

CHAPTER 2

Investigating associations between two variables

LESSONS

- 2A** Associations between two categorical variables
- 2B** Associations between numerical and categorical variables
- 2C** Associations between two numerical variables
- 2D** Correlation and causation

KEY KNOWLEDGE

- response and explanatory variables and their role in investigating associations between variables
- contingency (two-way) frequency tables, their associated bar charts (including percentage segmented bar charts) and their use in identifying and describing associations between two categorical variables
- back-to-back stem plots, parallel dot plots and boxplots and their use in identifying and describing associations between a numerical variable and a categorical variable
- scatterplots and their use in identifying and qualitatively describing the association between two numerical variables in terms of direction (positive/negative), form (linear/non-linear) and strength (strong/moderate/weak)
- answering statistical questions that require a knowledge of the associations between pairs of variables
- Pearson correlation coefficient, r , and its calculation and interpretation
- cause and effect; the difference between observation and experimentation when collecting data and the need for experimentation to definitively determine cause and effect.

2A Associations between two categorical variables

STUDY DESIGN DOT POINTS

- contingency (two-way) frequency tables, their associated bar charts (including percentage segmented bar charts) and their use in identifying and describing associations between two categorical variables
- answering statistical questions that require a knowledge of the associations between pairs of variables



KEY SKILLS

During this lesson, you will be:

- displaying bivariate data using two-way frequency tables
- displaying bivariate data using grouped bar charts
- displaying bivariate data using percentage segmented bar charts
- describing the association between two categorical variables.

KEY TERMS

- Two-way frequency table
- Two-way percentage frequency table
- Grouped bar chart

Data displays such as two-way frequency tables, grouped bar charts, and percentage segmented bar charts, can help visually compare the distributions of two categorical variables. If the distributions differ between categories, this indicates that there may be an association between the two variables.

Displaying bivariate data using two-way frequency tables

A **two-way frequency table** is a data display used to summarise bivariate data. They allow two sets of categorical data to be compared. The columns of the two-way frequency table are defined by the explanatory variable, and the rows are defined by the response variable. The explanatory variable is used to explain or predict a change in the response variable.

A difference in sample size between the categories of the explanatory variable can cause misleading patterns. As a result, further calculations are needed to accurately compare the distributions between each data set.

Data can be recorded within a frequency table as either a frequency or percentage frequency. In a **two-way percentage frequency table**, the percentages represent the proportion of times each value or category occurs.

Worked example 1

A group of people were surveyed on their opinion towards offering a student discount for public transport. Their *studying status* was recorded as 'studying' or 'not studying' along with their *opinion* 'for' or 'against' the student discount.

Complete the percentage frequency table by filling out the missing columns with the percentage frequency of each category, rounded to two decimal places.

	<i>studying status</i>		
<i>opinion</i>	studying	not studying	
for	20	5	
against	16	15	
total	36	20	

Continues →

Explanation**Step 1:** Calculate the percentage frequency for 'studying'.

The total percentage frequency should add up to 100%.

<i>opinion</i>	<i>studying status</i>			
	studying		not studying	
for	20	$\frac{20}{36} \times 100 \approx 55.56\%$	5	
against	16	$\frac{16}{36} \times 100 \approx 44.44\%$	15	
total	36	100%	20	

Step 2: Calculate the percentage frequency for 'not studying'.

The total percentage frequency should add up to 100%.

<i>opinion</i>	<i>studying status</i>			
	studying		not studying	
for	20	55.56%	5	$\frac{5}{20} \times 100 \approx 25.00\%$
against	16	44.44%	15	$\frac{15}{20} \times 100 \approx 75.00\%$
total	36	100%	20	100%

Answer

<i>opinion</i>	<i>studying status</i>			
	studying		not studying	
for	20	55.56%	5	25.00%
against	16	44.44%	15	75.00%
total	36	100%	20	100%

Displaying bivariate data using grouped bar charts

A **grouped bar chart** is a type of bar chart that visually displays two categorical variables.

It displays the categories of one variable on the horizontal axis, while the categories of the other variable are represented with different bars for each category.

The frequency of each category is represented by the height of the columns. These graphs should follow the same rules as a regular bar chart. The only difference is that there will be multiple columns grouped together, based on the categories. Spaces are included between groups of columns to indicate separate categories.

Data sets for grouped bar charts are usually presented in a frequency table. The chart typically has frequency plotted on the vertical axis, and the explanatory variable on the horizontal axis. The response variable is then represented by each of the columns. Colours or patterns can be used to distinguish between one category and another, and a legend indicates which bars relate to which categories.

Worked example 2

A group of individuals between the ages of 20 and 59 were asked how much coffee they drink a day. Their *caffeine intake* was categorised as 'low', 'medium' and 'high'. Their *age* was also recorded and grouped as '20–29', '30–39', '40–49' and '50–59'. The results are displayed in a two-way frequency table.

Use the frequency table to construct a grouped bar chart.

<i>caffeine intake</i>	<i>age</i>			
	20–29	30–39	40–49	50–59
low	22	16	10	2
medium	5	10	14	17
high	3	4	6	11

Explanation**Step 1:** Identify the explanatory variable.

Since *age* is more likely to dictate the level of *caffeine intake*, this would make it the explanatory variable.

The variable *age* will be plotted on the horizontal axis.

Step 2: Construct a set of axes with 'frequency' on the vertical axis and *age* on the horizontal axis.

The frequency should at least extend to the maximum value, which is 22.

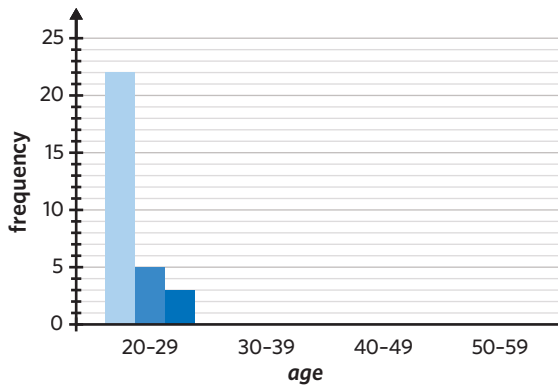
The horizontal axis should include labels for each of the *age* categories.

Continues →

Step 3: Plot the frequencies for the '20–29' age group.

Draw vertical columns for each *caffeine intake* category according to their frequency.

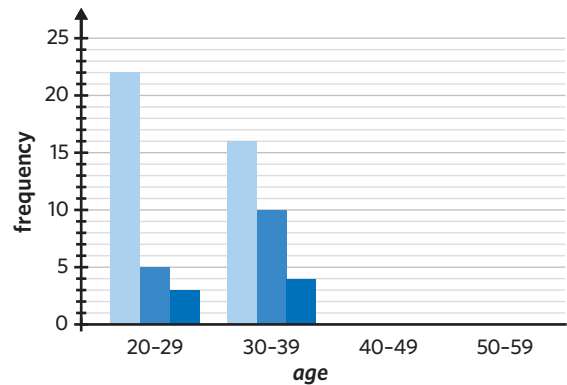
The categories will be graphed starting from 'low' to 'medium' to 'high'.



Step 4: Plot the frequencies for the '30–39' age group.

Draw vertical columns for each *caffeine intake* category according to their value in the frequency table.

The categories will be graphed starting from 'low' to 'medium' to 'high'.

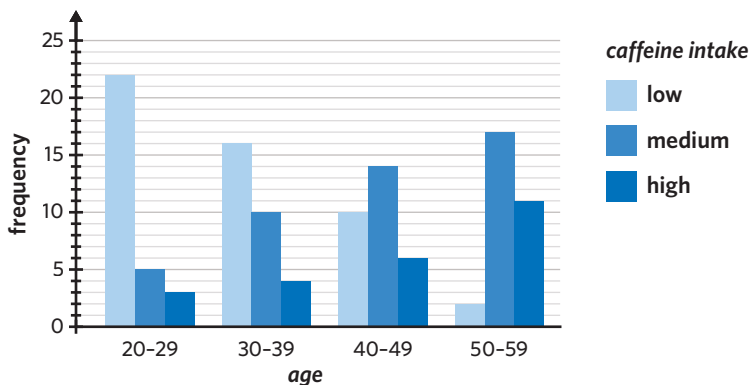


Step 5: Repeat this process for the remaining categories.

The remaining categories that need to be plotted are 40–49 and 50–59.

Step 6: Add a legend.

Answer



Displaying bivariate data using percentage segmented bar charts

Percentage segmented bar charts that display categorical variables follow the same rules as regular percentage segmented bar charts. The only difference is that there are multiple columns side by side. Previously, segmented bar charts were used to display one categorical variable, but they can be used to display data for two categorical variables by adding more columns.

There should be a gap between the columns and each should have a height of 100%.

The percentage frequency is plotted on the vertical axis, and the explanatory variable on the horizontal axis. The categories of the response variable are then represented by each of the segments.

Worked example 3

A group of students in Year 11 and Year 12 were asked to describe their *height* as 'short', 'average' or 'tall'. Their responses, as well as their *year level*, were recorded in the following percentage frequency table.

Use the frequency table to construct a percentage segmented bar chart.

<i>height</i>	<i>year level</i>	
	Year 11	Year 12
short	15%	20%
average	20%	30%
tall	65%	50%
total	100%	100%

Explanation

Step 1: Identify the explanatory variable.

Since *year level* is more likely to dictate the *height*, this would make it the explanatory variable. The variable *year level* will be plotted on the horizontal axis.

Step 2: Construct a set of axes with 'frequency (%)' on the vertical axis and *year level* on the horizontal axis.

The vertical axis should extend to the total percentage frequency, 100%.

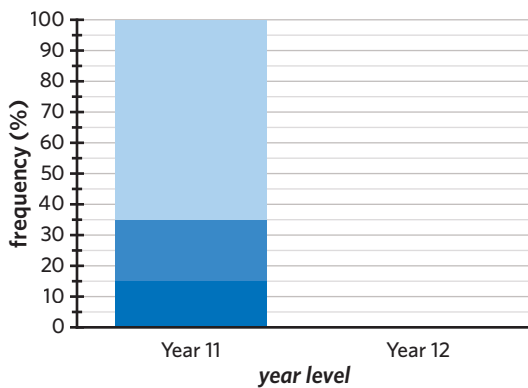
Step 3: Construct the column for 'Year 11' by adding the value of each segment.

The percentages will be graphed starting from 'short' to 'average' to 'tall'.

The 'short' segment should end at 15.

The 'average' segment should end at $15 + 20 = 35$.

The 'tall' segment should end at $35 + 65 = 100$.



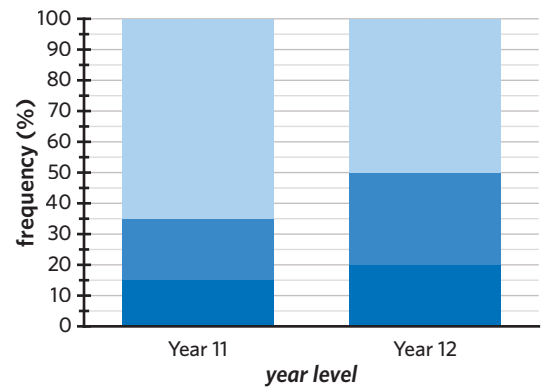
Step 4: Construct the column for 'Year 12' by adding the value of each segment.

The percentages will be graphed from 'short', to 'average', to 'tall'.

The 'short' segment should end at 20.

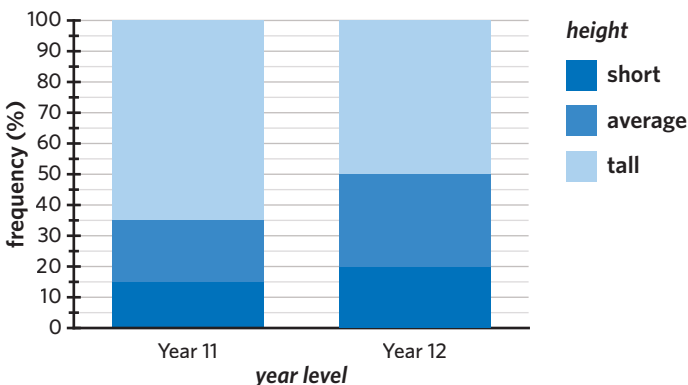
The 'average' segment should end at $20 + 30 = 50$.

The 'tall' segment should end at $50 + 50 = 100$.



Step 5: Add a legend.

Answer



Describing the association between two categorical variables

When two categorical variables are compared, they can be analysed for any associations or patterns that may exist. A brief report describing an association, or lack of association, can then be written for a two-way percentage frequency table. The report should always include whether or not an association may exist, and appropriate percentages that support the finding.

Associations between two sets of categorical data can also be analysed from percentage segmented bar charts.

See worked example 4

See worked example 5

Worked example 4

A group of people were surveyed on their *preference* towards reading hardcopy or digital books. Their *age* (under 20, 20 and over) was recorded along with their book *preference* (hardcopy, digital).

The results were displayed in a two-way percentage frequency table.

Is there an association between *age* and *preference*? Justify your answer by quoting appropriate percentages.

<i>preference</i>	<i>age</i>			
	under 20		20 and over	
hardcopy	20	56%	15	75%
digital	16	44%	5	25%
total	36	100%	20	100%

Explanation

Step 1: Consider whether there is a large percentage difference in *preference* for the different *age* categories.

- 56% of people under the age of 20 prefer hardcopy books.
- 75% of people aged 20 and over prefer hardcopy books.

There is a 19% difference between each group, which is a significant difference.

Step 2: Determine if there is an association.

An association can be determined by comparing one of the *preference* categories for both *age* categories.

Answer

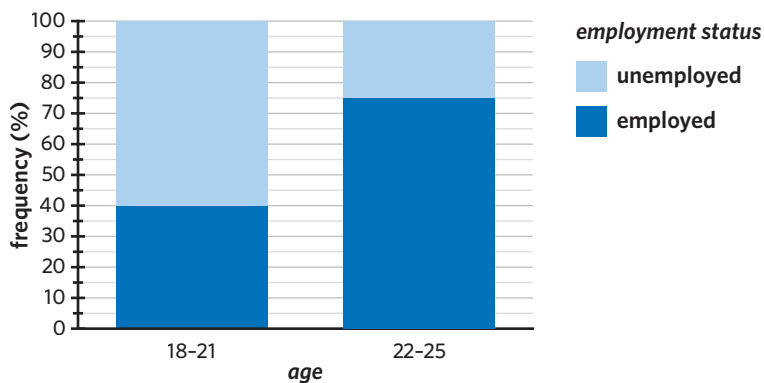
Yes, there is an association between *preference* and *age*. The percentage of people that prefer hardcopy books differs depending on their age, since only 56% of those under 20 prefer it while 75% of those aged 20 years and over do.

Answers may vary.

Worked example 5

A large group of people between the ages of 18 and 25 were asked what their current *employment status* was. They were classified as either 'employed' or 'unemployed'. Their *age* was also recorded and classified as '18–21' or '22–25'.

The results have been displayed in a percentage segmented bar chart.



Is there an association between *employment status* and *age*? Justify your answer by quoting appropriate percentages.

Continues →

Explanation

Step 1: Consider whether *employment status* is significantly different between the categories '18–21' and '22–25'.

- 40% of people between the ages of 18–21 are employed.
- 75% of people between the ages of 22–25 are employed.

There is a 35% difference between each age group. This is quite a large gap.

- 60% of people between the ages of 18–21 are unemployed.
- 25% of people between the ages of 22–25 are unemployed.

There is a 35% difference between each age group. This is quite a large gap.

Answer

Yes, there is an association between *employment status* and *age*. The percentage of people that are employed tends to increase with age, with 40% of 18–21 year olds employed and 75% of 22–25 year olds employed.

Answers may vary.

Step 2: Determine if there is an association.

An association can be determined by comparing one of the two categories for *employment status*.

Exam question breakdown

VCAA 2016 Exam 1 Data analysis Q1

The *blood pressure* (low, normal, high) and the *age* (under 50 years, 50 years or over) of 110 adults were recorded. The results are displayed in the two-way frequency table.

The **percentage** of adults under 50 years of age who have high blood pressure is closest to

- A. 11%
- B. 19%
- C. 26%
- D. 44%
- E. 58%

<i>blood pressure</i>	<i>age</i>	
	under 50 years	50 years or over
low	15	5
normal	32	24
high	11	23
total	58	52

Explanation

Step 1: Determine the number of adults under 50 years of age with high blood pressure.

There are 11 adults aged under 50 with high blood pressure.

Step 2: Represent this as a percentage frequency.

$$\begin{aligned} \text{percentage frequency} &= \frac{\text{frequency}}{\text{total frequency}} \times 100 \\ &= \frac{11}{58} \times 100 \\ &= 18.96\dots\% \end{aligned}$$

Answer

B

83% of students answered this question correctly.

12% of students incorrectly answered option A. These students likely did not convert the frequency to a percentage frequency.

2A Questions

Displaying bivariate data using two-way frequency tables

1. A group of people were surveyed on *commute time* from home to school. Their responses were classified as 'less than an hour' or 'an hour or more'. They were also grouped by the *location* of their home as either 'urban' or 'rural'. The results were displayed in the two-way percentage frequency table.

<i>commute time</i>	<i>location</i>	
	urban	rural
less than an hour	97%	62%
an hour or more	3%	38%
total	100%	100%

- a. What percentage of respondents living in an 'urban' *location* are 'less than an hour' away from their school?
- A. 3%
B. 38%
C. 62%
D. 97%
- b. What percentage of respondents living in a 'rural' *location* are 'less than an hour' away from their school?
- A. 3%
B. 38%
C. 62%
D. 97%

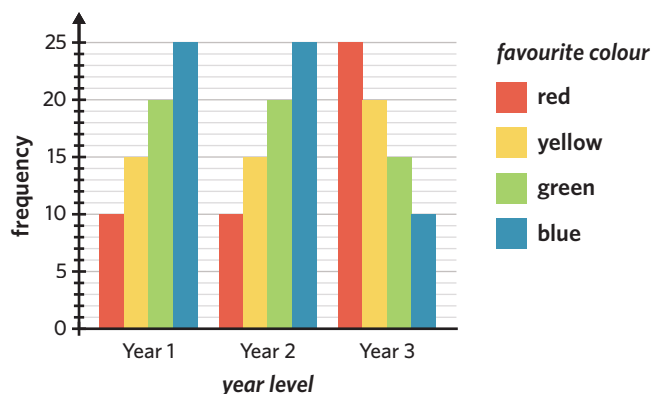
2. A survey was conducted on people's preferences for *reusable cup* brands (frank green, KeepCup, other) and their *coffee consumption* (often, rarely). The results were displayed in a two-way percentage frequency table, but some values are missing.

<i>reusable cup</i>	<i>coffee consumption</i>	
	often	rarely
frank green	41%	17%
KeepCup		11%
other	19%	
total	100%	100%

- a. Fill in the table with the missing values.
- b. What percentage of people who often drink coffee prefer to use a KeepCup?

Displaying bivariate data using grouped bar charts

3. A study was conducted to test the relationship between a student's *year level* (Year 1, Year 2, Year 3) and their *favourite colour* (red, yellow, green, blue). The results were plotted on a grouped bar chart.
- Determine the frequency table that matches the grouped bar chart.



A.

<i>favourite colour</i>	<i>year level</i>		
	Year 1	Year 2	Year 3
red	10	10	20
yellow	10	15	20
green	10	20	10
blue	10	25	20

B.

<i>favourite colour</i>	<i>year level</i>		
	Year 1	Year 2	Year 3
red	10	25	25
yellow	15	20	20
green	20	15	15
blue	25	10	10

C.

<i>favourite colour</i>	<i>year level</i>		
	Year 1	Year 2	Year 3
red	25	10	25
yellow	20	15	20
green	15	20	15
blue	10	25	10

D.

<i>favourite colour</i>	<i>year level</i>		
	Year 1	Year 2	Year 3
red	10	10	25
yellow	15	15	20
green	20	20	15
blue	25	25	10

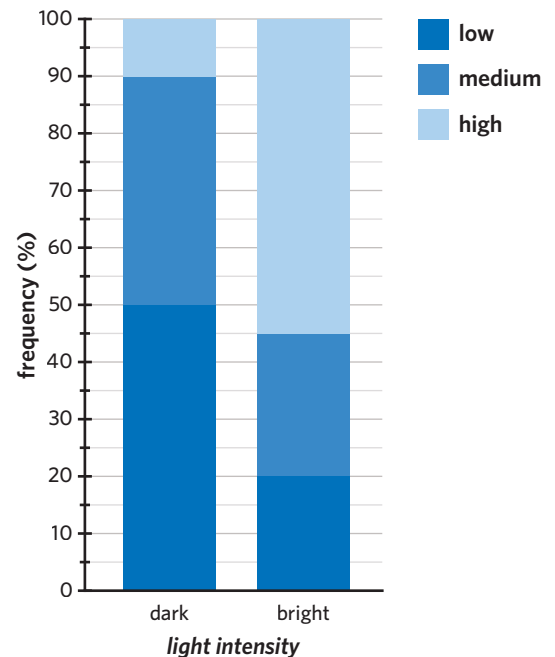
4. A survey was conducted for a group of students to see if their current *education level* (primary, secondary, tertiary) had a relationship with their *preferred sport* (soccer, football, volleyball). The results are displayed in a frequency table.

<i>preferred sport</i>	<i>education level</i>		
	primary	secondary	tertiary
soccer	5	7	9
football	8	4	1
volleyball	2	4	5
total	15	15	15

Display the data using a grouped bar chart.

Displaying bivariate data using percentage segmented bar charts

5. A scientific experiment was conducted to determine if there is a relationship between *light intensity* (dark, bright) and *respiration rate* (low, medium, high) for a small sample of spinach leaves. The results were displayed in a percentage segmented bar chart. Which of the following statements is correct?
- A. 25% of spinach leaves subjected to a bright light intensity had a low respiration rate.
- B. 55% of spinach leaves subjected to a dark light intensity had a low respiration rate.
- C. 35% of spinach leaves subjected to a dark light intensity had a medium respiration rate.
- D. 55% of spinach leaves subjected to a bright light intensity had a high respiration rate.



6. A chef is experimenting with *cooking time* (short, medium, extended) for various pasta types to see if there is a relationship with its *doneness* (undercooked, al dente, overcooked). The results were recorded in the percentage frequency table. Represent this information using a percentage segmented bar chart.

<i>doneness</i>	<i>cooking time</i>		
	short	medium	extended
undercooked	80%	20%	5%
al dente	15%	70%	20%
overcooked	5%	10%	75%
total	100%	100%	100%

Describing the association between two categorical variables

7. Each morning at the Edrolo office, a team member volunteers to do a coffee run. Everyone is able to choose either 'cappuccino', 'flat white' or 'latte'. Their *coffee preference* has been recorded alongside their *employment duration* (less than three years, three years or more) in a two-way percentage frequency table.

<i>coffee preference</i>	<i>employment duration</i>	
	less than three years	three years or more
cappuccino	13.1%	11.7%
flat white	23.7%	23.6%
latte	63.2%	64.7%
total	100%	100%

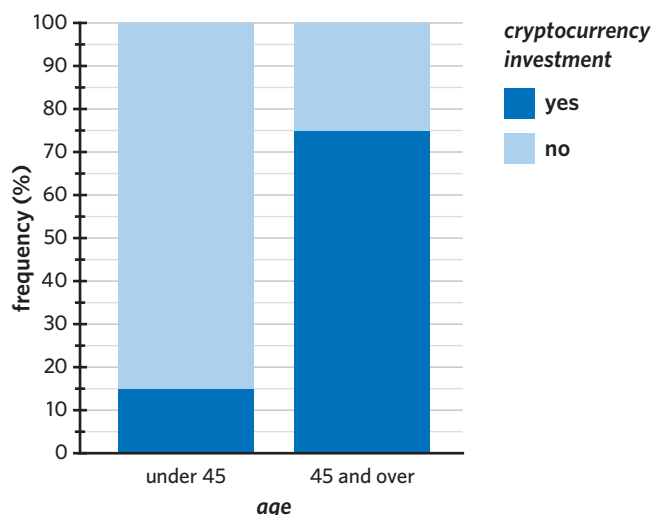
The data displayed supports the contention that there is no obvious association between *coffee preference* and *employment duration* because

- A. the coffee preferences between employment duration differ by a significant amount.
 B. 23.7% of team members who have been working for less than 3 years prefer a flat white.
 C. 63.2% of team members who have been working for less than 3 years prefer a latte and 64.7% of those who have been working for 3 or more years prefer a latte.
 D. 63.2% of team members who have been working for less than 3 years prefer a latte and 11.7% of those who have been working for 3 or more years prefer a cappuccino.

8. A study was conducted to see if individuals had a *cryptocurrency investment* (yes, no). Participants also had their *age* (under 45, 45 and over) recorded.

The results have been displayed in a percentage segmented bar chart.

Is there an association between *age* and *cryptocurrency investment*? Ensure to quote appropriate percentages.



Joining it all together

9. Mark wants to see if there is an association between the number of *hours studied* (0 to 9 hours, 10 to 19 hours, 20 to 29 hours, 30 or more hours) for an assessment and the *test score* (low, medium, high) that his students get.

- a. Which variable is most likely to be plotted on the horizontal axis of a percentage segmented bar chart?
 b. Mark fills out a table with the data he gathered from the test. However, Mark spilled his coffee on his paper on his drive to work, leaving a couple of stains. Fill out the rest of the table.

<i>test score</i>	<i>hours studied</i>			
	0 to 9 hours	10 to 19 hours	20 to 29 hours	30 or more hours
low		30%	20%	0%
medium	20%			10%
high	0%	10%	30%	90%
total	100%	100%	100%	100%

- c. Mark's colleague Julie has also filled out a two-way percentage frequency table for her General Mathematics class. Construct a percentage segmented bar chart from the following information.

<i>test score</i>	<i>hours studied</i>			
	0 to 9 hours	10 to 19 hours	20 to 29 hours	30 or more hours
low	50%	40%	40%	0%
medium	50%	50%	30%	10%
high	0%	10%	30%	90%
total	100%	100%	100%	100%

- d. Is there an association between the number of *hours studied* and the *test score* in Julie's class? Justify your answer by quoting appropriate percentages.

10. An astronomical observation was conducted to determine if there is a relationship between a star's *size* (small, large) and its *colour* (yellow, blue). The results are as follows:
- 12% of small stars observed were blue, with the rest being yellow.
 - 15% of large stars observed were yellow, with the rest being blue.
- a. Construct a percentage segmented bar chart to represent this information.
- b. Is there an association between a star's *size* and its *colour*? Justify your answer by quoting appropriate percentages.

Exam practice

11. The following data relates to the impact of traffic congestion in 2016 on travel times in 23 cities in the United Kingdom (UK).

<i>city</i>	<i>congestion level</i>	<i>size</i>	<i>increase in travel time (minutes per day)</i>
Belfast	high	small	52
Edinburgh	high	small	43
London	high	large	40
Manchester	high	large	44
Brighton and Hove	high	small	35
Bournemouth	high	small	36
Sheffield	medium	small	36
Hull	medium	small	40
Bristol	medium	small	39
Newcastle-Sunderland	medium	large	34
Leicester	medium	small	36
Liverpool	medium	large	29
Swansea	low	small	30
Glasgow	low	large	34
Cardiff	low	small	31
Nottingham	low	small	31
Birmingham-Wolverhampton	low	large	29
Leeds-Bradford	low	large	31
Portsmouth	low	small	27
Southampton	low	small	30
Reading	low	small	31
Coventry	low	small	30
Stoke-on-Trent	low	small	29

Data: TomTom International BV, <www.tomtom.com/en_gb/trafficindex>

The four variables in this data set are:

- *city* – name of city
- *congestion level* – traffic congestion level (high, medium, low)
- *size* – size of city (large, small)
- *increase in travel time* – increase in travel time due to traffic congestion (minutes per day).

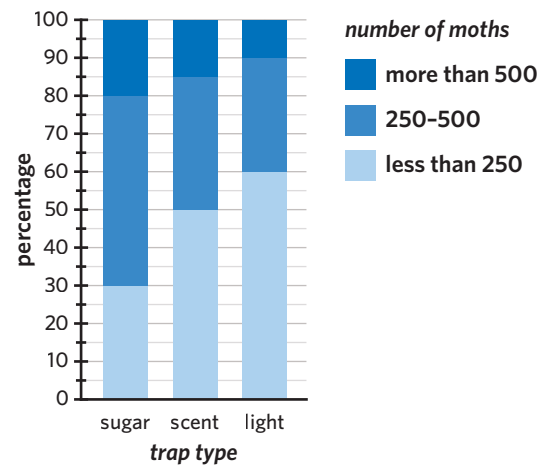
Use the data to complete the following two-way frequency table. (2 MARKS)

<i>congestion level</i>	<i>size</i>	
	small	large
high	4	
medium		
low		
total	16	

VCAA 2018 Exam 2 Data analysis Q1d

The average mark on this question was **1.9**.

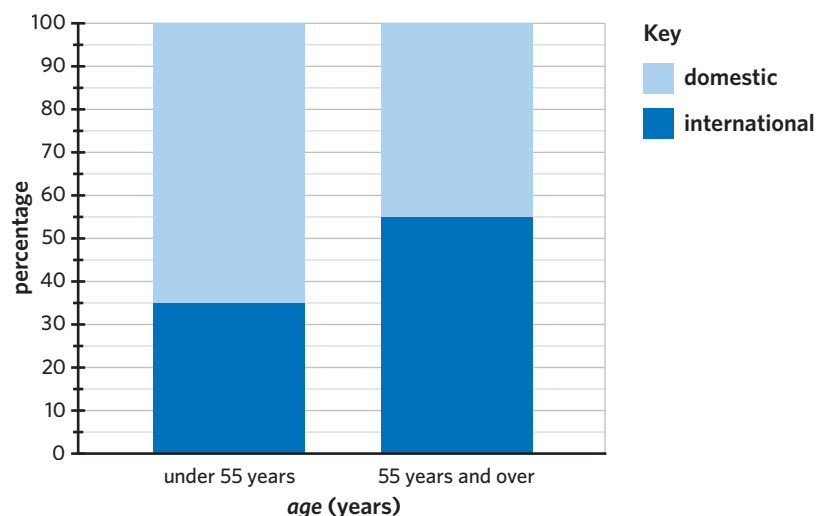
12. A study was conducted to investigate the association between the *number of moths* caught in a moth trap (less than 250, 250–500, more than 500) and the *trap type* (sugar, scent, light). The results are summarised in the percentage segmented bar chart.
- The data displayed in the percentage segmented bar chart supports the contention that there is an association between the *number of moths* caught in a moth trap and the *trap type* because
- most of the light traps contained less than 250 moths.
 - 15% of the scent traps contained 500 or more moths.
 - the percentage of sugar traps containing more than 500 moths is greater than the percentage of scent traps containing less than 500 moths.
 - 20% of sugar traps contained more than 500 moths while 50% of light traps contained less than 250 moths.
 - 20% of sugar traps contained more than 500 moths while 10% of light traps contained more than 500 moths.



71% of students answered this question correctly.

VCAA 2017 Exam 1 Data analysis Q6

13. The following percentage segmented bar chart shows the *age* (under 55 years, 55 years and over) of visitors at a travel convention, segmented by *preferred travel destination* (domestic, international).



The results could also be summarised in a two-way frequency table.

Which of the following frequency tables could match the percentage segmented bar chart?

A.

<i>preferred travel destination</i>	<i>age</i>	
	under 55 years	55 years and over
domestic	91	90
international	49	110
total	140	200

B.

<i>preferred travel destination</i>	<i>age</i>	
	under 55 years	55 years and over
domestic	65	35
international	45	55
total	110	90

C.

<i>preferred travel destination</i>	<i>age</i>	
	under 55 years	55 years and over
domestic	35	55
international	65	45
total	100	100

D.

<i>preferred travel destination</i>	<i>age</i>	
	under 55 years	55 years and over
domestic	50	70
international	100	50
total	150	120

E.

<i>preferred travel destination</i>	<i>age</i>	
	under 55 years	55 years and over
domestic	71	39
international	29	61
total	100	100

50% of students answered this question correctly.

VCAA 2021 Exam 1 Data analysis Q3

Questions from multiple lessons

Data analysis

14. The following data represents the height in centimetres of eighteen year 7 students.

144 162 131 156 165 171 149 148 159

182 171 165 167 166 158 151 158 132

What is the mean, \bar{x} , and the standard deviation, s_x , of the heights, in centimetres, of the year 7 class?

- A. $\bar{x} = 157.5$, $s_x = 12.92$
 B. $\bar{x} = 156.5$, $s_x = 12.92$
 C. $\bar{x} = 157.5$, $s_x = 13.29$
 D. $\bar{x} = 156.5$, $s_x = 13.29$
 E. $\bar{x} = 158.5$, $s_x = 12.92$

Adapted from VCAA 2018NH Exam 1 Data analysis Q7

Data analysis

15. Data was collected to investigate the association between the variables *height* (cm) and *weight* (kg).

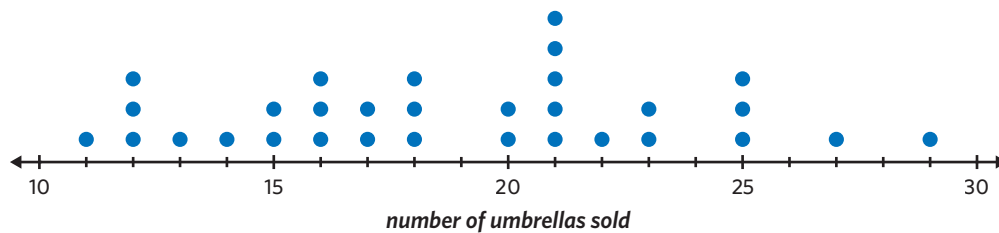
Which one of the following is most appropriate to display this data?

- A. Back-to-back stem plot
 B. Bar chart
 C. Parallel boxplots
 D. The coefficient of determination
 E. Scatterplot

Adapted from VCAA 2018 Exam 1 Data analysis Q6

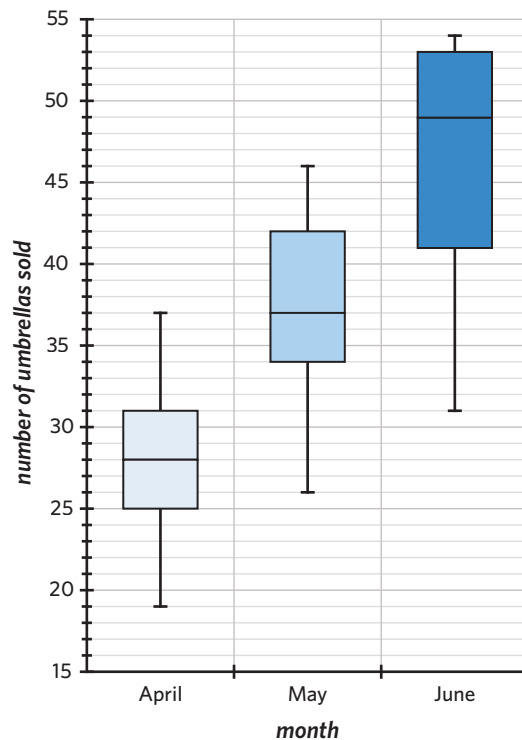
Data analysis Year 11 content

16. The owner of an umbrella shop records the number of umbrellas he sells each day. His sales for the month of January are shown in the following dot plot.



- a. Construct a boxplot based on the data in the dot plot. (2 MARKS)
- b. The number of umbrellas sold at the shop in April, May, and June are shown in the following parallel boxplots. The parallel boxplots indicate that the number of umbrellas sold could be associated with the month of the year. Explain, stating values of an appropriate statistic, why this conclusion could be made. (2 MARKS)

Adapted from VCAA 2017 Exam 2 Data analysis Q1c,d



2B Associations between numerical and categorical variables

STUDY DESIGN DOT POINTS

- back-to-back stem plots, parallel dot plots and boxplots and their use in identifying and describing associations between a numerical variable and a categorical variable
- answering statistical questions that require a knowledge of the associations between pairs of variables



KEY SKILLS

During this lesson, you will be:

- displaying data using back-to-back stem plots
- displaying data using parallel dot plots
- displaying data using parallel boxplots
- describing the association between numerical and categorical variables.

KEY TERMS

- Back-to-back stem plot
- Parallel dot plots
- Parallel boxplots

Data displays, such as back-to-back stem plots, parallel dot plots, and parallel boxplots help visually compare the distributions of two or more categories. If the distributions differ, this may signal that there is an association between the numerical and categorical variables.

Displaying data using back-to-back stem plots

A **back-to-back stem plot** is a stem plot that displays and compares the distribution of two categories.

The categories share the same stem, with the leaves of one category on the left and the leaves of the other category on the right. The leaves are ordered such that they get larger as they move away from the stem.

Back-to-back stem plots can be used to identify an association between two categories.

The following back-to-back stem plot shows the *number of points scored* in each game by two U18 basketball teams, the 'Crocodiles' and the 'Zebras'.

Key: 3 | 7 = 37 points

	Crocodiles		Zebras	
	4 0	2		
8	7 5 2	3	7	
7	4 4	4		
8	2	5	2 8	
3	1	6	0 1 5 9	
		7	0 3 5 8	
	0	8	1 3 4	

Worked example 1

The 'Abbotsford Avengers' soccer team beat the 'East Melbourne Eagles' in the final. In order for the 'East Melbourne Eagles' to see what they needed to improve on, they collected the *successful passes* (%) for each member of both teams.

Abbotsford Avengers: 64 68 70 53 57 62 74 77 36 41 66

East Melbourne Eagles: 59 90 83 87 88 76 82 83 91 78 75

- a. Construct a back-to-back stem plot to display this data, with the 'Abbotsford Avengers' on the left.

Explanation

Step 1: Consider the most appropriate stem.

The data values are two digit numbers.

The stems will refer to 'tens'.

The leaves will refer to 'ones'.

The data (from both teams) ranges from 36 to 91.

The appropriate stems are 3, 4, 5, 6, 7, 8 and 9.

3	
4	
5	
6	
7	
8	
9	

Step 2: Fill in the leaves for the left category.

They should increase as they move further from the stem. Make sure to add the category title.

**Abbotsford
Avengers**

6	3
1	4
7 3	5
8 6 4 2	6
7 4 0	7
	8
	9

Step 3: Fill in the leaves for the right category.

They should increase as they move further from the stem. Make sure to add the category title.

Abbotsford Avengers		East Melbourne Eagles
6	3	
1	4	
7 3	5	9
8 6 4 2	6	
7 4 0	7	5 6 8
	8	2 3 3 7 8
	9	0 1

Step 4: Construct a key.

As decided in step 1, the stems refer to 'tens' and the leaves refer to 'ones'.

Answer

Key: 5 | 9 = 59%

**Abbotsford
Avengers** **East Melbourne
Eagles**

6	3	
1	4	
7 3	5	9
8 6 4 2	6	
7 4 0	7	5 6 8
	8	2 3 3 7 8
	9	0 1

Continues →

- b. Which team had a higher median *successful passes*?

Explanation

Find the median *successful passes* for both teams.

Each team has 11 players so the median will be in the 6th position.

Key: $5 \mid 9 = 59\%$

Abbotsford Avengers	East Melbourne Eagles
6	3
1	4
7 3	5 9
8 6 4 2	6
7 4 0	7 5 6 8
	8 2 3 3 7 8
	9 0 1

Abbotsford Avengers: 64%

East Melbourne Eagles: 83%

Answer

East Melbourne Eagles

Displaying data using parallel dot plots

Parallel dot plots are a sequence of dot plots that display and compare the distribution of two or more categories.

A separate dot plot is constructed for each category, but they must use the same scale so the distribution can be compared. There is no restriction on the number of categories.

Parallel dot plots can be used to identify an association between two or more categories.

Worked example 2

A grade 5/6 class took a maths test on fractions. The *number of marks*, by grade, was recorded.

Grade 5: 2 4 5 5 5 6 6 7 7 7 7 9

Grade 6: 5 6 6 7 8 8 8 8 9 9 9 10

- a. Construct parallel dot plots to display this data.

Explanation

Step 1: Consider the most appropriate scale.

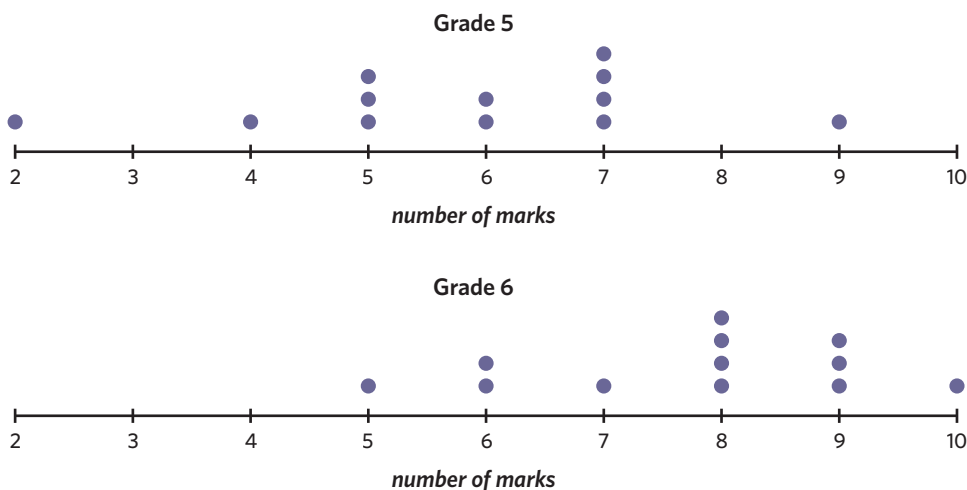
The data (from both grades) ranges from 2 to 10.

An appropriate scale ranges from 2 to 10 with tick marks for each whole number.

Step 2: Construct the parallel dot plots.

Create two identical axes that are vertically aligned. Mark a dot above the number on the number line each time a value appears in the data set of the associated grade.

Continues →

Answer

- b. Which grade has a larger range of *number of marks*?

Explanation

Find the range of *number of marks* for each grade.

Grade 5: $9 - 2 = 7$ marks

Grade 6: $10 - 5 = 5$ marks

Answer

Grade 5

Displaying data using parallel boxplots

Parallel boxplots are a sequence of boxplots that display and compare the distribution of two or more categories. They are best used for large data sets.

A separate boxplot is constructed for each category, however, the boxplots share the same axis so the distribution can be compared. There is no restriction on the number of categories.

Parallel boxplots can be used to identify an association between two or more categories.

Worked example 3

Daniel went to his local cafe every Sunday for a year and counted the *number of people* wearing Salomon shoes and the *number of people* wearing Crocs. The five-number summaries for both shoe brands are recorded.

Salomons: 2, 4, 5, 9, 15

Crocs: 0, 2, 4, 7, 8

- a. Construct parallel boxplots to display this data.

Explanation

Step 1: Consider the most appropriate scale.

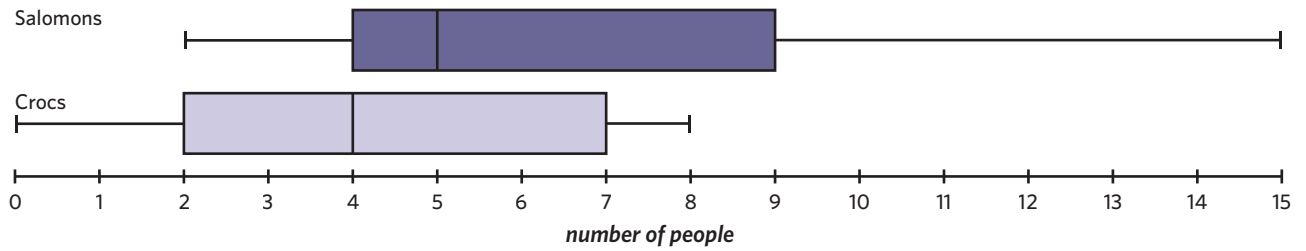
The data (for both shoe brands) ranges from 0 to 15.

An appropriate scale ranges from 0 to 15 with tick marks for each whole number.

Step 2: Construct the parallel boxplots.

Create one axis and use the five-number summary to construct the boxplots for each shoe brand. There are no outliers, so the whiskers will reach the minimum and maximum *number of people* in both boxplots.

Continues →

Answer

b. For which shoe brand was the *number of people* positively skewed?

Explanation

A positively skewed distribution trails off in a positive direction on the horizontal axis.

Only the distribution of the *number of people* for Salomons is positively skewed.

Answer

Salomons

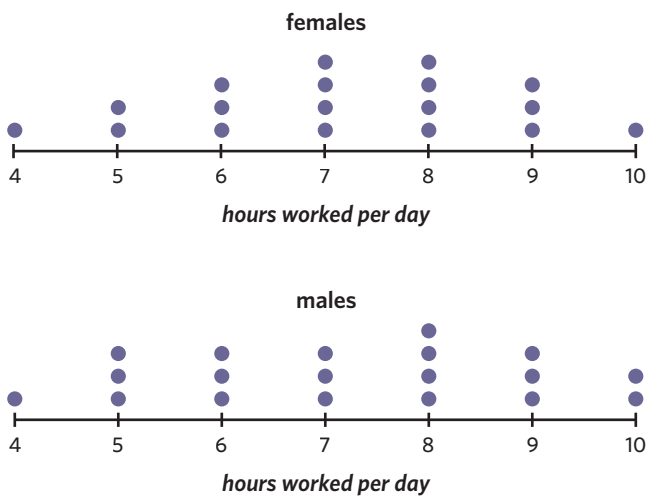
Describing the association between numerical and categorical variables

An association between the categories of the categorical variable can be identified and described by comparing:

- Shape (positively skewed, negatively skewed, approximately symmetrical)
- Centre (median)
- Spread (range, IQR)

If the distributions are similar between the categories of the categorical variable, there is likely to be no association between the variables.

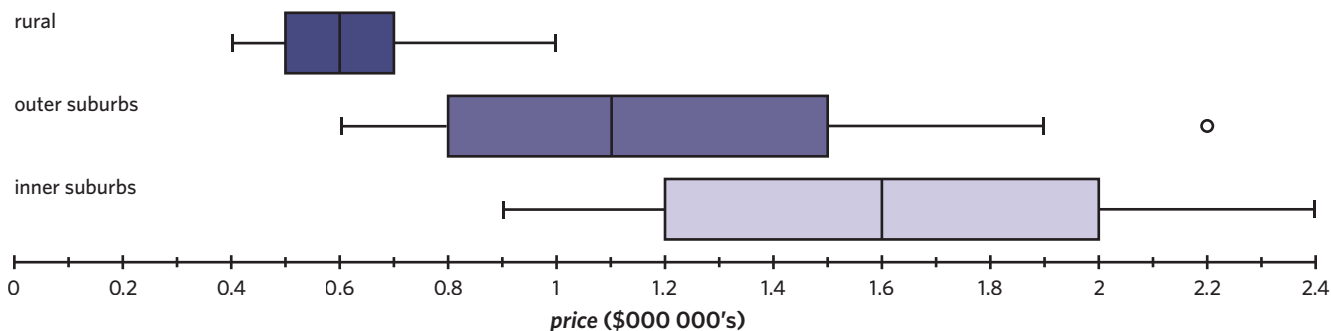
The following parallel dot plots show the number of *hours worked per day* for a group of males and females.



There is no significant difference between the dot plots. Therefore, it is likely that there is no association between *sex* and the number of *hours worked per day*.

Worked example 4

The following parallel boxplots display the distribution of prices for a 3-bedroom house in three different locations in NSW: 'rural', 'outer suburbs' and 'inner suburbs'.



Identify the association between *price* and *location*.

Write a report referencing shape, centre and spread.

Explanation

Step 1: Determine the shapes of the distributions.

'Rural': Positively skewed

'Outer suburbs': Positively skewed

'Inner suburbs': Approximately symmetric

Step 2: Determine the median *price* of each *location*.

'Rural': \$600 000

'Outer suburbs': \$1 100 000

'Inner suburbs': \$1 600 000

Step 3: Calculate the range and *IQR* of each *location*.

$$\text{range} = \text{maximum value} - \text{minimum value}$$

$$\text{'Rural': } 1\,000\,000 - 400\,000 = \$600\,000$$

$$\text{'Outer suburbs': } 2\,200\,000 - 600\,000 = \$1\,600\,000$$

$$\text{'Inner suburbs': } 2\,400\,000 - 900\,000 = \$1\,500\,000$$

$$\text{IQR} = Q_3 - Q_1$$

$$\text{'Rural': } 700\,000 - 500\,000 = \$200\,000$$

$$\text{'Outer suburbs': } 1\,500\,000 - 800\,000 = \$700\,000$$

$$\text{'Inner suburbs': } 2\,000\,000 - 1\,200\,000 = \$800\,000$$

Step 4: Write a report using shape, centre and spread to explain the association.

Answer

The parallel boxplots show that *price* and *location* are associated. The closer a 3-bedroom house is to a city in NSW, the higher the price of the house.

In relation to shape, the distribution of *price* is positively skewed in locations further from a city in NSW and approximately symmetrical in the inner suburbs.

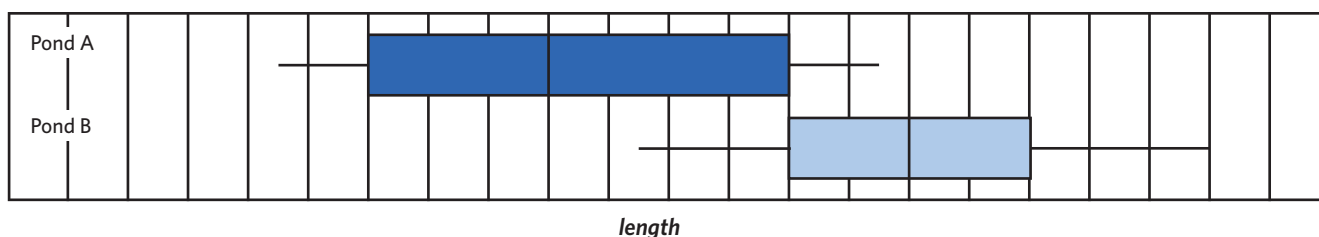
In relation to centre, the median *price* increases as the *location* is closer to a city in NSW.

In relation to spread, the range and *IQR* of *price* tends to be larger in locations closer to a city in NSW.

Exam question breakdown

VCAA 2021 Exam 1 Data analysis Q4

The following boxplots show the distribution of the *length* of fish caught in two different ponds, Pond A and Pond B.



Continues →

Based on the boxplots, it can be said that

- A. 50% of the fish caught in Pond A are the same length as the fish caught in Pond B.
- B. 50% of the fish caught in Pond B are longer than all of the fish caught in Pond A.
- C. 50% of the fish caught in Pond B are shorter than all of the fish caught in Pond A.
- D. 75% of the fish caught in Pond A are shorter than all of the fish caught in Pond B.
- E. 75% of the fish caught in Pond B are longer than all of the fish caught in Pond A.

Explanation

To solve this question, check whether each option is true or false.

A: This is false. Boxplots do not tell us the exact values of individual data points. ✗

B: This is true. The median of Pond B is larger than the maximum of Pond A. ✓

C: This is false. The median of Pond B is not smaller than the maximum of Pond A. ✗

D: This is false. Q_3 of Pond A is not smaller than the minimum of Pond B. ✗

E: This is false. Q_1 of Pond B is not larger than the maximum of Pond A. ✗

Answer

B

60% of students answered this question correctly.

15% of students incorrectly answered E. This is likely because they misinterpreted the Pond A boxplot and did not realise that the rightmost whisker contains 25% of the data.

2B Questions

Displaying data using back-to-back stem plots

- A popular local bakery has recently reduced the size of its workforce, and as such needs to reduce their menu. They are deciding between removing either 'muffins' or 'scones' from their selection based on the number of sales for each. They have recorded the number of sales each day for the month of June in the back-to-back stem plot.

Key: 1 | 9 = 19

muffins		scones
	1	9
	3	0 1 2 2 4 4
	9 8	2 6 6 6 7 8 8 9 9 9
5 4 4 3 3 2 1	3	0 0 1 1 1 2 2 3
9 9 8 8 7 6 6 5	3	5 5 6 7 8
4 4 3 2 2 2 1	4	
	9 8 7 6	4 5
	2	5

Use the stem plot to fill in the gaps in the following sentences.

The median number of 'scones' sold per day is _____ while the median number of 'muffins' sold per day is _____. In order to maximise revenue, the bakery should remove _____ from their selection.

2. The *points scored* by the NSW Blues and Queensland Maroons in the first 20 State of Origin rugby league games are recorded.

NSW: 20 7 5 12 10 22 12 2 22 18 21 6 22 24 18 20 6 8 26 6

Queensland: 16 11 10 24 6 43 29 14 12 2 14 20 16 20 16 16 12 10 18 16

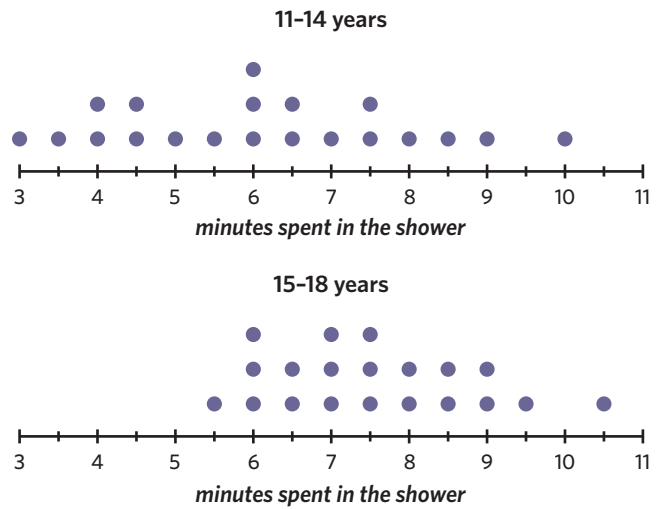
- Construct a back-to-back stem plot to display this data, with the NSW Blues on the left.
- Which team has the larger range of *points scored*?
- Which team has the lower median number of *points scored*?
- In rugby league, a try is worth 4 points, or 6 points if the place kick is accurate, a penalty kick is worth 2 points and a drop goal is worth 1 point. Why do you think most of the leaves in the back-to-back stem plot are even numbers?

Displaying data using parallel dot plots

3. The number of *minutes spent in the shower* by each participant at band camp was recorded correct to the nearest 30 seconds.

The participants were then categorised by age into two groups: '11–14 years' and '15–18 years'.

The results are displayed in the following dot plots.



- In this circumstance, *age* is classified as what kind of variable?
 - Discrete numerical variable
 - Nominal categorical variable
 - Continuous numerical variable
 - Ordinal categorical variable
- Which of these statements is not true?
 - The median number of *minutes spent in the shower* increases as age increases from 11–14 years to 15–18 years.
 - The modal number of *minutes spent in the shower* for 11–14 year olds is 6 minutes.
 - The middle 50% of data for 15–18 year olds is less variable than the middle 50% of data for 11–14 year olds.
 - The number of *minutes spent in the shower* is more variable for 15–18 year olds than for 11–14 year olds.

4. A fortnight worth of *daily sales* of Aesop and Sukin hand moisturiser brands at a local shop were recorded.

Aesop: 12 11 14 12 10 12 13 11 14 16 11 12 11 11

Sukin: 11 5 11 14 12 18 9 10 13 10 8 17 11 12

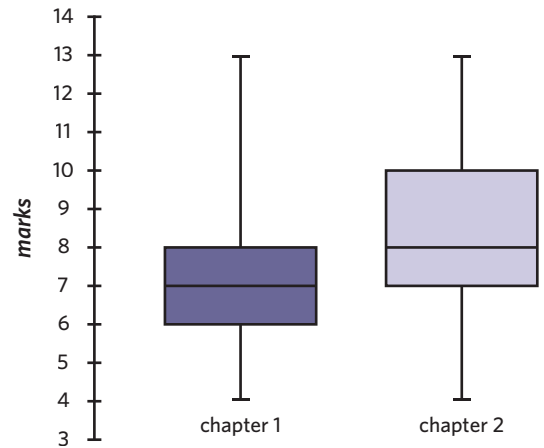
- Construct parallel dot plots to display this data.
- Which brand has *daily sales* that are positively skewed?
- Which brand has a greater *IQR*?

Displaying data using parallel boxplots

5. A Unit 3&4 General Mathematics teacher monitored the results of her class in the Edrolo AOS tests. At the end of the week, the teacher was expected to present the data to the Head of Mathematics. In her report, she displayed the data using parallel boxplots.

Which of the following statements relating to this data is not correct?

- The median chapter 2 result was higher than the median chapter 1 result.
- The lowest 25% of results in the chapter 2 test was better than the lowest 25% of results in the chapter 1 test.
- The *IQR* for the chapter 1 test was greater than the *IQR* for the chapter 2 test.
- The range for the chapter 1 test was the same as the range for the chapter 2 test.



6. The number of *wins* that Essendon and North Melbourne had in each of the seasons in the 2010's is represented in the following five-number summaries.

Essendon: 3, 7, 11.5, 12, 14

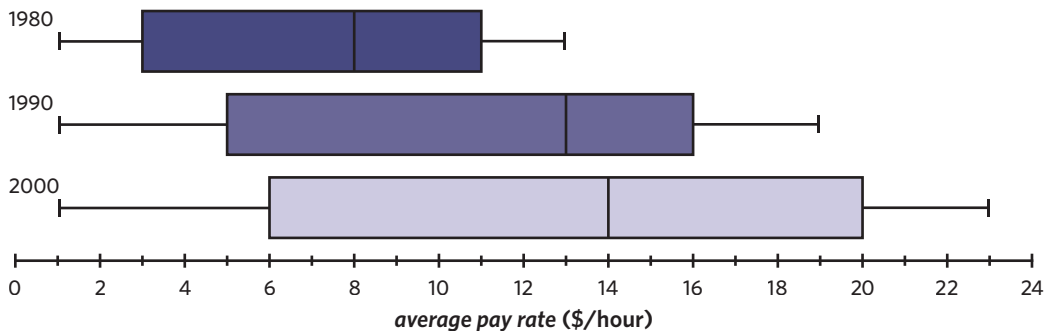
North Melbourne: 6, 10, 11.5, 13, 14

- Construct parallel boxplots to display this data.
- Which team has a greater *IQR* of *wins* in a season?
- Which team has a stronger negative skew in the number of *wins* in a season?

Describing the association between numerical and categorical variables

7. The *average pay rate* (\$/hour) of workers from 50 countries were tracked in the years '1980', '1990' and '2000'.

The results are displayed in the following parallel boxplots.

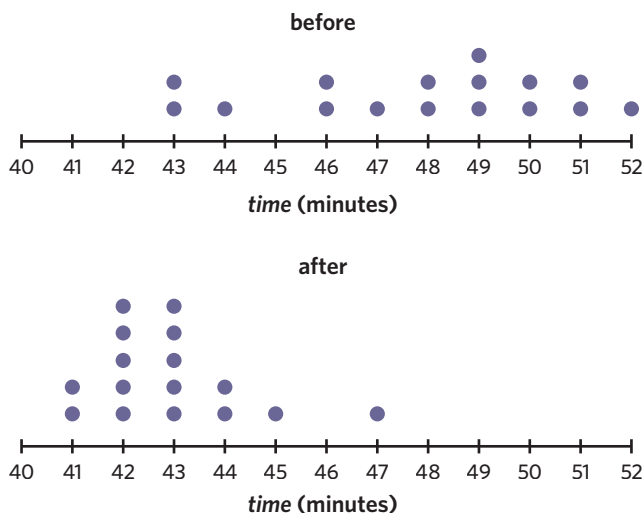


- Which of the following statements is not true?
 - In 1980, over 75% of countries had an *average pay rate* lower than the median *average pay rate* in 1990.
 - In 1990, 75% of the countries had an *average pay rate* lower than the median *average pay rate* in 2000.
 - In 1990, there was more variation in *average pay rate* in the middle 50% of countries than the middle 50% of countries in 1980.
 - In 2000, the top 50% of countries had an *average pay rate* higher than any of the countries in 1980.
- Is there an association between the *average pay rate* in these countries and *year*? Write a brief explanation, referencing centre.

8. Charlotte is a long distance runner. Her coach formulated an intense training program to prepare her for a big competition next year.

In order to see if the training program was effective, Charlotte tracked the *time* it took her to run 10 km (correct to the nearest minute) before and after she completed the program.

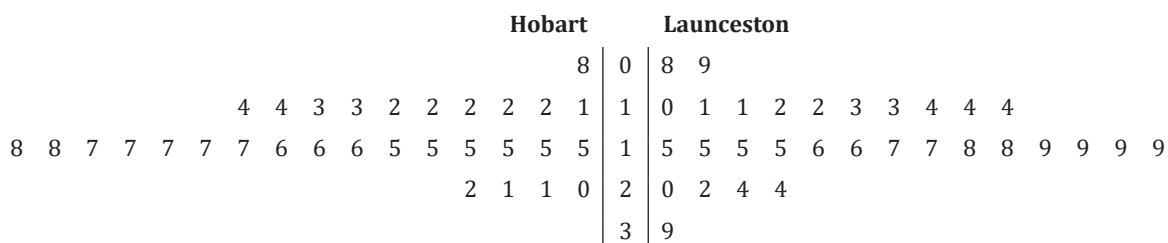
Her results are shown in the following parallel dot plots.



- Did the training program improve Charlotte's running time? Write a brief explanation, referencing centre.
- Did the training program improve Charlotte's consistency? Write a brief explanation, referencing spread.

9. The daily *maximum temperature* ($^{\circ}\text{C}$) was recorded in the month of October for Hobart and Launceston.

Key: 1 | 4 = 14°C



- Is there an association between the daily *maximum temperature* ($^{\circ}\text{C}$) and *location*? Write a brief explanation, referencing centre.
- Is there an association between the variability of the daily *maximum temperature* ($^{\circ}\text{C}$) and *location*? Write a brief explanation, referencing spread.
- Is there an association between the distribution of the daily *maximum temperature* ($^{\circ}\text{C}$) and *location*? Write a brief explanation, referencing shape.

Joining it all together

10. The number of *years of education* of 1000 random adults was recorded in Argentina, Germany and China, and the five-number summaries for each country are calculated.

Argentina: 3, 7, 10, 12, 18

Germany: 5, 11, 14, 16, 20

China: 3, 5, 8, 11, 18

- Construct parallel boxplots to display this data.
- Fill in the blanks:
75% of adults in Germany have more *years of education* than 75% of adults in _____.
The *IQR* of the *years of education* of adults in Argentina is equal to the *IQR* of adults in _____.
- Is there an association between *years of education* and *country*? Write a brief explanation, referencing centre.
- Is there an association between the variability of *years of education* and *country of residence*? Write a brief explanation, referencing spread.

Exam practice

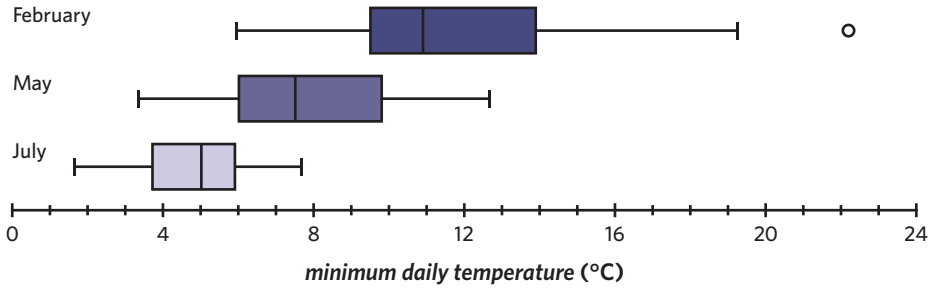
11. Parallel boxplots would be an appropriate graphical tool to investigate the association between the monthly median rainfall, in millimetres, and the
- A. monthly median wind speed, in kilometres per hour.
 - B. monthly median temperature, in degrees Celsius.
 - C. month of the year (January, February, March, etc.).
 - D. monthly sunshine time, in hours.
 - E. annual rainfall, in millimetres.

45% of students answered this question correctly.

VCAA 2016 Exam 1 Data analysis Q8

12. The five-number summary for the distribution of *minimum daily temperature* for the months of February, May and July in 2017 is shown in the table. The associated boxplots are shown following the table.

month	minimum	Q_1	median	Q_3	maximum
February	5.9	9.5	10.9	13.9	22.2
May	3.3	6.0	7.5	9.8	12.7
July	1.6	3.7	5.0	5.9	7.7



Explain why the information given supports the contention that *minimum daily temperature* is associated with the *month*. Refer to the values of an appropriate statistic in your response. (2 MARKS)

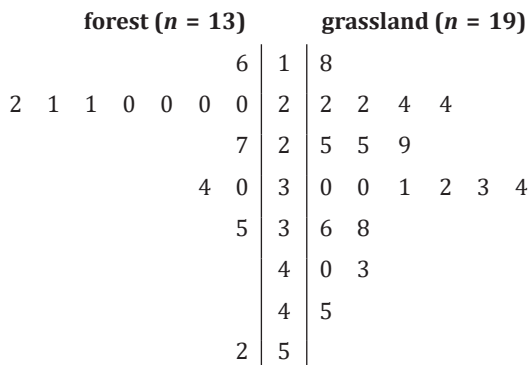
The average mark on this question was 1.

VCAA 2019 Exam 2 Data analysis Q3

13. The following back-to-back stem plot displays the *wingspan*, in millimetres, of 32 moths and their *place of capture* (forest or grassland).

Key: 1 | 8 = 18

wingspan (mm)



The back-to-back stem plot suggests that *wingspan* is associated with *place of capture*. Explain why, quoting the values of an appropriate statistic. (2 MARKS)

The average mark on this question was 1.

VCAA 2017 Exam 2 Data analysis Q2e

Questions from multiple lessons**Data analysis**

14. In 2022, Yohan Blake recorded his time running 100 m weekly, relative to 9.80 seconds. For example, a time of 9.85 seconds would be recorded as 0.05 seconds. Over the year, he had a mean time of 0.03 seconds, and a standard deviation of 0.16 seconds. What is the standardised value of one of his sprints, which was recorded as -0.10 seconds? Round your answer to two decimal places.
- A. 0.81 B. 0.72 C. -8.67 D. -0.72 E. -0.81

Adapted from VCAA 2017 Exam 1 Data analysis Q10

Recursion and financial modelling Year 11 content

15. The number of coins added to a coin collection each month follows a geometric sequence.
- 4 coins were added to the collection in the first month.
8 coins were added to the collection in the second month.
16 coins were added to the collection in the third month.
- Assuming this sequence continues, how many coins would there be in total after five months?
- A. 28 B. 60 C. 64 D. 124 E. 252

Adapted from VCAA 2015 Exam 1 Number patterns Q4

Data analysis

16. IQ is normally distributed with a mean of 100 and a standard deviation of 15.
- Calculate the standardised IQ (z -score) of an individual with an IQ of 73. (1 MARK)
 - What percentage of people are expected to have an IQ over 115? (1 MARK)
 - Estimate the number of people with an IQ between 85 and 130 in a sample of 850 people. Give your answer correct to the nearest whole number. (1 MARK)

Adapted from VCAA 2018NH Exam 2 Data analysis Q2

2C Associations between two numerical variables

STUDY DESIGN DOT POINTS

- response and explanatory variables and their role in investigating associations between variables
- scatterplots and their use in identifying and qualitatively describing the association between two numerical variables in terms of direction (positive/negative), form (linear/non-linear) and strength (strong/moderate/weak)



KEY SKILLS

During this lesson, you will be:

- identifying the response and explanatory variables
- using technology to construct scatterplots
- describing the relationship between two numerical variables.

KEY TERMS

- Response variable
- Explanatory variable
- Scatterplot
- Strength
- Direction
- Form

Once data has been collected for two numerical variables, it is useful to examine them graphically to determine if any associations exist. When doing this, it is important to identify which of the two variables is the response variable and which is the explanatory variable. From here, the data can be plotted on a scatterplot which allows conclusions to be drawn about the relationship between the two variables.

Identifying the response and explanatory variables

The **response variable**, *RV*, may be explained or predicted by changes in the explanatory variable. It can also be called the dependent variable.

The **explanatory variable**, *EV*, is used to explain or predict the changes observed in the response variable. It can also be called the independent variable.

For example, consider the *age* of children and their *height*.

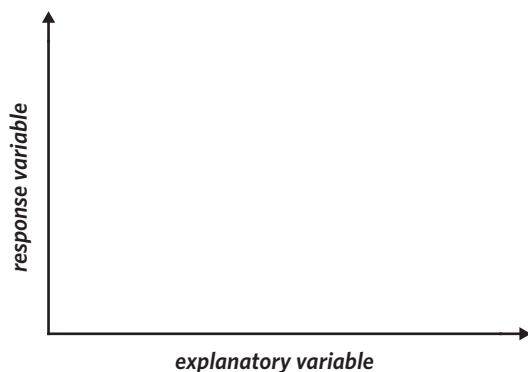
The *height* of children is the response variable, as increases in *height* may be predicted from increases in *age*.

The *age* of children is the explanatory variable, as increases in *age* can predict increases in *height*.

When comparing the relationship between a response variable and its explanatory variable, the data may be represented on a graph. In this case, the response variable is positioned on the vertical axis, and the explanatory variable on the horizontal axis.

See worked example 1

See worked example 2



Worked example 1

The following question was posed: 'Can the *number of ice creams sold* be predicted from the *temperature*?'
Identify the response variable, *RV*, and the explanatory variable, *EV*.

Explanation

Step 1: Assess whether each variable is predicting an outcome or being predicted.

The *number of ice creams sold* is being predicted from changes in *temperature*.

Step 2: Classify each variable as either response or explanatory.
The *RV* is predicted from changes in the *EV*.

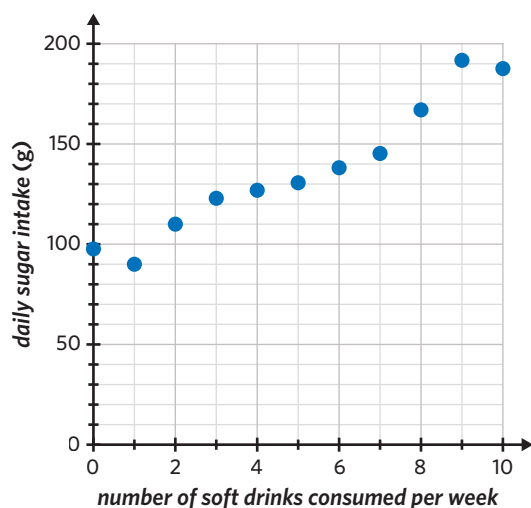
Answer

RV: number of ice creams sold

EV: temperature

Worked example 2

A study was conducted on the association between *daily sugar intake* and the *number of soft drinks consumed per week*. The following scatterplot displays the data that was collected.



Identify the response variable, *RV*, and the explanatory variable, *EV*.

Explanation

Step 1: Recall which variable is positioned on each axis for a scatterplot.

The *RV* is positioned on the vertical axis, and the *EV* on the horizontal axis.

Step 2: Identify the *RV* and the *EV* from the graph.

Answer

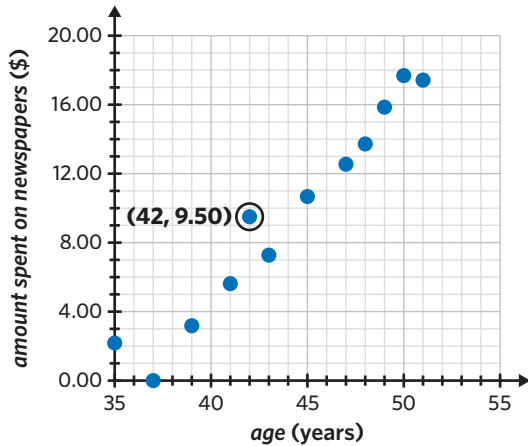
RV: daily sugar intake

EV: number of soft drinks consumed per week

Using technology to construct scatterplots

A **scatterplot** is a display used to represent data relating two numerical variables. Each point represents an individual data entry with the axes providing the numerical measurements. As mentioned previously, the response variable is positioned on the vertical axis, and the explanatory variable is positioned on the horizontal axis.

For example, the point circled on the following scatterplot represents a 42-year-old individual who spent \$9.50 on newspapers.



When provided with a table of data, calculators can be used to construct scatterplots.

Worked example 3

10 students were competing in the school cross-country. Each student's house would be awarded *points* based on their placing in the race. The *time* that it took each student to complete the race and the number of house *points* they received were recorded.

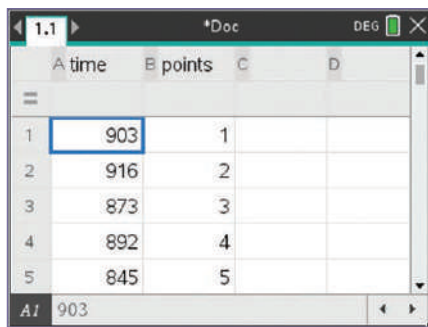
time (seconds)	903	916	873	892	845	823	794	702	740	715
points	1	2	3	4	5	6	7	8	9	10

With a calculator, use the data in the table to construct a scatterplot.

Explanation - Method 1: TI-Nspire

Step 1: From the home screen, select '1: New' → '4: Add Lists & Spreadsheet'.

Step 2: Name a list 'time' and another list 'points' and enter the data from the table.



Step 3: Identify the response and explanatory variables.

Students are awarded *points* based on their *time*, so *time* is used to predict *points*.

RV: points

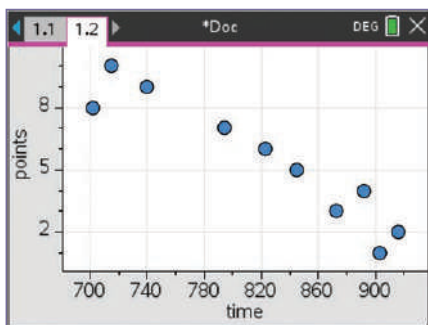
EV: time

Step 4: Press **ctrl** + **doc**, and select 'Add Data & Statistics'.

Step 5: Add the variables on each axis using the 'Click to add variable' function.

The *RV* will be positioned on the vertical axis and the *EV* will be positioned on the horizontal axis.

Answer

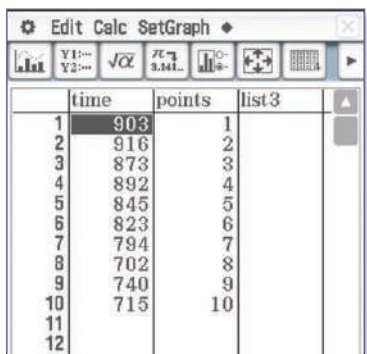


Continues →

Explanation – Method 2: Casio ClassPad

Step 1: From the main menu, tap  Statistics.

Step 2: Name a list 'time' and another list 'points' and enter the data from the table.



	time	points	list3
1	903	1	
2	916	2	
3	873	3	
4	892	4	
5	845	5	
6	823	6	
7	794	7	
8	702	8	
9	740	9	
10	715	10	
11			
12			

Step 3: Identify the response and explanatory variables.

Students are awarded *points* based on their *time*, so *time* is used to predict *points*.

RV: points


EV: time

Step 4: Configure the settings of the graph by tapping  in the icon bar.

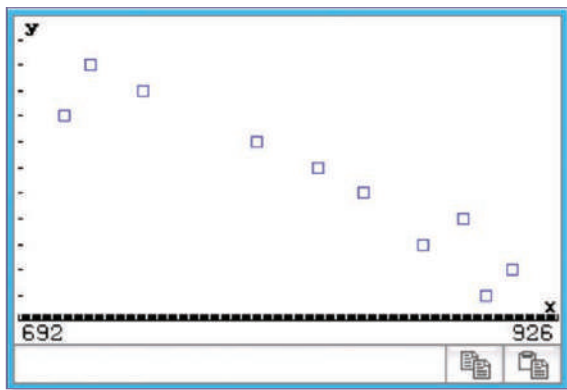
Create a scatterplot by keeping 'Type' as 'Scatter'.

Specify the data set by changing 'XList:' to 'main\time' and 'YList:' to 'main\points'. Tap 'Set' to confirm.

Step 5: Tap  icon bar to plot the graph.

To analyse the graph, tap  in the toolbar. Use the left and right arrow keys to navigate along the data points.

Answer



Describing the relationship between two numerical variables

When looking at scatterplots, the relationship between the two numerical variables can be described in terms of strength, direction and form.

Strength refers to how close the data points are to the general trend of the scatterplot. An association can be described as weak, moderate or strong.

Direction refers to the relationship between the two variables.

Positive relationships occur if the response variable increases as the explanatory variable increases.

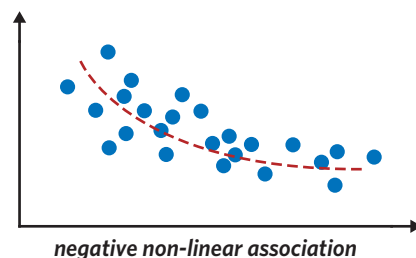
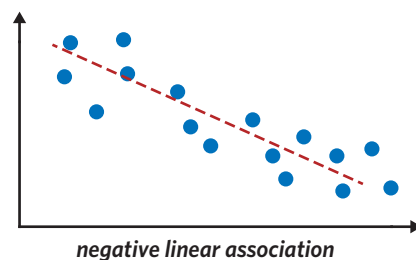
Negative relationships occur if the response variable decreases as the explanatory variable increases.

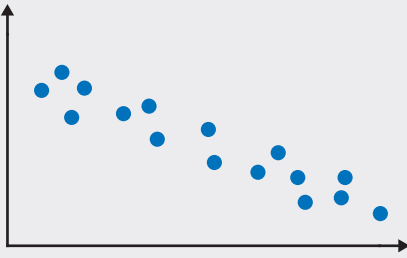
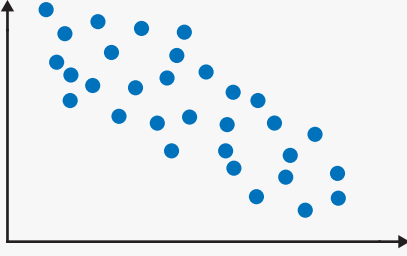
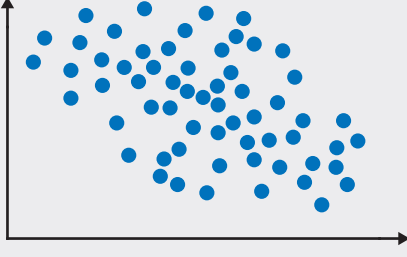
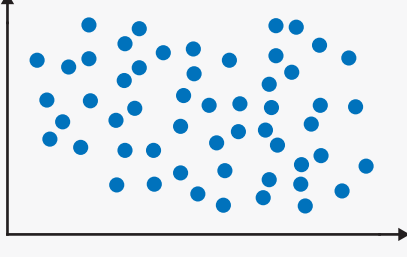
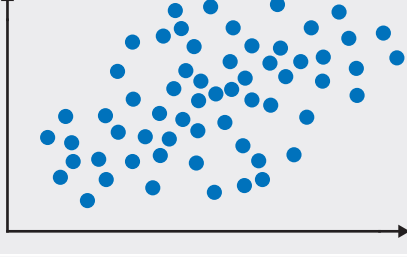
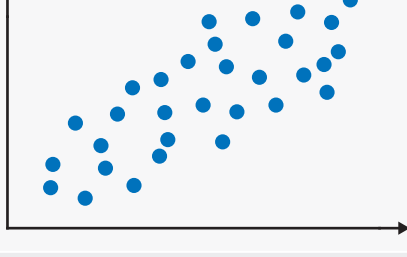
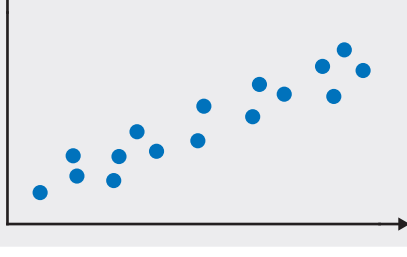
Form refers to whether the relationship is linear or non-linear.

Linear relationships occur when the distribution of data resembles a straight line.

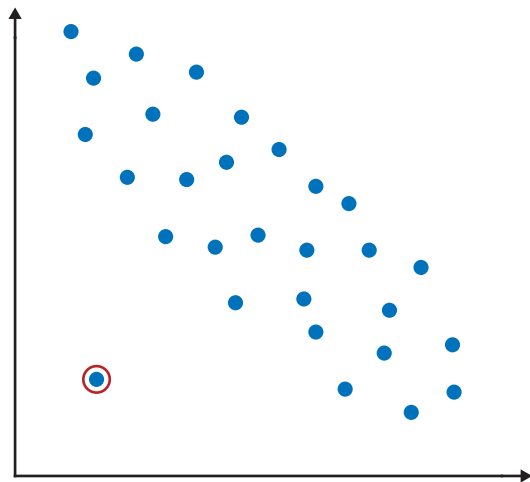
The distribution of data can be described as non-linear if there is a clear association that does not follow a straight line.

Linear relationships will be the primary focus of this chapter.



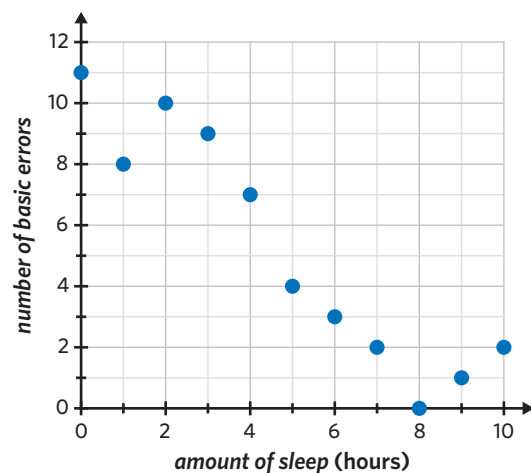
Strength	Direction	Form	Scatterplot
Strong	Negative	Linear	
Moderate	Negative	Linear	
Weak	Negative	Linear	
No association			
Weak	Positive	Linear	
Moderate	Positive	Linear	
Strong	Positive	Linear	

Outliers can also be identified visually, since they exist outside the normal range of the underlying trend.



Worked example 4

An investigation into the association between sleep and work performance was conducted. The following scatterplot displays the collected data.



Describe the relationship in terms of strength, direction, form and potential outliers.

Explanation

Step 1: Determine the strength.

All data points appear to be very close to the underlying trend. The association is strong.

Step 2: Determine the direction.

As the explanatory variable increases, the response variable decreases. The association is negative.

Answer

The scatterplot displays a strong, negative, linear relationship, with no visible outliers.

Step 3: Determine the form.

The distribution of data best resembles a straight line. The form is linear.

Step 4: Identify any outliers.

There are no data points that clearly lie outside the normal range of the underlying trend.

The *age*, in years, *top speed*, in kilometres per hour, and *weight*, in kilograms, of a sample of 12 pandas aged 13 to 15 years are shown in the following table.

<i>age</i> (years)	<i>top speed</i> (km/h)	<i>weight</i> (kg)
13	26.7	70.1
13	25.1	90.4
13	26.5	73.2
13	25.8	85.0
14	26.1	84.3
14	23.5	95.6
14	28.3	71.7
14	23.8	95.0
15	27.3	80.2
15	25.4	87.4
15	24.1	94.9
15	29.6	65.3

A line of best fit is to be fitted to the data with the aim of predicting *top speed* from *weight*.

Name the explanatory variable for this line of best fit. (1 MARK)

Explanation

Step 1: Assess whether each variable is predicting an outcome or being predicted.

The *top speed* is being predicted from changes in *weight*.

Answer

weight

Step 2: Classify each variable as either response or explanatory.

The *RV* is predicted from changes in the *EV*.

78% of students answered this type of question correctly.

Many of the students who answered this type of question incorrectly selected a variable unrelated to the question, such as *age*.

2C Questions

Identifying the response and explanatory variables

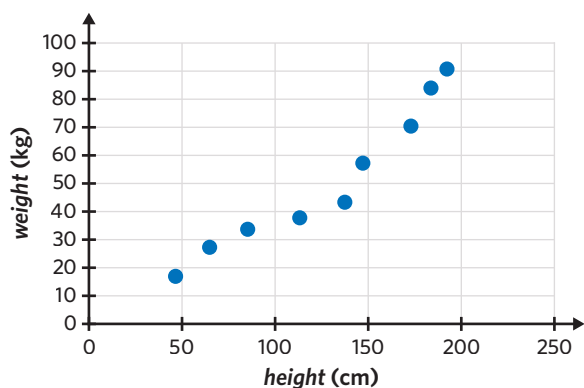
1. Fill in the blanks for the following statement.

The _____ variable is used to predict the changes observed in the _____ variable.
When representing the variables on a scatterplot, the response variable is positioned on the _____ axis and the explanatory variable is positioned on the _____ axis.

- A. response, explanatory, horizontal, vertical
- B. response, explanatory, vertical, horizontal
- C. explanatory, response, horizontal, vertical
- D. explanatory, response, vertical, horizontal

2. Identify the response variable, *RV*, and the explanatory variable, *EV*, in each of the following questions.
- Can the *number of umbrellas sold* on a day be predicted from the *amount of rainfall*?
 - Can the *age* of a second-hand TV predict its *selling price*?
 - Does the *price* of a watermelon depend on the *season*?
 - Do *years of experience* affect an individual's *salary*?
-
3. Identify the response variable, *RV*, and the explanatory variable, *EV*, in each of the following pairs of variables.
- high score* on a game and *time spent playing*
 - distance from work* and *money spent* on petrol
 - amount of rainfall* and *month*
 - age* and *number of homes owned*
-

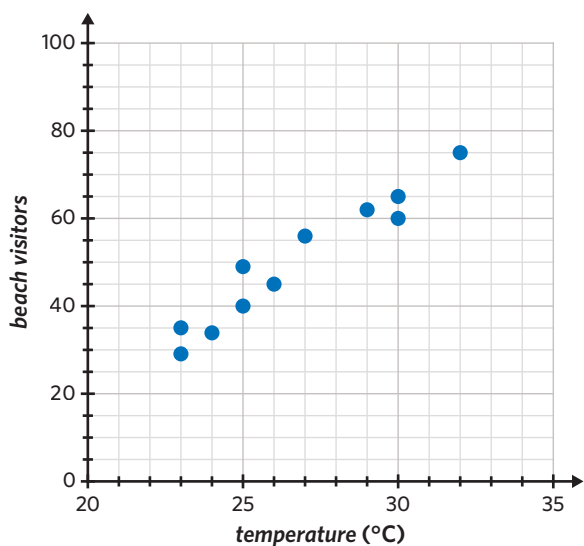
4. From the following graph, identify the response variable and explanatory variable.



5. Darren plans on collecting data to determine whether the *amount of sleep* can be predicted from *time spent watching Grey's Anatomy*.
- Which variable will be positioned on the horizontal axis if Darren wants to display the data on a scatterplot?

Using technology to construct scatterplots

6. Consider the following scatterplot.



The table used to construct this scatterplot is

- A.

<i>temperature (°C)</i>	23	23	24	25	25	26	27	29	30	30	32
<i>beach visitors</i>	40	30	35	50	50	45	55	75	60	65	75
- B.

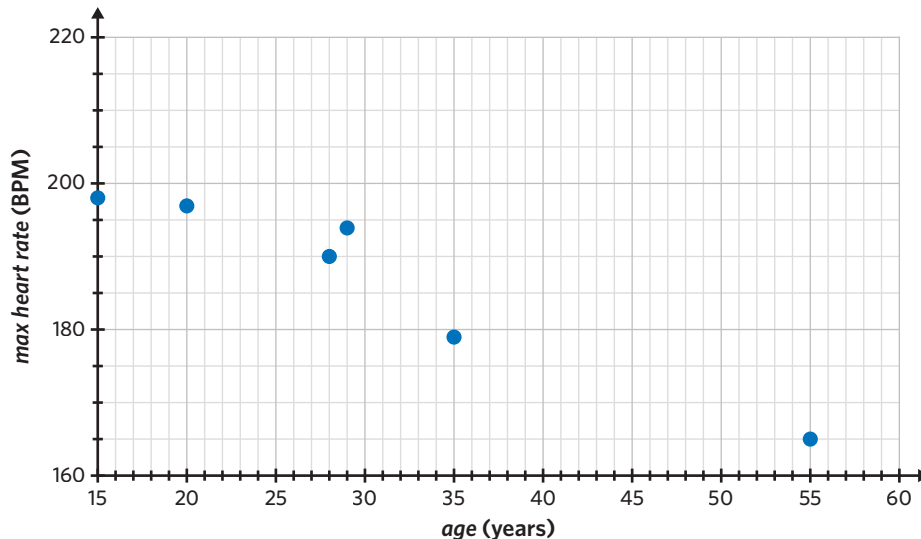
<i>temperature (°C)</i>	23	23	24	25	25	26	27	29	30	30	32
<i>beach visitors</i>	35	29	34	40	49	45	56	62	60	65	75
- C.

<i>temperature (°C)</i>	35	29	34	40	49	45	56	62	60	65	75
<i>beach visitors</i>	23	23	24	25	25	26	27	29	30	30	32
- D.

<i>temperature (°C)</i>	25	25	30	30	27	30	35	30	30	30	35
<i>beach visitors</i>	30	30	35	43	55	40	62	65	70	65	75

7. The following table was used to make the scatterplot shown. Fill in the missing data points that haven't been added to the scatterplot.

<i>age (years)</i>	18	15	42	35	55	29	28	20	31
<i>max heart rate (BPM)</i>	205	198	186	179	165	194	190	197	188



8. The *musical performance* and *mathematical performance* of 10 students were measured on two 100-mark tests. A teacher wanted to determine whether a student's *musical performance* can be predicted from their *mathematical performance*.
Use a CAS to construct a scatterplot from the following data showing the association between *musical performance* and *mathematical performance*. The explanatory variable is *mathematical performance*.

<i>mathematical performance</i>	43	62	84	97	39	53	67	71	78	91
<i>musical performance</i>	57	64	91	93	37	60	69	79	75	82

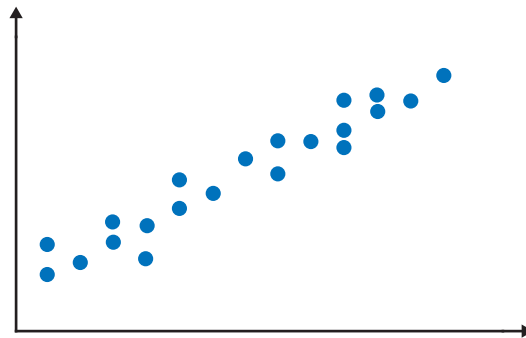
9. An English teacher wished to reduce the number of spelling mistakes her students made. She documented the *number of spelling mistakes* made in each student's essay and asked each student how many books they had read over the past year. Use a CAS to construct a scatterplot from the following data.

<i>number of spelling mistakes</i>	1	2	7	4	9	10	14	19	23	28
<i>number of books read</i>	8	10	7	5	4	3	4	1	2	0

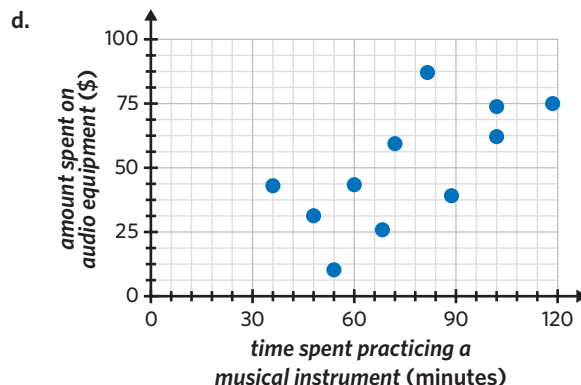
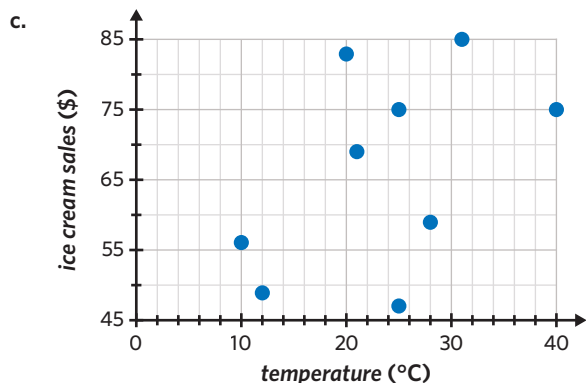
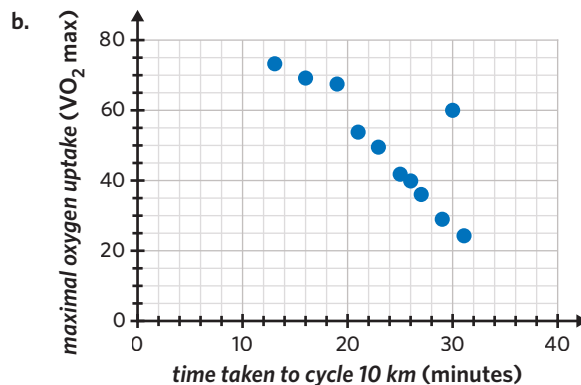
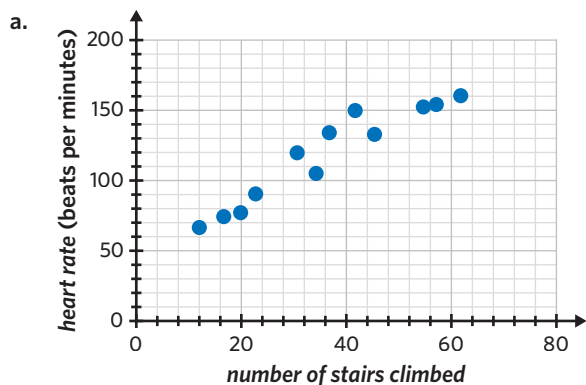
Describing the relationship between two numerical variables

10. The relationship in the scatterplot shown can be described as

- A. weak, positive and linear.
- B. strong, positive and linear.
- C. weak, negative and linear.
- D. strong, negative and linear.

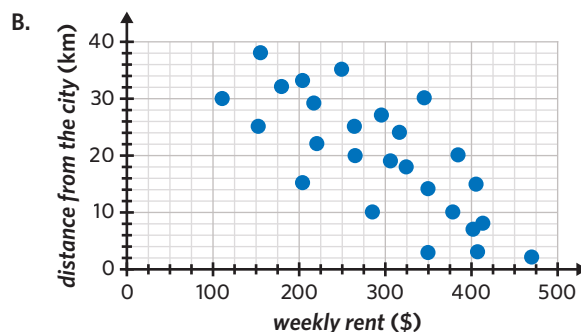
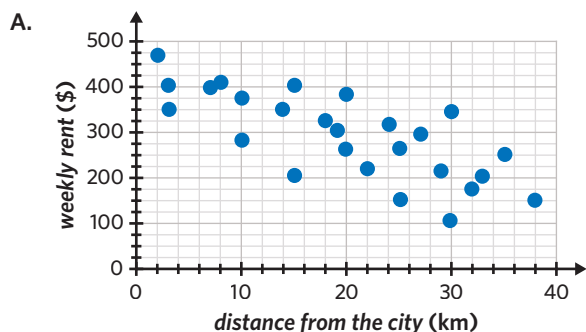


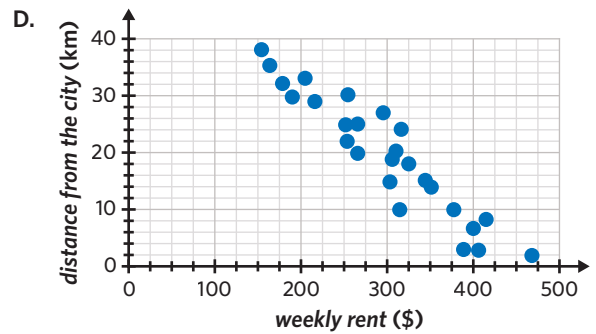
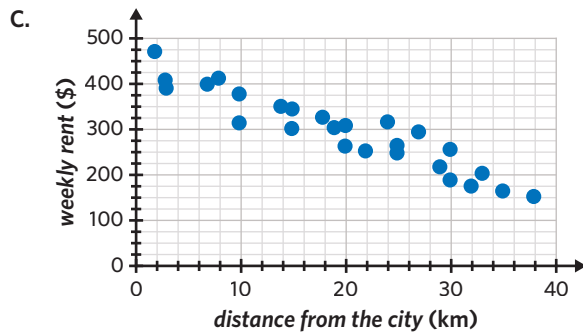
11. For each of the following scatterplots, describe the association between the variables in terms of strength, direction, form and potential outliers.



Joining it all together

12. Mavis collected data on how close her friends live to the city and how much they pay in *weekly rent*, and found that *distance from the city*, in km, was a useful predictor of *weekly rent*, in \$. She constructed a scatterplot which displayed a moderate, negative, linear relationship. Which of the following graphs could be the scatterplot that Mavis constructed from her data?





13. In Victoria, if an individual is under 21 years old, they need to have 120 hours of logged practice with a qualified driver before attempting the driving test to obtain their Probationary Driver's Licence (P's). VicRoads wants to review this legislation and has conducted a study.

The study tested ten 18-year-olds who had no prior driving experience, but were then permitted to do up to 150 hours of practice before taking the driving test. Their number of mistakes on the test were recorded.

- The association between *hours of practice* and *number of mistakes* was investigated. Identify the response and explanatory variables.
- VicRoads has provided the following data. Use a CAS to construct a scatterplot representing this data.

<i>hours of practice</i>	80	137	22	77	61	110	6	51	150	30
<i>number of mistakes</i>	3	1	5	3	4	2	8	4	0	6

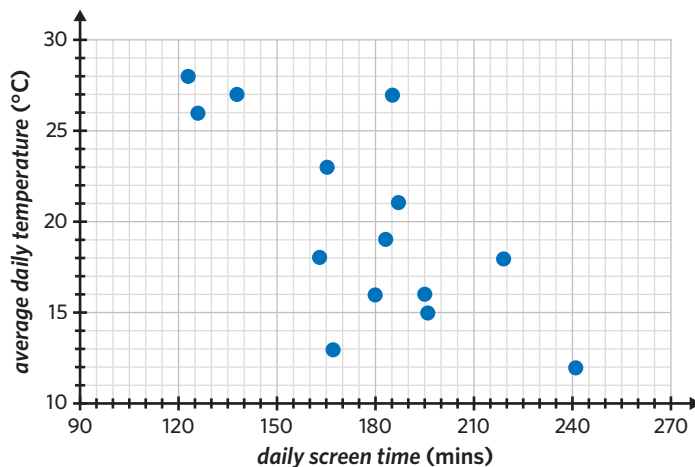
- Describe the association between the variables in terms of strength, direction and form.

14. Francis wants to determine whether his *daily screen time*, in minutes, on his phone can be predicted by the *average daily temperature*. He assumes that he will spend more time on his phone if it is colder outside.

Francis collects two weeks of data and presents them in the following table.

<i>average daily temperature (°C)</i>	16	23	18	21	28	12	13	16	27	26	27	19	15	18
<i>daily screen time (min)</i>	195	165	219	187	123	241	167	180	138	126	185	183	196	163

He uses the data to construct the following scatterplot.



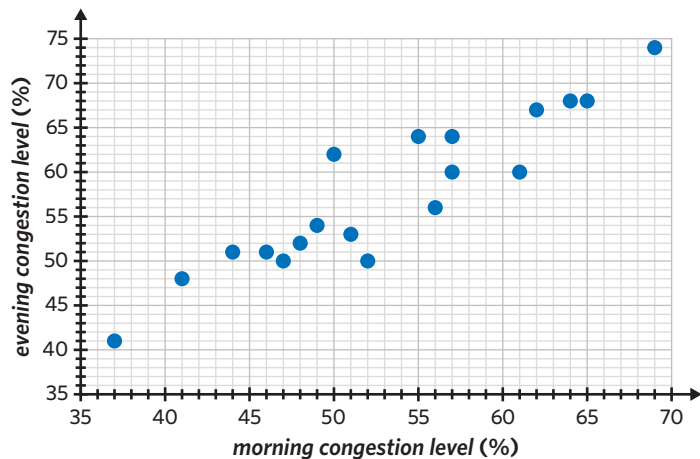
- Identify Francis' mistake in the scatterplot he has constructed.
- Use your CAS to reconstruct the scatterplot from the table.
- Describe the association between the variables in terms of strength, direction and form.

Exam practice

15. The congestion level in a city can be recorded as the percentage increase in travel time due to traffic congestion in peak periods (compared to non-peak periods).

This is called the percentage congestion level.

The percentage congestion levels for the morning and evening peak periods for 19 large cities are plotted on the following scatterplot.



A line of good fit is to be fitted to the data with the aim of predicting *evening congestion level* from *morning congestion level*.

Name the response variable in this line of good fit. (1 MARK)

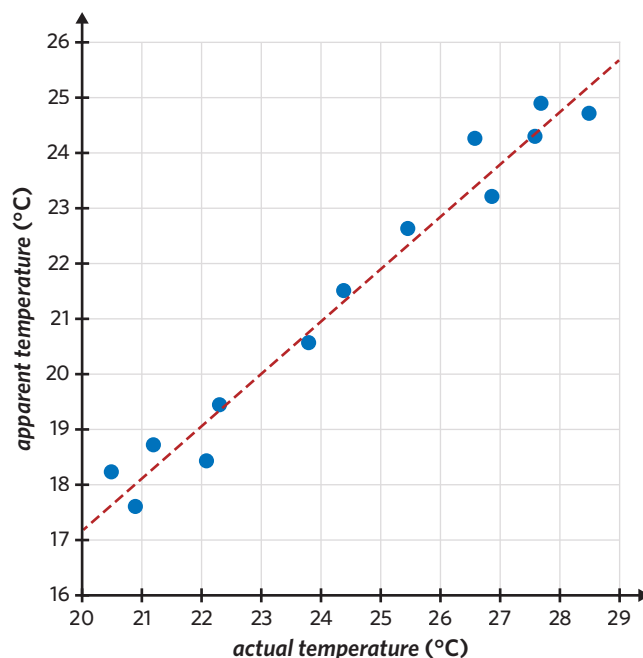
Adapted from VCAA 2018 Exam 2 Data analysis Q2b

90% of students answered this type of question correctly.

16. The data in the following table shows a sample of actual temperatures and apparent temperatures recorded at the weather station. A scatterplot of the data is also shown.

The data will be used to investigate the association between the variables *apparent temperature* and *actual temperature*.

<i>apparent temperature</i> (°C)	<i>actual temperature</i> (°C)
24.7	28.5
24.3	27.6
24.9	27.7
23.2	26.9
24.2	26.6
22.6	25.5
21.5	24.4
20.6	23.8
19.4	22.3
18.4	22.1
17.6	20.9
18.7	21.2
18.2	20.5



Use the scatterplot to describe the association between *apparent temperature* and *actual temperature* in terms of strength, direction and form. (1 MARK)

Adapted from VCAA 2016 Exam 2 Data analysis Q3a

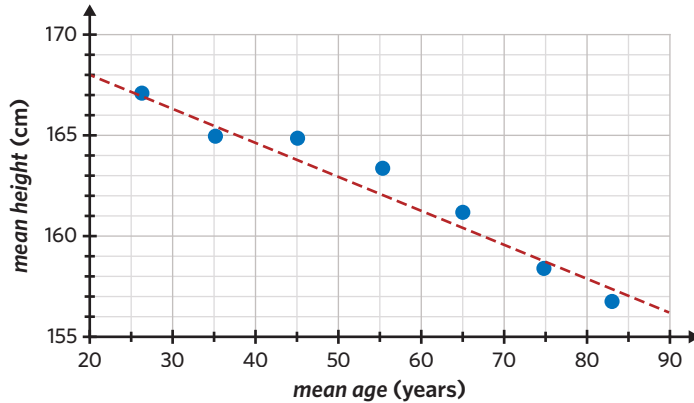
78% of students answered this type of question correctly.

17. The following table shows the *mean age*, in years, and the *mean height*, in centimetres, of 648 women from seven different age groups.

	<i>age group</i>						
	Twenties	Thirties	Forties	Fifties	Sixties	Seventies	Eighties
<i>mean age</i> (years)	26.3	35.2	45.2	55.3	65.1	74.8	83.1
<i>mean height</i> (cm)	167.1	164.9	164.8	163.4	161.2	158.4	156.7

Data: J Sorkin et al., 'Longitudinal change in height of men and women: Implications for interpretation of the body mass index', *American Journal of Epidemiology*, vol. 150, no. 9, 1999, p. 971

A scatterplot displaying this data shows an association between the *mean height* and *mean age* of these women. In an initial analysis of the data, a line is fitted by eye, as shown.



Describe this association in terms of strength and direction. (1 MARK)

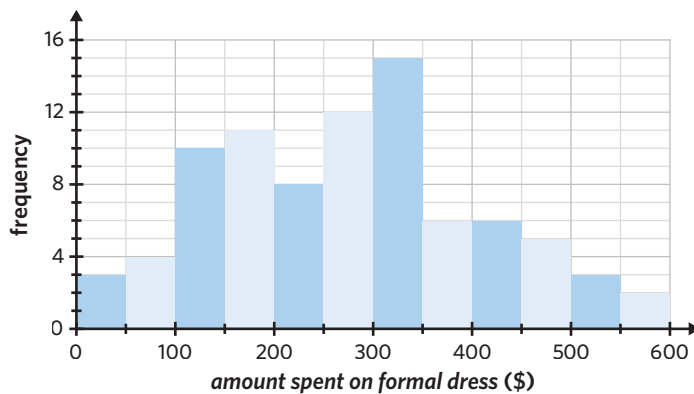
VCAA 2020 Exam 2 Data analysis Q6b

51% of students answered this question correctly.

Questions from multiple lessons

Data analysis Year 11 content

18. An elite private school, Edrollington Ladies' Academy has just had their Year 12 Formal. The following histogram displays the distribution of the *amount spent on formal dress*, in dollars, of a sample of 85 students at the school.



The *amount spent on formal dress* for this sample is most frequently

- greater than or equal to \$150 and less than \$200.
- greater than or equal to \$200 and less than \$250.
- greater than or equal to \$250 and less than \$300.
- greater than or equal to \$300 and less than \$350.
- greater than or equal to \$350 and less than \$400.

Adapted from VCAA 2018NH Exam 1 Data analysis Q3

Data analysis

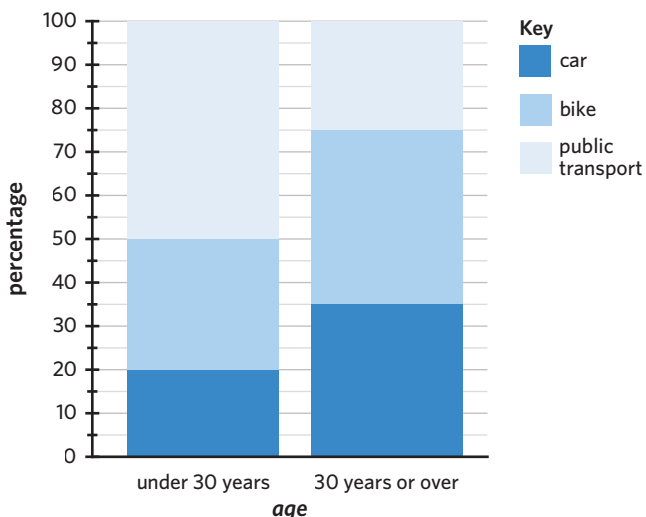
19. A survey was conducted with a sample of workers in a Melbourne office to investigate the association between *age* (under 30 years, 30 years or over) and *transport to work* (car, bike, public transport). The distribution is shown in the following percentage segmented bar chart.

There are 80 people under 30 years of age and 120 people 30 years or over at the office.

What is the number of workers aged 30 years or older that ride their bike to work?

- A. 20
B. 30
C. 40
D. 42
E. 48

Adapted from VCAA 2017 Exam 1 Data analysis Q5



Recursion and financial modelling Year 11 content

20. Roger recently purchased a commercial coffee machine.

The value of his coffee machine is depreciated with each coffee made using the unit cost method of depreciation.

The value of his coffee machine, in dollars, after n coffees are made, V_n , can be modelled by the recurrence relation $V_0 = 8000$, $V_{n+1} = V_n - 0.45$

- a. What is the value of the machine after one coffee has been made? (1 MARK)
b. What is the value of the machine after 58 coffees have been made? (1 MARK)

Adapted from VCAA 2017 Exam 2 Recursion and financial modelling Q5a

2D Correlation and causation

STUDY DESIGN DOT POINTS

- Pearson correlation coefficient, r , its calculation and interpretation
- cause and effect; the difference between observation and experimentation when collecting data and the need for experimentation to definitively determine cause and effect
- answering statistical questions that require a knowledge of the associations between pairs of variables

2A

2B

2C

2D

KEY SKILLS

During this lesson, you will be:

- calculating and interpreting the Pearson correlation coefficient
- distinguishing between correlation and causation.

KEY TERMS

- Pearson's correlation coefficient
- Causation
- Common response
- Confounding variables
- Coincidence
- Observation
- Experimentation

When looking at associations between two numerical variables, it can be helpful to use statistics such as the Pearson correlation coefficient to interpret the data. This is done to make definitive statements about observed associations. While the correlation coefficient provides information on associations, it is important to use appropriate experimentation to determine if an observed association demonstrates causation.

Calculating and interpreting the Pearson correlation coefficient

Pearson's correlation coefficient, r (also known simply as the correlation coefficient), is used to determine the strength and direction of a linear relationship between two numerical variables. When calculating and interpreting the correlation coefficient, the following assumptions are made:

- Data distribution is linear
- Data is numeric
- No outliers are present

The correlation coefficient can be calculated from the following formula:

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

where x and y represent the two numerical variables and n is the number of data values. However it is more efficient to calculate using a calculator.

The Pearson correlation coefficient ranges from -1 to 1 . The value of the correlation coefficient determines the strength and direction of a linear association.

A positive r value indicates a positive association, while a negative r value indicates a negative association. The closer r is to -1 or 1 , the stronger the association. The closer r is to 0 , the weaker the association.

$0.75 \leq r \leq 1$	Strong, positive, linear association
$0.5 \leq r < 0.75$	Moderate, positive, linear association
$0.25 \leq r < 0.5$	Weak, positive, linear association
$-0.25 < r < 0.25$	No association
$-0.5 < r \leq -0.25$	Weak, negative, linear association
$-0.75 < r \leq -0.5$	Moderate, negative, linear association
$-1 \leq r \leq -0.75$	Strong, negative, linear association

Worked example 1

A survey followed the study habits of a group of first-year uni students with the aim of predicting their *grade*, as a percentage, from the number of *lectures attended*. The results are shown in the following table.

<i>grade (%)</i>	75	61	92	53	47	86	74
<i>lectures attended</i>	11	6	10	5	2	8	9

- a. Is the Pearson correlation coefficient, r , an appropriate statistic for this data set?

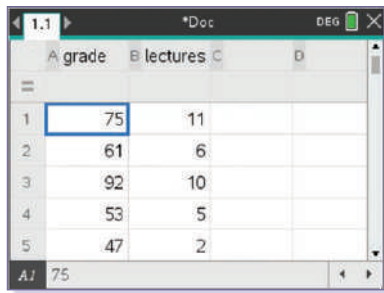
Explanation - Method 1: TI-Nspire

Step 1: From the home screen, select '1: New' → '4: Add Lists & Spreadsheet'.

Step 2: Name column A 'grade' and column B 'lectures'.

Enter the *grade* values into column A, starting from row 1.

Enter the *lectures attended* values into column B, starting from row 1.



Step 3: Determine the response and explanatory variables.

The variable *grade* is being predicted from the variable *lectures attended*, so *grade* is the response variable.

RV: *grade*

EV: *lectures attended*

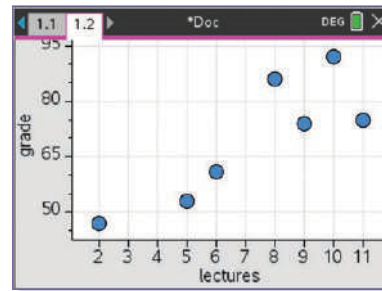
Step 4: Press **ctrl** + **doc**, and select '5: Add Data & Statistics'.

Move the cursor to the horizontal axis and select 'Click to add variable'.

Select 'lectures'.

Move the cursor to the vertical axis and select 'Click to add variable'.

Select 'grade'.



Step 5: Determine whether the scatterplot shows a linear or non-linear relationship.

There is no definitive curve or non-linear relationship in this scatterplot. Hence, r is an appropriate statistic.

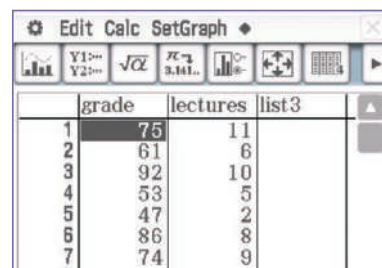
Explanation - Method 2: Casio ClassPad

Step 1: From the main menu, tap Statistics.

Step 2: Name the first list 'grade' and the second list 'lectures'.

Enter the *grade* values into list 'grade', starting from row 1.

Enter the *lectures attended* values into list 'lectures', starting from row 1.



Continues →

Step 3: Determine the response and explanatory variables.

The variable *grade* is being predicted from the variable *lectures attended*, so *grade* is the response variable.

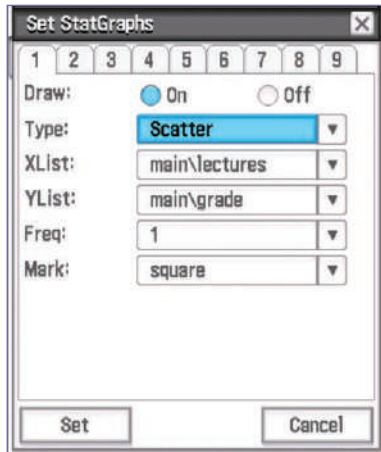
RV: *grade*

EV: *lectures attended*

Step 4: Configure the settings of the graph by tapping .

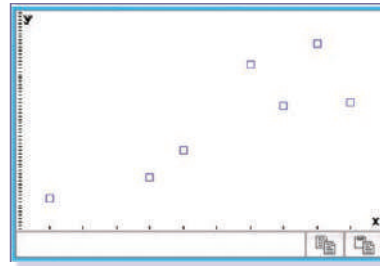
Create a scatterplot by changing 'Type' to 'Scatter'.

Specify the data set by changing 'XList:' to 'main\lectures' and 'YList:' to 'main\grade'.



Tap 'Set' to confirm.

Step 5: Tap  in the icon bar to plot the graph.



Step 6: Determine whether the scatterplot shows a linear or non-linear relationship.


There is no definitive curve or non-linear relationship in this scatterplot. Hence, r is an appropriate statistic.

Answer - Method 1 and 2

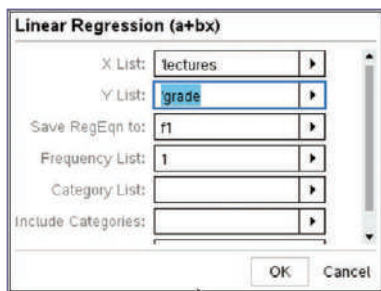
Yes, the Pearson correlation coefficient is an appropriate statistic for this data set.

- b. Determine the value of the correlation coefficient. Give the value correct to two decimal places.

Explanation - Method 1: TI-Nspire

Step 1: From the 'Lists & Spreadsheet' page, press  and select '4: Statistics' → '1: Stat Calculations' → '4: Linear Regression (a+bx)'.

The variable *lectures attended* is the explanatory variable and the variable *grade* is the response variable, so select 'lectures' in 'X List:' and 'grade' in 'Y List:'



Select 'OK'.

Step 2: Read the r value from the screen and round to two decimal places.

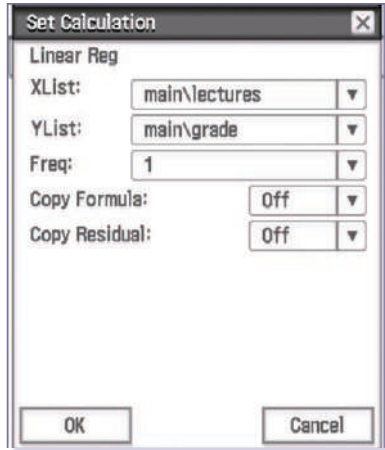
Row	Column	Value
2	RegEqn	a+b*x
3	a	36.7885
4	b	4.51923
5	r ²	0.722708
6	r	0.850122
E6		=0.8501222740023

Continues →

Explanation - Method 2: Casio ClassPad

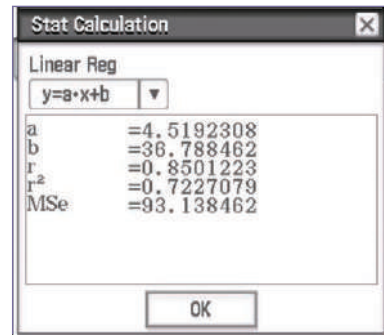
Step 1: Tap 'Calc' → 'Regression' → 'Linear Reg'.

The variable *lectures attended* is the explanatory variable and the variable *grade* is the response variable, so select 'main\lectures' in 'XList:' and 'main\grade' in 'YList:'.



Tap 'OK' to confirm.

Step 2: Read the r value from the screen and round to two decimal places.



Answer - Method 1 and 2

$$r = 0.85$$

- c. If possible, describe the association between the variables in terms of strength, direction, and form.

Explanation

Step 1: Recall the r value from part b.

$$r = 0.85$$

Step 2: Determine the strength of the association.

$$0.75 \leq 0.85 \leq 1$$

This indicates a strong association.

Step 3: Determine the direction of the association.

0.85 is positive, indicating a positive association.

Answer

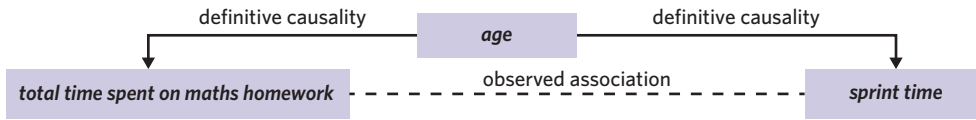
Strong, positive, linear association

Distinguishing between correlation and causation

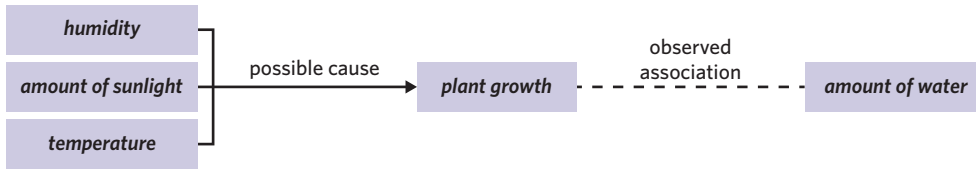
While two variables may have a strong correlation, this doesn't necessarily mean that the association implies causation. **Causation** occurs when a change in the explanatory variable definitively causes the observed change in the response variable. A correlation between two variables that do not have a causal relationship can occur due to three different circumstances.

A **common response** to a third variable may be present in the two variables with the observed association. For example, in a primary school, there is a negative correlation between *total time spent on maths homework* and *sprint time*. This means that children who have spent more time on maths homework are associated with faster sprint times. However, it is unlikely that increasing the *total time spent on maths homework* actually decreases a child's *sprint time*.

A common response to a third variable, *age*, is the likely cause of the correlation. Older students will have accumulated more hours on maths homework and they will also have become faster, resulting in lower sprint times.



There could be **confounding variables**, or external variables, that also produce a change in the response variable. For example, a positive association is found between *plant growth* and the *amount of water* poured on the plant per day. Although the *plant growth* could have been affected by a change in *amount of water*, other related factors such as *amount of sunlight*, *temperature* and *humidity* could also have an effect.



Lastly, a correlation may exist by pure **coincidence**. For example, a positive association between *number of siblings* and *number of mistakes* on a spelling test could be purely coincidental.

While **observation** (collecting data to identify an association) is sufficient to identify a correlation, more investigation is required to determine if causation between the variables exist. **Experimentation** is required in order to determine causation in an association. This involves observing or modifying changes in the explanatory variable and recording the changes in the response variable, while all possible external variables remain constant.

If the changes in the explanatory variable result in a persistent change in the response variable, it is reasonable to conclude causation.

Worked example 2

Gunther runs a cafe and collects data throughout the year on the daily number of *hot chocolates purchased*. He finds that there is a negative association between the daily number of *hot chocolates purchased* and the number of *daylight hours* on a particular day.

- a. Identify a potential third variable that could have caused this association.

Explanation

During colder seasons, the sun rises later and sets earlier, causing a decreased number of *daylight hours*.

During colder seasons, hot chocolate is more popular, causing an increased daily number of *hot chocolates purchased*.

Therefore, a potential third variable is the *season*.

Answer

season

- b. Has Gunther conducted an appropriate experiment to determine causation? Why or why not?

Explanation

Step 1: Recall the requirements for an experiment to determine causation.

All possible external variables must remain constant for an experiment to determine causation.

Step 2: Identify any external variables in the experiment.

From part **a**, it was determined that the *season* could cause a common response. Gunther has collected data across the whole year, so this variable does not remain constant.

Answer

No. The variable *season* does not remain constant across the experiment.

Exam question breakdown

VCAA 2016 Exam 1 Data analysis Q12

There is a strong positive association between a country's Human Development Index and its carbon dioxide emissions.

From this information, it can be concluded that

- A. increasing a country's carbon dioxide emissions will increase the Human Development Index of the country.
- B. decreasing a country's carbon dioxide emissions will increase the Human Development Index of the country.
- C. this association must be a chance occurrence and can be safely ignored.
- D. countries that have higher human development indices tend to have higher levels of carbon dioxide emissions.
- E. countries that have higher human development indices tend to have lower levels of carbon dioxide emissions.

Explanation

Step 1: Interpret the association.

A strong positive association means that an increase in the explanatory variable generally results in an increase in the response variable. Similarly, a decrease in the explanatory variable also generally results in a decrease in the response variable. The only options that follow this relationship are A and D.

Step 2: Determine if causation is present.

An experiment needs to be conducted to determine causation instead of just observing an association. In this situation, no experiment has been conducted, so we are unable to conclude that an increase in the explanatory variable definitively causes an increase in the response variable.

Answer

D

67% of students answered this question correctly.

11% of students incorrectly answered A. These students interpreted the association correctly but didn't identify that the association doesn't necessarily mean causation. A further 11% of students incorrectly answered E. These students recognised that observing the association only shows correlation rather than causation, but misinterpreted the direction of the association.

2D Questions

Calculating and interpreting the Pearson correlation coefficient

1. An r value of 1.00 indicates an association that is
 - A. strong, positive, and linear
 - B. moderate, positive, and linear
 - C. moderate, negative, and linear
 - D. strong, negative, and linear
2. In which of the following circumstances would calculating the correlation coefficient, r , be most useful?
 - A. Bradley is collecting data on the *favourite sport* of the students in his class.
 - B. The *weight* (g) and *leg length* (mm) of 10 different mice was measured.
 - C. Susie records how long it takes her to get to school on each day of the school week.
 - D. Raj measures the distance between his house and the five closest shopping centres.

3. Max and Penelope collect data on the *goals conceded* by their soccer team each week and the *average sleep*, in hours, of all members of the soccer team the night before each game. They calculated the value of the correlation coefficient to be $r = -0.21$.

Max says that *goals conceded* and *average sleep* have a negative association.

Penelope says there is no association.

Who is correct and why?

4. A group of people were surveyed on the number of hours they spent playing Pokémon GO over one weekend and how many Pokémon they caught in that time. The results are shown in the following table.

<i>hours spent playing Pokémon GO</i>	7	3	4	8	16	9	11	10	9
<i>Pokémon caught</i>	13	2	6	10	38	11	23	16	12

- Is the Pearson correlation coefficient, r , an appropriate statistic for this data set? Assume that *hours spent playing Pokémon GO* is the explanatory variable.
- Determine the value of Pearson's correlation coefficient, r , for the data shown correct to two decimal places.
- If possible, describe the association between the variables in terms of strength, direction, and form.

5. A select group of rural towns were investigated through the census. For each rural town, the *male population* and *average height* was recorded, with *average height* as the response variable. The data is shown in the following table.

<i>male population</i>	3158	7562	9901	1875	2238	10 027	8723	5267	4547
<i>average height (m)</i>	1.76	1.83	1.79	1.86	1.78	1.76	1.84	1.80	1.77

- Determine the value of Pearson's correlation coefficient, r , for the data shown correct to two decimal places.
- If possible, describe the association in terms of strength, direction, and form.

Distinguishing between correlation and causation

6. A strong, positive, linear association was observed between two variables that are seemingly unrelated. A circumstance that could have caused this association is
- a common response to a third variable.
 - an external variable(s) that also causes a change in the response variable.
 - a coincidence.
 - all of the above.
7. The total number of *Nobel Prize winners* from each country was found to have a strong, positive correlation ($r = 0.84$) with the number of *space rock impacts* per year within the country. This strong correlation is likely due to
- a common response to a third variable.
 - an external variable(s) that also causes a change in the response variable.
 - a coincidence.
 - a direct causal relationship.
8. Axel runs a bakery with a wide range of products and has found a positive correlation between the number of *cinnamon doughnuts sold* and the *total profit* each day. As a result, he decides to discontinue some popular items to make more cinnamon doughnuts and further increase his profit. Has Axel made a logical decision? If not, identify another variable that may also influence the *total profit* to justify your answer.

9. In a study of American cities, there was found to be a positive correlation between the number of *gyms* and the number of *crimes committed* in each city. What variable could cause a common response that would explain this association?
-
10. Robbie looks at the screen time statistics on his phone to determine whether there is an association between the *time spent on Facebook* and *time spent on Instagram* per day, both in minutes. He collects the data every day for 6 months, and finds that there is a strong association.
- Identify three confounding variables that could explain the association.
 - How could Robbie redesign his experiment so that he would be better able to determine causation?

Joining it all together

11. An association has been observed between the weekly *cream cheese sales* in Forest Hill and the weekly number of *parking tickets* handed out in Sunshine. Eight weeks of data are shown in the following table.

<i>cream cheese sales</i>	46	55	53	61	78	64	95	79
<i>parking tickets</i>	25	34	28	29	21	27	19	18

- Assuming *cream cheese sales* is the explanatory variable, is the Pearson correlation coefficient, r , an appropriate statistic for this data set?
 - Determine the value of the correlation coefficient, correct to two decimal places.
 - If possible, describe the association between *cream cheese sales* and *parking tickets* in terms of strength, direction and form.
 - Is it reasonable to conclude that there is a causal relationship between *cream cheese sales* and *parking tickets*? If not, identify what might be causing the observed association.
-
12. Mikayla plays netball and is trying to determine whether her recovery during the week impacts positively on her performance. For the first 13 games of the season, she collects data on the *time spent on recovery*, in minutes, between games, and the number of *goals* she scores each week. Her results are shown in the following table.

<i>time spent on recovery (mins)</i>	45	62	35	105	96	137	48	56	71	105	60	29	135
<i>goals</i>	5	9	8	9	13	9	8	7	9	10	12	7	15

- Is the Pearson correlation coefficient, r , an appropriate statistic for this data set? Why or why not?
- Determine the value of the correlation coefficient. Give the value correct to two decimal places.
- If possible, describe the association between *time spent on recovery* and number of *goals* in terms of strength, direction, and form.
- Identify five confounding variables that could also influence the *number of goals* that Mikayla scores each week.
- Has Mikayla conducted an appropriate experiment to determine causation? Why or why not?

Exam practice

13. Data collected over a period of 10 years indicated a strong, positive association between the number of stray cats and the number of stray dogs reported each year ($r = 0.87$) in a large, regional city. A positive association was also found between the population of the city and both the number of stray cats ($r = 0.61$) and the number of stray dogs (0.72). During the time that the data was collected, the population of the city grew from 34 564 to 51 055. From this information, we can conclude that
- if cat owners paid more attention to keeping dogs off their property, the number of stray cats reported would decrease.
 - the association between the number of stray cats and stray dogs reported cannot be causal because only a correlation of +1 or -1 shows causal relationships.
 - there is no logical explanation for the association between the number of stray cats and stray dogs reported in the city so it must be a chance occurrence.
 - because larger populations tend to have both a larger number of stray cats and stray dogs, the association between the number of stray cats and the number of stray dogs can be explained by a common response to a third variable, which is the increasing population size of the city.
 - more stray cats were reported because people are no longer as careful about keeping their cats properly contained on their property as they were in the past.

81% of students answered this question correctly.

VCAA 2017 Exam 1 Data analysis Q12

14. The relative humidity (%) at 9 am and 3 pm on 14 days in November 2017 is shown in the following table.

<i>relative humidity (%)</i>	9 am	100	99	95	63	81	94	96	81	73	53	57	77	51	41
	3 pm	87	75	67	57	57	74	71	62	53	54	36	39	30	32

Data: Australian Government, Bureau of Meteorology, <www.bom.gov.au/>

A least squares line is to be fitted to the data with the aim of predicting the relative humidity at 3 pm (*humidity 3 pm*) from the relative humidity at 9 am (*humidity 9 am*).

Determine the value of the correlation coefficient for this data set. Round your answer to three decimal places. (1 MARK)

76% of students answered this question correctly.

VCAA 2019 Exam 2 Data analysis Q4c

Questions from multiple lessons

Data analysis Year 11 content

15. Parallel boxplots are to be used to display the association between two chosen variables. One of these variables is *height* (cm). The second variable could be
- weight* (kg).
 - arm span* (cm).
 - ATAR*.
 - hair colour* (brown, blonde, black, red, other).
 - foot length* (cm).

Adapted from VCAA 2016 Exam 1 Data analysis Q8

Recursion and financial modelling Year 11 content

16. Arlo borrowed a decent sum of money from his friend Sid. They work out that to cover the amount of money Arlo owes Sid, he should pay Sid \$235 per month for one year. However, Arlo agrees to pay an extra 25% on top of what he already owes Sid for taking so long to pay him back. How much money in total does Arlo end up giving Sid at the end of the one-year period?

A. \$4875 B. \$3981.25 C. \$3525 D. \$3102 E. \$2878.75

Adapted from VCAA 2014 Exam 1 Business-related mathematics Q4

Data analysis Year 11 content

17. The following back-to-back stem plot displays the *weight* (kg) of 38 cavoodles and their *size*, toy ($n = 17$) or mini ($n = 21$).

Key: 9 | 3 = 9.3

<i>weight</i> (kg)	
<i>toy cavoodle</i>	<i>mini cavoodle</i>
	4
9	5
9 4	6
5 4 2	7
6 3 1	8 8
5 3 2 2	9 3 5 7
6 4	10 1 4 9
3	11 2 3 8 8 8
5	12 1 1 4
	13 5 6 6 8
	14 1 2

- a. Which variable, *weight* or *size*, is a categorical variable? (1 MARK)
- b. What is the modal weight of the mini cavoodles? (1 MARK)
- c. Find the values of a and b in the following table. (2 MARKS)

<i>size</i>	<i>weight</i> (kg)				
	<i>minimum</i>	Q_1	<i>median</i>	Q_3	<i>maximum</i>
<i>toy cavoodle</i>	5.9	7.3	8.6	9.95	a
<i>mini cavoodle</i>	8.8	10.25	b	13.55	14.2

Adapted from VCAA 2017 Exam 2 Data analysis Q2