2

Comparing data sets

2.1 Kick off with CAS

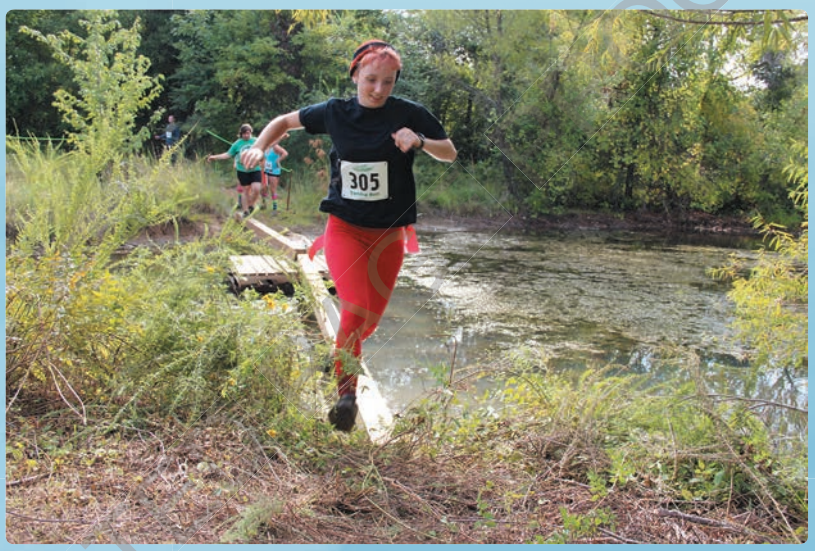
Exploring parallel boxplots with CAS

Parallel boxplots can be used to compare and contrast key information about two different numerical data sets.

1 Use CAS to draw parallel boxplots of the following two data sets, which detail the time it takes for two groups of individuals to complete an obstacle course (rounded to the nearest minute).

Group A: 18, 22, 24, 17, 22, 27, 15, 20, 25, 19, 26, 19, 23, 26, 18, 20, 27, 24, 16

Group B: 21, 22, 19, 21, 17, 21, 18, 24, 21, 20, 18, 24, 35, 22, 19, 17, 23, 20, 19



2 Use your parallel boxplots from question **1** to answer the following questions.

- **a** Which group has the larger range of data?
- **b** Which group has the larger interquartile range of data?
- **c** Which group has the higher median value?

3 One of the data sets has an outlier.

- **a** How is this marked on the parallel boxplot?
- **b** State the value of the outlier.

4 a Use CAS to draw a parallel boxplot of the following two data sets.

Set A: 41, 46, 38, 44, 49, 39, 50, 47, 47, 42, 53, 44, 46, 35, 39

Set B: 35, 31, 39, 41, 37, 43, 29, 40, 36, 38, 42, 33, 34, 30, 37

- **b** Which data set has the largest range?
- **c** Which data set has an interquartile range of 8?

Please refer to the Resources tab in the Prelims section of your **eBookPLUS** for a comprehensive step-by-step guide on how to use your CAS technology.

2.2 Back-to-back stem plots

studyon

Unit 3
AOS DA
Topic 6
Concept 2

Back-to-back stem plots
Concept summary
Practice questions

In topic 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 compares two sets of univariate 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 two sets of univariate data, involving a numerical variable and a categorical variable with 2 categories.

WORKED EXAMPLE 1

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 as shown.

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

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.

WRITE

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

2 Create an unordered stem plot first. Put the boys' scores on the left, and the girls' scores on the right.

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

eBook*plus*

Interactivity
Back-to-back stem plots
int-6252

The back-to-back stem plot allows us to make some visual comparisons of the two distributions. In Worked example 1, 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 topic 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 topic 1 that the calculation of these summary statistics is very straightforward using CAS.

WORKED EXAMPLE 2

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 as shown.

Labor	180	233	246	252	263	270	229	238	226	211
	193	202	210	222	257	247	234	226	214	204
Liberal	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 (Labor)	Stem	Leaf (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 CAS 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
Mean	227.9	257.5
Median	227.5	264.5
IQR	36	29.5
Standard deviation	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 more 'how to vote' cards than the Labor party volunteers did.

EXERCISE 2.2 Back-to-back stem plots

PRACTISE

1 WE1 Boys and girls submitted an assignment on the history of the ANZACs. The results out of 40 are shown.

Girls' results	30	35	31	32	38	33	35	30
Boys' results	34	33	37	39	31	32	39	36

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

2 The marks obtained out of 50 by students in Physics and Chemistry are shown. Display the data on a back-to-back stem plot.

Physics	32	45	48	32	24	30	41	29	44	45	36	34	28	49
Chemistry	46	31	38	28	45	49	34	45	47	33	30	21	32	28

3 WE2 The number of promotional pamphlets handed out for company A and company B by a number of their reps is shown.

Company A	144	156	132	138	148	160	141	134	132	142	132	134	168	149
Company B	146	131	138	155	145	153	134	153	138	133	130	162	148	160

Display the data using a back-to-back stem plot and use this, together with summary statistics, to compare the number of pamphlets handed out by each company.

4 A comparison of student achievements (out of 100) in History and English was recorded and the results shown.

History	75	78	42	92	59	67	78	82	84	64	77	98
English	78	80	57	96	58	71	74	87	79	62	75	100

a Draw a back-to-back stem plot.

b Use summary statistics and the stem plot to comment on the two subjects.

CONSOLIDATE

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

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

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

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

7 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 in the table.

A	11	15	20	25	12	16	21	27	16	17	17	22	23	24
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.

8 The marks out of 20 obtained by males and females for a science test in a Year 10 class are given.

Females	12	13	14	14	15	15	16	17
Males	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.

9 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 as shown.

2011	30	31	35	37	39	41	41	42	43	46
2012	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.

10 The age and gender of a group of people attending a fitness class are recorded.

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.

11 The scores on a board game for a group of kindergarten children and for a group of children in a preparatory school are given as shown.

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.

12 A 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

13 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

14 The score out of 100 a group of males and females received when going for their licence are shown. Construct a back-to-back stem plot of the data.

Male	86	92	100	90	94	82	72	90	88	94	76	80
Female	94	96	72	80	84	92	83	88	90	70	81	83

MASTER

15 A back-to-back stem plot is used to display two sets of data, involving which two variables?

A increasing variables
B discrete and numerical variables
C continuous and categorical variables
D numerical and categorical variables
E numerical and continuous variables

16 The study scores (out of 50) of students who studied both Mathematical Methods and Further Mathematics are shown.

Methods	28	34	41	36	33	39	44	40	39	42	36	31	29	44
Further	30	37	38	41	35	43	44	46	43	48	37	31	28	48

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

b Use the stem plot, together with some summary statistics, to compare the distributions of the scores for Mathematical Methods compared to Further Mathematics.

2.3 Parallel boxplots and dot plots

studyon
Unit 3
AOS DA
Topic 6
Concept 3
Parallel boxplots
Concept summary
Practice questions

eBookplus
Interactivity
Parallel boxplots
int-6248

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, **parallel boxplots** or **parallel dot plots** can be used.

Parallel boxplots are obtained by constructing individual boxplots for each distribution and positioning them on a common scale.

Parallel dot plots are obtained by constructing individual dot plots for each distribution and positioning them on a common scale.

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

WORKED EXAMPLE 3

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 as shown.

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 parallel boxplots. Use this to describe any similarities or differences in the distributions of the marks between the four classes.

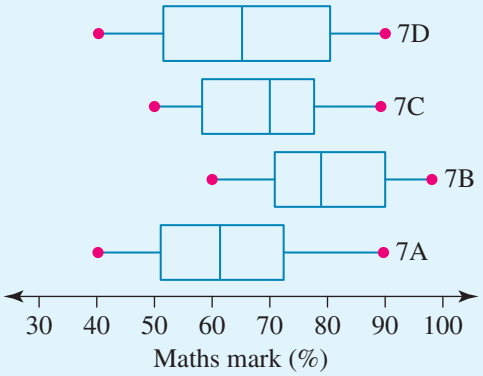
THINK

1 Use CAS 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 2.3 Parallel boxplots and dot plots

PRACTISE

1 WE3 The times run for a 100 m race in grade 6 are shown for both boys and girls. The times are expressed in seconds.

Boys	15.5	16.1	14.5	16.9	18.1	14.3	13.8	15.9	16.4	17.3	18.8	17.9	16.1
Girls	16.7	18.4	19.4	20.1	16.3	14.8	17.3	20.3	19.6	18.4	16.5	17.2	16.0

Display the data using parallel boxplots and use this to describe any similarities or differences between the boys' and girls' performances.

2 A teacher taught two Year 10 maths classes and wanted to see how they compared on the end of year examination. The marks are expressed as percentages.

10A	67	73	45	59	67	89	42	56	68	75	87	94	80	98
10D	76	82	62	58	40	55	69	71	89	95	100	84	70	66

Display the data using parallel boxplots and parallel dot plots. Use this to describe any similarities or differences between the two classes.

CONSOLIDATE

3 The heights (in cm) of students in 9A, 10A and 11A were recorded and are shown in the table.

9A	120	126	131	138	140	143	146	147	150
10A	140	143	146	147	149	151	153	156	162
11A	151	153	154	158	160	163	164	166	167

9A	156	157	158	158	160	162	164	165	170
10A	164	165	167	168	170	173	175	176	180
11A	169	169	172	175	180	187	189	193	199

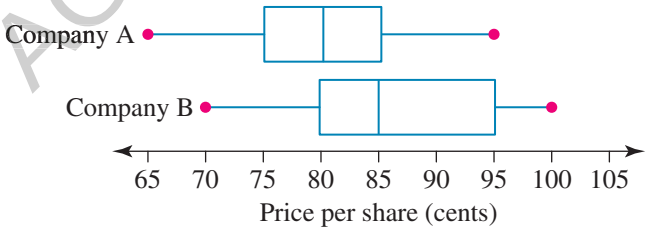
a Construct parallel boxplots to show the data.
b Use the boxplots to compare the distributions of height for the 3 classes.

4 The amounts of money contributed annually to superannuation schemes by people in 3 different age groups are as shown.

20–29	2000	3100	5000	5500	6200	6500	6700	7000	9200	10 000
30–39	4000	5200	6000	6300	6800	7000	8000	9000	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

a Construct parallel boxplots to show the data.
b Use the boxplots to comment on the distributions.

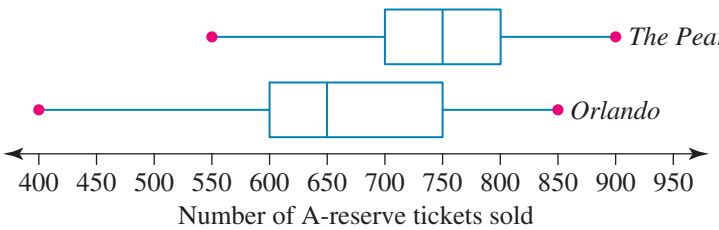
5 The daily share price of two companies was recorded over a period of one month. The results are presented as parallel boxplots.



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

a The distribution of share prices for Company A is symmetrical.
b On 25% of all occasions, share prices for Company B equalled or exceeded the highest price recorded for Company A.
c The spread of the share prices was the same for both companies.
d 75% of share prices for Company B were at least as high as the median share price for Company A.

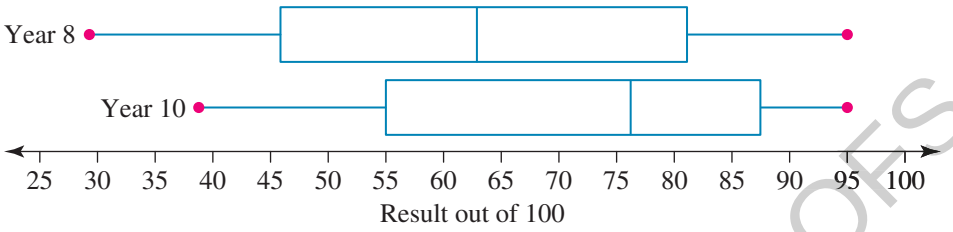
6 Last year, the spring season at the Sydney Opera House included two major productions: *The Pearlfishers* and *Orlando*. The number of A-reserve tickets sold for each performance of the two operas is shown as parallel boxplots.



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 production had a larger variability in the number of patrons purchasing A-reserve tickets? Support your answer with the necessary calculations.

7 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.



The percentage of Year 10 students who obtained a mark greater than 87 was:

A 2% **B** 5% **C** 20% **D** 25% **E** 75%

8 From the parallel boxplots in question **7**, 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

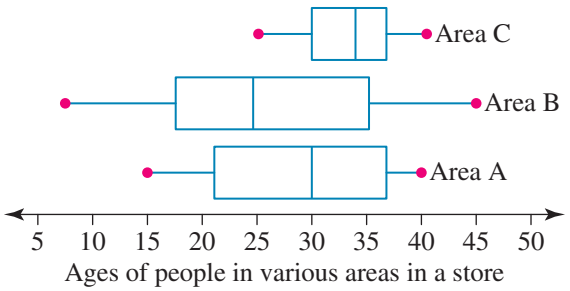
9 The scores of 10 competitors on two consecutive days of a diving competition are recorded below:

Day 1	5.4	4.1	5.4	5.6	4.9
	5.6	5.4	6.0	5.8	6.0
Day 2	4.9	5.1	5.3	5.8	5.7
	5.2	5.8	5.4	5.5	6.0

Construct parallel dot plots to show the data and comment on the divers' results over the two days.

The following figure relates to questions **10** to **12**.

The ages of customers in different areas of a department store are as shown.



10 Which area has the largest range from Q2 (the median) to Q3: Area A, Area B or Area C?

11 Which area has the largest range: Area A, Area B or Area C?

12 Which area has the highest median age: Area A, Area B or Area C?

MASTER

13 The numbers of jars of vitamin A, B, C and multi-vitamins sold per week by a local chemist are shown in the table.

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

14 Eleven golfers in a golf tournament play 18 holes each day. The scores for each of the golfers on the four days are given below. Display this data using parallel boxplots.

Thursday	Friday	Saturday	Sunday
70	77	81	70
71	78	83	81
75	81	84	81
79	82	84	88
80	83	86	88
81	83	86	89
83	85	87	90
83	85	87	90
84	85	87	91
85	88	88	93
90	89	89	94

2.4 Two-way (contingency) frequency tables and segmented bar charts

studyon

Unit 3
AOS DA
Topic 6
Concept 4

Two-way frequency tables and segmented bar charts
Concept summary
Practice questions

eBookplus

Interactivity
Two-way tables and segmented bar graphs
int-6249

When we are examining the relationship between two categorical variables, **two-way (or contingency) tables** are an excellent tool. Consider the following example.

Once the two-way table is formed, **marginal distributions** and **conditional distributions** can both be found. Marginal distributions are the sums (totals) of the row or the column and are found in the margins of the table. The conditional distribution is the sub-population (sample) and this is found in the middle of the table.

If we were to look at mobile phone preference as shown in the table the marginal distributions are the totals, as shown by the green highlighted numbers.

	Apple	Samsung	Nokia	Total
Men	13	9	3	25
Women	17	7	1	25
Total	30	16	4	50

The conditional distribution is the sub-population, so if we are looking at people who prefer Samsung, the conditional distribution is shown by the purple highlighted numbers.

	Apple	Samsung	Nokia	Total
Men	13	9	3	25
Women	17	7	1	25
Total	30	16	4	50

WORKED EXAMPLE 4

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 (contingency) table, and calculate the party preference for males and females.

THINK

1 Draw a table. Record the respondents' sex in the columns and party preference in the rows of the table.

WRITE

Party preference	Female	Male	Total
Labor			
Liberal			
Total			

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.

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

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: $34 - 18 = 16$.

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

4 Marginal distributions for party preference for males and females refers to percentage (probability) of each party. For Labor there are 30 out of a total of 57.

Labor: $\frac{30}{57} = 0.53$

5 For Liberal there are 27 out of a total of 57.

Liberal: $\frac{27}{57} = 0.47$

In Worked example 4, 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 5

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

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 also 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 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.

In the above example, the respondent's gender is referred to as the **explanatory variable**, and the party preference as the **response variable**.

As a general rule, when the explanatory variable is placed in the columns of the table, the percentages should be calculated in columns.

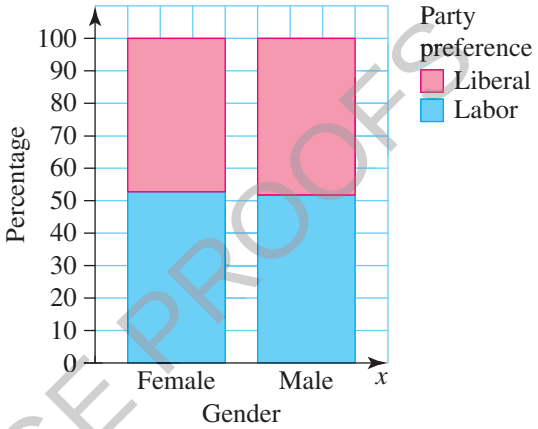
Comparing percentages in each row of a two-way table allows us to establish whether a relationship exists between the two categorical variables that are being examined. As we can see from the table in Worked example 5, the percentage of females who preferred Labor is about the same as that of males. 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 5 can be displayed using the segmented bar chart shown.



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

WORKED EXAMPLE 6

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.

Present these data in percentage form in a two-way 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.

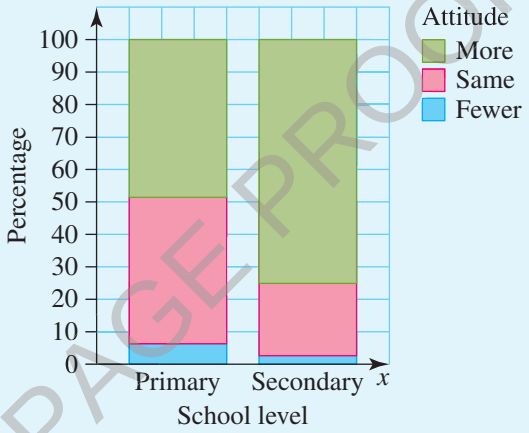
WRITE/DRAW

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

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

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 2.4 Two-way (contingency) frequency tables and segmented bar charts

PRACTISE

1 WE4 A group of 60 people, 38 females and 22 males, were asked whether they prefer an Apple or Samsung phone. Twenty-three females and 15 males said they preferred an Apple phone. Display this data in a two-way (contingency) table and calculate the marginal distribution for phone preference for males and females.

2 A group of 387 females and 263 males were asked their preference from Coke and Pepsi. Two hundred and twenty-one females preferred Coke, whereas 108 males preferred Pepsi. Display this data in a two-way (contingency) table and calculate the conditional distribution of drink preference among females.

3 WE5 A group of 60 people were asked their preferences on phones. The results are shown.

Convert the numbers in this table to percentages.

Phone	Females	Males	Total
Apple	23	15	38
Samsung	15	7	22
Total	38	22	60

4 A group of 650 people were asked their preferences on soft drink. The results are shown.

Convert the numbers in this table to percentages.

Drink	Females	Males	Total
Pepsi	221	155	376
Coke	166	108	274
Total	387	263	650

5 WE6 Sixty-one females and 57 males were asked which they prefer off the menu: entrée, main or dessert. Seven males and 18 females preferred entrée, while 31 males and 16 females said they preferred the main course, with the remainder having dessert as their preferred preference.

Present these data in percentage form in a two-way table and a segmented bar chart. Compare the opinions of the males and females on their preferences.

6 Ninety-three people less than 40 years of age and 102 people aged 40 and over were asked where their priority financially is, given the three options 'mortgage', 'superannuation' or 'investing'. Eighteen people in the 40 and over category and 42 people in the less than 40 years category identified mortgage as their priority, whereas 21 people under 40 years of age and 33 people aged 40 and over said investment was most important. The rest suggested superannuation was their most important priority.

Present these data in percentage form in a two-way table and segmented bar chart. Compare the opinions of the under 40s to the people aged 40 and over on their priority to their finances.

CONSOLIDATE

7 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), and calculate the approval or disapproval of the proposed freeway for men and women.

8 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), and calculate the conditional distribution on length of lessons and senior students.

9 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

10 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 **11** and **12**.

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

11 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

12 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

13 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. Comment on any differences between the reactions of the Liberal and Labor delegates.

14 Use the results in question **13** to draw a segmented bar chart.

The information in the following table relates to questions **15**–**18**.

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

15 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%

16 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

MASTER

17 Calculate the conditional distribution for amount of waste and large towns.

18 Calculate the percentage of small towns rated as having a high level of waste recycling.

eBookplus ONLINE ONLY 2.5 Review

www.jacplus.com.au

The Maths Quest Review is available in a customisable format for you to demonstrate your knowledge of this topic.

The Review contains:

- **Multiple-choice** questions — providing you with the opportunity to practise answering questions using CAS technology
- **Short-answer** questions — providing you with the opportunity to demonstrate the skills you have developed to efficiently answer questions using the most appropriate methods
- **Extended-response** questions — providing you with the opportunity to practise exam-style questions.

A summary of the key points covered in this topic is also available as a digital document.

REVIEW QUESTIONS

Download the Review questions document from the links found in the Resources section of your eBookPLUS.

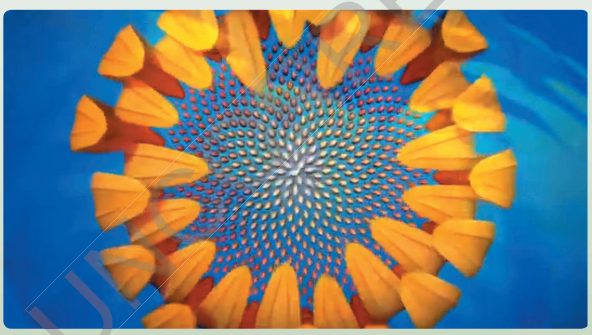
eBookplus ONLINE ONLY Activities

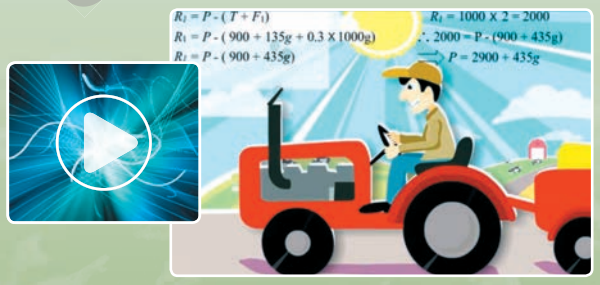
To access eBookPLUS activities, log on to

www.jacplus.com.au

Interactivities

A comprehensive set of relevant interactivities to bring difficult mathematical concepts to life can be found in the Resources section of your eBookPLUS.





studyON

studyON is an interactive and highly visual online tool that helps you to clearly identify strengths and weaknesses prior to your exams. You can then confidently target areas of greatest need, enabling you to achieve your best results.

2 Answers

EXERCISE 2.2

1 Key 3|1 = 31

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

2 Key 2|4 = 24

Physics Leaf	Stem	Chemistry Leaf
4	2	1
9 8	2*	8 8
4 2 2 0	3	0 1 2 3 4
6	3*	8
4 1	4	
9 8 5 5	4*	5 5 6 7 9

3 Key 13|0 = 130

Company A Leaf	Stem	Company B Leaf
4 4 2 2 2	13	0 1 3 4
8	13*	8 8
4 2 1	14	
9 8	14*	5 6 8
	15	3 3
6	15*	5
0	16	0 2
8	16*	

	Company A	Company B
Mean	143.57	144.71
Median	141.5	145.5
IQR	149 − 134 = 15	153 − 134 = 19
Standard deviation	11.42	10.87

They are both positively skewed. The median is a better indicator of the centre of the distribution than the mean. This shows Company B handing out more pamphlets, taking into account that the IQR and the standard deviations are quite similar.

4 a Key 5|7 = 57

History Leaf	Stem	English Leaf
2	4	
9	5	7 8
7 4	6	2
8 8 7 5	7	1 4 5 8 9
4 2	8	0 7
8 2	9	6
	10	0

b

	History	English
Mean	74.67	76.42
Median	77.5	76.5
IQR	83 − 65.5 = 17.5	83.5 − 66.5 = 17
Standard deviation	15.07	13.60

History has a slightly higher median; however, English has a slightly higher mean. Their standard deviations are similar, so overall the results are quite similar.

5 Key: 2|3 = 23

Leaf German	Stem	Leaf 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

6 Key: 2*|7 = 2.7 (kg)

Leaf Boys	Stem	Leaf 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	

7 a Key: 2*|5 = 25 trucks

Leaf A	Stem	Leaf 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.

8 a Key: 1|2 = 12 marks

Leaf Females	Stem	Leaf 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.

9 a Key: 2*|6 = 26 marks

Leaf 2011	Stem	Leaf 2012
	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 2011 and for 2012 are each symmetric.

For the 2011 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 2012 marks, the mean is 29.8, the median is 30.5, the standard deviation is 4.2 and the interquartile 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 2012 distribution lower. The work may have become a lot harder!

10 a Key: 3*|6 = 36 years old

Leaf Female	Stem	Leaf 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.

11 a Key: 5|0 = 50 points

Leaf Kindergarten	Stem	Leaf 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.

12 B

13 C

14 Key: 7|2 = 72

Male Leaf	Stem	Female Leaf
2	7	0 2
6	7*	
2 0	8	0 1 3 3 4
8 6	8*	8
4 4 2 0 0	9*	0 2 4
	9*	6
0	10	

15 D

16 a Key: 3|1 = 31

Mathematical Methods Leaf	Stem	Further Mathematics Leaf
9 8	2*	8
4 3 1	3	0 1
9 9 6 6	3*	5 7 7 8
4 4 2 1 0	4	1 3 3 4
	4*	6 8 8

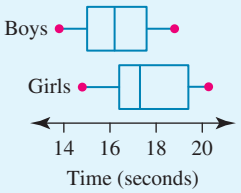
	Mathematical Methods	Further Mathematics
Mean	36.86	39.21
Median	37.5	39.5
IQR	41 − 33 = 8	44 − 35 = 9
Standard deviation	5.29	6.58

b Mathematical Methods has a slightly lower IQR and standard deviation. It was found that Further Mathematics had a greater mean (39.21) as compared to Mathematical Methods (39.5), as well as a greater median; 39.5 as compared to 36.86. This suggests that students do better in Further Mathematics as compared to Mathematical Methods by an average of two study scores.

EXERCISE 2.3

Note: When comparing and contrasting data sets, answers will naturally vary. It is good practice to discuss your conclusions in a group to consider different viewpoints.

1

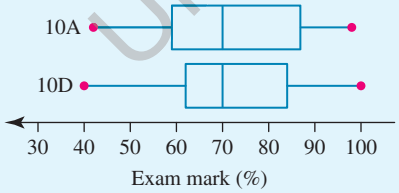


Boys: 13.8, 15, 16.1, 17.6, 18.8

Girls: 14.8, 16.4, 17.3, 19.5, 20.3

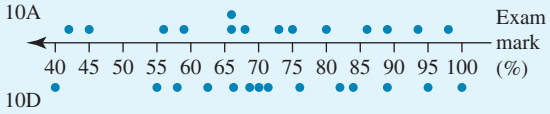
From the boxplots we can see that the boys have a significantly lower median. The boys' median is lower than Q_1 of the girls' time; that is; the lowest 25% of times for the girls is greater than the lowest 50% of times for the boys.

2



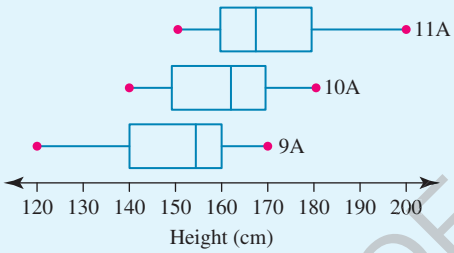
10A: 42, 59, 70.5, 87, 98

10D: 40, 62, 70.5, 84, 100



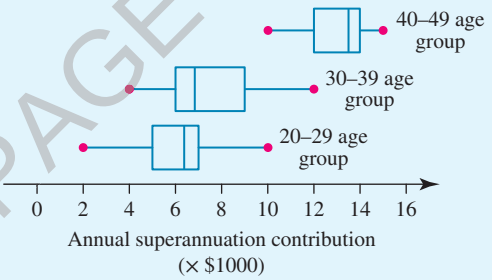
From the boxplots you can see the medians are the same but 10D has a higher mean. 10D also has the highest score of 100%, but 10D also has the lowest score. Since Q_1 and Q_3 are closer together for 10D their results are more consistent around the median. The parallel dot plot confirms this but doesn't give you any further information.

3 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.

4 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.

5 a True **b** True

c False **d** True

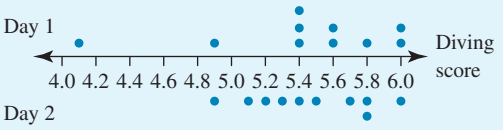
6 a *The Pearlfishers*, which had a significantly higher medium number of A-reserve tickets sold, as well as a higher minimum and maximum number of A-reserve tickets sold.

b *Orlando*, which had both a larger range and IQR of A-reserve tickets sold.

7 D

8 C

9



The dives on day 1 were more consistent than the dives on day 2 with most of the dives between 5.4 and 6.0

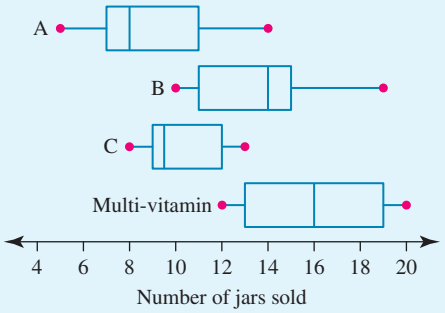
(inclusive), despite two lower dives. Day 2 was more spread with dives from 4.9 to 6.0 (inclusive). It must be noted that there were no very low scoring dives on the second day.

10 B

11 B

12 C

13

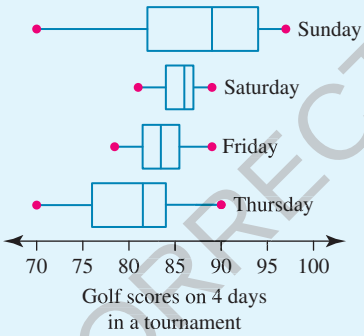


Overall, the biggest sales were of multi-vitamins, followed by vitamin B, then C and finally vitamin A.

14 For all four days, the median is the 6th score.

For all four days, Q_1 is the 3rd score. For all four days, Q_3 is the 9th score.

Day	Min.	Max.	Range	Median
Thursday	70	90	20	81
Friday	77	89	12	83
Saturday	81	89	8	86
Sunday	70	94	24	89



EXERCISE 2.4

1 Note that black data is given in the question; red data are the answers.

Phone	Female	Male	Total
Apple	23	15	38
Samsung	15	7	22
Total	38	22	60

Marginal distribution: Apple = 0.63 Samsung = 0.37

2 Note that black data is given in the question; red data are the answers.

Drink	Female	Male	Total
Coke	221	155	376
Pepsi	166	108	274
Total	387	263	650

Conditional distribution: Females who prefer Coke = 0.57

Females who prefer Pepsi = 0.43

3

Phone	Female	Male
Apple	60.5%	68.2%
Samsung	39.5%	31.8%
Total	100%	100%

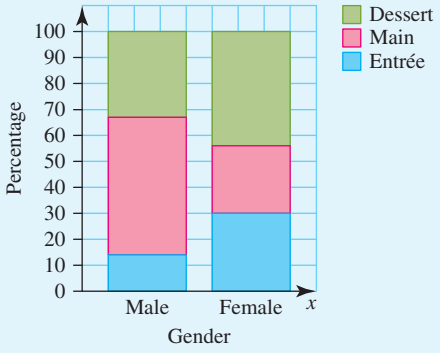
4

Drink	Female	Male
Coke	57.1%	58.9%
Pepsi	42.9%	41.1%
Total	100%	100%

5

Choice	Male	Female	Total
Entrée	8	18	26
Main	31	16	47
Dessert	19	27	46
Total	58	61	118

Choice	Male	Female
Entrée	14	30
Main	53	26
Dessert	33	44
Total	100	100

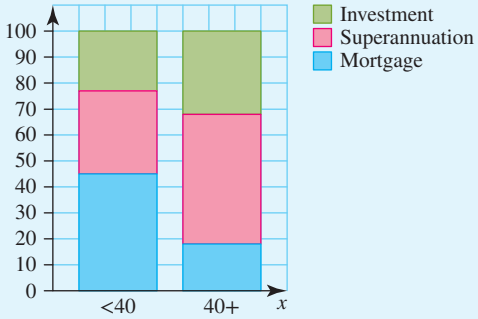


Males enjoy main meal the most compared to females who prefer their dessert the most.

6

Choice	<40	40+	Total
Mortgage	42	18	60
Superannuation	30	51	81
Investment	21	33	54
Total	93	102	195

Choice	<40	40+
Mortgage	45	18
Superannuation	32	50
Investment	23	32
Total	100	100



The under 40s have a focus on their mortgage, whereas the 40 and overs prioritise their superannuation.

7

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

Marginal distribution: For = 0.48, Against = 0.52

8

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

Conditional distribution:

Senior students prefer 45 mins = 0.35,
Senior students prefer an hour = 0.65

9 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%

10

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

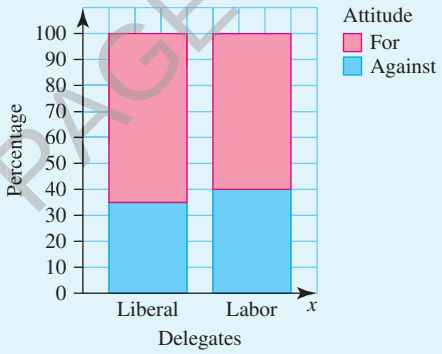
11 D

12 C

13

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

14



There is not a lot of difference in the reactions.

15 C

16 A

17 Conditional distribution:

Large town and no waste = 0.15

Large town and medium waste = 0.19

Large town and high waste = 0.67

Note: rounding causes the total to be greater than 100%.

18 26.32%