

# CHAPTER 2

# Bivariate data

eBookplus

DIGITAL DOC  
doc-9409  
10 Quick Questions

## CHAPTER CONTENTS

- 2A Dependent and independent variables
- 2B Back-to-back stem plots
- 2C Parallel boxplots
- 2D Two-way frequency tables and segmented bar charts
- 2E Scatterplots
- 2F Pearson's product-moment correlation coefficient
- 2G Calculating  $r$  and the coefficient of determination

## 2A Dependent and independent variables

In this chapter we will study sets of data that contain two variables. These are known as *bivariate data*. We will look at ways of displaying the data and of measuring relationships between the two variables.

The methods we employ to do this depend on the type of variables we are dealing with; that is, they depend on whether the data are numerical or categorical.

We will discuss the ways of measuring the relationship between the following pairs of variables:

1. a numerical variable and a categorical variable (for example, height and nationality)
2. two categorical variables (for example, gender and religious denomination)
3. two numerical variables (for example, height and weight).

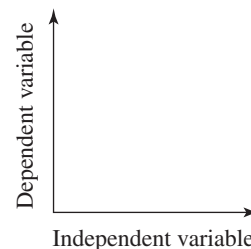
In a relationship involving two variables, if the values of one variable 'depend' on the values of another variable, then the former variable is referred to as the *dependent variable* and the latter variable is referred to as the *independent variable*.

When a relationship between two sets of variables is being examined, it is important to know which one of the two variables depends on the other. Most often we can make a judgement about this, although sometimes it may not be possible.

Consider the case where a study compared the heights of company employees against their annual salaries. Common sense would suggest that the height of a company employee would not depend on the person's annual salary nor would the annual salary of a company employee depend on the person's height. In this case, it is not appropriate to designate one variable as independent and one as dependent.

In the case where the ages of company employees are compared with their annual salaries, you might reasonably expect that the annual salary of an employee would depend on the person's age. In this case, the age of the employee is the independent variable and the salary of the employee is the dependent variable.

It is useful to identify the independent and dependent variables where possible, since it is the usual practice when displaying data on a graph to place the independent variable on the horizontal axis and the dependent variable on the vertical axis.



## study on

Units: 3 & 4

AOS: DA

Topic: 6

Concept: 1

 **Concept summary**  
Read a summary of this concept.

### WORKED EXAMPLE 1

For each of the following pairs of variables, identify the independent variable and the dependent variable. If it is not possible to identify this, then write 'not appropriate'.

- a The number of visitors at a local swimming pool and the daily temperature
- b The blood group of a person and his or her favourite TV channel

#### THINK

- a It is reasonable to expect that the number of visitors at the swimming pool on any day will depend on the temperature on that day (and not the other way around).
- b Common sense suggests that the blood type of a person does not depend on the person's TV channel preferences. Similarly, the choice of a TV channel does not depend on a person's blood type.

#### WRITE

- a Daily temperature is the independent variable; number of visitors at a local swimming pool is the dependent variable.
- b Not appropriate

## Exercise 2A Dependent and independent variables

- 1 **WE1** For each of the following pairs of variables, identify the independent variable and the dependent variable. If it is not possible to identify this, then write 'not appropriate'.
  - a The age of an AFL footballer and his annual salary
  - b The growth of a plant and the amount of fertiliser it receives
  - c The number of books read in a week and the eye colour of the readers
  - d The voting intentions of a woman and her weekly consumption of red meat
  - e The number of members in a household and the size of the house
  - f The month of the year and the electricity bill for that month
  - g The mark obtained for a maths test and the number of hours spent preparing for the test
  - h The mark obtained for a maths test and the mark obtained for an English test
  - i The cost of grapes (in dollars per kilogram) and the season of the year
- 2 **MC** In a scientific experiment, the independent variable was the amount of sleep (in hours) a new mother got per night during the first month following the birth of her baby. The dependent variable would most likely have been:
  - A the number of times (per night) the baby woke up for a feed
  - B the blood pressure of the baby
  - C the mother's reaction time (in seconds) to a certain stimulus
  - D the level of alertness of the baby
  - E the amount of time (in hours) spent by the mother on reading
- 3 **MC** A paediatrician investigated the relationship between the amount of time children aged two to five spend outdoors and the annual number of visits to his clinic. Which one of the following statements is not true?
  - A When graphed, the amount of time spent outdoors should be shown on the horizontal axis.
  - B The annual number of visits to the paediatric clinic is the dependent variable.
  - C It is impossible to identify the independent variable in this case.
  - D The amount of time spent outdoors is the independent variable.
  - E The annual number of visits to the paediatric clinic should be shown on the vertical axis.
- 4 **MC** Alex works as a personal trainer at the local gym. He wishes to analyse the relationship between the number of weekly training sessions and the weekly weight loss of his clients. Which one of the following statements is correct?
  - A When graphed, the number of weekly training sessions should be shown on the vertical axis, as it is the dependent variable.
  - B When graphed, the weekly weight loss should be shown on the vertical axis, as it is the independent variable.

- C When graphed, the weekly weight loss should be shown on the horizontal axis, as it is the independent variable.
- D When graphed, the number of weekly training sessions should be shown on the horizontal axis, as it is the independent variable.
- E It is impossible to identify the dependent variable in this case.

## 2B Back-to-back stem plots

In chapter 1, we saw how to construct a stem plot for a set of univariate data. We can also extend a stem plot so that it displays bivariate data. Specifically, we shall create a stem plot that displays the relationship between a numerical variable and a categorical variable. We shall limit ourselves in this section to categorical variables with just two categories, for example, gender. The two categories are used to provide two back-to-back leaves of a stem plot.

**A back-to-back stem plot is used to display bivariate data, involving a numerical variable and a categorical variable with 2 categories.**

### WORKED EXAMPLE 2

The girls and boys in Grade 4 at Kingston Primary School submitted projects on the Olympic Games. The marks they obtained out of 20 are given below.

Girls' marks	16	17	19	15	12	16	17	19	19	16
Boys' marks	14	15	16	13	12	13	14	13	15	14

eBook plus

TUTORIAL  
eles-1259  
Worked example 2

study on

Units: 3 & 4

AOS: DA

Topic: 6

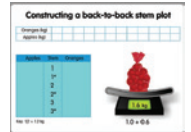
Concept: 2

 **Concept summary**

Read a summary of this concept.

 **See more**

Watch a video about constructing back-to-back stem plots.



Display the data on a back-to-back stem plot.

#### THINK

- 1 Identify the highest and lowest scores in order to decide on the stems.
- 2 Create an unordered stem plot first. Put the boys' scores on the left, and the girls' scores on the right.

#### WRITE

Highest score = 19  
Lowest score = 12  
Use a stem of 1, divided into fifths.

Leaf Boys	Stem	Leaf Girls
	1	
3 2 3 3	1	2
4 5 4 5 4	1	5
6	1	6 7 6 7 6
	1	9 9 9

Key:  $1|2 = 12$

- 3 Now order the stem plot. The scores on the left should increase in value from right to left, while the scores on the right should increase in value from left to right.

Leaf Boys	Stem	Leaf Girls
3 3 3 2	1	2
5 5 4 4 4	1	5
6	1	6 6 6 7 7
	1	9 9 9

Key:  $1|2 = 12$

The back-to-back stem plot allows us to make some visual comparisons of the two distributions. In the previous example, the centre of the distribution for the girls is higher than the centre of the distribution for the boys. The spread of each of the distributions seems to be about the same. For the boys, the scores are grouped around the 12–15 mark; for the girls, they are grouped around the 16–19 mark. On the whole, we can conclude that the girls obtained better scores than the boys did.

To get a more precise picture of the centre and spread of each of the distributions, we can use the summary statistics discussed in chapter 1. Specifically, we are interested in:

1. the mean and the median (to measure the centre of the distributions), and
2. the interquartile range and the standard deviation (to measure the spread of the distributions).

We saw in chapter 1 that the calculation of these summary statistics is very straightforward and rapid using a CAS calculator.



### WORKED EXAMPLE 3

The number of 'how to vote' cards handed out by various Australian Labor Party and Liberal Party volunteers during the course of a polling day is shown below.

<b>Labor</b>	180	233	246	252	263	270	229	238	226	211
	193	202	210	222	257	247	234	226	214	204
<b>Liberal</b>	204	215	226	253	263	272	285	245	267	275
	287	273	266	233	244	250	261	272	280	279



Display the data using a back-to-back stem plot and use this, together with summary statistics, to compare the distributions of the number of cards handed out by the Labor and Liberal volunteers.

#### THINK

- 1 Construct the stem plot.

#### WRITE

Leaf	Stem	Leaf
Labor		Liberal
0	18	
3	19	
4 2	20	4
4 1 0	21	5
9 6 6 2	22	6
8 4 3	23	3
7 6	24	4 5
7 2	25	0 3
3	26	1 3 6 7
0	27	2 2 3 5 9
	28	0 5 7

Key: 18|0 = 180

- 2 Use a calculator to obtain summary statistics for each party. Record the mean, median, IQR and standard deviation in the table. (IQR =  $Q_3 - Q_1$ )

	Labor	Liberal
<b>Mean</b>	227.9	257.5
<b>Median</b>	227.5	264.5
<b>IQR</b>	36	29.5
<b>Standard deviation</b>	23.9	23.4

- 3 Comment on the relationship.

From the stem plot we see that the Labor distribution is symmetric and therefore the mean and the median are very close, whereas the Liberal distribution is negatively skewed.

Since the distribution is skewed, the median is a better indicator of the centre of the distribution than the mean.

Comparing the medians therefore, we have the median number of cards handed out for Labor at 228 and for Liberal at 265, which is a big difference.

The standard deviations were similar, as were the interquartile ranges. There was not a lot of difference in the spread of the data.

In essence, the Liberal party volunteers handed out many more 'how to vote' cards than the Labor party volunteers did.

## Exercise 2B Back-to-back stem plots

- 1 **WE2** The marks out of 50 obtained for the end-of-term test by the students in German and French classes are given below. Display the data on a back-to-back stem plot.

<b>German</b>	20	38	45	21	30	39	41	22	27	33	30	21	25	32	37	42	26	31	25	37
<b>French</b>	23	25	36	46	44	39	38	24	25	42	38	34	28	31	44	30	35	48	43	34

- 2 The birth masses of 10 boys and 10 girls (in kilograms, to the nearest 100 grams) are recorded in the table below. Display the data on a back-to-back stem plot.

<b>Boys</b>	3.4	5.0	4.2	3.7	4.9	3.4	3.8	4.8	3.6	4.3
<b>Girls</b>	3.0	2.7	3.7	3.3	4.0	3.1	2.6	3.2	3.6	3.1

- 3 **WE3** The number of delivery trucks making deliveries to a supermarket each day over a 2-week period was recorded for two neighbouring supermarkets — supermarket A and supermarket B. The data are shown below.

<b>A</b>	11	15	20	25	12	16	21	27	16	17	17	22	23	24
<b>B</b>	10	15	20	25	30	35	16	31	32	21	23	26	28	29

- a Display the data on a back-to-back stem plot.  
 b Use the stem plot, together with some summary statistics, to compare the distributions of the number of trucks delivering to supermarkets A and B.
- 4 The marks out of 20 obtained by males and females for a science test in a Year 10 class are given below.

<b>Females</b>	12	13	14	14	15	15	16	17
<b>Males</b>	10	12	13	14	14	15	17	19

- a Display the data on a back-to-back stem plot.  
 b Use the stem plot, together with some summary statistics, to compare the distributions of the marks of the males and the females.
- 5 The end-of-year English marks for 10 students in an English class were compared over 2 years. The marks for 2011 and for the same students in 2012 are shown below.

<b>2011</b>	30	31	35	37	39	41	41	42	43	46
<b>2012</b>	22	26	27	28	30	31	31	33	34	36

- a Display the data on a back-to-back stem plot.  
 b Use the stem plot, together with some summary statistics, to compare the distributions of the marks obtained by the students in 2011 and 2012.

- 6 The age and gender of a group of people attending a fitness class are recorded below.

Female	23	24	25	26	27	28	30	31
Male	22	25	30	31	36	37	42	46

- a Display the data on a back-to-back stem plot.  
 b Use the stem plot, together with some summary statistics, to compare the distributions of the ages of the female members to the male members of the fitness class.



- 7 The scores on a board game for a group of kindergarten children and for a group of children in a preparatory school are given below.

Kindergarten	3	13	14	25	28	32	36	41	47	50
Prep. school	5	12	17	25	27	32	35	44	46	52

- a Display the data on a back-to-back stem plot.  
 b Use the stem plot, together with some summary statistics, to compare the distributions of the scores of the kindergarten children compared to the preparatory school children.
- 8 **MC** The pair of variables that could be displayed on a back-to-back stem plot is:  
 A the height of a student and the number of people in the student's household  
 B the time put into completing an assignment and a pass or fail score on the assignment  
 C the weight of a businessman and his age  
 D the religion of an adult and the person's head circumference  
 E the income of an employee and the time the employee has worked for the company
- 9 **MC** A back-to-back stem plot is a useful way of displaying the relationship between:  
 A the proximity to markets in kilometres and the cost of fresh foods on average per kilogram  
 B height and head circumference  
 C age and attitude to gambling (for or against)  
 D weight and age  
 E the money spent during a day of shopping and the number of shops visited on that day

### study on

Units:	3 & 4
AOS:	DA
Topic:	6
Concept:	3

 **Concept summary**  
 Read a summary of this concept.

## 2C Parallel boxplots

We saw in the previous section that we could display relationships between a numerical variable and a categorical variable with just two categories, using a back-to-back stem plot.

When we want to display a relationship between a numerical variable and a categorical variable with two or *more* categories, a *parallel boxplot* can be used.

A parallel boxplot is obtained by constructing individual boxplots for each distribution and positioning them on a common scale.

Construction of individual boxplots was discussed in detail in chapter 1 on univariate data. In this section we concentrate on comparing distributions represented by a number of boxplots (that is, on the interpretation of parallel boxplots).

### WORKED EXAMPLE 4

The four Year 7 classes at Western Secondary College complete the same end-of-year maths test. The marks, expressed as percentages for the four classes, are given below.

7A	40	43	45	47	50	52	53	54	57	60	69	63	63	68	70	75	80	85	89	90
7B	60	62	63	64	70	73	74	76	77	77	78	82	85	87	89	90	92	95	97	97
7C	50	51	53	55	57	60	63	65	67	69	70	72	73	74	76	80	82	82	85	89
7D	40	42	43	45	50	53	55	59	60	61	69	73	74	75	80	81	82	83	84	90

Display the data using a parallel boxplot and use this to describe any similarities or differences in the distributions of the marks between the four classes.

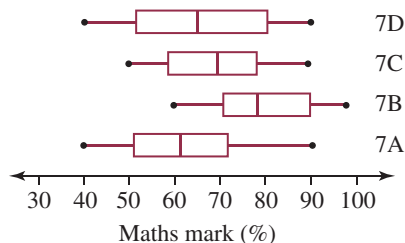
#### THINK

- 1 Use your CAS calculator to determine the five number summary for each data set.

#### WRITE/DRAW

	7A	7B	7C	7D
Min	40	60	50	40
$Q_1$	51	71.5	58.5	51.5
Median = $Q_2$	61.5	77.5	69.5	65
$Q_3$	72.5	89.5	78	80.5
Max	90	97	89	90

- 2 Draw the boxplots, labelling each class. All four boxplots share a common scale.



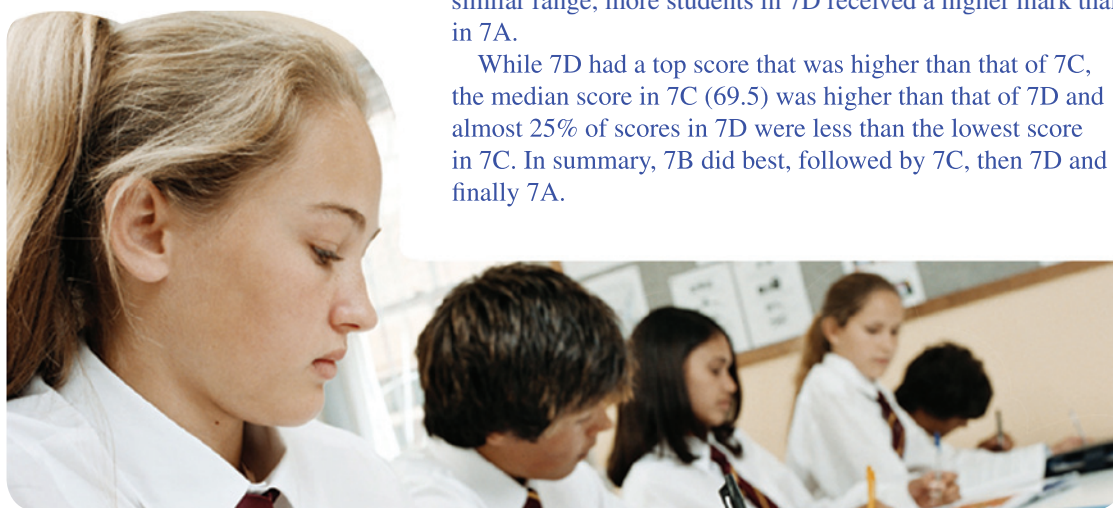
- 3 Describe the similarities and differences between the four distributions.

Class 7B had the highest median mark and the range of the distribution was only 37. The lowest mark in 7B was 60.

We notice that the median of 7A's marks is 61.5. So, 50% of students in 7A received less than 61.5. This means that about half of 7A had scores that were less than the lowest score in 7B.

The range of marks in 7A was the same as that of 7D with the highest scores in each equal (90), and the lowest scores in each equal (40). However, the median mark in 7D (65) was slightly higher than the median mark in 7A (61.5) so, despite a similar range, more students in 7D received a higher mark than in 7A.

While 7D had a top score that was higher than that of 7C, the median score in 7C (69.5) was higher than that of 7D and almost 25% of scores in 7D were less than the lowest score in 7C. In summary, 7B did best, followed by 7C, then 7D and finally 7A.



## Exercise 2C Parallel boxplots

eBookplus

DIGITAL DOC  
doc-9410  
Spreadsheet  
Parallel boxplots

- 1 **WE4** The heights (in cm) of students in 9A, 10A and 11A were recorded and are shown in the table below.

9A	120	126	131	138	140	143	146	147	150	156	157	158	158	160	162	164	165	170
10A	140	143	146	147	149	151	153	156	162	164	165	167	168	170	173	175	176	180
11A	151	153	154	158	160	163	164	166	167	169	169	172	175	180	187	189	193	199

- Construct a parallel boxplot to show the data.
  - Use the boxplot to compare the distributions of height for the 3 classes.
- 2 The amounts of money contributed annually to superannuation schemes by people in 3 different age groups are shown below.

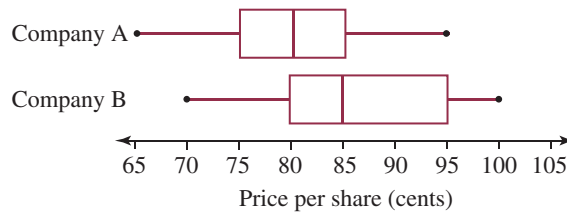
20–29	2 000	3 100	5 000	5 500	6 200	6 500	6 700	7 000	9 200	10 000
30–39	4 000	5 200	6 000	6 300	6 800	7 000	8 000	9 000	10 300	12 000
40–49	10 000	11 200	12 000	13 300	13 500	13 700	13 900	14 000	14 300	15 000

- Construct a parallel boxplot to show the data.
  - Use the boxplot to comment on the distributions.
- 3 The numbers of jars of vitamin A, B, C and multi-vitamins sold per week by a local chemist are shown below.

Vitamin A	5	6	7	7	8	8	9	11	13	14
Vitamin B	10	10	11	12	14	15	15	15	17	19
Vitamin C	8	8	9	9	9	10	11	12	12	13
Multi-vitamins	12	13	13	15	16	16	17	19	19	20

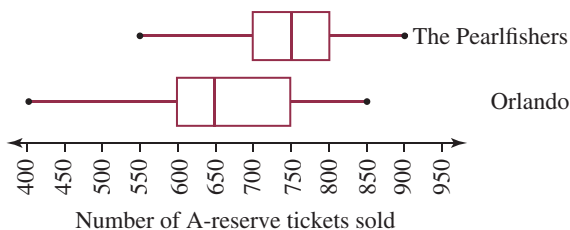
Construct a parallel boxplot to display the data and use it to compare the distributions of sales for the 4 types of vitamin.

- 4 The daily share price of two companies was recorded over a period of one month. The results are presented below as parallel boxplots.



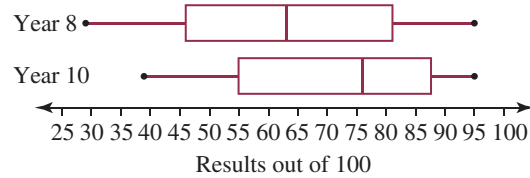
State whether each of the following statements is true or false.

- The distribution of share prices for company A is symmetrical.
  - On 25% of all occasions, share prices for company B equalled or exceeded the highest price recorded for company A.
  - The spread of the share prices was the same for both companies.
  - 75% of share prices for company B were at least as high as the median share price for company A.
- 5 Last year, the spring season of the Australian Opera included two major productions staged at the Sydney Opera House: The Pearlfishers and Orlando. The number of A-reserve tickets sold for each performance of the two operas is shown below as parallel boxplots.





- a Which of the two productions proved to be more popular with the public, assuming A-reserve ticket sales reflect total ticket sales? Explain your answer.
  - b Which opera had a larger variability in the number of patrons purchasing A-reserve tickets? Support your answer with the necessary calculations.
- 6 **MC** The results for a maths test given to classes in two different year levels, one in Year 8 and the other in Year 10, are given by the parallel boxplots below.



- a The percentage of Year 10 students who obtained a mark greater than 87 was:
  - A 2%
  - B 5%
  - C 20%
  - D 25%
  - E 75%
- b From the parallel boxplots, it can be concluded that:
  - A the Year 8 results were similar to the Year 10 results
  - B the Year 8 results were lower than the Year 10 results and less variable
  - C the Year 8 results were lower than the Year 10 results and more variable
  - D the Year 8 results were higher than the Year 10 results and less variable
  - E the Year 8 results were higher than the Year 10 results and more variable



DIGITAL DOC  
doc-9411  
WorkSHEET 2.1

## 2D Two-way frequency tables and segmented bar charts

When we are examining the relationship between two categorical variables, the two-way frequency table is an excellent tool. Consider the following example.

### WORKED EXAMPLE 5

At a local shopping centre, 34 females and 23 males were asked which of the two major political parties they preferred. Eighteen females and 12 males preferred Labor. Display these data in a two-way frequency table.

#### THINK

- 1 Draw a table. Record the respondents' sex in the columns and party preference in the rows of the table.
- 2 We know that 34 females and 23 males were asked. Put this information into the table and fill in the total.  
We also know that 18 females and 12 males preferred Labor. Put this information in the table and find the total of people who preferred Labor.
- 3 Fill in the remaining cells. For example, to find the number of females who preferred the Liberals, subtract the number of females preferring Labor from the total number of females asked:  $34 - 18 = 16$ .

#### WRITE

Party preference	Female	Male	Total
Labor			
Liberal			
Total			

Party preference	Female	Male	Total
Labor	18	12	30
Liberal			
Total	34	23	57

Party preference	Female	Male	Total
Labor	18	12	30
Liberal	16	11	27
Total	34	23	57

### study on

Units: 3 & 4

AOS: DA

Topic: 6

Concept: 4

**Concept summary**  
Read a summary of this concept.

**Do more**  
Interact with segmented graphs.



In Worked example 5, we have a very clear breakdown of data. We know how many females preferred Labor, how many females preferred the Liberals, how many males preferred Labor and how many males preferred the Liberals.

If we wish to compare the number of females who prefer Labor with the number of males who prefer Labor, we must be careful. While 12 males preferred Labor compared to 18 females, there were fewer males than females being asked. That is, only 23 males were asked for their opinion, compared to 34 females.

To overcome this problem, we can express the figures in the table as percentages.

### WORKED EXAMPLE 6

**Fifty-seven people in a local shopping centre were asked whether they preferred the Australian Labor Party or the Liberal Party. The results are given at right.**

**Convert the numbers in this table to percentages.**

Party preference	Female	Male	Total
Labor	18	12	30
Liberal	16	11	27
Total	34	23	57

#### THINK

Draw the table, omitting the 'total' column. Fill in the table by expressing the number in each cell as a percentage of its column's total. For example, to obtain the percentage of males who prefer Labor, divide the number of males who prefer Labor by the total number of males and multiply by 100%.  
 $\frac{12}{23} \times 100\% = 52.2\%$  (correct to 1 decimal place)

#### WRITE

Party preference	Female	Male
Labor	52.9	52.2
Liberal	47.1	47.8
Total	100.0	100.0

We could have calculated percentages from the table rows, rather than columns. To do that we would, for example, have divided the number of females who preferred Labor (18) by the total number of people who preferred labor (30) and so on. The table below shows this:

Party preference	Female	Male	Total
Labor	60.0	40.0	100
Liberal	59.3	40.7	100

By doing this we have obtained the percentage of people who were female and preferred Labor (60%), and the percentage of people who were male and preferred Labor (40%), and so on. This highlights facts different from those shown in the previous table. In other words, different results can be obtained by calculating percentages from a table in different ways.

**As a general rule, when the independent variable (in this case the respondent's gender) is placed in the columns of the table, the percentages should be calculated in columns.**

Comparing percentages in each row of a two-way table allows us to establish whether a relationship exists between the two categorical variables that are being examined. As we can see from the table in Worked example 6, the percentage of females who preferred Labor is about the same as that of males. Likewise, the percentage of females and males preferring the Liberal Party are almost equal. This indicates that for the group of people participating in the survey, party preference is not related to gender.

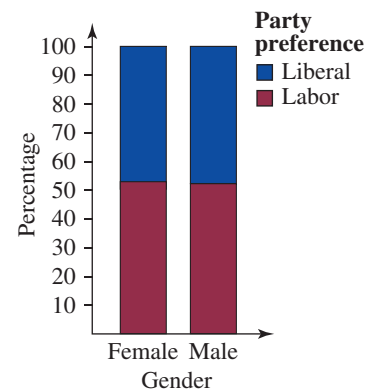
## Segmented bar charts

When comparing two categorical variables, it can be useful to represent the results from a two-way table (in percentage form) graphically. We can do this using segmented bar charts.

A segmented bar chart consists of two or more columns, each of which matches one column in the two-way table. Each column is subdivided into segments, corresponding to each cell in that column.

For example, the data from Worked example 6 can be displayed using the segmented bar chart shown at right.

The segmented bar chart is a powerful visual aid for comparing and examining the relationship between two categorical variables.



**WORKED EXAMPLE 7**

**eBookplus**

Sixty-seven primary and 47 secondary school students were asked about their attitude to the number of school holidays which should be given. They were asked whether there should be fewer, the same number, or more school holidays. Five primary students and 2 secondary students wanted fewer holidays, 29 primary and 9 secondary students thought they had enough holidays (that is, they chose the same number) and the rest thought they needed to be given more holidays.

**TUTORIAL**  
eles-1260  
Worked example 7

Present these data in percentage form in a two-way frequency table and a segmented bar chart. Compare the opinions of the primary and the secondary students.

**THINK**

1 Put the data in a table. First, fill in the given information, then find the missing information by subtracting the appropriate numbers from the totals.

**WRITE/DRAW**

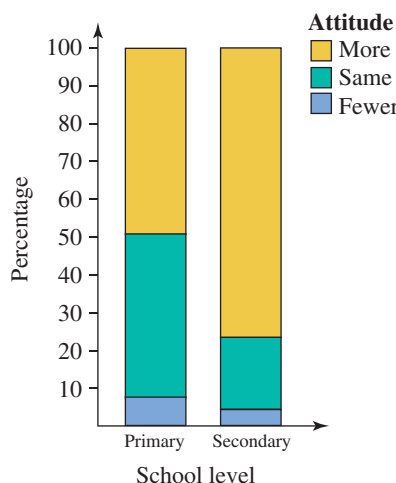
Attitude	Primary	Secondary	Total
Fewer	5	2	7
Same	29	9	38
More	33	36	69
Total	67	47	114

2 Calculate the percentages. Since the independent variable (the level of the student: primary or secondary) has been placed in the columns of the table, we calculate the percentages in columns. For example, to obtain the percentage of primary students who wanted fewer holidays, divide the number of such students by the total number of primary students and multiply by 100%.

Attitude	Primary	Secondary
Fewer	7.5	4.3
Same	43.3	19.1
More	49.2	76.6
Total	100.0	100.0

That is,  $\frac{5}{67} \times 100\% = 7.5\%$ .

3 Rule out the set of axes. (The vertical axis shows percentages from 0 to 100, while the horizontal axis represents the categories from the columns of the table.) Draw two columns to represent each category — primary and secondary. Columns must be the same width and height (up to 100%). Divide each column into segments so that the height of each segment is equal to the percentage in the corresponding cell of the table. Add a legend to the graph.



4 Comment on the results.

Secondary students were much keener on having more holidays than were primary students.

**Exercise 2D Two-way frequency tables and segmented bar charts**

1 **WE5** In a survey, 139 women and 102 men were asked whether they approved or disapproved of a proposed freeway. Thirty-seven women and 79 men approved of the freeway. Display these data in a two-way table (not as percentages).

**eBookplus**

**DIGITAL DOC**  
doc-9413  
Spreadsheet  
Two-way frequency table

- 2 Students at a secondary school were asked whether the length of lessons should be 45 minutes or 1 hour. Ninety-three senior students (Years 10–12) were asked and it was found 60 preferred 1-hour lessons, whereas of the 86 junior students (Years 7–9), 36 preferred 1-hour lessons. Display these data in a two-way table (not as percentages).
- 3 For each of the following two-way frequency tables, complete the missing entries.

a

Attitude	Female	Male	Total
For	25	i	47
Against	ii	iii	iv
Total	51	v	92

b

Attitude	Female	Male	Total
For	i	ii	21
Against	iii	21	iv
Total	v	30	63

c

Party preference	Female	Male
Labor	i	42%
Liberal	53%	ii
Total	iii	iv

- 4 **WE6** Sixty single men and women were asked whether they prefer to rent by themselves, or to share accommodation with friends. The results are shown below.

Preference	Men	Women	Total
Rent by themselves	12	23	35
Share with friends	9	16	25
Total	21	39	60

Convert the numbers in this table to percentages.

The information in the following two-way frequency table relates to questions 5 and 6.

The data show the reactions of administrative staff and technical staff to an upgrade of the computer systems at a large corporation.

Attitude	Administrative staff	Technical staff	Total
For	53	98	151
Against	37	31	68
Total	90	129	219

- 5 **MC** From the previous table, we can conclude that:
- A 53% of administrative staff were for the upgrade
  - B 37% of administrative staff were for the upgrade
  - C 37% of administrative staff were against the upgrade
  - D 59% of administrative staff were for the upgrade
  - E 54% of administrative staff were against the upgrade
- 6 **MC** From the previous table, we can conclude that:
- A 98% of technical staff were for the upgrade
  - B 65% of technical staff were for the upgrade
  - C 76% of technical staff were for the upgrade
  - D 31% of technical staff were against the upgrade
  - E 14% of technical staff were against the upgrade
- 7 **WE7** Delegates at the respective Liberal Party and Australian Labor Party conferences were surveyed on whether or not they believed that marijuana should be legalised. Sixty-two Liberal delegates were surveyed and 40 of them were against legalisation. Seventy-one Labor delegates were surveyed and 43 were against legalisation.

Present the data in percentage form in a two-way frequency table and a segmented bar chart. Comment on any differences between the reactions of the Liberal and Labor delegates.

- 8 **MC** The amount of waste recycled by 100 townships across Australia was rated as low, medium or high and the size of the town as small, mid-sized or large. The results of the ratings are:

Amount of waste recycled	Type of town		
	Small	Mid-sized	Large
Low	6	7	4
Medium	8	31	5
High	5	16	18



DIGITAL DOC  
doc-9412  
SkillSHEET 2.1  
Expressing one number  
as a percentage of  
another

- a The percentage of mid-sized towns rated as having a high level of waste recycling is closest to:  
 A 41%      B 25%      C 30%      D 17%      E 50%
- b The variables, *Amount of waste recycled* and *Type of town*, as used in this rating are:  
 A both categorical variables      B both numerical variables  
 C numerical and categorical respectively      D categorical and numerical respectively  
 E neither categorical nor numerical variables

## 2E Scatterplots

We often want to know if there is a relationship between two numerical variables. A scatterplot, which gives a visual display of the relationship between two variables, provides a good starting point.

Consider the data obtained from last year's 12B class at Northbank Secondary College. Each student in this class of 29 students was asked to give an estimate of the average number of hours they studied per week during Year 12. They were also asked for the ATAR score they obtained.

Average hours of study	ATAR score
18	59
16	67
22	74
27	90
15	62
28	89
18	71
19	60
22	84
30	98

Average hours of study	ATAR score
14	54
17	72
14	63
19	72
20	58
10	47
28	85
25	75
18	63
19	61

Average hours of study	ATAR score
17	59
16	76
14	59
29	89
30	93
30	96
23	82
26	35
22	78

The figure at right shows the data plotted on a scatterplot.

It is reasonable to think that the number of hours of study put in each week by students would affect their ATAR scores and so the number of hours of study per week is the independent variable and appears on the horizontal axis. The ATAR score is the dependent variable and appears on the vertical axis.

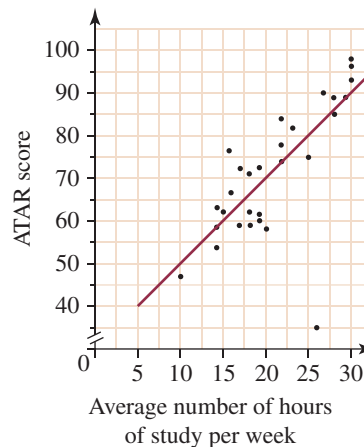
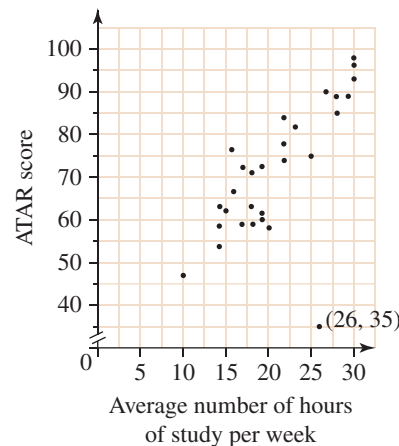
There are 29 points on the scatterplot. Each point represents the number of hours of study and the ATAR score of one student.

In analysing the scatterplot we look for a pattern in the way the points lie. Certain patterns tell us that certain relationships exist between the two variables. This is referred to as *correlation*. We look at what type of correlation exists and how strong it is.

In the figure above right we see some sort of pattern: the points are spread in a rough corridor from bottom left to top right. We refer to data following such a direction as having a *positive relationship*. This tells us that as the average number of hours studied per week increases, the ATAR score increases.

The point (26, 35) is an outlier. It stands out because it is well away from the other points and clearly is not part of the 'corridor' referred to previously. This outlier may have occurred because a student exaggerated the number of hours he or she worked in a week or perhaps there was a recording error. This needs to be checked.

We could describe the rest of the data as having a *linear* form as the straight line in the diagram at right indicates.



### study on

Units: 3 & 4

AOS: DA

Topic: 6

Concept: 5



**Concept summary**  
 Read a summary of this concept.

When describing the relationship between two variables displayed on a scatterplot, we need to comment on:

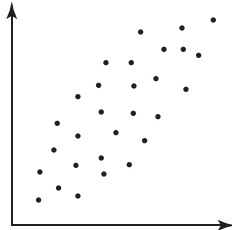
- the direction — whether it is positive or negative
- the form — whether it is linear or non-linear
- the strength — whether it is strong, moderate or weak
- possible outliers.

Below is a gallery of scatterplots showing the various patterns we look for.

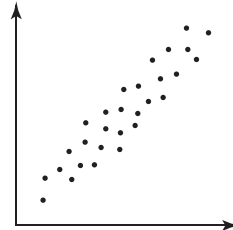
### study on

Units:	3 & 4
AOS:	DA
Topic:	6
Concept:	6

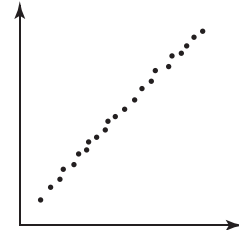
 **Concept summary**  
Read a summary of this concept.



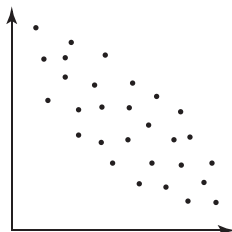
Weak, positive linear relationship



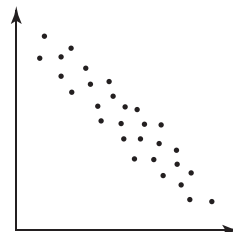
Moderate, positive linear relationship



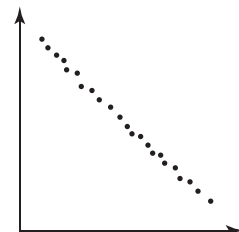
Strong, positive linear relationship



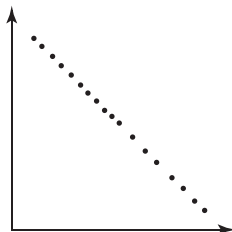
Weak, negative linear relationship



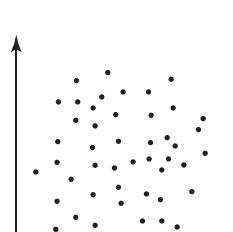
Moderate, negative linear relationship



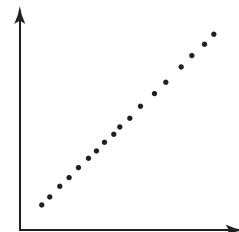
Strong, negative linear relationship



Perfect, negative linear relationship



No relationship

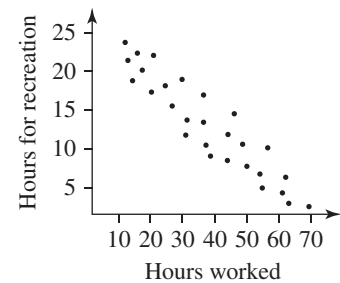


Perfect, positive linear relationship

### WORKED EXAMPLE 8

The scatterplot at right shows the number of hours people spend at work each week and the number of hours people get to spend on recreational activities during the week.

Decide whether or not a relationship exists between the variables and, if it does, comment on whether it is positive or negative; weak, moderate or strong; and whether or not it has a linear form.



#### THINK

- The points on the scatterplot are spread in a certain pattern, namely in a rough corridor from the top left to the bottom right corner. This tells us that as the work hours increase, the recreation hours decrease.
- The corridor is straight (that is, it would be reasonable to fit a straight line into it).

#### WRITE

- 3 The points are neither too tight nor too dispersed.
- 4 The pattern resembles the central diagram in the gallery of scatterplots shown previously.

There is a moderate, negative linear relationship between the two variables.

**WORKED EXAMPLE 9**

Data showing the average weekly number of hours studied by each student in 12B at Northbank Secondary College and the corresponding height of each student (to the nearest tenth of a metre) are given in the table below.

**eBookplus**



**TUTORIAL**  
eles-1261  
Worked example 9

Average hours of study	18	16	22	27	15	28	18	20	10	28	25	18	19	17
Height (m)	1.5	1.9	1.7	2.0	1.9	1.8	2.1	1.9	1.9	1.5	1.7	1.8	1.8	2.1

Average hours of study	19	22	30	14	17	14	19	16	14	29	30	30	23	22
Height (m)	2.0	1.9	1.6	1.5	1.7	1.8	1.7	1.6	1.9	1.7	1.8	1.5	1.5	2.1

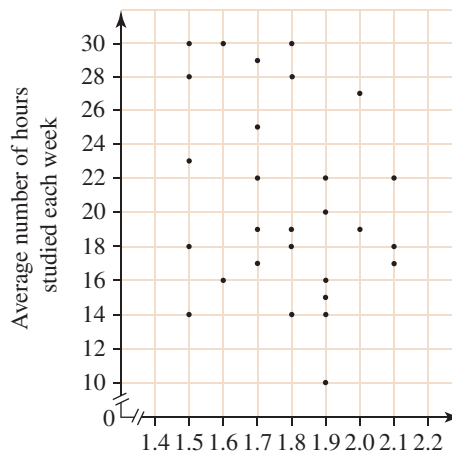


Construct a scatterplot for the data and use it to comment on the direction, form and strength of any relationship between the number of hours studied and the height of the students.

**THINK**

- 1 A calculator can be used to assist you in drawing a scatterplot.

**WRITE**



- 2 Comment on the direction of any relationship.

There is no relationship; the points appear to be randomly placed.





- 3 Comment on the form of the relationship.
- 4 Comment on the strength of any relationship.
- 5 Draw a conclusion.

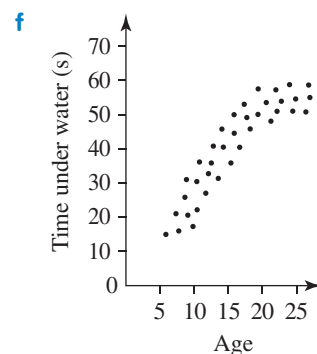
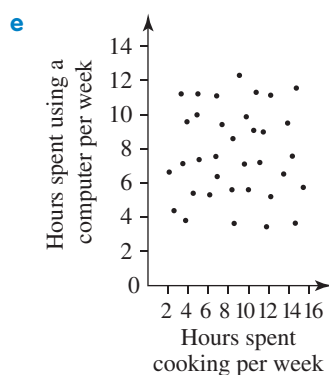
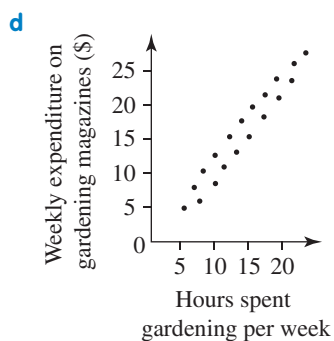
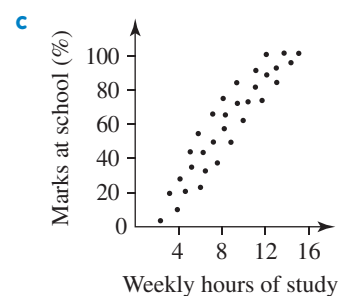
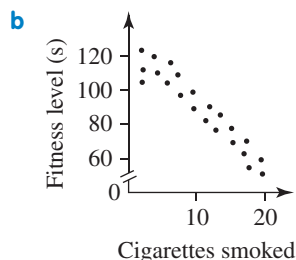
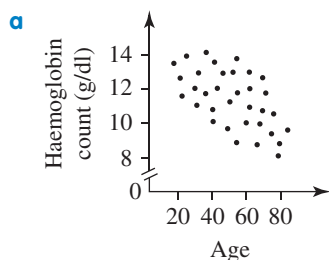
There is no form, no linear trend, no quadratic trend, just a random placement of points.

Since there is no relationship, strength is not relevant.

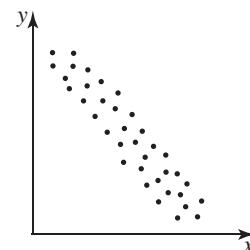
Clearly, from the graph, the number of hours spent studying for VCE has no relation to how tall you might be.

## Exercise 2E Scatterplots

- 1 For each of the following pairs of variables, write down whether or not you would reasonably expect a relationship to exist between the pair and, if so, comment on whether it would be a positive or negative association.
  - a Time spent in a supermarket and money spent
  - b Income and value of car driven
  - c Number of children living in a house and time spent cleaning the house
  - d Age and number of hours of competitive sport played per week
  - e Amount spent on petrol each week and distance travelled by car each week
  - f Number of hours spent in front of a computer each week and time spent playing the piano each week
  - g Amount spent on weekly groceries and time spent gardening each week
- 2 **WEB** For each of the scatterplots below, describe whether or not a relationship exists between the variables and, if it does, comment on whether it is positive or negative, whether it is weak, moderate or strong and whether or not it has a linear form.



- 3 **MC** From the scatterplot shown at right, it would be reasonable to observe that:
  - A as the value of  $x$  increases, the value of  $y$  increases
  - B as the value of  $x$  increases, the value of  $y$  decreases
  - C as the value of  $x$  increases, the value of  $y$  remains the same
  - D as the value of  $x$  remains the same, the value of  $y$  increases
  - E there is no relationship between  $x$  and  $y$





- 4 **WE9** The population of a municipality (to the nearest ten thousand) together with the number of primary schools in that particular municipality is given below for 11 municipalities.

<b>Population (× 1000)</b>	110	130	130	140	150	160	170	170	180	180	190
<b>Number of primary schools</b>	4	4	6	5	6	8	6	7	8	9	8

Construct a scatterplot for the data and use it to comment on the direction, form and strength of any relationship between the population and the number of primary schools.

- 5 The table below contains data for the time taken to do a paving job and the cost of the job.

Construct a scatterplot for the data. Comment on whether a relationship exists between the time taken and the cost. If there is a relationship, describe it.

<b>Time taken (hours)</b>	<b>Cost of job (\$)</b>
5	1000
7	1000
5	1500
8	1200
10	2000
13	2500
15	2800
20	3200
18	2800
25	4000
33	3000



- 6 The table below shows the time of booking (how many days in advance) of the tickets for a musical performance and the corresponding row number in A-reserve seating.

<b>Time of booking</b>	<b>Row number</b>
5	15
6	15
7	15
7	14
8	14
11	13
13	13

<b>Time of booking</b>	<b>Row number</b>
14	12
14	10
17	11
20	10
21	8
22	5
24	4

<b>Time of booking</b>	<b>Row number</b>
25	3
28	2
29	2
29	1
30	1
31	1

Construct a scatterplot for the data. Comment on whether a relationship exists between the time of booking and the number of the row and, if there is a relationship, describe it.

## 2F Pearson's product-moment correlation coefficient

In the previous section, we estimated the strength of association by looking at a scatterplot and forming a judgment about whether the correlation between the variables was positive or negative and whether the correlation was weak, moderate or strong.

A more precise tool for measuring correlation between two variables is Pearson's product-moment correlation coefficient. This coefficient is used to measure the strength of *linear relationships* between variables.

**eBookplus**

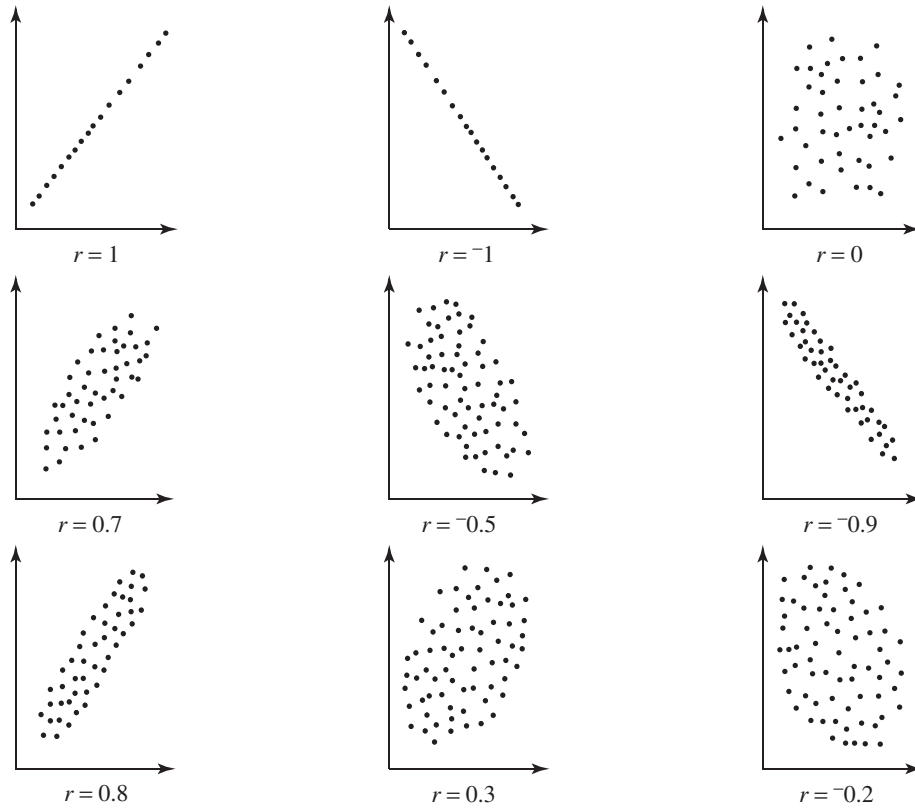
**DIGITAL DOC**  
doc-9414  
Spreadsheet  
Scatterplot

**eBookplus**

**INTERACTIVITY**  
int-0183  
Pearson's  
product-moment  
correlation coefficient

The symbol for Pearson's product-moment correlation coefficient is  $r$ . The value of  $r$  ranges from  $-1$  to  $1$ ; that is,  $-1 \leq r \leq 1$ .

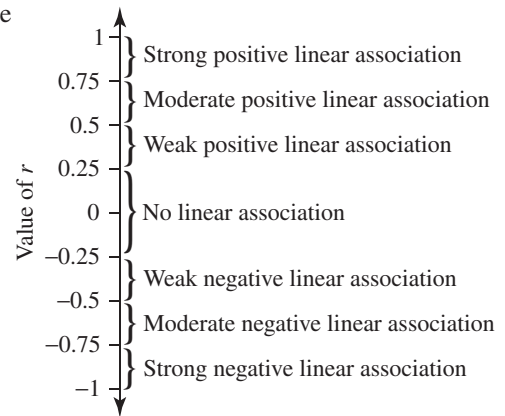
Following is a gallery of scatterplots with the corresponding value of  $r$  for each.



The two extreme values of  $r$  ( $1$  and  $-1$ ) are shown in the first two diagrams respectively.

From these diagrams we can see that a value of  $r = 1$  or  $-1$  means that there is perfect linear association between the variables.

The value of the Pearson's product-moment correlation coefficient indicates the strength of the linear relationship between two variables. The diagram below gives a rough guide to the strength of the correlation based on the value of  $r$ .



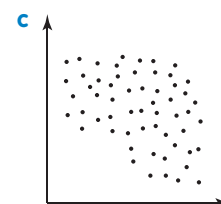
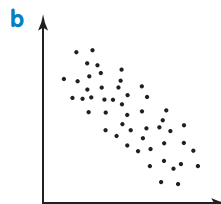
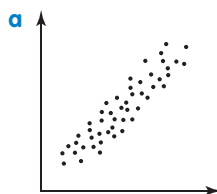
### WORKED EXAMPLE 10

For each of the following:

- i Estimate the value of Pearson's product-moment correlation coefficient ( $r$ ) from the scatterplot.
- ii Use this to comment on the strength and direction of the relationship between the two variables.

**eBookplus**

**TUTORIAL**  
eles-1262  
Worked example 10



**THINK**

- a** **i** Compare these scatterplots with those in the gallery of scatterplots shown previously and estimate the value of  $r$ .
- ii** Comment on the strength and direction of the relationship.
- b** Repeat parts **i** and **ii** as in **a**.
- c** Repeat parts **i** and **ii** as in **a**.

**WRITE**

- a** **i**  $r \approx 0.9$
- ii** The relationship can be described as a strong, positive, linear relationship.
- b** **i**  $r \approx -0.7$
- ii** The relationship can be described as a moderate, negative, linear relationship.
- c** **i**  $r \approx -0.1$
- ii** There is almost no linear relationship.

Note that the symbol  $\approx$  means ‘approximately equal to’. We use it instead of the  $=$  sign to emphasise that the value (in this case  $r$ ) is only an estimate.

In completing the worked example above, we notice that estimating the value of  $r$  from a scatterplot is rather like making an informed guess. In the next section of work, we will see how to obtain the actual value of  $r$ .

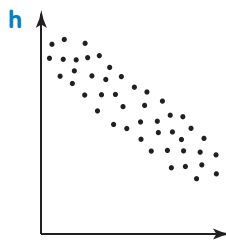
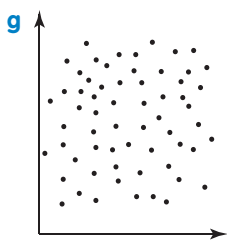
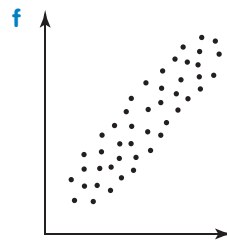
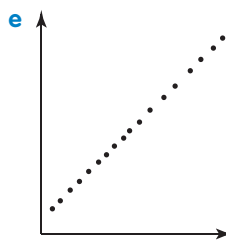
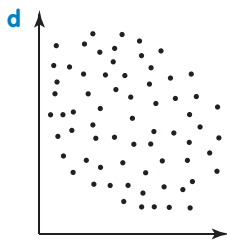
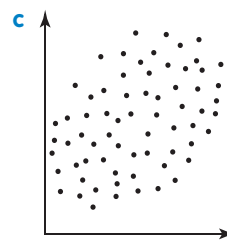
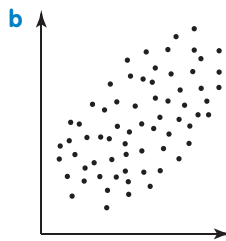
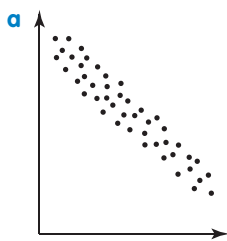
## Exercise 2F Pearson’s product–moment correlation coefficient

1 What type of linear relationship does each of the following values of  $r$  suggest?

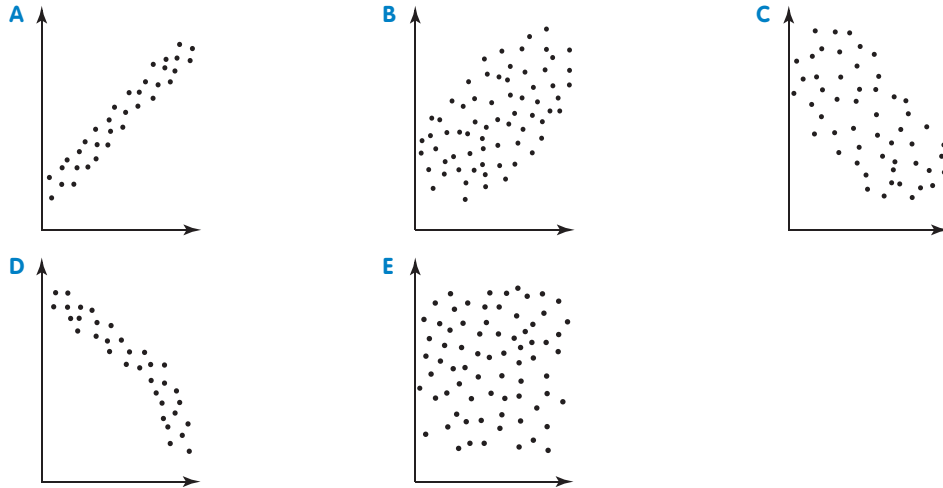
- |               |               |                |                |
|---------------|---------------|----------------|----------------|
| <b>a</b> 0.21 | <b>b</b> 0.65 | <b>c</b> -1    | <b>d</b> -0.78 |
| <b>e</b> 1    | <b>f</b> 0.9  | <b>g</b> -0.34 | <b>h</b> -0.1  |

2 **WE10** For each of the following:

- i** Estimate the value of Pearson’s product–moment correlation coefficient ( $r$ ), from the scatterplot.
- ii** Use this to comment on the strength and direction of the relationship between the two variables.



- 3 **MC** A set of data relating the variables  $x$  and  $y$  is found to have an  $r$  value of 0.62. The scatterplot that could represent the data is:



- 4 **MC** A set of data relating the variables  $x$  and  $y$  is found to have an  $r$  value of  $-0.45$ . A true statement about the relationship between  $x$  and  $y$  is:
- A There is a strong linear relationship between  $x$  and  $y$  and when the  $x$ -values increase, the  $y$ -values tend to increase also.
  - B There is a moderate linear relationship between  $x$  and  $y$  and when the  $x$ -values increase, the  $y$ -values tend to increase also.
  - C There is a moderate linear relationship between  $x$  and  $y$  and when the  $x$ -values increase, the  $y$ -values tend to decrease.
  - D There is a weak linear relationship between  $x$  and  $y$  and when the  $x$ -values increase, the  $y$ -values tend to increase also.
  - E There is a weak linear relationship between  $x$  and  $y$  and when the  $x$ -values increase, the  $y$ -values tend to decrease.

**eBookplus**

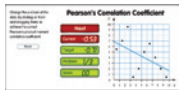
**DIGITAL DOC**  
doc-9415  
WorkSHEET 2.2

### study on

Units:	3 & 4
AOS:	DA
Topic:	6
Concept:	7

 **Concept summary**  
Read a summary of this concept.

 **Do more**  
Interact with  $r$ .



## 2G Calculating $r$ and the coefficient of determination

### Pearson's product-moment correlation coefficient ( $r$ )

The formula for calculating Pearson's correlation coefficient  $r$  is as follows:

$$r = \frac{1}{n-1} \sum_{i=1}^n \left( \frac{x_i - \bar{x}}{s_x} \right) \left( \frac{y_i - \bar{y}}{s_y} \right)$$

where  $n$  is the number of pairs of data in the set  
 $s_x$  is the standard deviation of the  $x$ -values  
 $s_y$  is the standard deviation of the  $y$ -values  
 $\bar{x}$  is the mean of the  $x$ -values  
 $\bar{y}$  is the mean of the  $y$ -values.

The calculation of  $r$  is often done using a CAS calculator.

There are two important limitations on the use of  $r$ . First, since  $r$  measures the strength of a linear relationship, it would be inappropriate to calculate  $r$  for data which are not linear — for example, data which a scatterplot shows to be in a quadratic form.

Second, outliers can bias the value of  $r$ . Consequently, if a set of linear data contains an outlier, then  $r$  is not a reliable measure of the strength of that linear relationship.

**The calculation of  $r$  is applicable to sets of bivariate data which are known to be linear in form and which do not have outliers.**

With those two provisos, it is good practice to draw a scatterplot for a set of data to check for a linear form and an absence of outliers before  $r$  is calculated. Having a scatterplot in front of you is also useful because it enables you to estimate what the value of  $r$  might be — as you did in the previous exercise, and thus you can check that your workings on the calculator are correct.

**WORKED EXAMPLE 11**



The heights (in centimetres) of 21 football players were recorded against the number of marks they took in a game of football. The data are shown in the following table.

**TUTORIAL**  
eles-1244  
Worked example 11



- a Construct a scatterplot for the data.
- b Comment on the correlation between the heights of players and the number of marks that they take, and estimate the value of  $r$ .
- c Calculate  $r$  and use it to comment on the relationship between the heights of players and the number of marks they take in a game.

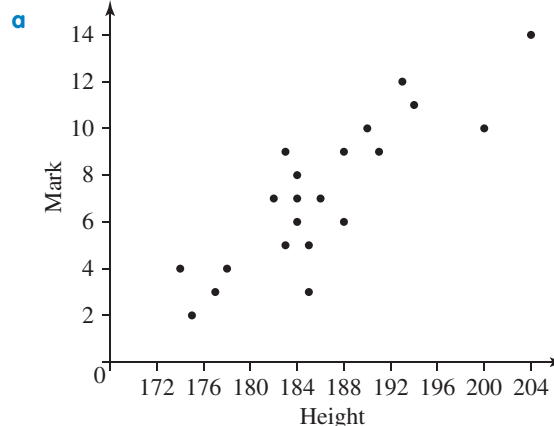
Height (cm)	Number of marks taken
184	6
194	11
185	3
175	2
186	7
183	5
174	4
200	10
188	9
184	7
188	6

Height (cm)	Number of marks taken
182	7
185	5
183	9
191	9
177	3
184	8
178	4
190	10
193	12
204	14

**THINK**

- a Height is the independent variable, so plot it on the  $x$ -axis; the number of marks is the dependent variable, so show it on the  $y$ -axis.

**WRITE/DRAW**





- b** Comment on the correlation between the variables and estimate the value of  $r$ .
- c** **1** Because there is a linear form and there are no outliers, the calculation of  $r$  is appropriate.
- 2** Use a calculator to find the value of  $r$ . Round to 2 decimal places.
- 3** The value of  $r = 0.86$  indicates a strong positive linear relationship.

**b** The data show what appears to be a linear form of moderate strength.  
We might expect  $r \approx 0.8$ .

**c**

$$r = 0.859\ 311\ \dots$$


$$\approx 0.86$$

$r = 0.86$ . This indicates there is a strong positive linear association between the height of a player and the number of marks he takes in a game. That is, the taller the player, the more marks we might expect him to take.

### study on

Units:	3 & 4
AOS:	DA
Topic:	6
Concept:	9

 **Concept summary**  
Read a summary of this concept.

 **See more**  
Watch a video about correlation and causation.



## Correlation and causation

In Worked example 11 we saw that  $r = 0.86$ . While we are entitled to say that there is a strong association between the height of a footballer and the number of marks he takes, we cannot assert that the height of a footballer causes him to take a lot of marks. Being tall might assist in taking marks, but there will be many other factors which come into play; for example, skill level, accuracy of passes from teammates, abilities of the opposing team, and so on.

So, while establishing a high degree of correlation between two variables may be interesting and can often flag the need for further, more detailed investigation, it in no way gives us any basis to comment on whether or not one variable *causes* particular values in another variable.

## The coefficient of determination ( $r^2$ )

The coefficient of determination is given by  $r^2$ . It is very easy to calculate — we merely square Pearson's product-moment correlation coefficient ( $r$ ). The value of the coefficient of determination ranges from 0 to 1; that is,  $0 \leq r^2 \leq 1$ .

- The coefficient of determination is useful when we have two variables which have a linear relationship. It tells us the proportion of variation in one variable which can be explained by the variation in the other variable.**
- The coefficient of determination provides a measure of how well the linear rule linking the two variables ( $x$  and  $y$ ) predicts the value of  $y$  when we are given the value of  $x$ .**

### WORKED EXAMPLE 12

A set of data giving the number of police traffic patrols on duty and the number of fatalities for the region was recorded and a correlation coefficient of  $r = -0.8$  was found. Calculate the coefficient of determination and interpret its value.

#### THINK

- Calculate the coefficient of determination by squaring the given value of  $r$ .
- Interpret your result.

#### WRITE

$$\begin{aligned} \text{Coefficient of determination} &= r^2 \\ &= (-0.8)^2 \\ &= 0.64 \end{aligned}$$

We can conclude from this that 64% of the variation in the number of fatalities can be explained by the variation in the number of police traffic patrols on duty. This means that the number of police traffic patrols on duty is a major factor in predicting the number of fatalities.

**eBookplus**

**TUTORIAL**  
eles-1263  
Worked example 12

### study on

Units:	3 & 4
AOS:	DA
Topic:	6
Concept:	10

 **Concept summary**  
Read a summary of this concept.

Note: In the previous worked example, 64% of the variation in the number of fatalities was due to the variation in the number of police cars on duty and 36% was due to other factors; for example, days of the week or hour of the day.

## Exercise 2G Calculating $r$ and the coefficient of determination

- 1 **WE11** The yearly salary ( $\times \$1000$ ) and the number of votes polled in the Brownlow medal count are given below for 10 footballers.

<b>Yearly salary (<math>\times \\$1000</math>)</b>	180	200	160	250	190	210	170	150	140	180
<b>Number of votes</b>	24	15	33	10	16	23	14	21	31	28

- Construct a scatterplot for the data.
  - Comment on the correlation of salary and the number of votes and make an estimate of  $r$ .
  - Calculate  $r$  and use it to comment on the relationship between yearly salary and number of votes.
- 2 **WE12** A set of data, obtained from 40 smokers, gives the number of cigarettes smoked per day and the number of visits per year to the doctor. The Pearson's correlation coefficient for these data was found to be 0.87. Calculate the coefficient of determination for the data and interpret its value.
- 3 Data giving the annual advertising budgets ( $\times \$1000$ ) and the yearly profit increases (%) of 8 companies are shown below.

<b>Annual advertising budget (<math>\times \\$1000</math>)</b>	11	14	15	17	20	25	25	27
<b>Yearly profit increase (%)</b>	2.2	2.2	3.2	4.6	5.7	6.9	7.9	9.3

- Construct a scatterplot for these data.
  - Comment on the correlation of the advertising budget and profit increase and make an estimate of  $r$ .
  - Calculate  $r$ .
  - Calculate the coefficient of determination.
  - Write the proportion of the variation in the yearly profit increase that can be explained by the variation in the advertising budget.
- 4 Data showing the number of tourists visiting a small country in a month and the corresponding average monthly exchange rate for the country's currency against the American dollar are given below.

<b>Number of tourists (<math>\times 1000</math>)</b>	2	3	4	5	7	8	8	10
<b>Exchange rate</b>	1.2	1.1	0.9	0.9	0.8	0.8	0.7	0.6

- Construct a scatterplot for the data.
  - Comment on the correlation between the number of tourists and the exchange rate and give an estimate of  $r$ .
  - Calculate  $r$ .
  - Calculate the coefficient of determination.
  - Write the proportion of the variation in the number of tourists that can be explained by the exchange rate.
- 5 Data showing the number of people in 9 households against weekly grocery costs are given below.

<b>Number of people in household</b>	2	5	6	3	4	5	2	6	3
<b>Weekly grocery costs (\$)</b>	60	180	210	120	150	160	65	200	90

- Construct a scatterplot for the data.
- Comment on the correlation of the number of people in a household and the weekly grocery costs and give an estimate of  $r$ .

**eBookplus**

**DIGITAL DOC**  
doc-9416  
Spreadsheet  
Pearson's  
product-moment  
correlation

- c Calculate  $r$ .
  - d Calculate the coefficient of determination.
  - e Write the proportion of the variation in the weekly grocery costs that can be explained by the variation in the number of people in a household.
- 6 Data showing the number of people on 8 fundraising committees and the annual funds raised are given below.

<b>Number of people on committee</b>	3	6	4	8	5	7	3	6
<b>Annual funds raised (\$)</b>	4500	8500	6100	12 500	7200	10 000	4700	8800

- a Construct a scatterplot for these data.
- b Comment on the correlation between the number of people on a committee and the funds raised and make an estimate of  $r$ .
- c Calculate  $r$ .
- d Based on the value of  $r$  obtained in part c, would it be appropriate to conclude that the increase in the number of people on the fundraising committee causes the increase in the amount of funds raised?
- e Calculate the coefficient of determination.
- f Write the proportion of the variation in the funds raised that can be explained by the variation in the number of people on a committee.

The following information applies to questions 7 and 8. A set of data was obtained from a large group of women with children under 5 years of age. They were asked the number of hours they worked per week and the amount of money they spent on child care. The results were recorded and the value of Pearson's correlation coefficient was found to be 0.92.

- 7 **MC** Which of the following is not true?
- A The relationship between the number of working hours and the amount of money spent on child care is linear.
  - B There is a positive correlation between the number of working hours and the amount of money spent on child care.
  - C The correlation between the number of working hours and the amount of money spent on child care can be classified as strong.
  - D As the number of working hours increases, the amount spent on child care increases as well.
  - E The increase in the number of hours worked causes the increase in the amount of money spent on child care.
- 8 **MC** Which of the following is not true?
- A The coefficient of determination is about 0.85.
  - B The number of working hours is the major factor in predicting the amount of money spent on child care.
  - C About 85% of the variation in the number of hours worked can be explained by the variation in the amount of money spent on child care.
  - D Apart from number of hours worked, there could be other factors affecting the amount of money spent on child care.
  - E About  $\frac{17}{20}$  of the variation in the amount of money spent on child care can be explained by the variation in the number of hours worked.





- 9 An investigation is undertaken with people following the Certain Slim diet to explore the link between weeks of dieting and total weight loss. The data are shown below.

Total weight loss (kg)	Number of weeks on the diet
1.5	1
4.5	5
9	8
3	3
6	6
8	9
3.5	4
3	2
6.5	7
8.5	10
4	4
6.5	6
10	9
2.5	2
6	5



- Display the data on a scatterplot.
- Describe the association between the two variables in terms of direction, form and strength.
- Is it appropriate to use Pearson's correlation coefficient to explain the link between the number of weeks on the Certain Slim diet and total weight loss?
- Estimate the value of Pearson's correlation coefficient from the scatterplot.
- Calculate the value of this coefficient.
- Is the total weight loss affected by the number of weeks staying on the diet?
- Calculate the value of the coefficient of determination.
- What does the coefficient of determination say about the relationship between total weight loss and the number of weeks on the Certain Slim diet?

# Summary

## Dependent and independent variables

- Bivariate data are data with two variables.
- In a relationship involving two variables, if the values of one variable depend on the values of another variable, then the former variable is referred to as the *dependent variable* and the latter variable is referred to as the *independent variable*.
- When data are displayed on a graph, the independent variable is placed on the horizontal axis and the dependent variable is placed on the vertical axis.

## Back-to-back stem plots

- A back-to-back stem plot displays bivariate data involving a numerical variable and a categorical variable with two categories.
- Together with summary statistics, back-to-back stem plots can be used to compare the two distributions.

## Parallel boxplots

- To display a relationship between a numerical variable and a categorical variable with two or *more* categories, we can use a parallel boxplot.
- A parallel boxplot is obtained by constructing individual boxplots for each distribution and positioning them on a common scale.

## Two-way frequency tables and segmented bar charts

- The two-way frequency table is a tool for examining the relationship between two categorical variables.
- If the total number of scores in each of the two categories is unequal, percentages should be calculated to analyse the table properly.
- When the independent variable is placed in the columns of the table, the numbers in each column should be expressed as a percentage of that column's total.
- The data in a two-way frequency table in percentage form can be represented graphically as a segmented bar chart.
- The columns in a segmented bar chart match the columns in a two-way frequency table. Each segment corresponds to each cell in the table.

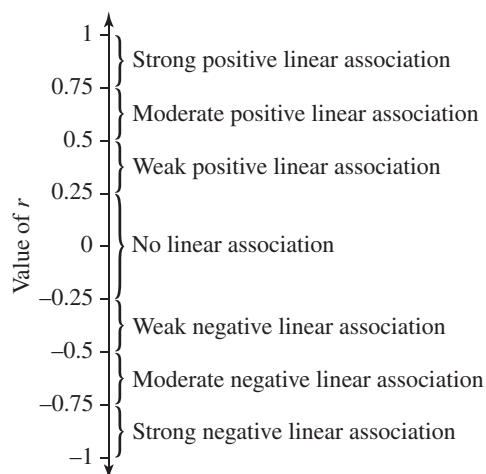
## Scatterplots

- A scatterplot gives a visual display of the relationship between two numerical variables.
- In analysing the scatterplot we look for a pattern in the way the points lie. Certain patterns tell us that certain relationships exist between the two variables. This is referred to as a *correlation*. We look at what type of correlation exists and how strong it is.
- When describing the relationship between two variables displayed on a scatterplot, we need to comment on:
  - (a) the direction — whether it is positive or negative
  - (b) the form — whether it is linear or non-linear
  - (c) the strength — whether it is strong, moderate or weak
  - (d) possible outliers.

## Pearson's product-moment correlation coefficient

- Pearson's product-moment correlation coefficient is used to measure the strength of a linear relationship between two variables.
- The symbol for Pearson's product-moment correlation coefficient is  $r$ .
- The calculation of  $r$  is applicable to sets of bivariate data which are known to be linear in form and which don't have outliers.

- The value of  $r$  can be estimated from the scatterplot;  $-1 \leq r \leq 1$ .



- The formula for calculating Pearson's correlation coefficient  $r$  is as follows:

$$r = \frac{1}{n-1} \sum_{i=1}^n \left( \frac{x_i - \bar{x}}{s_x} \right) \left( \frac{y_i - \bar{y}}{s_y} \right)$$

where  $n$  is the number of pairs of data in the set

$s_x$  is the standard deviation of the  $x$ -values

$s_y$  is the standard deviation of the  $y$ -values

$\bar{x}$  is the mean of the  $x$ -values

$\bar{y}$  is the mean of the  $y$ -values.

- The calculation of  $r$  is often done using a CAS calculator.
- Even if we find that two variables have a very high degree of correlation, for example  $r = 0.95$ , we cannot say that the value of one variable is *caused* by the value of the other variable.
- If  $r = 1$  or  $-1$  there is a perfect linear relationship.

### Calculating $r$ and the coefficient of determination

- The coefficient of determination =  $r^2$ ;  $0 \leq r^2 \leq 1$ .
- The coefficient of determination is useful when we have two variables which have a linear relationship. It tells us the proportion of variation in one variable which can be explained by the variation in the other variable.
- To change the value of  $r^2$  to a percentage, multiply it by 100.

# Chapter review

## MULTIPLE CHOICE

- 1 In a study on the effectiveness of vitamin C, a researcher asked a group of people with cold and flu symptoms to record the number of days these symptoms persisted and their daily dosage (in mg) of vitamin C. If the researcher wishes to represent these data graphically, which of the following should she do?
- A Show the number of days the symptoms persisted on the  $x$ -axis, as this is the independent variable and the daily dosage of vitamin C on the  $y$ -axis, as this is the dependent variable.
  - B Show the daily dosage of vitamin C on the  $x$ -axis, as this is the dependent variable and the number of days the symptoms persisted on the  $y$ -axis, as this is the independent variable.
  - C Show the number of days the symptoms persisted on the  $x$ -axis, as this is the dependent variable and the daily dosage of vitamin C on the  $y$ -axis, as this is the independent variable.
  - D Show the daily dosage of vitamin C on the  $x$ -axis, as this is the independent variable and the number of days the symptoms persisted on the  $y$ -axis, as this is the dependent variable.
  - E It is impossible to decide which of the two variables is dependent and which one is independent, so it does not matter which axes we use.
- 2 A back-to-back stem plot is a useful way of displaying the relationship between:
- A the number of children attending a day care centre and whether or not the centre has federal funding
  - B height and wrist circumference
  - C age and weekly income
  - D weight and the number of takeaway meals eaten each week
  - E the age of a car and amount spent each year on servicing it

The following information relates to questions 3 and 4.

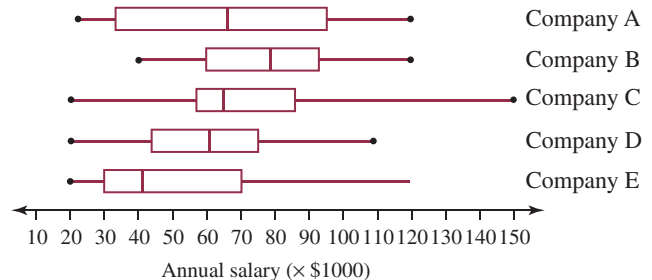
The salaries of people working at five different advertising companies are shown on the following parallel boxplots.

- 3 The company with the largest interquartile range is:

- A Company A
- B Company B
- C Company C
- D Company D
- E Company E

- 4 The company with the lowest median is:

- A Company A
- B Company B
- C Company C
- D Company D
- E Company E



Questions 5 and 6 relate to the following information.

Data showing reactions of junior staff and senior staff to a relocation of offices are given below in a two-way frequency table.

Attitude	Junior staff	Senior staff	Total
For	23	14	37
Against	31	41	72
Total	54	55	109

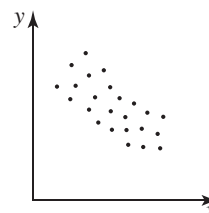
- 5 From this table, we can conclude that:
- A 23% of junior staff were for the relocation
  - B 42.6% of junior staff were for the relocation
  - C 31% of junior staff were against the relocation
  - D 62.1% of junior staff were for the relocation
  - E 28.4% of junior staff were against the relocation

- 6 From this table, we can conclude that:
- A 14% of senior staff were for the relocation
  - B 37.8% of senior staff were for the relocation
  - C 12.8% of senior staff were for the relocation
  - D 72% of senior staff were against the relocation
  - E 74.5% of senior staff were against the relocation

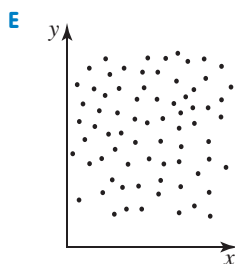
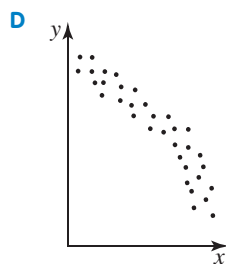
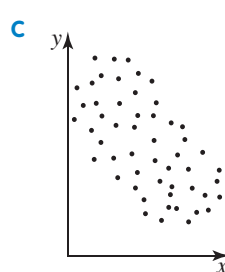
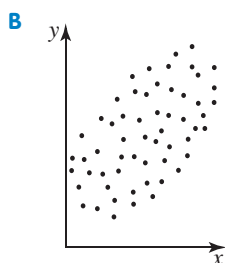
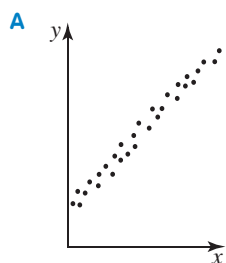
- 7 The relationship between the variables  $x$  and  $y$  is shown on the scatterplot at right.

The correlation between  $x$  and  $y$  would be best described as:

- A a weak positive association
- B a weak negative association
- C a strong positive association
- D a strong negative association
- E non-existent



- 8 A set of data relating the variables  $x$  and  $y$  is found to have an  $r$  value of  $-0.83$ . The scatterplot that would best represent this data set is:



- 9 A set of data comparing age with blood pressure is found to have a Pearson's correlation coefficient of 0.86. The coefficient of determination for these data would be closest to:

- A -0.86
- B -0.74
- C -0.43
- D 0.43
- E 0.74

- 1 For each of the following, write down which is the dependent and which is the independent variable or whether it is appropriate to classify the variables as such.

- a The number of injuries in a netball season and the age of a netball player
- b The suburb and the size of a home mortgage
- c IQ and weight

- 2 The number of hours of counselling received by a group of 9 full-time firefighters and 9 volunteer firefighters after a serious bushfire is given below.

<b>Full-time</b>	2	4	3	5	2	4	6	1	3
<b>Volunteer</b>	8	10	11	11	12	13	13	14	15

- a Construct a back-to-back stem plot to display the data.
- b Comment on the distributions of the number of hours of counselling of the full-time firefighters and the volunteers.

**SHORT ANSWER**

- 3 The IQ of 8 players in 3 different football teams were recorded and are shown below.

<b>Team A</b>	120	105	140	116	98	105	130	102
<b>Team B</b>	110	104	120	109	106	95	102	100
<b>Team C</b>	121	115	145	130	120	114	116	123

Display the data in parallel boxplots.

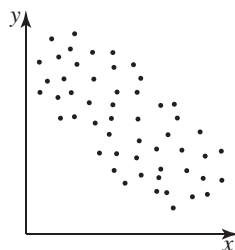
- 4 Delegates at the respective Liberal and Labor Party conferences were surveyed on whether or not they believed that uranium mining should continue. Forty-five Liberal delegates were surveyed and 15 were against continuation. Fifty-three Labor delegates were surveyed and 43 were against continuation.
- Present the data as percentages in a two-way frequency table and a segmented bar chart.
  - Comment on any difference between the reactions of the Liberal and Labor delegates.



- 5 a Construct a scatterplot for the data given in the table below.

<b>Age</b>	15	17	18	16	19	19	17	15	17
<b>Pulse rate</b>	79	74	75	85	82	76	77	72	70

- Use the scatterplot to comment on any relationship which exists between the variables.
- 6 For the variables shown on the scatterplot below, give an estimate of the value of  $r$  and use it to comment on the nature of the relationship between the two variables.



- 7 The table at right gives data relating the percentage of lectures attended by students in a semester and the corresponding mark for each student in the exam for that subject.
- Construct a scatterplot for these data.
  - Comment on the correlation between the lectures attended and the examination results and make an estimate of  $r$ .
  - Calculate  $r$ .
  - Calculate the coefficient of determination.
  - Write the proportion of the variation in the examination results that can be explained by the variation in the number of lectures attended.

<b>Lectures attended (%)</b>	<b>Exam result (%)</b>
70	80
59	62
85	89
93	98
78	84
85	91
84	83
69	72
70	75
82	85

1 An investigation into the relationship between age and salary bracket among some employees of a large computer company is made and the results are shown at right.

- a State which is the independent variable and which is the dependent variable.
- b State which of the following you *could* use to display the data:
  - i back-to-back stem plot
  - ii parallel boxplot
  - iii scatterplot
  - iv two-way frequency table in percentage form.
- c State which of the following you could calculate in order to find out more about the relationship between age and salary bracket:
  - i  $r$ , the Pearson product-moment correlation coefficient
  - ii the coefficient of determination.

Salary bracket (× \$1000)	Age
20–39	32 21 43 23 22 27 37
40–59	29 31 37 26 33 37
60–79	41 29 39 42 47 45 43 38
80–99	43 48 38 37 49 51 53 59
100–120	48 37 55 61

2 For marketing purposes, the administration of the Arts Centre needs to compare the ages of people attending two different concerts: a symphony orchestra concert and a jazz concert. Twenty people were randomly selected from each audience and their ages were recorded as shown.

Event	Ages of people attending the event
Symphony orchestra concert	20, 23, 30, 35, 39, 42, 45, 45, 47, 48, 48, 49, 49, 50, 53, 54, 56, 58, 58, 60
Jazz concert	16, 18, 19, 19, 20, 23, 24, 27, 29, 30, 33, 34, 38, 39, 40, 42, 43, 45, 46, 62

- a Display the data on a back-to-back stem plot.
- b For each category calculate the following statistics:
  - i  $X_{\min}$
  - ii  $Q_1$
  - iii median
  - iv  $Q_3$
  - v  $X_{\max}$
  - vi mean
  - vii interquartile range (IQR)
  - viii standard deviation.



- c Use the stem plot together with some summary statistics to compare the distributions of the ages of patrons attending the two concerts.

One month later, at the beginning of the opera season, twenty people were again selected (this time from the opera audience) and their ages were recorded as shown.

Event	Ages of people attending the event
Opera	12, 18, 29, 30, 33, 35, 38, 39, 42, 46, 49, 50, 54, 56, 56, 57, 59, 63, 65, 68

The administration of the Arts Centre now wishes to compare all three distributions of the ages.

- d Explain why it is not possible to use a back-to-back stem plot for this task.  
 e Calculate the eight summary statistics for the ages of the opera-goers (as in part b above).  
 f Display the data for the three events using parallel boxplots.  
 g Use the boxplots and some summary statistics to compare the three distributions.
- 3 In one study, 380 Year 12 students were asked how often they were engaged in any sporting activity outside school. Students were also asked to classify their stress level in relation to their VCE studies. The results at right were obtained.

Level of stress	Engaged in sporting activity outside school		
	Regularly	Sometimes	Never
Low	16	32	36
Medium	12	40	56
High	6	52	130

- a In this study, which would be the independent variable: stress level, or the amount of sporting activity?  
 b How many students in this study reported a high level of stress?  
 c How many students were engaged in sport activity outside of school?  
 d Represent the data in a two-way frequency table in percentage form.  
 e Display the data from part d using a segmented bar chart.  
 f Comment on any relationship between the stress level and the amount of sporting activity for this group of Year 12 students.
- 4 The data in the table below show the number of hours spent by students learning to touch-type and their corresponding speed in words per minute (wpm).

Time (h)	20	33	22	39	40	37	46	44	24	36	50	48	29
Speed (wpm)	34	46	38	53	52	49	60	58	36	42	65	63	40

- a State which variable is independent and which is dependent.  
 b Represent the data on a scatterplot.  
 c Use the scatterplot to comment on the relationship between the two variables.  
 d Is it appropriate to use these data to calculate the value of Pearson's product-moment correlation coefficient? Explain.  
 e Estimate the value of  $r$  from the scatterplot.  
 f Calculate the value of  $r$  using a CAS calculator. Does the value of  $r$  support the observations you made in part c?  
 g Calculate the coefficient of determination and interpret the result.

**eBookplus**

DIGITAL DOC  
 doc-9417  
 Test Yourself  
 Chapter 2

**study on**

Units: 3 & 4

AOS: DA



Practice  
 VCE exam  
 questions

Use StudyON to  
 access all exam  
 questions on this  
 topic since 2002.



**Chapter opener****DIGITAL DOC**

- 10 Quick Questions doc-9409: Warm up with a quick quiz on bivariate data. (page 57)

**2B Back-to-back stem plots****TUTORIAL**

- **WE2** eles-1259: Watch a tutorial on displaying data on a back-to-back stem plot. (page 59)

**2C Parallel boxplots****DIGITAL DOCS**

- Spreadsheet doc-9410: Compare two sets of data using parallel boxplots. (page 64)
- WorkSHEET 2.1 doc-9411: Identify independent and dependent variables and construct parallel boxplots and back-to-back stem plots. (page 65)

**2D Two-way frequency tables and segmented bar charts****DIGITAL DOCS**

- Spreadsheet doc-9413: Construct a two-way frequency table. (page 67)
- SkillSHEET 2.1 doc-9412: Practise expressing one number as a percentage of another. (page 68)

**TUTORIAL**

- **WE7** eles-1260: Learn how to present data in a two-way table. (page 67)

**2E Scatterplots****DIGITAL DOC**

- Spreadsheet doc-9414: Investigate the relationship between two variables by constructing a scatterplot (page 73)

**TUTORIAL**

- **WE9** eles-1261: Watch a worked example on constructing a scatterplot to determine the relationship between the heights of students and the number of hours they study. (page 71)

**2F Pearson's product-moment correlation coefficient****DIGITAL DOC**

- WorkSHEET 2.2 doc-9415: Displaying data using scatterplots and recognising linear and non-linear relationships. (page 76)

**TUTORIAL**

- **WE10** eles-1262: Watch a worked example on estimating  $r$  and using it to comment on the relationship between two variables. (page 74)

**INTERACTIVITY**

- Pearson's product-moment correlation coefficient int-0183: Use the interactivity to consolidate your knowledge of Pearson's product-moment correlation coefficient and how it relates to bivariate data. (page 73)

**2G Calculating  $r$  and the coefficient of determination****DIGITAL DOC**

- Spreadsheet doc-9416: Investigate Pearson's product-moment correlation coefficient and the coefficient of determination. (page 79)

**TUTORIALS**

- **WE11** eles-1244: Watch a worked example on how to construct a scatterplot and use it to estimate the value of  $r$ . (page 77)
- **WE12** eles-1263: Watch a worked example on calculating the coefficient of determination and how it is used to interpret the relationship between two variables. (page 78)

**Chapter review****DIGITAL DOC**

- Test Yourself doc-9417: Take the end-of-chapter test to test your progress. (page 88)

To access eBookPLUS activities, log on to [www.jacplus.com.au](http://www.jacplus.com.au)

# Answers CHAPTER 2

## BIVARIATE DATA

### Exercise 2A Dependent and independent variables

- 1 a Independent — age, dependent — salary  
 b Independent — amount of fertiliser, dependent — growth  
 c Not appropriate  
 d Not appropriate  
 e Independent — number in household, dependent — size of house  
 f Independent — month of the year, dependent — size of electricity bill  
 g Independent — number of hours, dependent — mark on the test  
 h Not appropriate  
 i Independent — season, dependent — cost
- 2 C  
 3 C  
 4 D

### Exercise 2B Back-to-back stem plots

- 1 Key: 2|3 = 23

Leaf	Stem	Leaf
German		French
2 1 1 0	2	3 4
7 6 5 5	2*	5 5 8
3 2 1 0 0	3	0 1 4 4
9 8 7 7	3*	5 6 8 8 9
2 1	4	2 3 4 4
5	4*	6 8

- 2 Key: 2\*|7 = 2.7 (kg)

Leaf	Stem	Leaf
Boys		Girls
	2*	6 7
4 4	3	0 1 1 2 3
8 7 6	3*	6 7
3 2	4	0
9 8	4*	
0	5	

- 3 a Key: 2\*|5 = 25 trucks

Leaf	Stem	Leaf
A		B
2 1	1	0
7 7 6 6 5	1*	5 6
4 3 2 1 0	2	0 1 3
7 5	2*	5 6 8 9
	3	0 1 2
	3*	5

- b For supermarket A the mean is 19, the median is 18.5, the standard deviation is 4.9 and the interquartile range is 7. The distribution is symmetric. For supermarket B the mean is 24.4, the median is 25.5, the standard deviation is 7.2 and the interquartile range is 10. The distribution is symmetric. The centre and spread of the distribution of supermarket B is higher than that of supermarket A. There is greater variation in the number of trucks arriving at supermarket B.

- 4 a Key: 1|2 = 12 marks

Leaf	Stem	Leaf
Females		Males
	1	0
3 2	1	2 3
5 5 4 4	1	4 4 5
7 6	1	7
	1	9

- b For the marks of the females, the mean is 14.5, the median is 14.5, the standard deviation is 1.6 and the interquartile range is 2. The distribution is symmetric. For the marks of the males, the mean is 14.25, the median is 14, the standard deviation is 2.8 and the interquartile range is 3.5. The distribution is symmetric. The centre of each distribution is about the same. The spread of marks for the boys is greater, however. This means that there is a wider variation in the abilities of the boys compared to the abilities of the girls.

- 5 a Key: 2\*|6 = 26 marks

Leaf	Stem	Leaf
2007		2008
	2	2
	2*	6 7 8
1 0	3	0 1 1 3 4
9 7 5	3*	6
3 2 1 1	4	
6	4*	

- b The distribution of marks for 2007 and for 2008 are each symmetric. For the 2007 marks, the mean is 38.5, the median is 40, the standard deviation is 5.2 and the interquartile range is 7. The distribution is symmetric. For the 2008 marks, the mean is 29.8, the median is 30.5, the standard deviation is 4.2 and the inter-quartile range is 6. The spread of each of the distributions is much the same, but the centre of each distribution is quite different with the centre of the 2008 distribution lower. The work may have become a lot harder!

- 6 a Key: 3\*|6 = 36 years old

Leaf	Stem	Leaf
Female		Male
4 3	2	2
8 7 6 5	2*	5
1 0	3	0 1
	3*	6 7
	4	2
	4*	6

- b For the distribution of the females, the mean is 26.75, the median is 26.5, the standard deviation is 2.8 and the interquartile range is 4.5. For the distribution of the males, the mean is 33.6, the median is 33.5,

the standard deviation is 8.2 and the interquartile range is 12. The centre of the distributions is very different: it is much higher for the males. The spread of the ages of the females who attend the fitness class is very small but very large for males.

- 7 a Key: 5|0 = 50 points

Leaf	Stem	Leaf
Kindergarten		Prep.
3	0	5
4 3	1	2 7
8 5	2	5 7
6 2	3	2 5
7 1	4	4 6
0	5	2

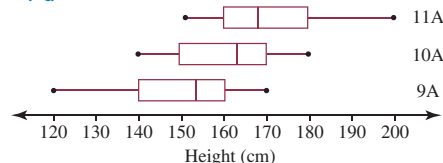
- b For the distribution of scores of the kindergarten children, the mean is 28.9, the median is 30, the standard deviation is 15.4 and the interquartile range is 27. For the distribution of scores for the prep. children, the mean is 29.5, the median is 29.5, the standard deviation is 15.3 and the interquartile range is 27. The distributions are very similar. There is not a lot of difference between the way the kindergarten children and the prep. children scored.

8 B

9 C

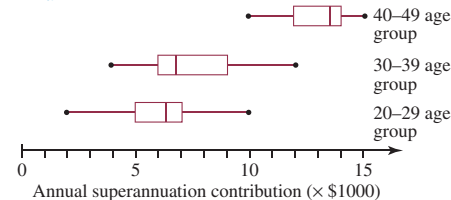
### Exercise 2C Parallel boxplots

- 1 a

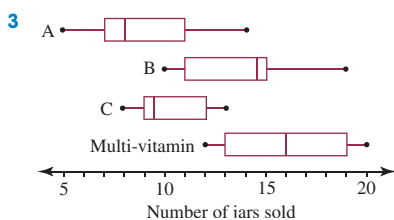


- b Clearly, the median height increases from Year 9 to Year 11. There is greater variation in 9A's distribution than in 10A's. There is a wide range of heights in the lower 25% of the distribution of 9A's distribution. There is a greater variation in 11A's distribution than in 10A's, with a wide range of heights in the top 25% of the 11A distribution.

- 2 a



- b Clearly, there is a great jump in contributions to superannuation for people in their 40s. The spread of contributions for that age group is smaller than for people in their 20s or 30s, suggesting that a high proportion of people in their 40s are conscious of superannuation. For people in their 20s and 30s, the range is greater, indicating a range of interest in contributing to super.



Overall, the biggest sales were of multi-vitamins, followed by vitamin B, then C and finally vitamin A.

- 4 a True                                      b True  
 c False                                      d True  
 5 a The Pearlfishers                      b Orlando  
 6 a D    b C

**Exercise 2D Two-way frequency tables and segmented bar charts**

1

Attitude	Female	Male	Total
For	37	79	116
Against	102	23	125
Total	139	102	241

2

Lesson length	Junior	Senior	Total
45 minutes	50	33	83
1 hour	36	60	96
Total	86	93	179

- 3 a i 22                                      ii 26                                      iii 19  
 iv 45                                      v 41  
 b i 12                                      ii 9                                      iii 21  
 iv 42                                      v 33  
 c i 47%                                      ii 58%  
 iii 100%                                      iv 100%

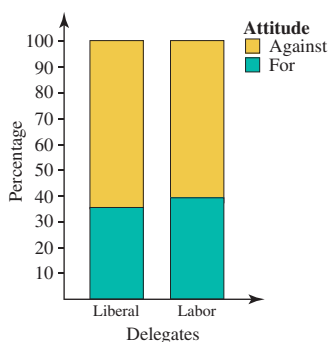
4

Preference	Men	Women
Rent by themselves	57%	59%
Share with friends	43%	41%
Total	100%	100%

5 D    6 C

7

Attitude	Liberal	Labor
For	35.5	39.4
Against	64.5	60.6
Total	100.0	100.0

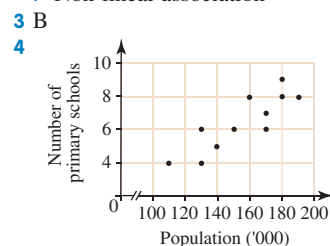


There is not a lot of difference in the reactions.

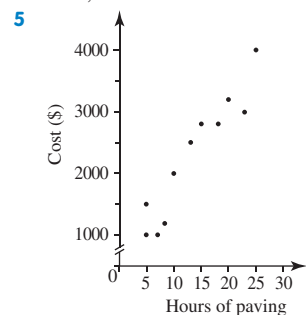
- 8 a C    b A

**Exercise 2E Scatterplots**

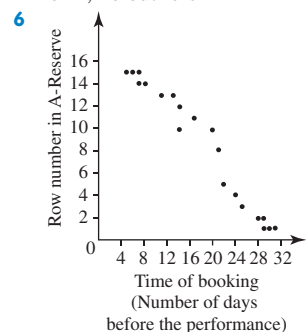
- 1 a Yes — positive association  
 b Yes — positive association  
 c Yes — positive association  
 d Yes — negative association  
 e Yes — positive association  
 f Yes — negative association  
 g No — no association  
 2 a Weak, negative association of linear form  
 b Strong, negative association of linear form  
 c Moderate, positive association of linear form  
 d Strong, positive association of linear form  
 e No association  
 f Non-linear association



Moderate positive association of linear form, no outliers



Strong positive association of linear form, no outliers



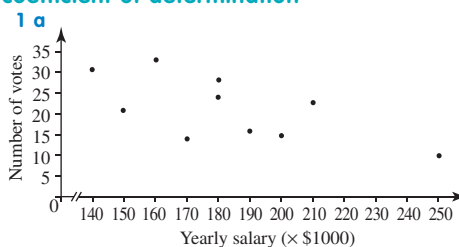
Strong negative association of linear form, no outliers

**Exercise 2F Pearson's product-moment correlation coefficient**

- 1 a No association  
 b Moderate positive  
 c Strong negative  
 d Strong negative  
 e Strong positive

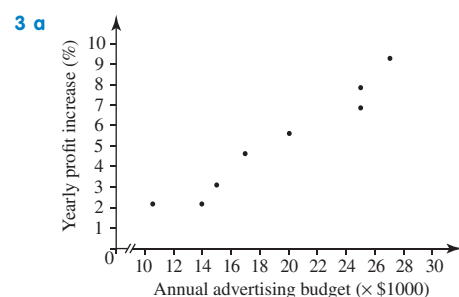
- f Strong positive  
 g Weak negative  
 h No association  
 2 a i  $r \approx -0.8$   
 ii Strong, negative, linear association  
 b i  $r \approx 0.6$   
 ii Moderate, positive, linear association  
 c i  $r \approx 0.2$   
 ii No linear association  
 d i  $r \approx -0.2$   
 ii No linear association  
 e i  $r = 1$   
 ii Perfect, positive, linear association  
 f i  $r \approx 0.8$   
 ii Strong, positive, linear association  
 g i  $r \approx 0$   
 ii No linear association  
 h i  $r \approx -0.7$   
 ii Moderate, negative, linear association

**Exercise 2G Calculating  $r$  and the coefficient of determination**

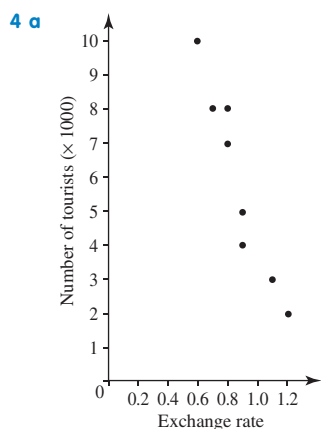


- b There is moderate, negative linear association.  $r$  is approximately  $-0.6$ .  
 c  $r = -0.66$ . There is a moderate negative linear association between the yearly salary and the number of votes. That is, the larger the yearly salary of the player, the fewer the number of votes we might expect to see.

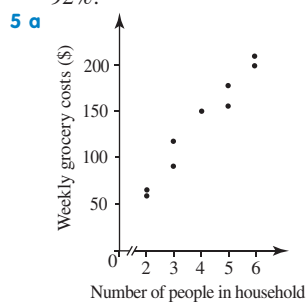
- 2 Coefficient of determination is 0.7569. The portion of variation in the number of visits to the doctor that can be explained by the variation in the number of cigarettes smoked is about 76%.



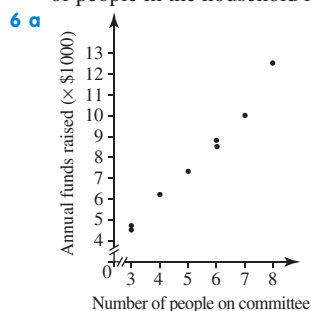
- b There is strong, positive linear association.  $r$  is approximately 0.8.  
 c  $r = 0.98$   
 d Coefficient of determination is 0.96.  
 e The proportion of the variation in the yearly profit increase that can be explained by the variation in the advertising budget is 96%.



- b** There is strong, negative association of a linear form and  $r$  is approximately  $-0.9$ .  
**c**  $r = -0.96$   
**d** Coefficient of determination is  $0.92$ .  
**e** The proportion of the variation in the number of tourists that can be explained by the variation in the exchange rate is  $92\%$ .

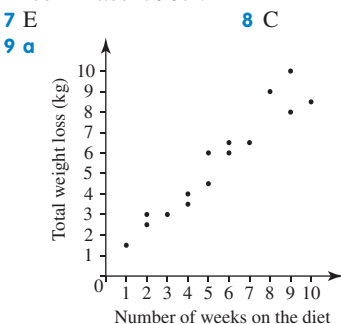


- b** There is strong, positive association of a linear form and  $r$  is approximately  $0.9$ .  
**c**  $r = 0.98$   
**d** Coefficient of determination is  $0.96$ .  
**e** The proportion of the variation in the weekly grocery costs that can be explained by the variation in the number of people in the household is  $96\%$ .



- b** There is almost perfect positive association of a linear form and  $r$  is nearly  $1$ .  
**c**  $r = 0.99$   
**d** No. High degree of correlation does not mean we can comment on whether one variable causes particular values in another.  
**e** Coefficient of determination is  $0.98$ .

- f** The proportion of the variation in the funds raised that can be explained by the variation in the number of people on the committee is  $98\%$ .



- b** The scatterplot shows strong, positive association of linear form.  
**c** It is appropriate since the scatterplot indicates association showing linear form and there are no outliers.  
**d**  $r \approx 0.9$  **e**  $r = 0.96$   
**f** We cannot say whether total weight loss is affected by the number of weeks people stayed on the Certain Slim diet. We can only note the degree of correlation.  
**g**  $r^2 = 0.92$   
**h** The coefficient of determination tells us that  $92\%$  of the variation in total weight loss can be explained by the variation in the number of weeks on the Certain Slim diet.

## CHAPTER REVIEW

### MULTIPLE CHOICE

- 1** D   **2** A   **3** A   **4** E   **5** B  
**6** E   **7** D   **8** D   **9** E

### SHORT ANSWER

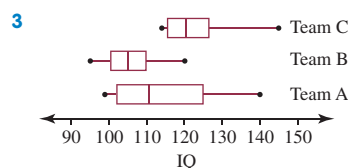
- 1 a** Number of injuries — dependent, age of player — independent  
**b** Suburb — independent, size of mortgage — dependent  
**c** It is not appropriate to designate one or other as independent or dependent.

**2 a**

Leaf	Stem	Leaf
Full-time		Volunteer
1	0	
2	2	
4 4 3 3	0	
6 5	0	
	0	8
	1	0 1 1
	1	2 3 3
	1	4 5
	1	
	1	

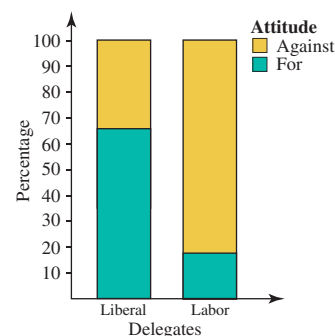
Key:  $0|3 = 3$  hours

- b** Both distributions are symmetric with the same spread. The centre of the volunteers' distribution is much higher than that of the full-time firefighters' distribution. Clearly, the volunteers needed more counselling.

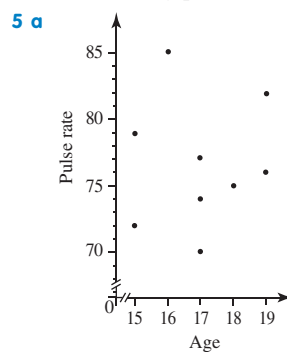


**4 a**

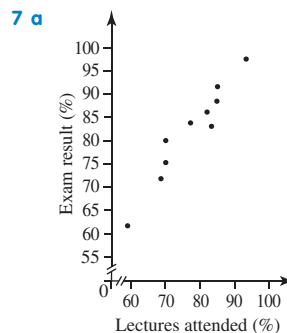
Attitude	Liberal	Labor
For	66.7	18.9
Against	33.3	81.1
Total	100.0	100.0



- b** Clearly, the reaction to uranium mining is affected by political affiliation.



- b** There appears to be an extremely weak or no association between the variables.  
**6**  $r$  is approximately equal to  $-0.7$ . There is a moderate, negative linear association between the variables  $x$  and  $y$ .



- b** There is strong, positive correlation of a linear form between the variables and  $r$  is approximately equal to  $0.8$ .  
**c**  $r = 0.96$

- d The coefficient of determination is 0.93.  
 e The proportion of the variation in the exam results that can be explained by the variation in the number of lectures attended is 93%.

**EXTENDED RESPONSE**

- 1 a Age is independent and salary bracket is the dependent variable.  
 b Parallel boxplot  
 c Neither, since we have categorical data versus numerical data and not numerical data versus numerical data.  
 2 a Key: 1|6 = 16 years old

Leaf	Stem	Leaf
Jazz concert		Symphony concert
9 9 8 6	1	
9 7 4 3 0	2	0 3
9 8 4 3 0	3	0 5 9
6 5 3 2 0	4	2 5 5 7 8 8 9 9
	5	0 3 4 6 8 8
	6	0

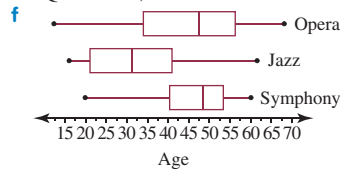
b

	Summary	Symphony concert	Jazz concert
i	$X_{\min}$	20	16
ii	$Q_1$	40.5	21.5
iii	Median	48	31.5
iv	$Q_3$	53.5	41
v	$X_{\max}$	60	62
vi	Mean	45.45	32.35
vii	IQR	13	19.5
viii	Standard deviation	11.20	12.04

- c Overall, it appears that people who attended the symphony concert were older than those who attended the jazz concert. The spread of ages is nearly

the same (slightly higher for the jazz audience).

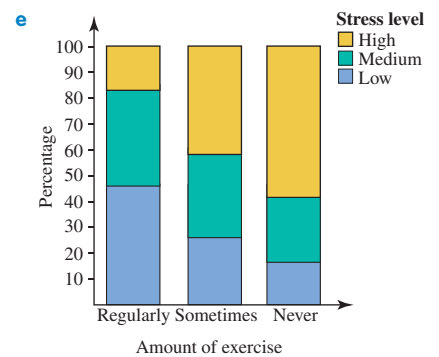
- d Back-to-back stem plots can be used only for data with two categories. Since there are three events, parallel boxplots should be used.  
 e  $X_{\min} = 12$ ,  $Q_1 = 34$ , median = 47.5,  $Q_3 = 56.5$ ,  $X_{\max} = 68$ , mean = 44.95, IQR = 22.5, standard deviation = 15.55



- g Overall, the people who went to the symphony concert and to the opera were of similar ages and older than those who went to the jazz concert. The ages of people who went to the opera are the most spread out, while the ages of people who attended the symphony concert are the least spread out.

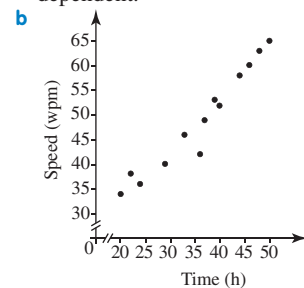
- 3 a The amount of sporting activity is independent; the level of stress is dependent.  
 b 188  
 c 158  
 d

Level of stress	Engaged in sport activity outside school		
	Regularly	Sometimes	Never
Low	47.1%	25.8%	16.2%
Medium	35.3%	32.3%	25.2%
High	17.6%	41.9%	58.6%
Total	100%	100%	100%



- f Overall, it appears that for this group of students, stress levels are related to the amount of physical activity they are engaged in outside of school.

- 4 a Hours spent touch-typing — independent, speed of touch-typing — dependent.



- c Strong, positive, linear relationship between the two variables  
 d Yes, since the scatterplot shows a linear relationship with no outliers.  
 e  $r$  is about 0.9  
 f  $r = 0.97$ ; yes  
 g  $r^2 = 0.94$ . This means that 94% of variation in the speed of touch-typing can be explained by variation in the number of hours spent touch-typing.

