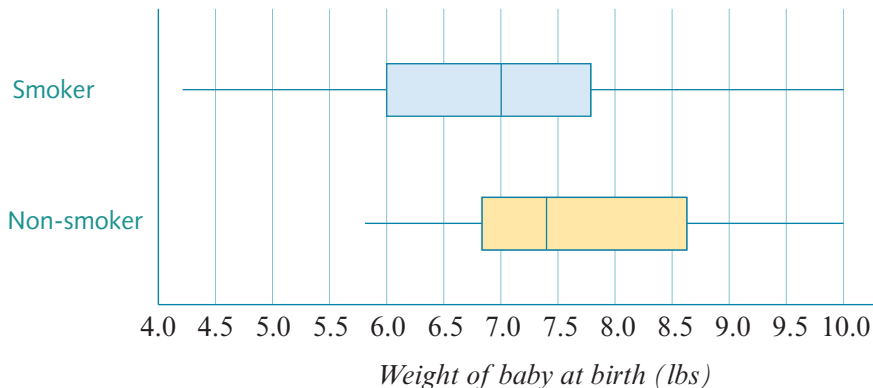
- a** The two variables displayed here are battery *lifetime* and battery *price* (low, medium, high). Which is the numerical and which is the categorical variable?
- b** Do the parallel boxplots support the contention that battery lifetime depends on price? Write a brief explanation.



## Exam 1 style questions

Use the following information to answer Questions 7 and 8

The data in the following boxplots was collected to investigate the association between smoking and the birth weight of babies.



Use the information in the boxplots to answer the following questions.

- 7** Which of the following statements is true:
- A** 75% of babies born to non-smokers weigh more than the lightest 50% of babies born to smokers.
- B** 50% of babies born to non-smokers weigh more than the heaviest 25% of babies born to smokers.
- C** 25% of the babies born to smokers weigh less than all the babies born to non-smokers
- D** All of the babies born to non-smokers weigh more than the heaviest 75% of the babies born to smokers
- E** The range of baby weights for smokers is less than the range of baby weights for non smokers.

- 8 The information in the boxplots supports the contention that there is an association between *smoking* and *weight of baby at birth* because:
- A The *IQRs* of birthweight for both groups are approximately the same.
  - B The median birthweight for smokers is more than the median birthweight for non-smokers.
  - C The *IQRs* of birthweight for both groups are very different.
  - D The median birthweight for smokers is less than the median birthweight for non-smokers.
  - E Both distributions are approximately symmetric.

## 2D Investigating associations between two numerical variables

### Learning intentions

- ▶ To be able to introduce the scatterplot for displaying data from two numerical variables.
- ▶ To be able to construct a scatterplot using a CAS calculator.

In this section, we will learn to identify and describe associations between two numerical variables. Suppose, for example, we wish to investigate the association between university *participation rate* (the EV) and average *hours worked* (the RV) in nine countries. The starting point for this investigation is again a graphical display of the data. Here our options are to construct a scatterplot. The data for 9 countries are shown below.

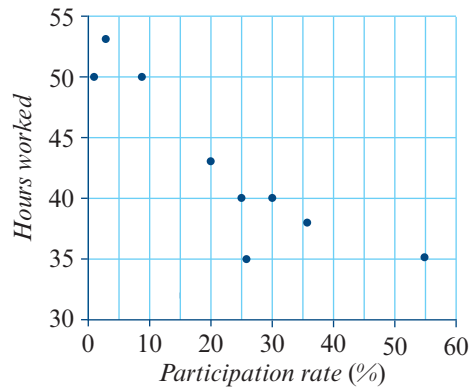
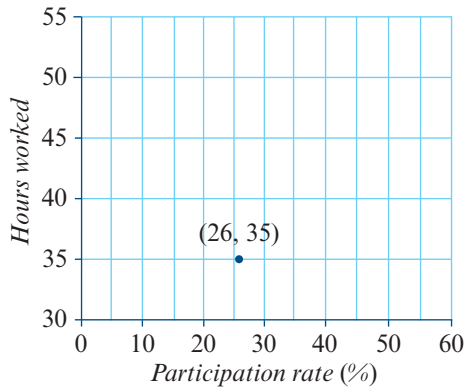
<i>Participation rate (%)</i>	26	20	36	1	25	9	30	3	55
<i>Hours worked</i>	35	43	38	50	40	50	40	53	35

The first step in investigating an association between two numerical variables is to construct a visual display of the data, which we call a **scatterplot**.

### The scatterplot

- A scatterplot is a plot which enables us to display bivariate data when **both of the variables are numerical**.
- In a scatterplot, each point represents a single case.
- When constructing a scatterplot, it is conventional to use the **vertical** or **y-axis** for the response variable (RV) and the **horizontal** or **x-axis** for the explanatory variable (EV).

The scatterplot below left shows the point for a country for which the university participation rate is 26% and average hours worked is 35, and the scatterplot below right is the completed scatterplot when each of the remaining countries are plotted.



### CAS 1: How to construct a scatterplot using the TI-Nspire CAS

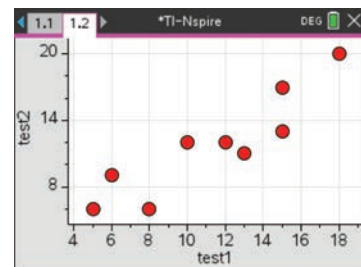
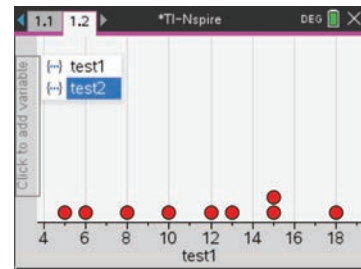
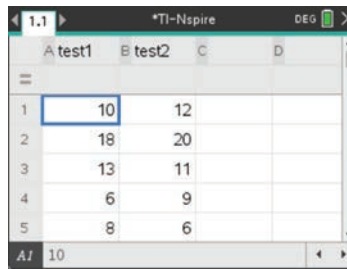
Construct a scatterplot for the set of test scores given below.

Treat *test 1* as the explanatory (i.e.  $x$ ) variable.

<i>Test 1</i>	10	18	13	6	8	5	12	15	15
<i>Test 2</i>	12	20	11	9	6	6	12	13	17

#### Steps

- 1 Start a new document by pressing **ctrl** + **N**.
- 2 Select **Add Lists & Spreadsheet**. Enter the data into lists named *test1* and *test2*.
- 3 Press **ctrl** + **I** and select **Add Data & Statistics**.
- 4 **a** Click on **Click to add variable** on the  $x$ -axis and select the explanatory variable *test1*.  
**b** Click on **Click to add variable** on the  $y$ -axis and select the response variable *test2*. A scatterplot is displayed. The plot is scaled automatically.



### CAS 1: How to construct a scatterplot using the ClassPad

Construct a scatterplot for the set of test scores given below.


Treat *test 1* as the explanatory (i.e.  $x$ ) variable.



<i>Test 1</i>	10	18	13	6	8	5	12	15	15
<i>Test 2</i>	12	20	11	9	6	6	12	13	17


**Steps**


1 Open the **Statistics** application and enter the data into the columns named *test1* and *test2*.

2 Tap  to open the **Set StatGraphs** dialog box and complete as given below.

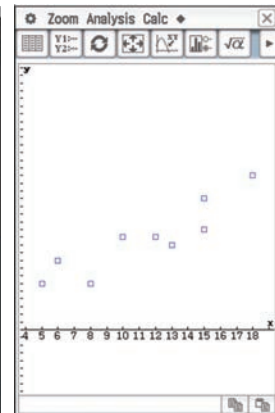
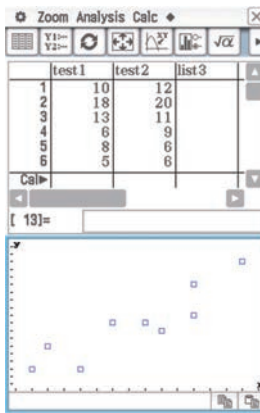
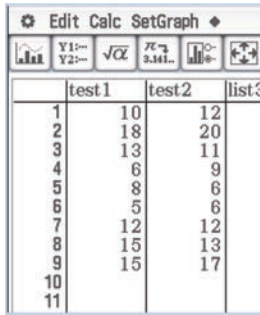
- **Draw:** select **On**.
- **Type:** select **Scatter** (▼).
- **XList:** select **main\test1** (▼).
- **YList:** select **main\test2** (▼).
- **Freq:** leave as **1**.
- **Mark:** leave as **square**.

Tap **Set** to confirm your selections.

3 Tap  in the toolbar at the top of the screen to plot the scatterplot in the bottom half of the screen.

4 To obtain a full-screen plot, tap  from the icon panel.

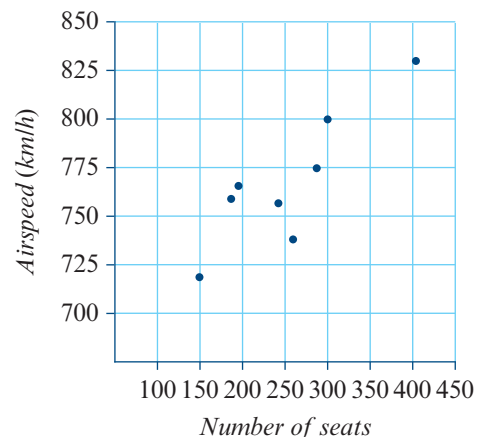
**Note:** If you have more than one graph on your screen, tap the data screen, select StatGraph and turn off any unwanted graphs.



**Exercise 2D**

1 The scatterplot opposite has been constructed to investigate the association between the airspeed (in km/h) of commercial aircraft and the number of passenger seats. Use the scatterplot to answer the following questions.

- a Which is the explanatory variable?
- b What type of variable is airspeed?
- c How many aircraft were investigated?
- d What was the airspeed of the aircraft that has 300 seats?



2

<i>Minimum temperature (x)</i>	17.7	19.8	23.3	22.4	22.0	22.0
<i>Maximum temperature (y)</i>	29.4	34.0	34.5	35.0	36.9	36.4

The table above shows the maximum and minimum temperatures (in °C) during a hot week in Melbourne.

- a** Enter the data into your calculator, naming the variables *mintemp* and *maxtemp*.  
**b** Construct a scatterplot with minimum temperature as the EV.

3

<i>Balls faced</i>	29	16	19	62	13	40	16	9	28	26	6
<i>Runs scored</i>	27	8	21	47	3	15	13	2	15	10	2

The table above shows the number of runs scored and the number of balls faced by batsmen in a 1-day international cricket match. Use a calculator to construct an appropriate scatterplot.

4

<i>Temperature (°C)</i>	0	10	50	75	100	150
<i>Diameter (cm)</i>	2.00	2.02	2.11	2.14	2.21	2.28

The table above shows the changing diameter of a metal ball as it is heated. Use a calculator to construct an appropriate scatterplot, with temperature as the EV.

5

<i>Number in theatre</i>	87	102	118	123	135	137
<i>Time (minutes)</i>	0	5	10	15	20	25

The table above shows the number of people in a movie theatre at 5-minute intervals after the advertisements started. Use a calculator to construct an appropriate scatterplot.

### Exam 1 style questions

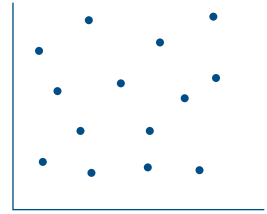
- 6 For which one of the following pairs of variables would it be appropriate to construct a scatterplot?
- A** *eye colour* (blue, green, brown, other) and *hair colour* (black, brown, blonde, other)  
**B** *test score* and *sex* (male, female)  
**C** *political party preference* (Labor, Liberal, Other) and *age* in years  
**D** *age* in years and *blood pressure* in mmHg  
**E** *height* in cm and *sex* (male, female)

## 2E How to interpret a scatterplot

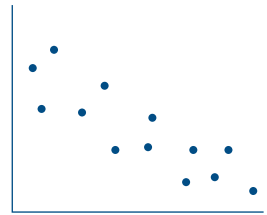
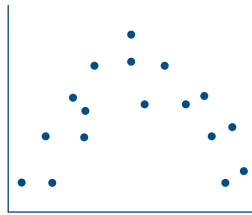
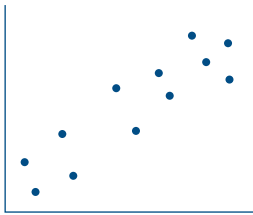
### Learning intentions

- ▶ To be able to use a scatterplot to identify an association between two variables.
- ▶ From the the scatterplot, be able to classify an association according to:
  - ▷ Direction, which may be positive or negative.
  - ▷ Form, which may be linear or non-linear.
  - ▷ Strength, which may be weak, moderate or strong.

What features do we look for in a scatterplot to help us identify and describe any associations present? First we look to see if there is a **clear pattern** in the scatterplot. In the scatterplot opposite, there is **no clear pattern** in the points. The points are **randomly scattered** across the plot, so we conclude that there is **no association**.



For the three examples below, there is a clear (but different) pattern in each set of points, so we conclude that there is an association in each case.

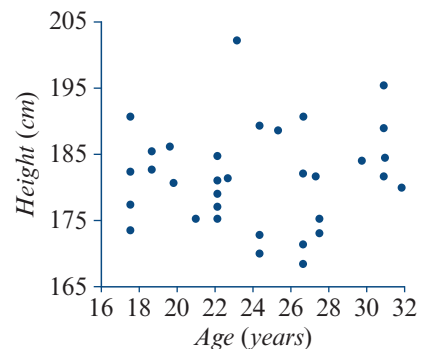


Having found a clear pattern, we need to be able to describe these associations clearly, as they are obviously quite different. The three features we look for in the pattern of points are **direction**, **form** and **strength**. Having found a clear pattern, there are several things we look for in the pattern of points. These are:

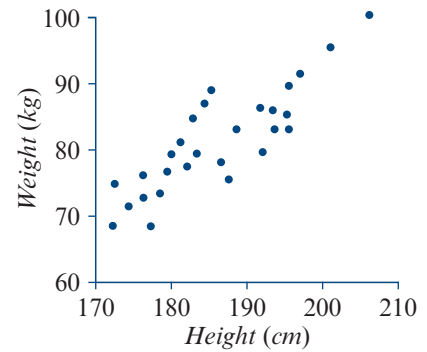
- direction and outliers (if any)
- form
- strength.

### Direction and outliers

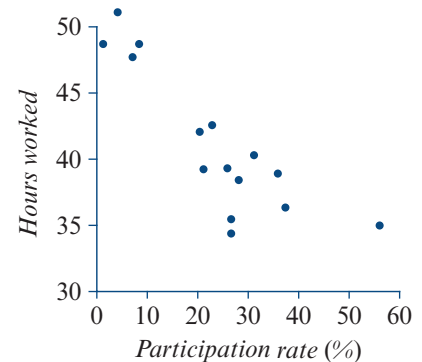
The scatterplot of height against age for a group of footballers (shown opposite) is just a random scatter of points. This suggests that there is **no association** between the variables *height* and *age* for this group of footballers. However, there is a possible outlier for *height*; a footballer who is 201 cm tall.



In contrast, there is a clear pattern in the scatterplot of *weight* against *height* for these footballers (shown opposite). The two variables are associated. If the points in the scatterplot trend upwards as we go from left to right we say there is a **positive association** between the variables. In this example the positive association means that taller players tend to be heavier. In this scatterplot, there are no outliers.



Likewise, the scatterplot of working hours against university participation rates for 15 countries shows a clear pattern. The two variables are associated. If the points in the scatterplot trend downwards as we go from left to right we say there is a **negative association** between the variables. In this example the negative association means that countries with university participation rate tend to work fewer hours. In this scatterplot, there are no outliers.



In general terms, we can classify the direction of an association as follows.

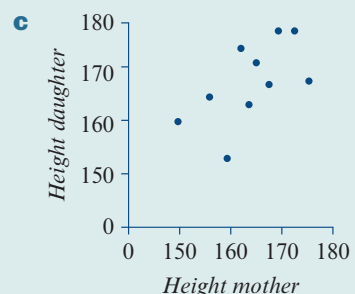
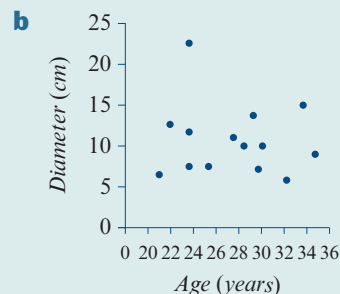
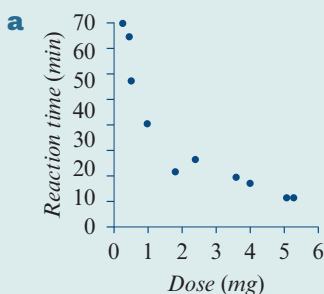
### Direction of an association

- Two variables have a **positive association** when the value of the response variable tends to increase as the value of the explanatory variable increases.
- Two variables have a **negative association** when the value of the response variable tends to decrease as the value of the explanatory variable increases.
- Two variables have **no association** when there is no consistent change in the value of the response variable when the values of the explanatory variable increase.



### Example 13 Direction of association

Classify each of the following scatterplots as exhibiting positive, negative or no association. Where there is an association, describe the direction of the association in terms of the variables in the scatterplot and what it means in terms of the variables involved.



**Explanation**

- a** There is a clear pattern in the scatterplot. The points in the scatterplot trend *downwards* from left to right.
- b** There is no pattern in the scatterplot of *diameter* against *age*.
- c** There is a clear pattern in the scatterplot. The points in the scatterplot trend *upwards* from left to right.

**Solution**

The direction of the association is **negative**. Reaction times tend to decrease as the drug dose increases.

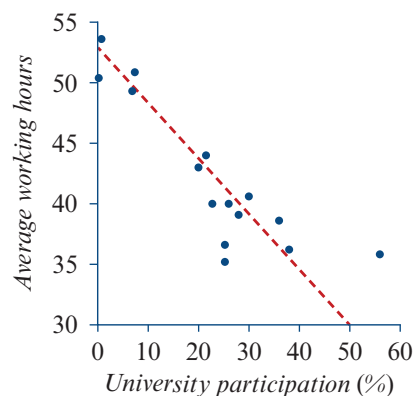
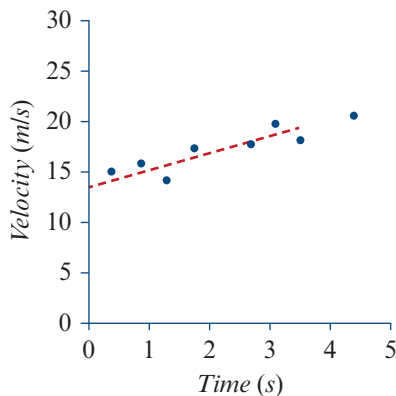
There is no association between diameter and age.

The direction of the association is **positive**. Taller mothers tend to have taller daughters.

**Form**

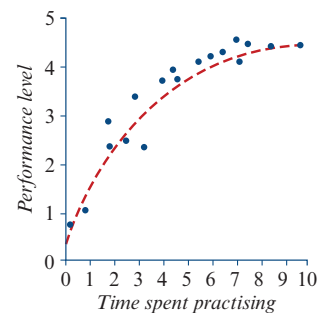
The next feature that interests us in an association is its general form. Do the points in a scatterplot tend to follow a linear pattern or a curved pattern? If the scatterplot has a linear form then we say that the association between the variables is **linear**.

For example, both of the scatterplots below can be described as having a linear form; that is, the scatter in the points can be thought of as scattered around a straight line. (The dotted lines have been added to the graphs to make it easier to see the linear form.)



By contrast, consider the scatterplot opposite, plotting performance level against time spent practising a task. There is an association between performance level and time spent practising, but it is clearly non-linear.

This scatterplot shows that while level of performance on a task increases with practice, there comes a time when the performance level will no longer improve substantially with extra practice.



While non-linear relationships exist (and we must always check for their presence by examining the scatterplot), many of the relationships we meet in practice are linear or can be made linear by transforming the data (a technique you will meet in Chapter 4). For this reason we will restrict ourselves to the analysis of scatterplots with linear forms for now.

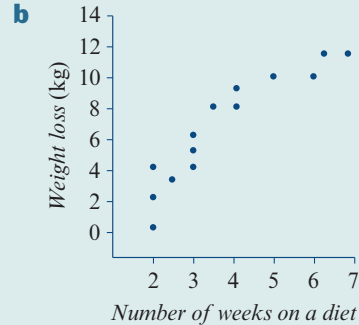
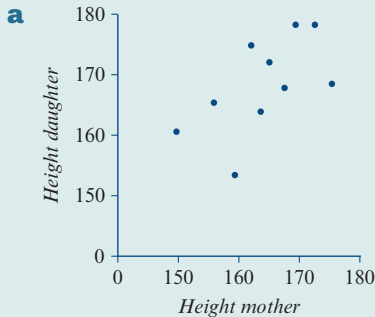
In general terms, we can describe the **form of an association** as follows.

**Form**

A scatterplot is said to have a **linear form** when the points tend to follow a straight line. A scatterplot is said to have a **non-linear form** when the points tend to follow a curved line.

**Example 14** Form of an association

Classify the form of the association in each of scatterplot as linear or non-linear.

**Explanation**

- a** There is a clear pattern.  
The points in the scatterplot can be imagined to be scattered around a straight line.
- b** There is a clear pattern.  
The points in the scatterplot can be imagined to be scattered around a curved line rather than a straight line.

**Solution**

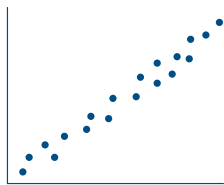
The association is linear.

The association is non-linear.

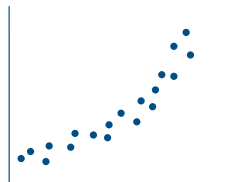
**Strength**

The **strength of an association** is a measure of how much scatter there is in the scatterplot.

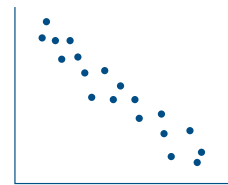
When there is a **strong association** between the variables, there is only a small amount of scatter in the plot, and a pattern is clearly seen.



Strong positive association

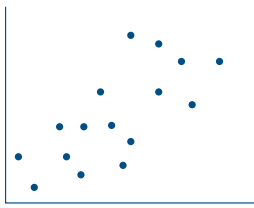


Strong positive association

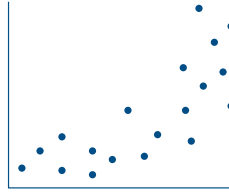


Strong negative association

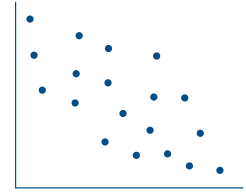
As the amount of scatter in the plot increases, the pattern becomes less clear. This indicates that the association is less strong. In the examples below, we might say that there is a **moderate association** between the variables.



Moderate positive association

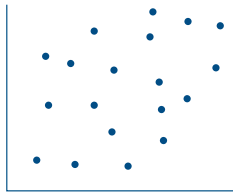


Moderate positive association

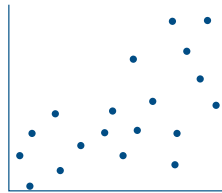


Moderate negative association

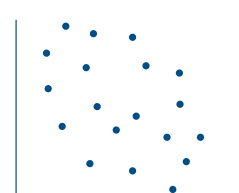
As the amount of scatter increases further, the pattern becomes even less clear. The scatterplots below are examples of **weak association** between the variables.



Weak positive association



Weak positive association



Weak negative association

In general terms, we can describe the **strength of an association** as follows.

### Strength

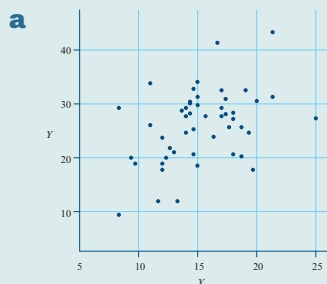
An association is classified as:

- **Strong** if the points on the scatterplot tend to be tightly clustered about a trend line.
- **Moderate** if the points on the scatterplot tend to be broadly clustered about a trend line.
- **Weak** if the points on the scatterplot tend to be loosely clustered about a trend line.
- When no pattern can be seen we say that there is **no association**.



### Example 15 Strength of an association

Classify the **strength** of the association in each of the following scatterplots.



#### Explanation

- a** The points are loosely clustered.
- b** The points are tightly clustered.



#### Solution

- The association is weak.
- The association is strong.

## Exercise 2E

### Assessing the direction of an association from the variables

- 1 For each of the following pairs of variables:
  - a Indicate whether you expect an association to exist between the variables.
  - b If associated, say which variable you would expect to be the EV and which would be the RV, and whether you would expect the variables to be positively or negatively associated.
    - i *intelligence and height*
    - ii *level of education and salary level*
    - iii *salary and tax paid*
    - iv *frustration and aggression*
    - v *population density and distance from the city centre*
    - vi *time using social media and time spent studying*

### Using a scatterplot to assess the direction, form and strength of an association

Example 13

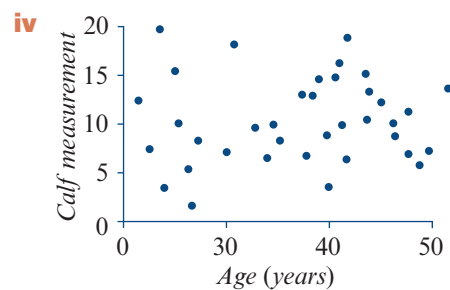
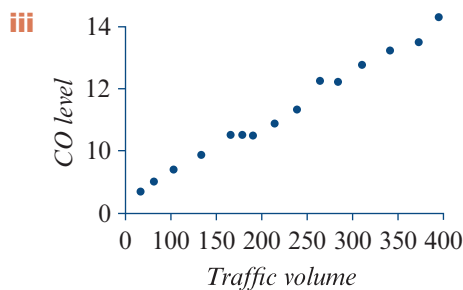
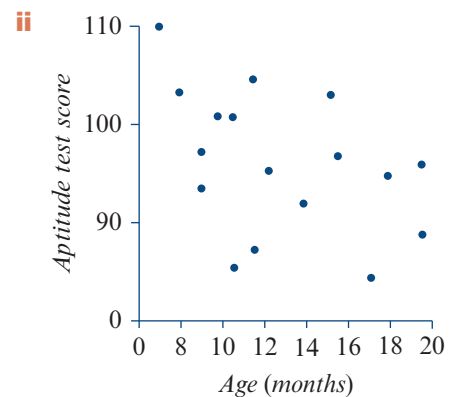
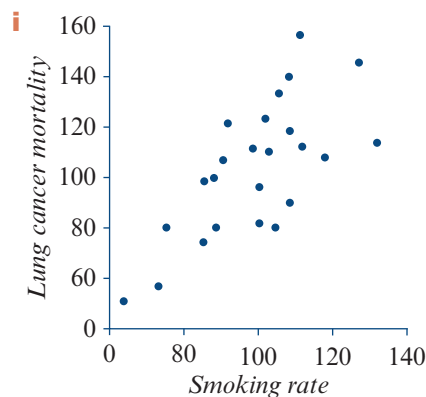
- 2 For each of the following scatterplots, state whether the variables appear to be associated. If the variables appear to be associated:

Example 14

- a Describe the association in terms of its direction (positive/negative), form (linear/non-linear) and strength (strong/moderate/weak).

Example 15

- b Write a sentence describing the direction of the association in terms of the variables in the scatterplot.





## 2F Strength of a linear relationship: the correlation coefficient

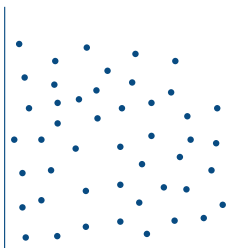
### Learning intentions

- ▶ To introduce Pearson's correlation coefficient  $r$  as a measure of the strength of a linear association between two variables.
- ▶ To be able to use technology to determine the value of Pearson's correlation coefficient  $r$ .
- ▶ To be able to classify the strength of a linear association as weak, moderate or strong based on the value of Pearson's correlation coefficient  $r$ .

The strength of a linear association is an indication of how closely the points in the scatterplot fit a straight line. If the points in the scatterplot lie exactly on a straight line, we say that there is a perfect linear association. If there is no fit at all we say there is no association. In general, we have an imperfect fit, as seen in all of the scatterplots to date.

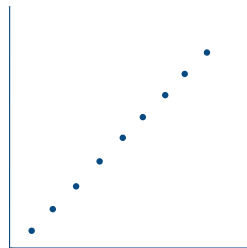
To measure the **strength of a linear relationship**, a statistician called Carl Pearson developed a **correlation coefficient**,  $r$ , which has the following properties.

■ If there is *no linear* association,  $r = 0$ .



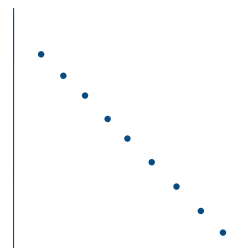
$r = 0$

■ If there is a *perfect positive linear* association,  $r = +1$ .



$r = +1$

■ If there is a *perfect negative linear* association,  $r = -1$ .

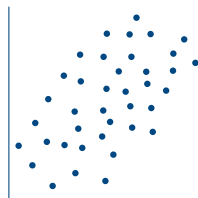


$r = -1$

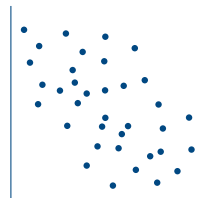
If there is a less than perfect linear association, then the correlation coefficient,  $r$ , has a value between  $-1$  and  $+1$ , or  $-1 < r < +1$ . The scatterplots below show approximate values of  $r$  for linear associations of varying strengths.



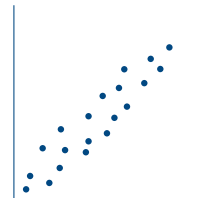
$r = -0.7$



$r = +0.5$



$r = -0.3$



$r = +0.9$

### Pearson's correlation coefficient

The Pearson's correlation coefficient,  $r$ :

- measures the **strength** of a **linear association**, with larger values indicating stronger relationships
- has a value between  $-1$  and  $+1$
- is positive if the direction of the linear association is positive.
- is negative if the direction of the linear association is negative.
- is close to zero if there is no association.

## Calculating the correlation coefficient

Pearson's correlation coefficient,  $r$ , gives a numerical measure of the degree to which the points in the scatterplot tend to cluster around a straight line.

Formally, if we call the two variables we are working with  $x$  and  $y$ , and we have  $n$  observations, then  $r$  is given by:

$$r = \frac{\sum(x - \bar{x})(y - \bar{y})}{(n - 1)s_x s_y}$$

In this formula,  $\bar{x}$  and  $s_x$  are the mean and standard deviation of the  $x$ -values, and  $\bar{y}$  and  $s_y$  are the mean and standard deviation of the  $y$ -values.

## Calculating $r$ using the formula (optional)

In practice, you can always use your calculator to determine the value of the correlation coefficient. However, to understand what is involved when you use your calculator, it is best that you know how to calculate the correlation coefficient from the formula first.

### How to calculate the correlation coefficient using the formula

Use the formula to calculate the correlation coefficient,  $r$ , for the following data.

$x$	1	3	5	4	7
$y$	2	5	7	2	9

$$\bar{x} = 4, \quad s_x = 2.236$$

$$\bar{y} = 5, \quad s_y = 3.082$$

Give the values rounded to two decimal places.

#### Steps

- 1 Write down the values of the means, standard deviations and  $n$ .

$$\bar{x} = 4 \quad s_x = 2.236$$

$$\bar{y} = 5 \quad s_y = 3.082 \quad n = 5$$

- 2** Set up a table like that shown opposite to calculate  $\sum(x - \bar{x})(y - \bar{y})$ .

$x$	$(x - \bar{x})$	$y$	$(y - \bar{y})$	$(x - \bar{x}) \times (y - \bar{y})$
1	-3	2	-3	9
3	-1	5	0	0
5	1	7	2	2
4	0	2	-3	0
7	3	9	4	12
<i>Sum</i>	0		0	23

$$\therefore \sum(x - \bar{x})(y - \bar{y}) = 23$$

- 3** Write down the formula for  $r$ .  
Substitute the appropriate values and evaluate, rounding the answer to two decimal places.

$$r = \frac{\sum(x - \bar{x})(y - \bar{y})}{(n - 1)s_x s_y}$$

$$\therefore r = \frac{23}{(5 - 1) \times 2.236 \times 3.082}$$

$$= 0.834\dots = 0.83 \text{ (2 d.p.)}$$

### CAS 2: How to calculate the correlation coefficient using the TI-Nspire CAS

The following data show the per capita income (in \$'000) and the per capita carbon dioxide emissions (in tonnes) of 11 countries.

Determine the value of Pearson's correlation coefficient rounded to two decimal places.

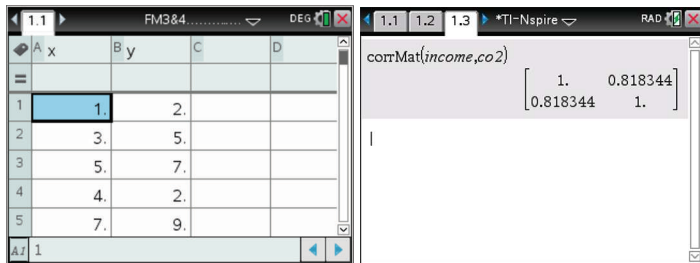
<i>Income (000)</i>	8.9	23.0	7.5	8.0	18.0	16.7	5.2	12.8	19.1	16.4	21.7
<i>CO<sub>2</sub> (tonnes)</i>	7.5	12.0	6.0	1.8	7.7	5.7	3.8	5.7	11.0	9.7	9.9

#### Steps

- 1** Start a new document by pressing **ctrl** + **N**.

- 2** Select **Add Lists & Spreadsheet**.

Enter the data into lists named *income* and *co2*.



- 3** Press **ctrl** + **I** and select **Add Calculator**.

Using the **correlation matrix** command: type in **corrmat(income, co2)** and press **enter**.

Alternatively: Press **2nd** **1** **C** to access the **Catalog**, scroll down to **corrMat(** and press **enter**. Complete the command by typing in **income, co2** and press **enter**.

The value of the correlation coefficient is  $r = 0.8342\dots$  or  $0.83$  (2 d.p.)

## CAS 2: How to calculate the correlation coefficient using the ClassPad

The following data show the per capita income (in \$'000) and the per capita carbon dioxide emissions (in tonnes) of 11 countries.

Determine the value of Pearson's correlation coefficient rounded to two decimal places.

<i>Income (\$'000)</i>	8.9	23.0	7.5	8.0	18.0	16.7	5.2	12.8	19.1	16.4	21.7
<i>CO<sub>2</sub> (tonnes)</i>	7.5	12.0	6.0	1.8	7.7	5.7	3.8	5.7	11.0	9.7	9.9

### Steps

- 1 Open the **Statistics** application



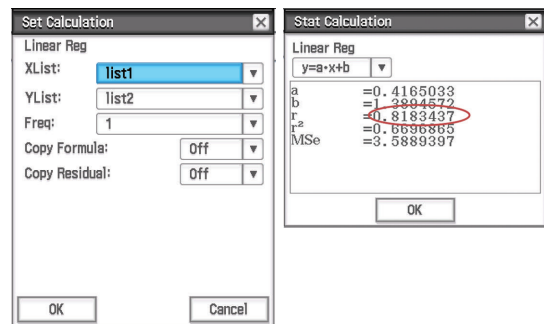
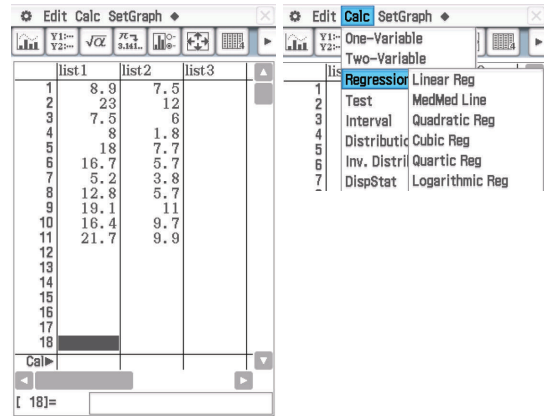
- 2 Enter the data into the columns:
  - *Income* in List1
  - *CO<sub>2</sub>* in List2.

- 3 Select **Calc>Regression>Linear Reg** from the menu bar.

- 4 Press **EXE**.

- 5 Tap **OK** to confirm your selections.

The value of the correlation coefficient is  
 $r = 0.818 \dots$  or  $0.82$  (to 2 d.p.).



## Classifying the strength of a linear association

Pearson's correlation coefficient,  $r$ , can be used to classify the strength of a linear associations as follows:

$0.75 \leq r \leq 1$	<b>strong</b> positive association
$0.5 \leq r < 0.75$	<b>moderate</b> positive association
$0.25 \leq r < 0.5$	<b>weak</b> positive association
$-0.25 < r < 0.25$	<b>no</b> association
$-0.5 < r \leq -0.25$	<b>weak</b> negative association
$-0.75 < r \leq -0.5$	<b>moderate</b> negative association
$-1 \leq r \leq -0.75$	<b>strong</b> negative association


**Example 16** Classifying the strength of a linear association

Classify the strength of each of the following linear associations using the previous table:

**a**  $r = 0.35$

**b**  $r = -0.507$

**c**  $r = 0.992$

**d**  $r = -0.159$

**Explanation**

- a** The value 0.35 is more than 0.25 and less than 0.5. That is,  $0.25 \leq r < 0.5$
- b** The value  $-0.507$  is more than  $-0.75$  and less than  $-0.5$ . That is,  $-0.75 < r \leq -0.5$
- c** The value 0.992 is more than 0.75 and less than 1. That is,  $0.75 \leq r \leq 1$
- d** The value  $-0.159$  is more than  $-0.25$  and less than 0.25. That is,  $-0.25 < r < 0.25$

**Solution**

weak, positive

moderate, negative

strong, positive

no association

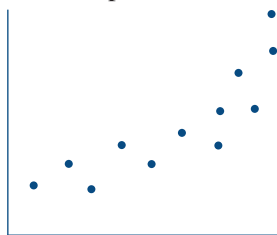
**Warning!**

If you use the value of the correlation coefficient as a measure of the strength of an association, you should ensure that:

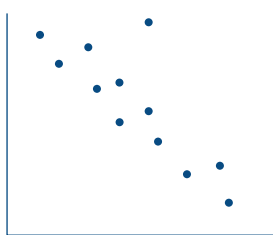
- 1** the variables are **numeric**
- 2** the association is **linear**
- 3** there are **no outliers** in the data (the correlation coefficient can give a misleading indication of the strength of the linear association if there are outliers present)


**Exercise 2F**
**Basic ideas**

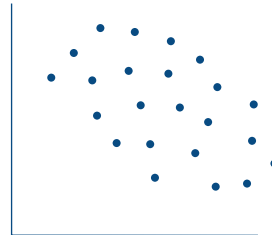
- 1** The scatterplots of three sets of related variables are shown.



Scatterplot A



Scatterplot B



Scatterplot C

- a** For each scatterplot, describe the association in terms of strength, direction, form and outliers (if any).

- b** For which of these scatterplots would it be inappropriate to use the correlation coefficient,  $r$ , to give a measure of the strength of the association between the variables? Give reasons.

### Calculating $r$ using the formula (optional)

- 2** Use the formula to calculate the correlation coefficient,  $r$ , correct to two decimal places.

$x$	2	3	6	3	6
$y$	1	6	5	4	9

$$\bar{x} = 4, s_x = 1.871$$

$$\bar{y} = 5, s_y = 2.915$$

### Calculating $r$ using a CAS calculator

- 3 a** The table below shows the maximum and minimum temperatures during a heat-wave. The *maximum* and *minimum* temperature each day are linearly associated. Use your calculator to show that  $r = 0.818$ , correct to three decimal places.

Day	Sunday	Monday	Tuesday	Wednesday	Thursday	Friday
Maximum ( $^{\circ}\text{C}$ )	29.4	34.0	34.5	35.0	36.9	36.4
Minimum ( $^{\circ}\text{C}$ )	17.7	19.8	23.3	22.4	22.0	22.0

- b** This table shows the number of runs scored and balls faced by batsmen in a cricket match. *Runs scored* and *balls faced* are linearly associated. Use your calculator to show that  $r = 0.8782$ , correct to four decimal places.

Batsman	1	2	3	4	5	6	7	8	9	10	11
Runs scored	27	8	21	47	3	15	13	2	15	10	2
Balls faced	29	16	19	62	13	40	16	9	28	26	6

- c** This table shows the hours worked and university participation rate (%) in six countries. *Hours worked* and *university participation rate* are linearly associated. Use your calculator to show that  $r = -0.6727$ , correct to four decimal places.

Country	Australia	Britain	Canada	France	Sweden	US
Hours worked	35.0	43.0	38.2	39.8	35.6	34.8
Participation rate (%)	26	20	36	25	37	55

### Classifying the strength of the association based on the value of $r$

#### Example 16

- 4** Use the guidelines on page 143 to classify the strength of the linear associations for each of the linear associations in Question 3.

## 2G The coefficient of determination

### Learning intentions

- ▶ To be able to calculate the value of the coefficient of determination.
- ▶ To be able to use the coefficient of determination to assess the strength of the association in terms of the explained variation.

If two variables are associated, it is possible to estimate the value of one variable from that of the other. For example, people's weights and heights are associated. Thus, given a person's height, we can roughly predict their weight. The degree to which we can make such predictions depends on the value of  $r$ . If there is a perfect linear association ( $r = 1$ ) between two variables, we can make an exact prediction.

For example, when you buy cheese by the gram there is an exact association between the weight of the cheese and the amount you pay ( $r = 1$ ). At the other end of the scale, there is no association between an adult's height and their IQ ( $r \approx 0$ ). So knowing an adult's height will not enable you to predict their IQ any better than guessing.

### The coefficient of determination

The degree to which one variable can be predicted from another linearly related variable is given by a statistic called the **coefficient of determination**.

The coefficient of determination is calculated by squaring the correlation coefficient:

$$\text{coefficient of determination} = r^2$$



### Example 17 Calculating the coefficient of determination

If the correlation between weight and height is  $r = 0.8$ , find the value of the coefficient of determination. Express your answer as a percentage.

#### Solution

The coefficient of determination  $= r^2 = 0.8^2 = 0.64 = 64\%$

**Note:** We have converted the coefficient of determination into a percentage (64%) as this is the most useful form when we come to interpreting the coefficient of determination.

We now know how to calculate the coefficient of determination, but what does it tell us?

### Interpreting the coefficient of determination

The coefficient of determination (as a percentage) tells us the **variation in the response variable** that is **explained** by the **variation in the explanatory variable**.

**Example 18** Interpreting the coefficient of determination

In the previous example we found the coefficient of determination between height and weight to be 0.64 (or 64%). Interpret this value in terms of the variables *weight* and *height*.

**Solution**

The coefficient of determination tells us that 64% of the variation in people's *weight* is explained by the variation in their *height*.

**What do we mean by 'explained'?**

If we take a group of people, their weights and heights will vary. One explanation is that taller people tend to be heavier and shorter people tend to be lighter. The coefficient of determination tells us that 64% of the variation in people's weights can be explained by the variation in their heights. The rest of the variation (36%) in their weights will be explained by other factors, such as diet, lifestyle, build. We could say that 36% of the variation in weight is NOT explained by the variation in height.

**Example 19** Calculating and interpreting the coefficient of determination

The level of carbon monoxide (CO) in the air measured at the roadside, and the traffic volume at the same location are linearly related, with  $r = +0.985$ . Determine the value of the coefficient of determination, write it in percentage terms and interpret. In this relationship, *traffic volume* is the explanatory variable.

**Solution**

The coefficient of determination is:

$$r^2 = (0.985)^2 = 0.9702$$

Written as a percentage:  $0.9702 \times 100 = 97.0\%$  rounded to one decimal place.

Therefore, 97.0% of the variation in carbon monoxide levels in the air can be explained by the variation in traffic volume.

Clearly, traffic volume is a very good predictor of carbon monoxide levels in the air. Thus, knowing the traffic volume enables us to predict carbon monoxide levels with a high degree of accuracy. This is not the case with the next example.

**Example 20** Calculating and interpreting the coefficient of determination

Scores on tests of verbal and mathematical ability are linearly related with correlation coefficient  $r = +0.275$ . Determine the value of the coefficient of determination, write it in percentage terms, and interpret. In this relationship, *verbal ability* is the explanatory variable.



**Solution**

The coefficient of determination is:

$$r^2 = (0.275)^2 = 0.0756$$

Written as a percentage:  $0.0756 \times 100 = 7.6\%$  rounded to one decimal place.

Therefore, only 7.6% of the variation observed in scores on the mathematical ability test can be explained by the variation in scores obtained on the verbal ability test.

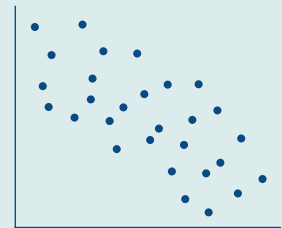
Clearly, scores on the verbal ability test are not good predictors of the scores on the mathematical ability test; 92.4% of the variation in mathematical ability is explained by other factors.

Given the value of the coefficient of determination we can reverse the calculation and find the value of the correlation coefficient. However, since the square root of a number can be positive or negative, we need more information to be able to do this correctly, such as a scatterplot.


**Example 21** Calculating the correlation coefficient from the coefficient of determination

For the relationship described by this scatterplot, the coefficient of determination = 0.5210.

Determine the value of the correlation coefficient,  $r$ , rounded to four decimal places.

**Explanation**

- 1 Since we know the value of the coefficient of determination ( $= r^2$ ), we need to find the square root of this value to find  $r$ .
- 2 There are two solutions, one positive and the other negative. Use the scatterplot to decide which applies.
- 3 Write down your answer.

**Solution**

$$r^2 = 0.5210$$

$$\therefore r = \pm\sqrt{0.5210} = \pm 0.7218$$

Scatterplot indicates a negative association.

$$\therefore r = -0.7218$$

## Exercise 2G

Calculating the coefficient of determination from  $r$

**Example 17**

- 1 For each of the following values of  $r$ , calculate the value of the coefficient of determination and convert to a percentage (correct to one decimal place).

**a**  $r = 0.675$     **b**  $r = 0.345$     **c**  $r = -0.567$     **d**  $r = -0.673$     **e**  $r = 0.124$

### Calculating and interpreting the coefficient of determination

**Example 18**

- 2** For each of the following, determine the value of the coefficient of determination, write it in percentage terms, and interpret.

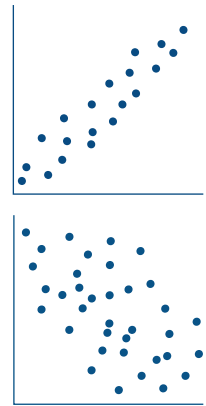
**Example 19**

- a** Scores on a hearing test and age (EV) are linearly related, with  $r = -0.611$ .
- b** Mortality rate and smoking rate (EV) are linearly related, with  $r = 0.716$ .
- c** Life expectancy and birth rate (EV) are linearly related, with  $r = -0.807$ .
- d** Daily maximum (RV) and minimum temperatures are linearly related, with  $r = 0.818$ .
- e** Runs scored (RV) and balls faced by a batsman are linearly related, with  $r = 0.8782$ .

### Calculating $r$ from the coefficient of determination given a scatterplot

**Example 21**

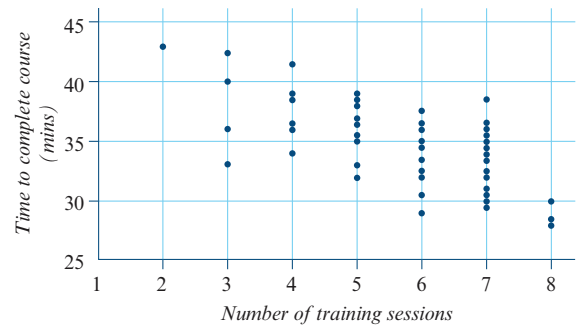
- 3 a** For the relationship described by the scatterplot shown, the coefficient of determination,  $r^2 = 0.8215$ . Determine the value of the correlation coefficient,  $r$  (rounded to three decimal places).
- b** For the relationship described by the scatterplot shown, the coefficient of determination  $r^2 = 0.1243$ . Determine the value of the correlation coefficient,  $r$  (rounded to three decimal places).



### Exam 1 style questions

Use the following information to answer Questions 4 to 6

The association between the *number of training sessions* attended by participants before undertaking an obstacle course, and the *time* in minutes it took them to complete the course, is described by the scatterplot shown. The coefficient of determination is 0.3969.



- 4** The value of the correlation coefficient,  $r$  (rounded to two decimal places) is closest to.
- A** 0.16      **B** 0.40      **C** 0.63      **D** -0.40      **E** -0.63

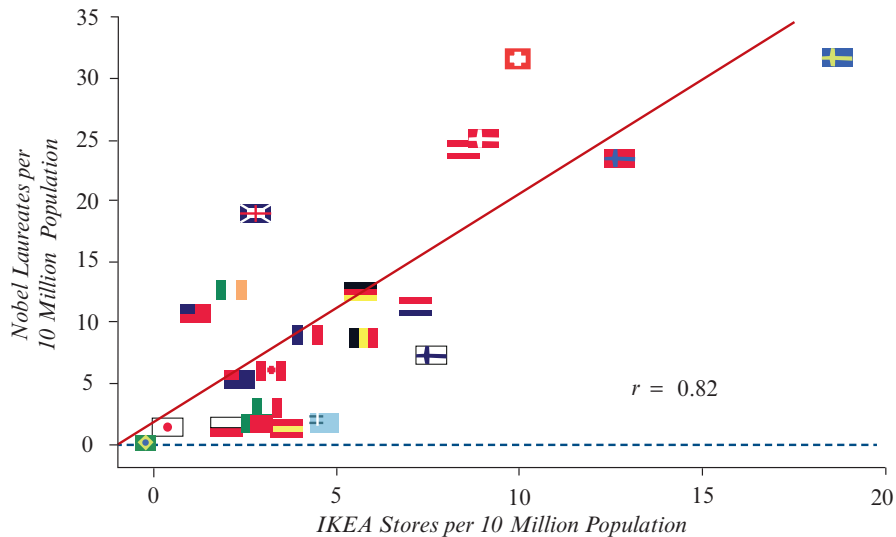
- 5 The percentage of variation in *time* explained by the variation in the *number of training sessions* is closest to:  
A 39.7%      B 63.0%      C 15.8%      D 37.0%      E 60.3%
- 6 The percentage of variation in *time* NOT explained by the variation in the *number of training sessions* is closest to:  
A 39.7%      B 63.0%      C 15.8%      D 37.0%      E 60.3%
- 7 Suppose that in a certain industry the correlation between *years spent studying* and *income* for employees is 0.73, and the correlation between *age* and *income* is 0.45. Given this information, which one of the following statements is true?  
A Older employees tend to have spent more years studying.  
B The correlation between *age* and *years spent studying* is 0.32.  
C *Age* explains a higher percentage of the variation in *income* than *years spent studying*.  
D *Years spent studying* explains a higher percentage of the variation in *income* than *age*.  
E Together *age* and *years spent studying* explain 100% of the variation in *income*.
- 8 Which of the following statements could be true?  
A The correlation coefficient between *height* (in centimetres) and *weight* (1 = light, 2 = medium, 3 = heavy) was found to be 0.68.  
B The correlation coefficient between *height* (in centimetres) and *head circumference* (in centimetres) was found to be 1.45.  
C The correlation coefficient between *blood pressure* (in mmHg) and *weight* (in kg) was found to be -0.3, and the coefficient of determination was found to be  $r^2 = -0.09$ .  
D The correlation coefficient between *age* (in years) and *salary* (in \$000's) was found to be 0.68.  
E The correlation coefficient between *height* (in centimetres) and *head circumference* (in centimetres) was found to be 0.49, and the coefficient of determination was found to be 70%.

## 2H Correlation and causality

### Learning intentions

- ▶ To be able to define and differentiate the concepts of association and causation.

Recently there has been interest in the strong association between the number of Nobel prizes a country has won and the number of IKEA stores in that country ( $r = 0.82$ ). This strong association is evident in the scatterplot below. Here country flags are used to represent the data points.



Does this mean that one way to increase the number of Australian Nobel prize winners is to build more IKEA stores?

Almost certainly not, but this association highlights the problem of assuming that a strong correlation between two variables indicates the association between them is **causal**.

### Correlation does not imply causality

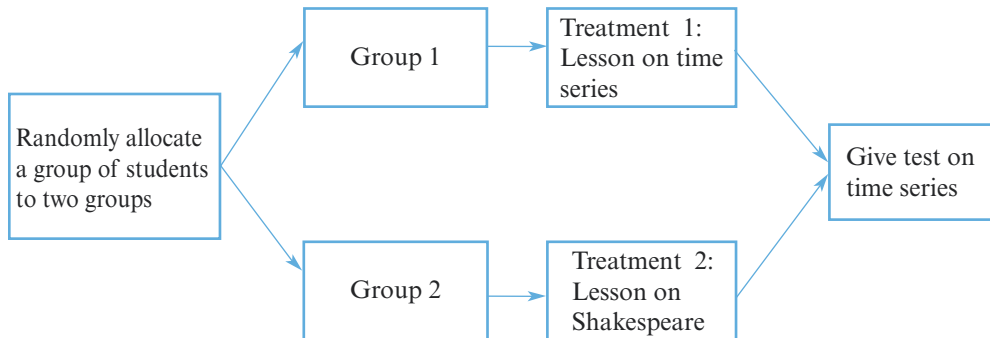
A correlation tells you about the strength of the association between the variables, but no more. It tells you nothing about the source or cause of the association.

### Video

To help you with this concept, you should watch the video ‘The Question of Causation’, which can be accessed through the link below. It is well worth 15 minutes of your time.  
<http://cambridge.edu.au/redirect/?id=6103>

## Establishing causality

To establish causality, you need to conduct an **experiment**. In an experiment, the value of the explanatory variable is deliberately manipulated, while all other possible explanatory variables are kept constant or controlled. A simplified version of an experiment is displayed below.



In this experiment, a class of students is randomly allocated into two groups. Random allocation ensures that both groups are as similar as possible.

Next, group 1 is given a lesson on time series (treatment 1), while group 2 is given a lesson on Shakespeare (treatment 2). Both lessons are given under the same classroom conditions. When both groups are given a test on time series the next day, group 1 does better than group 2.

We then conclude that this was because the students in group 1 were given a lesson on time series.

### Is this conclusion justified?

In this experiment, the students' test score is the response variable and the type of lesson they were given is the explanatory variable. We randomly allocated the students to each group while ensuring that all other possible explanatory variables were controlled by giving the lessons under the same classroom conditions. In these circumstances, the observed difference in the response variable (*test score*) can reasonably be attributed to the explanatory variable (*lesson type*).

Unfortunately, it is extremely difficult to conduct properly controlled experiments, particularly when the people involved are going about their everyday lives.

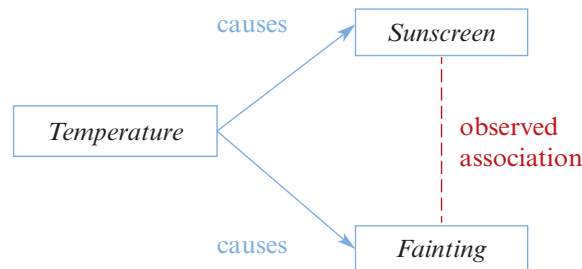
When data are collected through observation rather than experimentation, we must accept that strong association between two variables is insufficient evidence by itself to conclude that an observed change in the response variable has been caused by an observed change in the explanatory variable. It may be, but unless all of the relevant variables are under our control, there will always be alternative non-causal explanations to which we can appeal. We will now consider the various ways this might occur.

## Possible non-causal explanations for an association

### Common response

Consider the following. There is a strong positive association between the number of people using sunscreen and the number of people fainting. Does this mean that applying sunscreen causes people to faint?

Almost certainly not. On hot and sunny days, more people apply sunscreen and more people faint due to heat exhaustion. The two variables are associated because they are both strongly associated with a common third variable, *temperature*. This phenomenon is called a **common response**. See the diagram below.

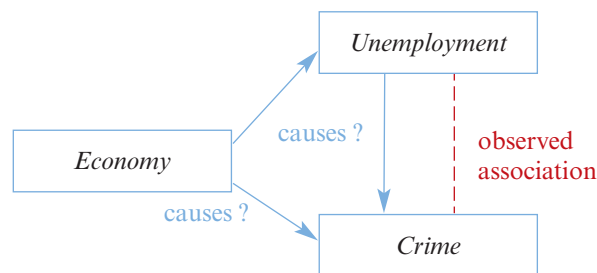


Unfortunately, being able to attribute an association to a single third variable is the exception rather than the rule. More often than not, the situation is more complex.

### Confounding variables

Statistics show that *crime* rates and *unemployment* rates in a city are strongly correlated. Can you then conclude that a decrease in unemployment will lead to a decrease in crime rates?

It might, but other possible causal explanations could be found. For example, these data were collected during an economic downturn. Perhaps the state of the economy caused the problem. See the diagram below.



In this situation, we have at least two possible causal explanations for the observed association, but we have no way of disentangling their separate effects. When this happens, the effects of the two possible explanatory variables are said to be **confounded**, because we have no way of knowing which is the actual cause of the association.

## Coincidence

It turns out that there is a strong correlation ( $r = 0.99$ ) between the consumption of margarine and the divorce rate in the American state of Maine. Can we conclude that eating margarine causes people in Maine to divorce?

A better explanation is that this association is purely coincidental.

Occasionally, it is almost impossible to identify any feasible confounding variables to explain a particular association. In these cases we often conclude that the association is ‘spurious’ and it has happened just happened by chance. We call this **coincidence**.

## Conclusion

However suggestive a strong association may be, this alone does not provide sufficient evidence for you to conclude that two variables are causally related. Unless the association is totally spurious and devoid of meaning, it will always be possible to find at least one variable ‘lurking’ in the background that could explain the association.

### Association (correlation) and causation

By itself, an observed association between two variables is never enough to justify the conclusion that two variables are causally related, no matter how obvious the causal explanation may appear to be.

## Exercise 2H

- 1 A study of primary school children aged 5 to 11 years finds a strong positive correlation between height and score on a test of mathematics ability. Does this mean that taller people are better at mathematics? What common cause might counter this conclusion?
- 2 There is a clear positive correlation between the number of churches in a town and the amount of alcohol consumed by its inhabitants. Does this mean that religion is encouraging people to drink? What common cause might counter this conclusion?
- 3 There is a strong positive correlation between the amount of ice-cream consumed and the number of drownings each day. Does this mean that eating ice-cream at the beach is dangerous? What common cause might explain this association?
- 4 The number of days a patient stays in hospital is positively correlated with the number of beds in the hospital. Can it be said that bigger hospitals encourage patients to stay longer than necessary just to keep their beds occupied? What common cause might counter this conclusion?
- 5 Suppose we found a high correlation between smoking rates and heart disease across a group of countries. Can we conclude that smoking causes heart disease? What confounding variable(s) could equally explain this correlation?

- 6** There is a strong correlation between cheese consumption and the number of people who died after becoming tangled in their bed sheets. What do you think is the most likely explanation for this correlation?
- 7** There is a strong positive correlation between the number of fire trucks attending a house fire and the amount of damage caused by the fire. Is the amount of damage in a house fire caused by the fire trucks? What common cause might explain this association?

### Exam 1 style questions

- 8** There is a positive correlation between the Gross Domestic Product (GDP), a measure of a country's wealth, and the country's carbon dioxide emissions. From this information it can be concluded that:
- A** increasing a country's GDP will increase the carbon dioxide emissions of that country.
  - B** decreasing a country's GDP will increase the carbon dioxide emissions of that country.
  - C** increasing a country's carbon dioxide emissions will increase the GDP of that country.
  - D** countries with higher GDP also tend to have lower carbon dioxide emissions.
  - E** countries with higher GDP also tend to have higher carbon dioxide emissions.

## 2I Which graph?

When investigating associations your first decision is choosing an appropriate graph to display and understand the data you have been given. This decision depends on the type of variables involved – that is, whether they are both categorical, one categorical and one numerical, or both numerical.

The following guidelines might help you make your decision. They are guidelines only, because in some instances there may be more than one suitable graph.

<i>Type of variables</i>		<i>Graph</i>
<i>Response variable</i>	<i>Explanatory variable</i>	
Categorical	Categorical	Segmented bar chart.
Numerical	Categorical	Parallel boxplots, parallel dot plots
Numerical	Categorical (two categories only)	Back-to-back stem plot, parallel dot plots or parallel boxplots
Numerical	Numerical	Scatterplot



**Exercise 2I**

- 1 Which graphical display (parallel boxplots, parallel dot plots, back-to-back stem plot, a segmented bar chart or a scatterplot) would be appropriate to display the relationships between the following? There may be more than one appropriate graph.
  - a vegetarian (yes, no) and sex (male, female)
  - b mark obtained on a statistics test and time spent studying (in hours)
  - c number of hours spent at the beach each year and state of residence
  - d number of CDs purchased per year and income (in dollars)
  - e runs scored in a cricket game and number of ‘overs’ faced
  - f attitude to compulsory sport in school (agree, disagree, no opinion) and school type (government, independent)
  - g income level (high, medium, low) and place of residence (urban, rural)
  - h number of cigarettes smoked per day and sex (male, female)
  
- 2 A back-to-back stem plot would be an appropriate graphical tool to investigate the association between a car’s *speed*, in kilometres per hour, and the
  - A driver’s *age*, in years
  - B car’s *colour* (white, red, grey, other)
  - C car’s *fuel consumption*, in kilometres per litre
  - D average *distance* travelled, in kilometres
  - E driver’s *type of licence* (probationary licence, full licence)

**Exam 1 style questions**

- 3 The relationship between *height* (in centimetres) and *weight* (1 = light, 2 = medium, 3 = heavy) is best displayed using:
  - A a histogram
  - B segmented bar charts
  - C a scatterplot
  - D parallel boxplots
  - E a percentaged two-way frequency table

## Key ideas and chapter summary



### Bivariate data

**Bivariate data** are generated when information about two variables is recorded for each subject.

### Explanatory and response variables

When investigating associations (relationships) between two variables, the **explanatory** variable (EV) is the variable we expect to explain or predict the value of the **response** variable (RV).

### Two-way frequency tables

**Two-way frequency tables** are used as the starting point for investigating the association between two categorical variables.

### Segmented bar charts

A **segmented bar chart** can be used to graphically display the information contained in a two-way frequency table. It is a useful tool for identifying relationships between two categorical variables.

### Identifying associations between two categorical variables

Associations between two categorical variables are described by comparing appropriate percentages in a **percentaged two-way frequency table** or **percentaged segmented bar chart**.

### Identifying associations between a numerical and a categorical variable

Associations between a numerical and a categorical variable are identified using **parallel dot plots**, **parallel boxplots** or a **back-to-back stem plot**. Associations between a numerical and a categorical variable are described by comparing the shape, centre and spread for the distributions.

### Scatterplots

A **scatterplot** is used to help identify and describe an association between two numerical variables. In a scatterplot, the **response variable (RV)** is plotted on the vertical axis and the **explanatory variable (EV)** is plotted on the horizontal axis.

### Identifying associations between two numerical variables

Associations between two numerical variables are identified using a **scatterplot**. Associations are classified according to:

- **Direction**, which may be positive or negative.
- **Form**, which may be linear or non-linear.
- **Strength**, which may be weak, moderate or strong.

### Correlation coefficient, $r$

The **correlation coefficient**,  $r$ , gives a measure of the strength of a linear association.

**The coefficient of determination****Coefficient of determination** =  $r^2$ 

The coefficient of determination gives the percentage of variation in the response variable that can be explained by the variation in the explanatory variable.

**Correlation and causation**

A correlation between two variables does not automatically imply that the association is causal. Alternative non-causal explanations for the association include a **common response** to a common third variable, a **confounded** variable or simply **coincidence**.

## Skills checklist



Download this checklist from the Interactive Textbook, then print it and fill it out to check your skills.

**2A****1** I can identify categorical and numerical variables in bivariate data. 

See Example 1, and Exercise 2A Question 1

**2A****2** I can identify explanatory and response variables. 

See Example 2, and Exercise 2A Question 2

**2B****3** I can construct a two-way frequency table. 

See Example 4, and Exercise 2B Question 1

**2B****4** I can percentage a two-way frequency table. 

See Example 5, and Exercise 2B Question 2

**2B****5** I can describe an association from a percentaged two-way frequency table. 

See Example 6, and Exercise 2B Question 3

**2B****6** I can construct a segmented bar chart from a percentaged two-way frequency table. 

See Example 7, and Exercise 2B Question 6

**2B****7** I can describe an association from a percentaged segmented bar chart. 

See Example 8, and Exercise 2B Question 7

**2C****8** I can use parallel dot plots to identify and describe the association between a numerical variable and a categorical variable. 

See Example 9, and Exercise 2C Question 1

- 2C** **9** I can use back-to-back stem plots to display and describe the association between a numerical variable and a categorical variable.   
See Example 10, and Exercise 2C Question 2
- 2C** **10** I can use parallel boxplots to display and describe the association between a numerical variable and a categorical variable.   
See Example 11, and Exercise 2C Question 4
- 2D** **11** I can construct a scatterplot using a CAS calculator.   
See CAS 1, and Exercise 2D Question 2
- 2E** **12** I can classify the direction, form and strength of an association from a scatterplot.   
See Example 13, Example 14, Example 15, and Exercise 2E Question 2
- 2F** **13** I can use technology to determine the value of the correlation coefficient  $r$ .   
See CAS 2, and Exercise 2F Question 3
- 2F** **14** I can classify the strength of a linear association as weak, moderate or strong based on the value of the correlation coefficient  $r$ .   
See Example 16, and Exercise 2F Question 4
- 2G** **15** I can calculate the value of the coefficient of determination.   
See Example 17, and Exercise 2G Question 1
- 2G** **16** I can use the coefficient of determination to assess the strength of the association in terms of the the explained variation.   
See Example 18, and Exercise 2G Question 2
- 2H** **17** I understand that correlation does not imply causation.   
See Exercise 2H Question 1

## Multiple-choice questions

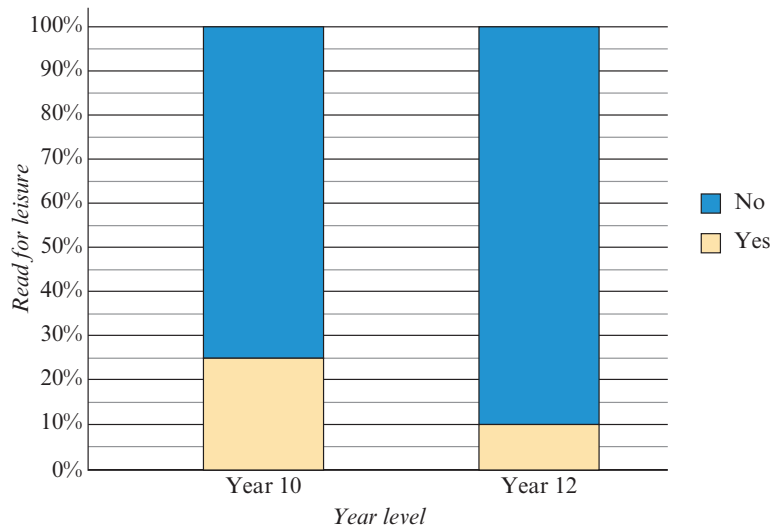
The information in the following frequency table relates to Questions 1 to 4.

Plays sport	Gender	
	Male	Female
Yes	68	79
No	34	
Total	102	175

- The variables *plays sport* and *gender* are:
  - both categorical variables
  - a categorical and a numerical variable, respectively
  - a numerical and a categorical variable, respectively
  - both numerical variables
  - neither numerical nor categorical variables
- The number of females who do not play sport is:
  - 21
  - 45
  - 79
  - 96
  - 175
- The percentage of males who do not play sport is:
  - 19.4%
  - 33.3%
  - 34.0%
  - 66.7%
  - 68.0%
- The variables *plays sport* and *gender* appear to be associated because:
  - more females play sport than males
  - fewer males play sport than females
  - a higher percentage of females play sport compared to males
  - a higher percentage of males play sport compared to females
  - both males and females play a lot of sport

Questions 5 to 7 relate to the following information

Students in Year 10 and Year 12 in a certain school were asked whether they read for leisure (*read*). Their responses are summarised in the percentaged segmented bar chart shown.



- 5 The percentage of Year 12 students who do not read for leisure is closest to:  
**A** 10%      **B** 25%      **C** 30%      **D** 75%      **E** 90%
- 6 The results could be summarised in a two-way frequency table. Which of the following frequency tables could match the percentage segmented bar chart?

**A**

Read	Year Level	
	Year 10	Year 12
Yes	31	45
No	47	66

**B**

Read	Year Level	
	Year 10	Year 12
Yes	45	11
No	135	99

**C**

Read	Year Level	
	Year 10	Year 12
Yes	75	90
No	25	10

**D**

Read	Year Level	
	Year 10	Year 12
Yes	75	25
No	90	10

**E**

Read	Year Level	
	Year 10	Year 12
Yes	40	8
No	38	5

- 7 The variables *read* and *year level* appear to be associated because:
- A** very few students in either year level read for leisure  
**B** 75% of Year 10 students do not read for leisure  
**C** only 10% of Year 12 students read for leisure  
**D** 25% of Year 10 students read for leisure, while only 10% of Year 12 students read for leisure  
**E** a higher percentage of Year 12 students read for leisure than Year 10 students

- 8 The stem plots displays the *time taken* (in minutes) for two groups of 12 people to solve a complex puzzle. Before commencing the puzzle the people were divided into two groups and assigned a different *activity*. Group A were asked to exercise vigorously for 10 minutes, while Group B were asked to meditate for 10 minutes.

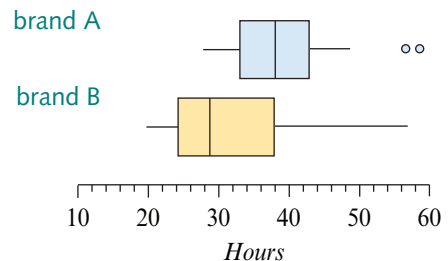
		<i>time</i>								
Group A					Group B					
1 2 = 21 minutes		9	0	5	6	7	1 6 = 16 minutes			
		3	1	1	2	4				
		8	6	5	1	5	5	6		
		4	4	3	1	2	0	1	2	3
		8	5	2						

The information in the stem plots supports the contention that there is an association between *time* and *activity* because:

- A The median time for Group A is more than the median time for Group B.
- B The range of times for both groups are approximately equal.
- C The median time for Group B is more than the median time for Group A.
- D Both distributions are approximately symmetric.
- E Both distributions are negatively skewed.

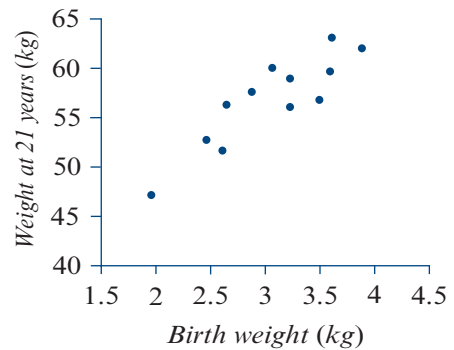
The information in the following parallel boxplots relates to Questions 9 and 10.

The parallel boxplots shown display the distribution of battery life (in hours) for two brands of batteries (brand A and brand B).



- 9 The variables *battery life* and *brand* are:
- A both categorical variables
  - B a categorical and a numerical variable respectively
  - C a numerical and a categorical variable respectively
  - D both numerical variables
  - E neither a numerical nor a categorical variable

- 10** Which of the following statements (there may be more than one) support the contention that *battery life* and *brand* are related?
- I** the median battery life for brand A is clearly higher than for brand B
  - II** battery lives for brand B are more variable than brand A
  - III** the distribution of battery lives for brand A is symmetrical with outliers but positively skewed for brand B
- A** I only      **B** II only      **C** III only      **D** I and II only      **E** I, II and III
- 11** The association between weight at age 21 (in kg) and weight at birth (in kg) is to be investigated. The variables *weight at age 21* and *weight at birth* are:
- A** both categorical variables
  - B** a categorical and a numerical variable respectively
  - C** a numerical and a categorical variable respectively
  - D** both numerical variables
  - E** neither numerical nor categorical variables
- 12** The scatterplot shows the *weights at age 21* and *weight at birth* of 12 women. The association is best described as a:
- A** weak positive linear
  - B** weak negative linear
  - C** moderate positive non-linear
  - D** strong positive non-linear
  - E** moderate positive linear
- 13** The association between *weight at age 21* and *weight at birth* for a group of males is found to be positive and linear, with a correlation coefficient of  $r = 0.58$ . For males, the percentage of variation in *weight at age 21* explained by the variation in *weight at birth* is closest to:
- A** 0.34%      **B** 24%      **C** 34%      **D** 58%      **E** 76%
- 14** The variables *response time* to a drug and *drug dosage* are linearly associated, with  $r = -0.9$ . From this information, we can conclude that:
- A** response times are  $-0.9$  times the drug dosage
  - B** response times decrease with decreased drug dosage
  - C** response times decrease with increased drug dosage
  - D** response times increase with increased drug dosage
  - E** response times are 81% of the drug dosage





- 15 The birth weight and weight at age 21 of eight women are given in the table below.

<i>Birth weight (kg)</i>	1.9	2.4	2.6	2.7	2.9	3.2	3.4	3.6
<i>Weight at 21 (kg)</i>	47.6	53.1	52.2	56.2	57.6	59.9	55.3	56.7

The value of the correlation coefficient is closest to:

- A 0.536      B 0.6182      C 0.7863      D 0.8232      E 0.8954
- 16 The value of a correlation coefficient is  $r = -0.7685$ . The value of the corresponding coefficient of determination is closest to:
- A  $-0.77$       B  $-0.59$       C 0.23      D 0.59      E 0.77

Use the following information to answer Questions 17 and 18.

The correlation coefficient between heart weight and body weight in a group of mice is  $r = 0.765$ .

- 17 Using body weight as the EV, we can conclude that:
- A 58.5% of the variation in heart weight is explained by the variation in body weights  
 B 76.5% of the variation in heart weight is explained by the variation in body weights  
 C heart weight is 58.5% of body weight  
 D heart weight is 76.5% of body weight  
 E 58.5% of the mice had heavy hearts
- 18 Given that heart weight and body weight of mice are strongly correlated ( $r = 0.765$ ), we can conclude that:
- A increasing the body weights of mice will decrease their heart weights  
 B increasing the body weights of mice will increase their heart weights  
 C increasing the body weights of mice will not change their heart weights  
 D heavier mice tend to have lighter hearts  
 E heavier mice tend to have heavier hearts
- 19 We wish to investigate the association between the variables *weight* (in kg) of young children and *level of nutrition* (poor, adequate, good). The most appropriate graphical display would be:
- A a histogram      B parallel boxplots      C a segmented bar chart  
 D a scatterplot      E a back-to-back stem plot
- 20 We wish to investigate the association between the variables *weight* (underweight, normal, overweight) of young children and *level of nutrition* (poor, adequate, good). The most appropriate graphical display would be:
- A a histogram      B parallel boxplots      C a segmented bar chart  
 D a scatterplot      E a back-to-back stem plot

- 21** There is a strong linear positive correlation ( $r = 0.85$ ) between the amount of *garbage recycled* and *salary level*.

From this information, we can conclude that:

- A** the amount of garbage recycled can be increased by increasing people's salaries
  - B** the amount of garbage recycled can be increased by decreasing people's salaries
  - C** increasing the amount of garbage you recycle will increase your salary
  - D** people on high salaries tend to recycle less garbage
  - E** people on high salaries tend to recycle more garbage
- 22** There is a strong linear positive correlation ( $r = 0.95$ ) between the marriage rate in Kentucky and the number of people who drown falling out of a fishing boat.
- From this information, the most likely conclusion we can draw from this correlation is:
- A** reducing the number of marriages in Kentucky will decrease the number of people who drown falling out of a fishing boat
  - B** increasing the number of marriages in Kentucky will increase the number of people who drown falling out of a fishing boat
  - C** this correlation is just coincidence, and changing the marriage rate will not affect the number of people drowning in Kentucky in any way
  - D** only married people in Kentucky drown falling out of a fishing boat
  - E** stopping people from going fishing will reduce the marriage rate in Kentucky

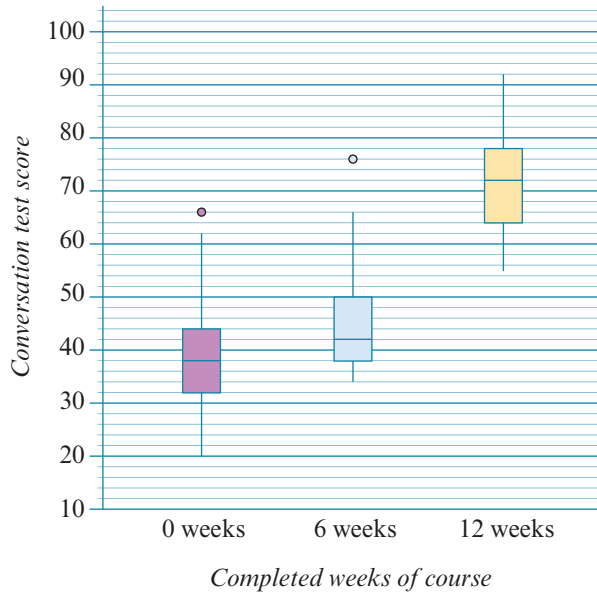
## Written response questions

- 1** One thousand drivers who had an accident during the past year were classified according to age and the number of accidents.

<i>Number of accidents</i>	<i>Age &lt; 30</i>	<i>Age ≥ 30</i>
At most one accident	130	170
More than one accident	470	230
Total	600	400

- a** What are the variables shown in the table? Are they categorical or numerical?
- b** Determine the response and explanatory variables.
- c** How many drivers under the age of 30 had more than one accident?
- d** Convert the table values to percentages by calculating the column percentages.
- e** Use these percentages to comment on the statement: 'Of drivers who had an accident in the past year, younger drivers (age < 30) are more likely than older drivers (age ≥ 30) to have had more than one accident.'

- 2 In order to improve their ability in French conversation a group of 50 students who were studying French participated in a 12 weeks intensive conversation course. The students were given a test to assess their conversation ability at the start of the course, midway through the course, and at the end of the course. Their results in each of the three tests are shown in the following boxplots.

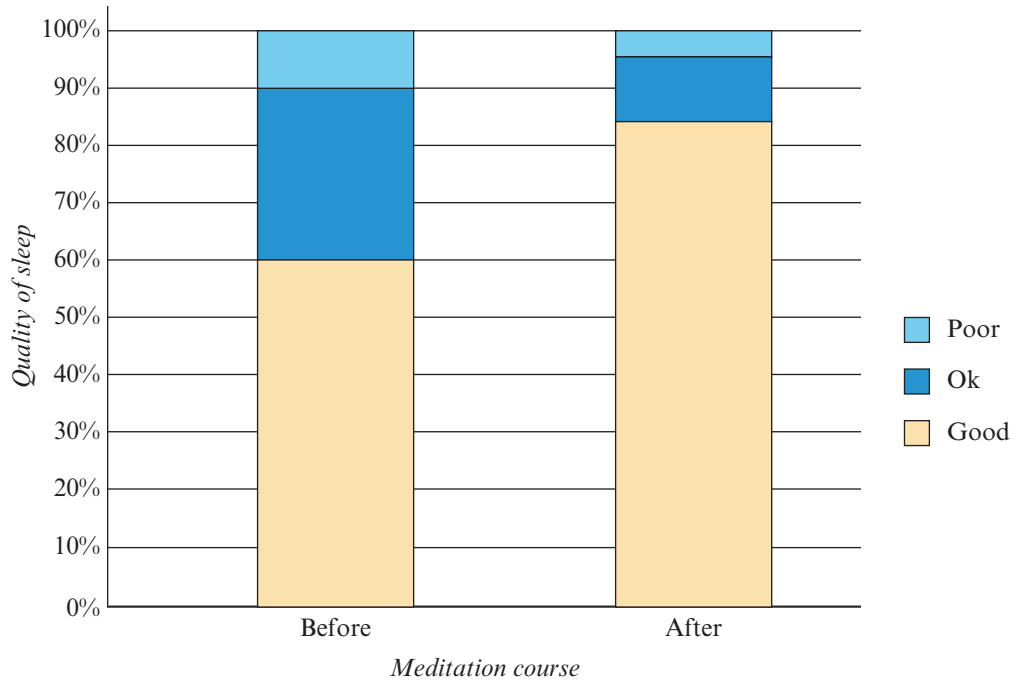


- a The two variables are *Completed weeks of course* and *Conversation test score*. Which is numerical and which is categorical?
- b Use the boxplots to compare these distributions, and draw an appropriate conclusion about the association between the number of weeks of the course completed and the score in the conversation test. Quote appropriate statistics in your response.
- 3 The data below give the hourly pay rates (in dollars per hour) of 10 production-line workers along with their years of experience on initial appointment.

Rate (\$/h)	22.57	25.78	28.84	27.37	27.23	24.64	28.95	33.35	29.68	33.99
Experience (years)	1	1	2	2	3	4	5	6	8	12

- a Determine which variable is the explanatory variable and which is the response variable.
- b Use a CAS calculator to construct a scatterplot of the data,
- c Comment on direction, outliers, form and strength of any association revealed.
- d Determine the value of the correlation coefficient ( $r$ ) rounded to three decimal places.
- e Determine the value of the coefficient of determination ( $r^2$ ), giving your answer as a percentage rounded to one decimal place, and interpret.

- 4 In a study of the effects of meditation on the quality of sleep a sample of 500 people were asked to rate the quality of their sleep as ‘good’, ‘OK’, or ‘poor’ before and after participating in the course. Their responses are shown in the segmented bar chart below.



- a What percentage of people rated the quality of their sleep as ‘good’ before they participated in the course?
- b Does the segmented bar chart support the contention that for these people their quality of sleep is associated with participation in the course? Justify your answer by quoting appropriate percentages.