CHAPTER 1
Univariate data

eBook*plus*

DIGITAL DOC
doc-9399
10 Quick Questions

CHAPTER CONTENTS

1A Types of data

Univariate data are data that contain one variable. That is, the information deals with only one quantity that changes. Therefore, the number of cars sold by a car salesman during one week is an example of *univariate* data. Sets of data that contain two variables are called *bivariate* data and those that contain more than two variables are called *multivariate* data. You will learn more about bivariate data in chapter 2.

Data can be numerical, categorical, discrete or continuous. The methods we use to display data depend on the type of information we are dealing with.

Numerical and categorical data

Examples of numerical data are:
1. the heights of a group of teenagers
2. the marks for a maths test
3. the number of universities in a country
4. ages
5. salaries.

As the name suggests, numerical data involve quantities which are, broadly speaking, *measurable* or *countable*.

Examples of categorical data are:
1. genders
2. AFL football teams
3. religions
4. finishing positions in the Melbourne Cup
5. municipalities
6. ratings of 1–5 to indicate preferences for 5 different cars
7. age groups, for example 0–9, 10–19, 20–29
8. hair colours.

studyon

Units:	3 & 4
AOS:	DA
Topic:	1
Concept:	1

Concept summary
Read a summary of this concept.

Do more
Interact with classification of data.

Such categorical data, as the name suggests, have categories like masculine, feminine and neuter for gender, or Christian, Islamic, Buddhist and so on for religious denomination, or 1st, 2nd, 3rd for finishing position in the Melbourne Cup.

Note: Some numbers may *look* like numerical data, but are actually *names* or *titles* (for example, ratings of 1 to 5 given to different samples of cake — 'This one's a 4'; the numbers on netball players' uniforms — 'she's number 7'). These 'titles' are *not countable*; they place the subject in a category (with a name), and so they are *categorical*.

Discrete and continuous data

Data are said to be *discrete* when a variable can take only certain fixed values. For example, if we counted the number of children per household in a particular suburb, the data obtained would always be whole numbers starting from zero. A value in between, such as 2.5, would clearly not be possible.

Other examples of variables that produce discrete data are the number of crayfish caught in a fisherman's pots each day and the number of people that attend a restaurant each day. If objects can be counted, then the data are discrete.

Continuous data are obtained when a variable takes any value between two values. If the heights of students in a school were obtained, then the data could consist of any values between the smallest and largest heights. The values recorded would be restricted only by the precision of the measuring instrument.

Other examples of variables that produce continuous data are weight, length and the time to complete a certain task. If variables can be measured, then the data are continuous.

Exercise 1A Types of data

1 Write whether each of the following represents numerical or categorical data.
- **a** The heights, in centimetres, of a group of children
- **b** The diameters, in millimetres, of a collection of ball-bearings
- **c** The numbers of visitors to an exibit each day
- **d** The modes of transport that students in Year 12 take to school
- **e** The 10 most-watched television programs in a week
- **f** The occupations of a group of 30-year-olds
- **g** The numbers of subjects offered to VCE students at various schools
- **h** Life expectancies
- **i** Species of fish
- **j** Blood groups
- **k** Years of birth
- **l** Countries of birth
- **m** Tax brackets

2 For each set of numerical data identified in question **1** above, state whether the data are discrete or continuous.

3 MC An example of a numerical variable is:
- **A** attitude to 4-yearly elections (for or against)
- **B** year level of students
- **C** the total attendance at Carlton football matches
- **D** position in a queue at the pie stall
- **E** television channel numbers shown on a dial

4 MC The weight of each truck-load of woodchips delivered to the wharf during a one-month period was recorded. This is an example of:
- **A** categorical and discrete data
- **B** discrete data
- **C** continuous and numerical data
- **D** continuous and categorical data
- **E** numerical and discrete data

eBook*plus*

DIGITAL DOC
doc-9400
WorkSHEET 1.1

1B Stem plots

A *stem-and-leaf plot*, or stem plot for short, is a way of displaying a set of data. It is best suited to data which contain up to about 50 observations (or records).

The stem plot below shows the ages of people attending an advanced computer class.

The ages of the members of the class are 16, 22, 22, 23, 30, 32, 34, 36, 42, 43, 46, 47, 53, 57 and 61.

A stem plot is constructed by splitting the numerals of a record into two parts — the stem, which in this case is the first digit, and the leaf, which is always the last digit.

Stem	Leaf
1	6
2	2 2 3
3	0 2 4 6
4	2 3 6 7
5	3 7
6	1

Key: 2|2 = 22 years old

WORKED EXAMPLE 1

The number of cars sold in a week at a large car dealership over a 20-week period is given below.

16	**12**	**8**	**7**	**26**	**32**	**15**	**51**	**29**	**45**
19	**11**	**6**	**15**	**32**	**18**	**43**	**31**	**23**	**23**

Construct a stem plot to display the number of cars sold in a week at the dealership.

THINK

1 In this example the observations are one- or two-digit numbers and so the stems will be the digits referring to the 'tens', and the leaves will be the digits referring to the units.

Work out the lowest and highest numbers in the data in order to determine what the stems will be.

WRITE

Lowest number = 6
Highest number = 51
Use stems from 0 to 5.

2 Before we construct an ordered stem plot, construct an unordered stem plot by listing the leaf digits in the order they appear in the data.

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

3 Now rearrange the leaf digits in numerical order to create an ordered stem plot.

Include a key so that the data can be understood by anyone viewing the stem plot.

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

Key: 2|3 = 23 cars

studyon

Units:	3 & 4
AOS:	DA
Topic:	2
Concept:	4

Concept summary
Read a summary of this concept.

WORKED EXAMPLE 2

The masses (in kilograms) of the members of an Under-17 football squad are given below.

70.3	**65.1**	**72.9**	**66.9**	**68.6**	**69.6**	**70.8**
72.4	**74.1**	**75.3**	**75.6**	**69.7**	**66.2**	**71.2**
68.3	**69.7**	**71.3**	**68.3**	**70.5**	**72.4**	**71.8**

Display the data in a stem plot.

THINK

1 In this case the observations contain 3 digits. The last digit always becomes the leaf and so in this case the digit referring to the tenths becomes the leaf and the two preceding digits become the stem.

Work out the lowest and highest numbers in the data in order to determine what the stems will be.

WRITE

Lowest number = 65.1
Highest number = 75.6
Use stems from 65 to 75.

2 Construct an unordered stem plot. Note that the decimal points are omitted since we are aiming to present a quick visual summary of data.

Stem	Leaf
65	1
66	9 2
67	
68	6 3 3
69	6 7 7
70	3 8 5
71	2 3 8
72	9 4 4
73	
74	1
75	3 6

3 Construct an ordered stem plot. Provide a key.

Stem	Leaf
65	1
66	2 9
67	
68	3 3 6
69	6 7 7
70	3 5 8
71	2 3 8
72	4 4 9
73	
74	1
75	3 6

Key: 74|1 = 74.1 kg

Sometimes data which are very bunched make it difficult to get a clear idea about the data variation. To overcome the problem, we can split the stems. Stems can be split into halves or fifths.

WORKED EXAMPLE 3

A set of golf scores for a group of professional golfers trialling a new 18-hole golf course is shown on the following stem plot.

Stem	Leaf
6	**1 6 6 7 8 9 9 9**
7	**0 1 1 2 2 3 7**

Key: 6|1 = 61

Produce another stem plot for these data by splitting the stems into:
a halves **b fifths.**

THINK

a By splitting the stem 6 into halves, any leaf digits in the range 0–4 appear next to the 6, and any leaf digits in the range 5–9 appear next to the 6*. Likewise for the stem 7.

WRITE

a

Stem	Leaf
6	1
6*	6 6 7 8 9 9 9
7	0 1 1 2 2 3
7*	7

Key: 6|1 = 61

b Alternatively, to split the stems into fifths, each stem would appear 5 times. Any 0s or 1s are recorded next to the first 6. Any 2s or 3s are recorded next to the second 6. Any 4s or 5s are recorded next to the third 6. Any 6s or 7s are recorded next to the fourth 6 and, finally, any 8s or 9s are recorded next to the fifth 6.

This process would be repeated for those observations with a stem of 7.

b

Stem	Leaf
6	1
6	
6	
6	6 6 7
6	8 9 9 9
7	0 1 1
7	2 2 3
7	
7	7
7	

Key: 6|1 = 61

Exercise 1B Stem plots

1 In each of the following, write down all the pieces of data shown on the stem plot.

a

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

b

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

c

Stem	Leaf
10	1 2
11	5 8
12	2 3 3
13	6 6 7
14	1 3 4
15	5 5 6 7

d

Stem	Leaf
5	0 1
5	3 3
5	4 5 5
5	6 6 7
5	9

e

Stem	Leaf
0	1 4
0*	5 8
1	0 2
1*	6 9 9
2	1 1
2*	5 9

The key used for each stem plot is 3|2 = 32.

2 **WE1** The money (to the nearest dollar) earned each week by a busker over an 18-week period is shown below. Construct a stem plot for the busker's weekly earnings. What can you say about the busker's earnings?

5	19	11	27	23	35	18	42	29
31	52	43	37	41	39	45	32	36

3 The ages of those attending an embroidery class are given below. Construct a stem plot for these data and draw a conclusion from it.

39	68	51	57	63	51	37	42
63	49	52	61	58	59	49	53

4 **MC** The observations shown on the stem plot at right are:

A 4 10 27 28 29 31 34 36 41
B 14 10 27 28 29 29 31 34 36 41 41
C 4 22 27 28 29 29 30 31 34 36 41 41
D 14 22 27 28 29 30 30 31 34 36 41 41
E 4 2 27 28 29 29 30 31 34 36 41

Stem	Leaf
0	4
1	
2	2 7 8 9 9
3	0 1 4 6
4	1 1

Key: 2|5 = 25

5 The ages of the mothers of a class of children attending an inner-city kindergarten are given below. Construct a stem plot for these data. Based on your display, comment on the statement 'Parents of kindergarten children are very young'.

32	30	19	28	25	29	32
28	29	34	32	35	39	30
37	33	29	35	38	33	

6 The number of hit outs made by each of the principal ruckmen in each of the AFL teams for Round 11 is recorded below. Construct a stem plot to display these data. Which teams had the three highest scoring ruckmen?

Team	Number of hit outs
Collingwood	19
Bulldogs	41
Kangaroos	29
Port Adelaide	24
Geelong	21
Sydney	31
Melbourne	40
Brisbane	25

Team	Number of hit outs
Adelaide	32
St Kilda	34
Essendon	31
Carlton	26
West Coast	29
Fremantle	22
Hawthorn	33
Richmond	28

7 **WE2** The heights of members of a squad of basketballers are given below in metres. Construct a stem plot for these data.

1.96	1.85	2.03	2.21	2.17	1.89	1.99	1.87
1.95	2.03	2.09	2.05	2.01	1.96	1.97	1.91

8 The 2008 median house price of a number of Melbourne suburbs is given below. Construct a stem plot for these data and comment on it.

Suburb	Price (× $1000)
Ashburton	670
Ashwood	600
Blackburn	670
Bulleen	628
Burwood	652
Caulfield East	653
Chadstone	608
Chettenham	576
Clayton	525
Cobury	526

Suburb	Price (× $1000)
Collingwood	583
Dancaster	620
Essendon	670
Highett	600
Huntingdale	517
Ivanhoe	633
Moonee Ponds	638
Newport	536
Oakleigh	548
Preston	515

9 **WE3** The data below give the head circumference (to the nearest cm) of 16 four-year-old girls.

48	49	47	52	51	50	49	48
50	50	53	52	43	47	49	50

Construct a stem plot for head circumference, using:

a the stems 4 and 5
b the stems 4 and 5 split into halves
c the stems 4 and 5 split into fifths.

10 A random sample of 20 screws is taken and the length of each is recorded to the nearest millimetre below.

23	15	18	17	17	19	22
19	20	16	20	21	19	23
17	19	21	23	20	21	

Construct a stem plot for screw length using:

a the stems 1 and 2
b the stems 1 and 2 split into halves
c the stems 1 and 2 split into fifths.

Use your plots to help you comment on the screw lengths.

1C Dot plots, frequency histograms and bar charts

Dot plots, frequency histograms and bar charts display data in graphical form.

studyon

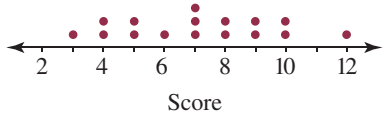
Units:	3 & 4
AOS:	DA
Topic:	2
Concept:	3

Concept summary
Read a summary of this concept.

Dot plots

In picture graphs, a single picture represents each data value. Similarly, in dot plots, a single dot represents each data value. Dot plots are used to display discrete data where values are not spread out very much. They are also used to display categorical data.

Dot plots have a scaled horizontal axis and each data value is indicated by a dot above this scale. The end result is a set of vertical 'lines' of evenly-spaced dots.



WORKED EXAMPLE 4

The number of hours per week spent on art by 18 students is given below.

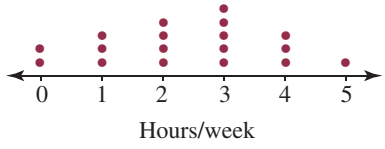
4	**0**	**3**	**1**	**3**	**4**	**2**	**2**	**3**
4	**1**	**3**	**2**	**5**	**3**	**2**	**1**	**0**

Display the data as a dot plot.

THINK

1. Determine the lowest and highest scores and then draw a suitable scale.
2. Represent each score by a dot on the scale.

DRAW



studyon

Units:	3 & 4
AOS:	DA
Topic:	2
Concept:	5

Concept summary
Read a summary of this concept.

Frequency histograms

A histogram is a useful way of displaying large data sets (say, over 50 observations). The vertical axis on the histogram displays the frequency and the horizontal axis displays class intervals of the variable (for example, height or income).

When data are given in raw form — that is, just as a list of figures in no particular order — it is helpful to first construct a frequency table.

WORKED EXAMPLE 5

The data below show the distribution of masses (in kilograms) of 60 students in Year 7 at Northwood Secondary College. Construct a frequency histogram to display the data more clearly.

45.7	**45.8**	**45.9**	**48.2**	**48.3**	**48.4**	**34.2**	**52.4**	**52.3**	**51.8**	**45.7**	**56.8**	**56.3**	**60.2**	**44.2**
53.8	**43.5**	**57.2**	**38.7**	**48.5**	**49.6**	**56.9**	**43.8**	**58.3**	**52.4**	**54.3**	**48.6**	**53.7**	**58.7**	**57.6**
45.7	**39.8**	**42.5**	**42.9**	**59.2**	**53.2**	**48.2**	**36.2**	**47.2**	**46.7**	**58.7**	**53.1**	**52.1**	**54.3**	**51.3**
51.9	**54.6**	**58.7**	**58.7**	**39.7**	**43.1**	**56.2**	**43.0**	**56.3**	**62.3**	**46.3**	**52.4**	**61.2**	**48.2**	**58.3**

THINK

1. First construct a frequency table. The lowest data value is 34.2 and the highest is 62.3. Divide the data into class intervals. If we started the first class interval at, say, 30 kg and ended the last class interval at 65 kg, we would have a range of 35. If each interval was 5 kg, we would then have 7 intervals which is a reasonable number of class intervals.

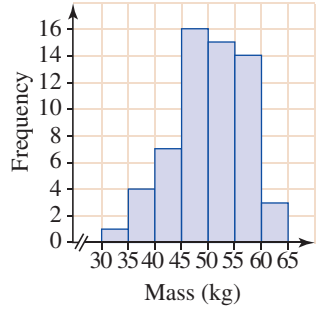
   While there are no set rules about how many intervals there should be, somewhere between about 5 and 15 class intervals is usual. So, in this example, we would have class intervals of 30–34.9 kg, 35–39.9 kg, 40–44.9 kg and so on. Complete a tally column using one mark for each value in the appropriate interval. Add up the tally marks and write them in the frequency column.

2. Check that the frequency column totals 60. The data are in a much clearer form now.

3. A histogram can be constructed.

WRITE/DRAW

Class interval	Tally	Frequency
30–34.9	I	1
35–39.9	IIII	4
40–44.9	~~IIII~~ II	7
45–49.9	~~IIII~~ ~~IIII~~ ~~IIII~~ I	16
50–54.9	~~IIII~~ ~~IIII~~ ~~IIII~~	15
55–59.9	~~IIII~~ ~~IIII~~ IIII	14
60–64.9	III	3
	Total	**60**



WORKED EXAMPLE 6

The marks out of 20 received by 30 students for a book-review assignment are given in the frequency table below.

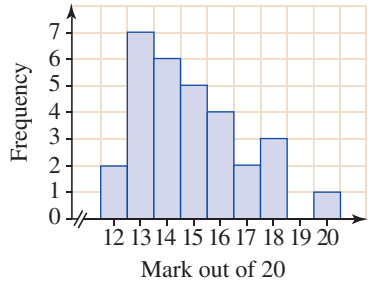
Mark	12	13	14	15	16	17	18	19	20
Frequency	2	7	6	5	4	2	3	0	1

Display these data on a histogram.

THINK

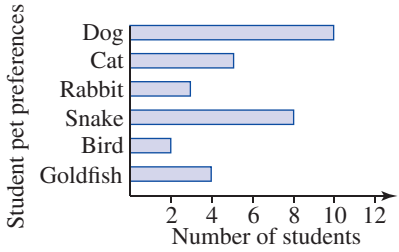
In this case we are dealing with integer values (discrete data). Since the horizontal axis should show a class interval, we extend the base of each of the columns on the histogram halfway either side of each score.

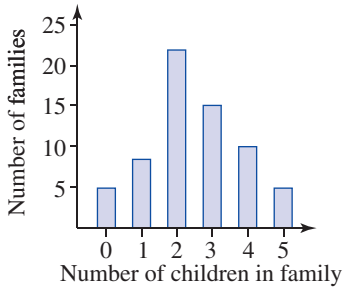
DRAW



Bar charts

A *bar chart* is similar to a histogram. However, it consists of bars of equal width *separated* by small, equal spaces and may be arranged either *horizontally* or *vertically*. Bar charts are often used to display categorical data.

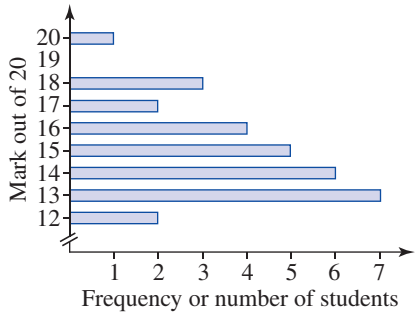




In bar charts, the frequency is graphed against a variable as shown in both figures above.

The variable may or may not be numerical. However, if it is, the variable should represent discrete data because the scale is broken by the gaps between the bars. The numerical values are generally close together and have little spread, for example, consecutive years.

The bar chart at right represents the data presented in Worked example 6. It could also have been drawn with vertical bars (columns).



Segmented bar charts

A segmented (divided) bar chart is a single bar which is used to represent all the data being studied. It is divided into segments, each segment representing a particular group of the data. Generally, the information is presented as percentages and so the total bar length represents 100% of the data.

Consider the following table, showing fatal road accidents in Australia.

Road traffic accidents involving fatalities									
Accidents involving fatalities									
Year	NSW	Vic.	Qld	SA	WA	Tas.	NT	ACT	**Aust.**
2003	483	294	284	136	155	39	44	10	**1445**
2004	471	313	288	128	162	52	34	10	**1458**
2005	469	316	294	127	151	48	51	25	**1481**
2006	453	309	314	104	183	42	39	12	**1456**
2007	405	289	338	107	214	39	49	14	**1453**
2008	376	278	293	87	189	38	67	14	**1342**
Persons killed									
Year	NSW	Vic.	Qld	SA	WA	Tas.	NT	ACT	**Aust.**
2003	539	330	310	157	180	41	53	11	**1621**
2004	522	343	311	139	178	58	35	10	**1596**
2005	518	348	328	148	163	50	55	26	**1636**
2006	500	337	336	117	202	54	42	13	**1601**
2007	435	332	360	124	235	45	58	14	**1603**
2008	397	303	327	99	209	40	75	14	**1464**

Source: Australian Bureau of Statistics 2010, *Year book Australia 2009–10*, cat. no. 1301.0, ABS, Canberra, table 24.20, p. 638.

It is appropriate to represent the number of accidents involving fatalities in all states and territories during 2008 as a segmented bar chart.

studyon

Units:	3 & 4
AOS:	DA
Topic:	2
Concept:	1

Concept summary
Read a summary of this concept.

Do more
Interact with bar charts.

eBook*plus*

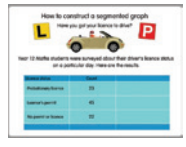
DIGITAL DOC
doc-9401
Spreadsheet
Segmented bar charts

studyon

Units:	3 & 4
AOS:	DA
Topic:	2
Concept:	2

Concept summary
Read a summary of this concept.

See more
Watch a video about segmented graphs.



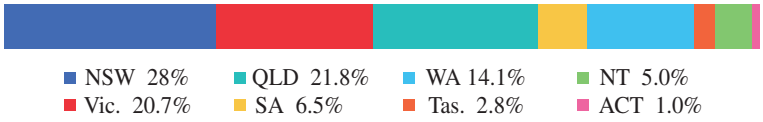
eBook*plus*

DIGITAL DOC
doc-9402
SkillSHEET 1.1
Converting a fraction to a percentage

First, we convert each state's proportion of accidents out of the total to a percentage.

State	Number of accidents	Percentage
NSW	376	$376 \div 1342 \times 100\% = 28.0\%$
Vic.	278	$278 \div 1342 \times 100\% = 20.7\%$
Qld	293	$293 \div 1342 \times 100\% = 21.8\%$
SA	87	$87 \div 1342 \times 100\% = 6.5\%$
WA	189	$189 \div 1342 \times 100\% = 14.1\%$
Tas.	38	$38 \div 1342 \times 100\% = 2.8\%$
NT	67	$67 \div 1342 \times 100\% = 5.0\%$
ACT	14	$14 \div 1342 \times 100\% = 1.0\%$

The segmented bar chart is drawn to scale. An appropriate scale would be constructed by drawing the total bar 100 mm long, so that 1 mm represents 1%. That is, accidents in NSW would be represented by a segment of 28 mm, those in Victoria by a segment of 20.7 mm and so on. Each segment is then labelled directly, or a key may be used.



Exercise 1C Dot plots, frequency histograms and bar charts

eBook*plus*

DIGITAL DOC
doc-9403
Spreadsheet
Frequency histograms

1 **WE5** Construct a frequency table for each of the following sets of data.
- **a** 4.3 4.5 4.7 4.9 5.1 5.3 5.5 5.6 5.2 3.6 2.5 4.3 2.5 3.7 4.5 6.3 1.3
- **b** 11 13 15 15 16 18 20 21 22 21 18 19 20 16 18 20 16 10 23 24 25 27 28 30 35 28 27 26 29 30 31 24 28 29 20 30 32 33 29 30 31 33 34
- **c** 0.4 0.5 0.7 0.8 0.8 0.9 1.0 1.1 1.2 1.0 1.3 0.4 0.3 0.9 0.6

2 Using the frequency tables from question **1**, construct a histogram for each set of data.

3 Using a CAS calculator, construct a histogram for each of the sets of data given in question **1**. Compare this histogram with the one drawn for question **2**.

4 **WE4** The data below represent the number of hours each week that 40 teenagers spent on household chores. Display these data on a bar chart and a dot plot.

2 5 2 0 8 7 8 5 1 0 2 1 8 0 4 2 2 9 8 5
7 5 4 2 1 2 9 8 1 2 8 5 8 10 0 3 4 5 2 8

5 Using the information provided in the table below:
- **a** calculate the proportion of residents who travelled in 2005 to each of the countries listed
- **b** draw a segmented bar graph showing the major destinations of Australians travelling abroad in 2005.

Short-term resident departures by major destinations					
	2004 (× 1000)	**2005 (× 1000)**	**2006 (× 1000)**	**2007 (× 1000)**	**2008 (× 1000)**
New Zealand	815.8	835.4	864.7	902.1	921.1
United States of America	376.1	426.3	440.3	479.1	492.3
United Kingdom	375.1	404.2	412.8	428.5	420.3
Indonesia	335.1	319.7	194.9	282.6	380.7
China (excluding Special Administrative Regions (SARs))	182.0	235.1	251.0	284.3	277.3

	2004 (× 1000)	2005 (× 1000)	2006 (× 1000)	2007 (× 1000)	2008 (× 1000)
Thailand	188.2	202.7	288.0	374.4	404.1
Fiji	175.4	196.9	202.4	200.3	236.2
Singapore	159.0	188.5	210.9	221.5	217.8
Hong Kong (SAR of China)	152.6	185.7	196.3	206.5	213.1
Malaysia	144.4	159.8	168.0	181.3	191.0

Source: Australian Bureau of Statistics 2010, *Year book Australia 2009–10*, cat. no. 1301.0, ABS, Canberra, table 23.12, p. 621.

6 Presented below is information about adult participation in sport and physical activities in 2005–06. Draw a segmented bar graph to compare the participation of all persons from various age groups. Comment on the statement, 'Only young people participate in sport and physical activities'.

Participation in sport and physical activities[(a)] — 2005–06

	Males		Females		Persons	
Age group (years)	**Number (× 1000)**	**Participation rate (%)**	**Number (× 1000)**	**Participation rate (%)**	**Number (× 1000)**	**Participation rate (%)**
18–24	735.2	73.3	671.3	71.8	1406.4	72.6
25–34	1054.5	76.3	1033.9	74.0	2088.3	75.1
35–44	975.4	66.7	1035.9	69.1	2011.2	68.0
45–54	871.8	63.5	923.4	65.7	1795.2	64.6
55–64	670.1	60.4	716.3	64.6	1386.5	62.5
65 and over	591.0	50.8	652.9	48.2	1243.9	49.4
Total	**4898**	**64.6**	**5033.7**	**64.4**	**9931.5**	**64.5**

[(a)] Relates to persons aged 18 years and over who participated in sport or physical activity as a player during the 12 months prior to interview.

Source: Participation in Sport and Physical Activities, Australia, 2005–06 (4177.0). Viewed 10 October 2008 <http://abs.gov.au/Ausstats>

1D Describing the shape of stem plots and histograms

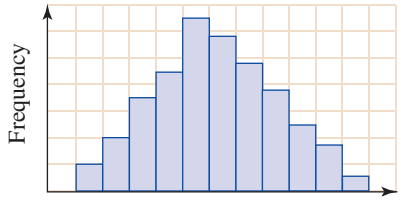
Symmetric distributions

The data shown in the histogram at right can be described as *symmetric*.

There is a single peak and the data trail off on both sides of this peak in roughly the same fashion.

Similarly in the stem plot at right, the distribution of the data could be described as symmetric.

The single peak for these data occur at the stem 3. On either side of the peak, the number of observations reduces in approximately matching fashion.



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

studyon

Units:	3 & 4
AOS:	DA
Topic:	2
Concept:	6

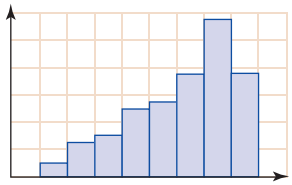
Concept summary
Read a summary of this concept.

Skewed distributions

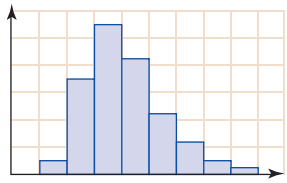
Each of the histograms shown below are examples of skewed distributions.

The figure below left shows data which are *negatively skewed*. The data in this case peak to the right and trail off to the left.

The figure below right shows *positively skewed* data. The data in this case peak to the left and trail off to the right.



Negatively skewed distribution



Positively skewed distribution

WORKED EXAMPLE 7

eBook*plus*

TUTORIAL
eles-1254
Worked example 7

The ages of a group of people who were taking out their first home loan is shown below.

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

Key: 1|9 = 19 years old

Describe the shape of the distribution of these data.

THINK

Check whether the distribution is symmetric or skewed. The peak of the data occurs at the stem 2. The data trail off as the stems increase in value. This seems reasonable since most people would take out a home loan early in life to give themselves time to pay it off.

WRITE

The data are positively skewed.

Exercise 1D Describing the shape of stem plots and histograms

1 **WE7** For each of the following stem plots, describe the shape of the distribution of the data.

a

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

Key: 1|2 = 12

b

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

Key: 2|6 = 2.6

c

Stem	Leaf
2	3 5 5 6 7 8 9 9
3	0 2 2 3 4 6 6 7 8 8
4	2 2 4 5 6 6 6 7 9
5	0 3 3 5 6
6	2 4
7	5 9
8	2
9	7
10	

Key: 10|4 = 104

d

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

Key: 2|4 = 24

e

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

Key: 4|3 = 0.43

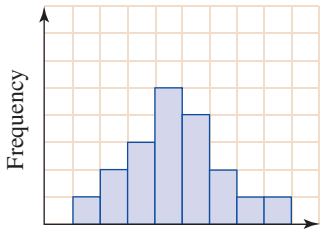
f

Stem	Leaf
60	2 5 8
61	1 3 3 6 7 8 9
62	0 1 2 4 6 7 8 8 9
63	2 2 4 5 7 8
64	3 6 7
65	4 5 8
66	3 5
67	4

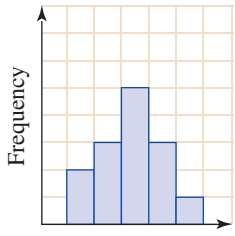
Key: 62|3 = 623

2 For each of the following histograms, describe the shape of the distribution of the data and comment on the existence of any outliers.

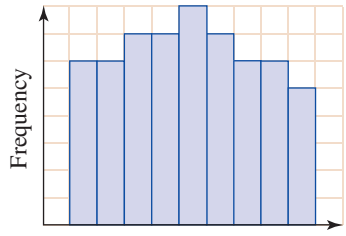
a



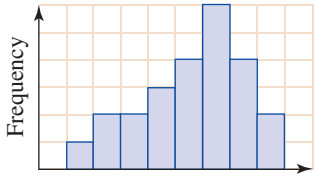
b



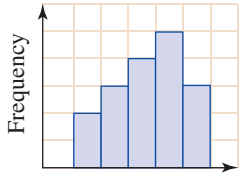
c



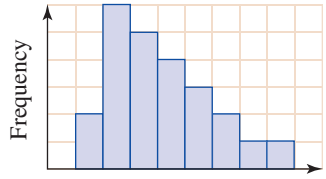
d



e



f



3 **MC** The distribution of the data shown in this stem plot could be described as:

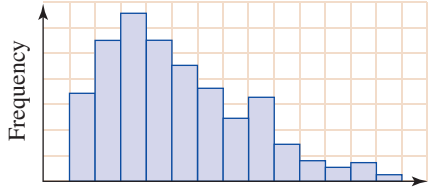
A negatively skewed
B negatively skewed and symmetric
C positively skewed
D positively skewed and symmetric
E symmetric

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

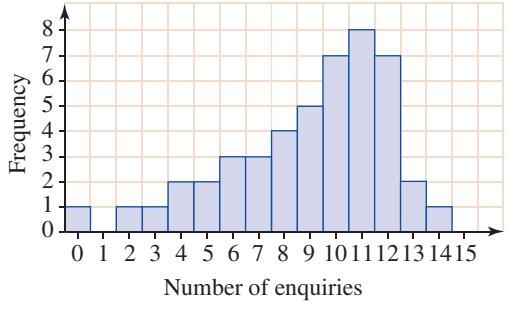
Key: 1|8 = 18

4 **MC** The distribution of the data shown in the histogram at right could be described as:

A negatively skewed
B negatively skewed and symmetric
C positively skewed
D positively skewed and symmetric
E symmetric



5 The average number of product enquiries per day received by a group of small businesses who advertised in the Yellow Pages telephone directory is given at right. Describe the shape of the distribution of these data.



6 The number of nights per month spent interstate by a group of flight attendants is shown on the stem plot at right. Describe the shape of distribution of these data and explain what this tells us about the number of nights per month spent interstate by this group of flight attendants.

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

Key: 1|4 = 14 nights

7 The mass (to the nearest kilogram) of each dog at a dog obedience school is shown on the stem plot below.

a Describe the shape of the distribution of these data.

b What does this information tell us about this group of dogs?

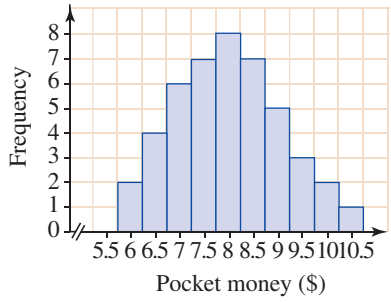
Stem	Leaf
0	4
0*	5 7 9
1	1 2 4 4
1*	5 6 6 7 8 9
2	1 2 2 3
2*	6 7

Key: |0 = 4 kg

8 The amount of pocket money (to the nearest 50 cents) received each week by students in a Grade 6 class is illustrated in the histogram below.

a Describe the shape of the distribution of these data.

b What conclusions can you reach about the amount of pocket money received weekly by this group of students?



9 Statistics were collected over 3 AFL games on the number of goals kicked by forwards over 3 weeks. This is displayed in the histogram below.

a Describe the shape of the histogram.

b Use the histogram to determine:

i the number of players who kicked 3 or more goals over the 3 weeks

ii the percentage of players who kicked between 2 and 6 goals over the 3 weeks.

eBook*plus*

DIGITAL DOC
doc-9404
WorkSHEET 1.2

1E The median, the interquartile range, the range and the mode

After displaying data using a histogram or stem plot, we can make even more sense of the data by calculating what are called *summary statistics*. Summary statistics are used because they give us an idea about:

1. where the centre of the distribution is
2. how the distribution is spread out.

We will look first at four summary statistics — the median, the interquartile range, the range and the mode — which require that the data be in ordered form before they can be calculated.

eBook*plus*

INTERACTIVITY
int-0084
The median, the interquartile range, the range and the mode

study on

Units:	3 & 4
AOS:	DA
Topic:	3
Concept:	1

Concept summary
Read a summary of this concept.

The median

The *median* is the midpoint of an ordered set of data. Half the data are less than or equal to the median.

Consider the set of data: 2 5 6 8 11 12 15. These data are in ordered form (that is, from lowest to highest). There are 7 observations. The median in this case is the middle or fourth score; that is, 8.

Consider the set of data: 1 3 5 6 7 8 8 9 10 12. These data are in ordered form also; however, in this case there is an even number of scores. The median of this set lies halfway between the 5th score (7) and the 6th score (8). So the median is 7.5. (Alternatively, median $= \frac{7+8}{2} = 7.5$.)

When there are n records in a set of ordered data, the median can be located at the $\left(\frac{n+1}{2}\right)$th position.

Checking this against our previous example, we have $n = 10$; that is, there were 10 observations in the set. The median was located at the $\left(\frac{10+1}{2}\right) = 5.5$th position; that is, halfway between the 5th and the 6th terms.

A stem plot provides a quick way of locating a median since the data in a stem plot are already ordered.

WORKED EXAMPLE 8

Consider the stem plot below which contains 22 observations. What is the median?

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

Key: 3|4 = 34

THINK	WRITE
1 Find the median position, where $n = 22$.	Median $= \left(\frac{n+1}{2}\right)$ th position $= \left(\frac{22+1}{2}\right)$ th position $= 11.5$th position
2 Find the 11th and 12th terms.	11th term = 35 12th term = 38
3 The median is halfway between the 11th and 12th terms.	Median = 36.5

study on

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

Concept summary
Read a summary of this concept.

The interquartile range

We have seen that the median divides a set of data in half. Similarly, quartiles divide a set of data in quarters. The symbols used to refer to these quartiles are Q_1, Q_2 and Q_3.

The middle quartile, Q_2, is the median.

The interquartile range IQR = $Q_3 - Q_1$.

The interquartile range gives us the range of the middle 50% of values in a set of data.

There are four steps to locating Q_1 and Q_3.

Step 1. Write down the data in ordered form from lowest to highest.
Step 2. Locate the median; that is, locate Q_2.
Step 3. Now consider just the lower half of the set of data. Find the middle score. This score is Q_1.
Step 4. Now consider just the upper half of the set of data. Find the middle score. This score is Q_3.

The four cases given below illustrate this method.

Case 1

Consider data containing the 6 observations: 3 6 10 12 15 21.
The data are already ordered. The median is 11.
Consider the lower half of the set, which is 3 6 10. The middle score is 6, so $Q_1 = 6$.
Consider the upper half of the set, which is 12 15 21. The middle score is 15, so $Q_3 = 15$.

Case 2

Consider a set of data containing the 7 observations: 4 9 11 13 17 23 30.
The data are already ordered. The median is 13.
Consider the lower half of the set, which is 4 9 11. The middle score is 9, so $Q_1 = 9$.
Consider the upper half of the set, which is 17 23 30. The middle score is 23, so $Q_3 = 23$.

Case 3

Consider a set of data containing the 8 observations: 1 3 9 10 15 17 21 26.
The data are already ordered. The median is 12.5.
Consider the lower half of the set, which is 1 3 9 10. The middle score is 6, so $Q_1 = 6$.
Consider the upper half of the set, which is 15 17 21 26. The middle score is 19, so $Q_3 = 19$.

Case 4

Consider a set of data containing the 9 observations: 2 7 13 14 17 19 21 25 29.
The data are already ordered. The median is 17.
Consider the lower half of the set, which is 2 7 13 14. The middle score is 10, so $Q_1 = 10$.
Consider the upper half of the set, which is 19 21 25 29. The middle score is 23, so $Q_3 = 23$.

WORKED EXAMPLE 9

eBook*plus*
TUTORIAL
eles-1255
Worked example 9

The ages of the patients who attended the casualty department of an inner-suburban hospital on one particular afternoon are shown below.

14	3	27	42	19	17	73
60	62	21	23	2	5	58
33	19	81	59	25	17	69

Find the interquartile range of these data.

THINK	WRITE
1 Order the data.	2 3 5 14 17 17 19 19 21 23 25 27 33 42 58 59 60 62 69 73 81
2 Find the median.	The median is 25 since ten scores lie below it and ten lie above it.

3 Find the middle score of the lower half of the data.	For the scores 2 3 5 14 17 17 19 19 21 23, the middle score is 17. So, $Q_1 = 17$.
4 Find the middle score of the upper half of the data.	For the scores 27 33 42 58 59 60 62 69 73 81, the middle score is 59.5. So, $Q_3 = 59.5$.
5 Calculate the interquartile range.	$\text{IQR} = Q_3 - Q_1$ $= 59.5 - 17$ $= 42.5$

A CAS or graphics calculator can be a fast way of locating quartiles and hence finding the value of the interquartile range.

WORKED EXAMPLE 10

Parents are often shocked at the amount of money their children spend. The data below give the amount spent (to the nearest whole dollar) by each child in a group that was taken on an excursion to the Royal Melbourne Show.

15	**12**	**17**	**23**	**21**	**19**	**16**
11	**17**	**18**	**23**	**24**	**25**	**21**
20	**37**	**17**	**25**	**22**	**21**	**19**

Calculate the interquartile range for these data.

THINK	WRITE
1 Enter the data into a calculator. (There is no need to order it.) Use the calculator to generate one-variable statistics. Copy down the values of the first and third quartiles.	$Q_1 = 17$ and $Q_3 = 23$
2 Calculate the interquartile range.	So, $\text{IQR} = Q_3 - Q_1$ $= 23 - 17$ $= 6$

The range

The *range* of a set of data is the difference between the highest and lowest values in that set.

It is usually not too difficult to locate the highest and lowest values in a set of data. Only when there is a very large number of observations might the job be made more difficult. In the previous worked example, the minimum and maximum values were 11 and 37, respectively. The range, therefore, can be calculated as:

$$\begin{aligned}\text{Range} &= \text{maxX} - \text{minX}\\ &= 37 - 11\\ &= 26.\end{aligned}$$

While the range gives us some idea about the spread of the data, it is not very informative since it gives us no idea of how the data are distributed between the highest and lowest values.

Now let us look at another measure of the centre of a set of data: the mode.

The mode

The mode is the score that occurs most often; that is, it is the score with the highest frequency. If there is more than one score with the highest frequency, then all scores with that frequency are the modes.

The mode is a weak measure of the centre of data because it may be a value that is close to the extremes of the data. If we consider the set of data in Worked example 8, the mode is 48 since it occurs three times and hence is the score with the highest frequency. In Worked example 9 there are two modes, 17 and 19, because they equally occur most frequently.

Exercise 1E The median, the interquartile range, the range and the mode

eBook*plus*

DIGITAL DOC
doc-9405
Spreadsheet
One-variable statistics

1 WE8 Write the median, the range and the mode of the sets of data shown in the following stem plots. The key for each stem plot is 3|4 = 34.

a

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

b

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

c

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

d

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

e

Stem	Leaf
60	2 5 8
61	1 3 3 6 7 8 9
62	0 1 2 4 6 7 8 8 9
63	2 2 4 5 7 8
64	3 6 7
65	4 5 8
66	3 5
67	4

2 For each of the following sets of data, write the median and the range.

a 2 4 6 7 9
b 12 15 17 19 21
c 3 4 5 6 7 8 9
d 3 5 7 8 12 13 15 16
e 12 13 15 16 18 19 21 23 24 26
f 3 8 4 2 1 6 5
g 16 21 14 28 23 15 11 19 25
h 7 4 3 4 9 5 10 4 2 11
i 29 23 22 33 26 18 37 22 16

3 a WE9 The number of cars that used the drive-in at a McBurger restaurant during each hour, from 7.00 am until 10.00 pm on a particular day, is shown below.

14 18 8 9 12 24 25 15 18 25 24 21 25 24 14

Find the interquartile range of this set of data.

b On the same day, the number of cars stopping during each hour that the nearby Kenny's Fried Chicken restaurant was open is shown below.

7 9 13 16 19 12 11 18 20 19 21 20 18 10 14

Find the interquartile range of these data.

c What do these values suggest about the two restaurants?

4 Write down a set of data for which $n = 5$, the median is 6 and the range is 7. Is this the only set of data with these parameters?

5 Is it possible to have a set of data in which the:

a lower quartile equals the lowest score? **b** IQR is zero?

Give an example of each.

6 **MC** The quartiles for a set of data are calculated and found to be $Q_1 = 13$, $Q_2 = 18$ and $Q_3 = 25$. Which of the following statements is true?

A The interquartile range of the data is 5.
B The interquartile range of the data is 7.
C The interquartile range of the data is 12.
D The median is 12.
E The median is 19.

7 **WE10** For each of the following sets of data find the median, the interquartile range, the range and the mode.

a	16	12	8	7	26	32	15	51	29	45
	19	11	6	15	32	18	43	31	23	23
b	22	25	27	36	31	32	39	29	20	30
	23	25	21	19	29	28	31	27	22	29
c	1.2	2.3	4.1	2.4	1.5	3.7	6.1	2.4	3.6	1.2
	6.1	3.7	5.4	3.7	5.2	3.8	6.3	7.1	4.9	

8 For each set of data shown on the stem plots, find the median, the interquartile range, the range and the mode. Compare these values for both data sets.

a

Stem	Leaf
2	3 5 5 6 7 8 9 9
3	0 2 2 3 4 6 6 7 8 8
4	2 2 4 5 6 6 6 7 9
5	0 3 3 5 6
6	2 4
7	5 9
8	2
9	7
10	
11	4

Key: 4|2 = 42

b

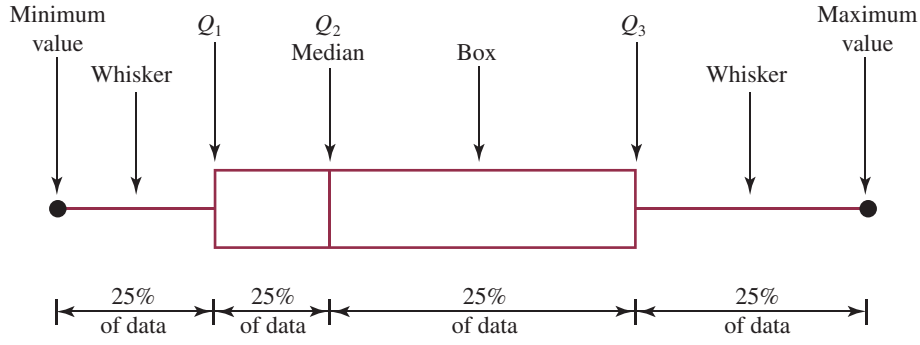
Stem	Leaf
1	4
1*	
2	1 4
2*	5 7 8 8 9
3	1 2 2 2 4 4 4 4
3*	5 5 5 6
4	3 4
4*	

Key: 2|1 = 21
2*|5 = 25

1F Boxplots

The five number summary statistics that we looked at in the previous section can be illustrated very neatly in a special diagram known as a *boxplot* (or *box-and-whisker* diagram). The diagram is made up of a box with straight lines (whiskers) extending from opposite sides of the box.

A boxplot displays the minimum and maximum values of the data together with the quartiles and is drawn with a labelled scale. The length of the box is given by the interquartile range. A boxplot gives us a very clear visual display of how the data are spread out.



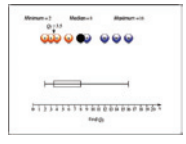
A boxplot

studyon

Units:	3 & 4
AOS:	DA
Topic:	3
Concept:	5

Concept summary
Read a summary of this concept.

See more
Watch a video about how to construct boxplots.



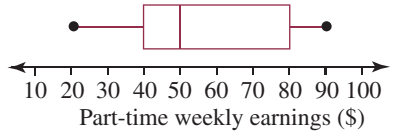
Boxplots can be drawn horizontally or vertically.

Horizontal boxplot

Vertical boxplot

WORKED EXAMPLE 11

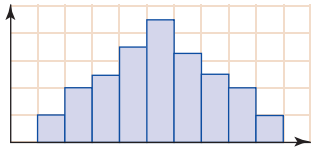
The boxplot at right shows the distribution of the part-time weekly earnings of a group of Year 12 students. Write down the range, the median and the interquartile range for these data.



THINK	WRITE
1 Range = Maximum value – Minimum value. The minimum value is 20 and the maximum value is 90.	Range = 90 – 20 = 70
2 The median is located at the bar inside the box.	Median = 50
3 The ends of the box are at 40 and 80. IQR = $Q_3 - Q_1$	$Q_1 = 40$ and $Q_3 = 80$ IQR = 80 – 40 = 40

Earlier, we noted three general types of shape for histograms and stem plots: symmetric, negatively skewed and positively skewed. It is useful to compare the corresponding boxplots of distributions with such shapes.

In the figures below, a symmetric distribution is represented in the histogram and in the boxplot. The characteristics of this boxplot are that the whiskers are about the same length and the median is located about halfway along the box.



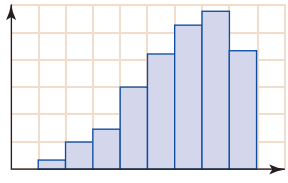
Symmetric histogram

Symmetric boxplot

The figures below show a negatively skewed distribution. In such a distribution, the data peak to the right on the histogram and trail off to the left.

In corresponding fashion on the boxplot, the bunching of the data to the right means that the left-hand whisker is longer and the right-hand whisker is shorter; that is, the lower 25% of data are sparse and spread out whereas the top 25% of data are bunched up.

The median occurs further towards the right end of the box.



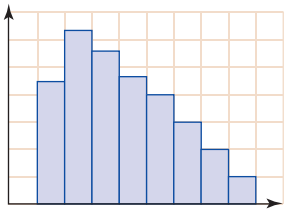
Negatively skewed histogram

Negatively skewed boxplot

In the figures below, we have a positively skewed distribution. In such a distribution, the data peak to the left on the histogram and trail off to the right.

In corresponding fashion on the boxplot, the bunching of the data to the left means that the left-hand whisker is shorter and the right-hand whisker is longer; that is, the upper 25% of data are sparse and spread out whereas the lower 25% of data are bunched up.

The median occurs further towards the left end of the box.

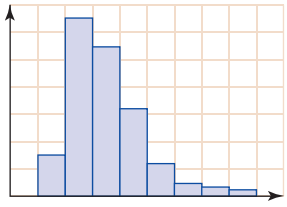


Positively skewed histogram

Positively skewed boxplot

WORKED EXAMPLE 12

Explain whether or not the histogram and the boxplot shown below could represent the same data.



THINK

The histogram shows a distribution which is positively skewed.

The boxplot shows a distribution which is approximately symmetric.

WRITE

The histogram and the boxplot could not represent the same data since the histogram shows a distribution that is positively skewed and the boxplot shows a distribution that is approximately symmetric.

WORKED EXAMPLE 13

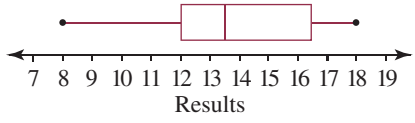
eBook*plus*

TUTORIAL
eles-1256
Worked example 13

The results (out of 20) of oral tests in a Year 12 Indonesian class are:

15 12 17 8 13 18 14 16 17 13 11 12

Display these data using a boxplot and discuss the shape obtained.

THINK	WRITE/DRAW
1 Find the lowest and highest scores, Q_1, the median (Q_2) and Q_3 by first ordering the data.	8 11 12 12 13 13 14 15 16 17 17 18 The median score is 13.5. The lower half of the scores are 8 11 12 12 13 13. So, $Q_1 = 12$ The upper half of the scores are 14 15 16 17 17 18. So, $Q_3 = 16.5$ The lowest score is 8. The highest score is 18.
2 Using these five number summary statistics, draw the boxplot.	
3 Consider the spread of each quarter of the data.	The scores are grouped around 12 and 13, as well as around 17 and 18 with 25% of the data in each section. The scores are more spread elsewhere.

CAS calculators can also be used to draw boxplots.

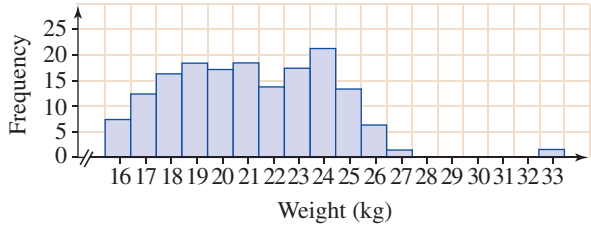
study on

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

Concept summary
Read a summary of this concept.

Outliers

When one observation lies well away from other observations in a set, we call it an *outlier*. Sometimes an outlier occurs because data have been incorrectly obtained or misread. For example, at right we see a histogram showing the weights of a group of 5-year-old boys.



The outlier, 33, may have occurred because a weight was incorrectly recorded as 33 rather than 23 or perhaps there was a boy in this group who, for some medical reason, weighed a lot more than his counterparts. When an outlier occurs, the reasons for its occurrence should be checked.

To identify possible outliers, we can apply a simple rule.

An outlier is a score, x, which lies *outside* the interval:

$$Q_1 - 1.5 \times \text{IQR} \le x \le Q_3 + 1.5 \times \text{IQR}.$$

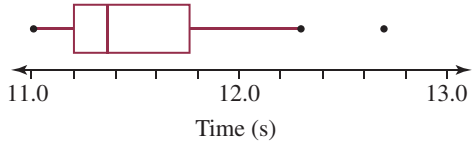
An outlier is not included in the boxplot but simply plotted as a point beyond the end of the whisker.

WORKED EXAMPLE 14

The times (in seconds) achieved by the 12 fastest runners in the 100-m sprint at a school athletics meeting are listed below.

11.2	**12.3**	**11.5**	**11.0**	**11.6**	**11.4**
11.9	**11.2**	**12.7**	**11.3**	**11.2**	**11.3**

Draw a boxplot to represent the data, describe the shape of the distribution and comment on the existence of any outliers.

THINK	WRITE/DRAW
1 Determine the five number summary statistics by first ordering the data and obtain the interquartile range.	11.0 11.2 11.2 11.2 11.3 11.3 11.4 11.5 11.6 11.9 12.3 12.7 Lowest score = 11.0 Highest score = 12.7 Median = Q_2 = 11.35 Q_1 = 11.2 Q_3 = 11.75 IQR = 11.75 − 11.2 = 0.55
2 Identify any outliers by applying the outlier rule.	$Q_1 - 1.5 \times \text{IQR} = 11.2 - 1.5 \times 0.55$ $= 10.375$ The lowest score lies above 10.375, so there is no outlier below. $Q_3 + 1.5 \times \text{IQR} = 11.75 + 1.5 \times 0.55$ $= 12.575$ The score 12.7 lies above 12.575, so it is an outlier and 12.3 becomes the end of the upper whisker.
3 Draw the boxplot with the outlier.	
4 Describe the shape of the distribution. Data peak to the left and trail off to the right with one outlier.	The data are positively skewed with 12.7 seconds being an outlier. This may be due to incorrect timing or recording but more likely the top eleven runners were significantly faster than the other competitors in the event.

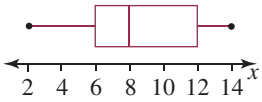
Exercise 1F Boxplots

eBook*plus*

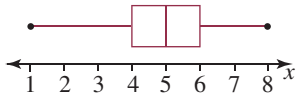
DIGITAL DOC
doc-9406
Spreadsheet
Boxplots

1 **WE11** For the boxplots shown, write down the range, the interquartile range and the median of the distributions which each one represents.

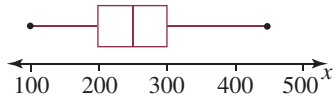
a



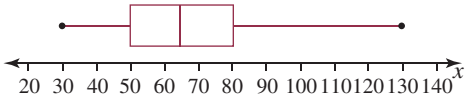
b



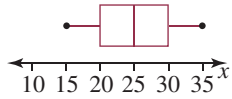
c



d

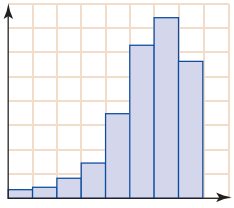


e

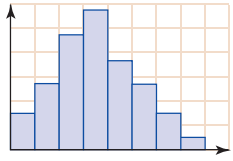


2 **WE12** Match each histogram below with the boxplot which could show the same distribution.

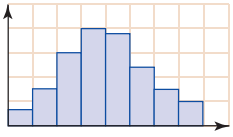
a



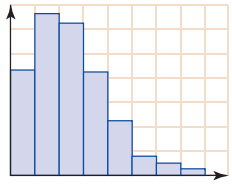
b



c



d



i

ii

iii

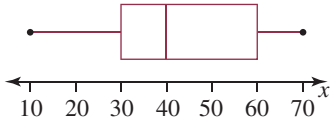
iv

3 **WE13** For each of the following sets of data, construct a boxplot.

a 3 5 6 8 8 9 12 14 17 18
b 3 4 4 5 5 6 7 7 7 8 8 9 9 10 10 12
c 4.3 4.5 4.7 4.9 5.1 5.3 5.5 5.6
d 11 13 15 15 16 18 20 21 22 21 18 19 20 16 18 20
e 0.4 0.5 0.7 0.8 0.8 0.9 1.0 1.1 1.2 1.0 1.3

4 **MC** For the distribution shown in the boxplot below, it is true to say that:

A the median is 30
B the median is 45
C the interquartile range is 10
D the interquartile range is 30
E the interquartile range is 60



5 The number of clients seen each day over a 15-day period by a tax consultant is:

3 5 2 7 5 6 4 3 4 5 6 6 4 3 4

Represent these data on a boxplot.

6 The maximum daily temperatures (in °C) for the month of October in Melbourne are:

18 26 28 23 16 19 21 27 31 23 24 26 21 18 26 27
23 21 24 20 19 25 27 32 29 21 16 19 23 25 27

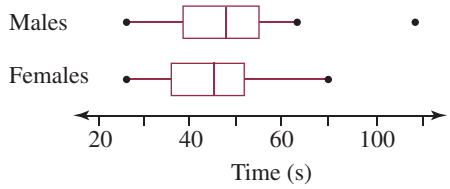
Represent these data on a boxplot.

7 **WE14** The number of rides that 16 children had at the annual show are listed below.

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

a Draw a boxplot to represent the data, describe the shape of the distribution and comment on the existence of any outliers.
b Use a CAS calculator to draw a boxplot for these data.

8 A concentration test was carried out on 40 students in Year 12 across Australia. The test involved the use of a computer mouse and the ability to recognise multiple images. The less time required to complete the activity, the better the student's ability to concentrate.
The data are shown by the parallel boxplots below.



a Identify two similar properties of the concentration spans for boys and girls.
b Find the interquartile range for boys and girls.
c Comment on the existence of an outlier in the boys' data.

study on

Units:	3 & 4
AOS:	DA
Topic:	3
Concept:	1

Concept summary
Read a summary of this concept.

1G The mean

The *mean* of a set of data is what is referred to in everyday language as the *average*.

For the set of data {4, 7, 9, 12, 18}:

$$\text{mean} = \frac{4+7+9+12+18}{5}$$
$$= 10.$$

The symbol we use to represent the mean is $\bar{x}$, that is, a lower-case x with a bar on top. So, in this case, $\bar{x} = 10$.

The formal definition of the mean is:

$$\bar{x} = \frac{\Sigma x}{n}$$

where Σx represents the sum of all of the observations in the data set and n represents the number of observations in the data set.

Note that the symbol, Σ, is the Greek letter, sigma, which represents 'the sum of'.

The mean is also referred to as a *summary statistic* and is a measure of the centre of a distribution. The mean is the point about which the distribution 'balances'.

Consider the masses of 7 potatoes, given in grams, in the photograph below.

The mean is 145 g. The observations 130 and 160 'balance' each other since they are each 15 g from the mean. Similarly, the observations 120 and 170 'balance' each other since they are each 25 g from the mean, as do the observations 100 and 190. Note that the median is also 145 g. That is, for this set of data the mean and the median give the same value for the centre. This is because the distribution is symmetric.

Now consider two cases in which the distribution of data is not symmetric.

Case 1

Consider the masses of a different set of 7 potatoes, given in grams below.

100 105 110 115 120 160 200

The median of this distribution is 115 g and the mean is 130 g. There are 5 observations that are less than the mean and only 2 that are more. In other words, the mean does not give us a good indication of the *centre* of the distribution. However, there is still a 'balance' between observations below the mean and those above, in terms of the spread of all the observations from the mean. Therefore, the mean is still useful to give a measure of the central tendency of the distribution but in cases where the distribution is skewed, the median gives a better indication of the centre. For a positively skewed distribution, as in the previous case, the mean will be greater than the median. For a negatively skewed distribution the mean will be less than the median.

Case 2

Consider the data below, showing the weekly income (to the nearest \$10) of 10 families living in a suburban street.

\$600 \$1340 \$1360 \$1380 \$1400 \$1420 \$1420 \$1440 \$1460 \$1500

In this case, $\bar{x} = \frac{13\,320}{10} = \$1332$, and the median is \$1410.

One of the values in this set, \$600, is clearly an outlier. As a result, the value of the mean is below the weekly income of the other 9 households. In such a case the mean is not very useful in establishing the centre; however, the 'balance' still remains for this negatively skewed distribution.

The mean is calculated by using the values of the observations and because of this it becomes a less reliable measure of the centre of the distribution when the distribution is skewed or contains an outlier. Because the median is based on the order of the observations rather than their value, it is a better measure of the centre of such distributions.

WORKED EXAMPLE 15

Calculate the mean of the set of data below.

10, 12, 15, 16, 18, 19, 22, 25, 27, 29

THINK	WRITE
1 Write the formula for calculating the mean, where $\sum x$ is the sum of all scores; n is the number of scores in the set.	$\bar{x} = \frac{\sum x}{n}$
2 Substitute the values into the formula and evaluate.	$= \frac{10 + 12 + 15 + 16 + 18 + 19 + 22 + 25 + 27 + 29}{10}$ $\bar{x} = 19.3$ The mean, $\bar{x}$, is 19.3.

When data are presented in a frequency table with class intervals and we don't know what the raw data are, we employ another method to find the mean of these grouped data. This other method is shown in the example that follows and uses the midpoints of the class intervals to represent the raw data.

Recall that the Greek letter sigma, Σ, represents 'the sum of'. So, Σf means the sum of the frequencies and is the total of all the numbers in the frequency column.

To find the mean for grouped data,

$$\bar{x} = \frac{\Sigma(f \times m)}{\Sigma f}$$

where f represents the frequency of the data and m represents the midpoint of the class interval of the grouped data.

WORKED EXAMPLE 16

eBook*plus*

TUTORIAL
eles-1257
Worked example 16

The ages of a group of 30 people attending a superannuation seminar are recorded in the frequency table below.

Age (class intervals)	Frequency f
20–29	1
30–39	6
40–49	13
50–59	6
60–69	3
70–79	1

Calculate the mean age of those attending the seminar.

THINK

1. Since we don't have individual raw ages, but rather a class interval, we need to decide on one particular age to represent each interval. We use the midpoint, m, of the class interval. Add an extra column to the table to display these.
   The midpoint of the first interval is $\frac{20+29}{2} = 24.5$, the midpoint of the second interval is 34.5 and so on.

2. Multiply each of the midpoints by the frequency and display these values in another column headed $f \times m$. For the first interval we have $24.5 \times 1 = 24.5$. For the second interval we have $34.5 \times 6 = 207$ and so on.

3. Sum the product of the midpoints and the frequencies in the $f \times m$ column.
   $24.5 + 207 + 578.5 + 327 + 193.5 + 74.5 = 1405$

4. Divide this sum by the total number of people attending the seminar (given by the sum of the frequency column).

WRITE

Age (class intervals)	Frequency f	Midpoint of class interval m	$f \times m$
20–29	1	24.5	24.5
30–39	6	34.5	207
40–49	13	44.5	578.5
50–59	6	54.5	327
60–69	3	64.5	193.5
70–79	1	74.5	74.5
	$\Sigma f = 30$		$\Sigma(f \times m) = 1405$

So, $\bar{x} = \frac{1405}{30}$
≈ 46.8 (correct to 1 decimal place).

Exercise 1G The mean

1 **WE15** Find the mean of each of the following sets of data.
- **a** 5 6 8 8 9
- **b** 3 4 4 5 5 6 7 7 7 8 8 9 9 10 10 12
- **c** 4.3 4.5 4.7 4.9 5.1 5.3 5.5 5.6
- **d** 11 13 15 15 16 18 20 21 22
- **e** 0.4 0.5 0.7 0.8 0.8 0.9 1.0 1.1 1.2 1.0 1.3

2 Calculate the mean of each of the following and explain whether or not it gives us a good indication of the centre of the data.
- **a** 0.7 0.8 0.85 0.9 0.92 2.3
- **b** 14 16 16 17 17 17 19 20
- **c** 23 24 28 29 33 34 37 39
- **d** 2 15 17 18 18 19 20

3 The number of people attending sculpture classes at the local TAFE college for each week during the first semester is given below.

15 12 15 11 14 8 14 15 11 10
7 11 12 14 15 14 15 9 10 11

What is the mean number of people attending each week? (Express your answer to the nearest whole number.)

4 **MC** The ages of a group of junior pilots joining an international airline are indicated on the stem plot at right.

The mean age of this group of pilots is:
- **A** 20
- **B** 28
- **C** 29
- **D** 29.15
- **E** 29.5

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

Key: 2|1 = 21 years

5 **MC** The number of people present each week at a 15-week horticultural course is given by the stem plot at right.

The mean number of people attending each week was closest to:
- **A** 17.7
- **B** 18
- **C** 19.5
- **D** 20
- **E** 21.2

Stem	Leaf
0	4
0*	7
1	2 4
1*	5 5 6 7 8
2	1 2 4
2*	7 7 7

Key: 2|4 = 24 people

6 For each of the following, write down whether the mean or the median would provide a better indication of the centre of the distribution.
- **a** A positively skewed distribution
- **b** A symmetric distribution
- **c** A distribution with an outlier
- **d** A negatively skewed distribution

7 **WE16** Find the mean of each set of data given below.

a

Class interval	Frequency, f
0–9	1
10–19	3
20–29	6
30–39	17
40–49	12
50–59	5

b

Class interval	Frequency, f
0–4	2
5–9	5
10–14	7
15–19	13
20–24	8
25–29	6

c

Class interval	Frequency, f
0–49	2
50–99	7
100–149	8
150–199	14
200–249	12
250–299	5

d

Class interval	Frequency, f
1–6	14
7–12	19
13–18	23
19–24	22
25–30	20
31–36	14

8 The ages of people attending a beginner's course in karate are indicated in the following frequency table.

a What is the mean age of those attending the course? (Express your answer correct to 1 decimal place.)

b Calculate the median. What does this value, compared to the mean, suggest about the shape of the distribution?

Age	Frequency, f
10–14	5
15–19	5
20–24	7
25–29	4
30–34	3
35–39	2
40–44	2
45–49	1

studyon

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

Concept summary
Read a summary of this concept.

Do more
Interact with standard deviations.

1H Standard deviation

The standard deviation gives us a measure of how data are spread around the mean. For the set of data {8, 10, 11, 12, 12, 13}, the mean, $\bar{x} = 11$.

The amount that each observation 'deviates' (that is, differs) from the mean is calculated and shown in the table below.

Particular observation, x	Deviation from the mean, $(x - \bar{x})$
8	$8 - 11 = {}^{-}3$
10	$10 - 11 = {}^{-}1$
11	$11 - 11 = 0$
12	$12 - 11 = 1$
12	$12 - 11 = 1$
13	$13 - 11 = 2$

The deviations from the mean are either positive or negative depending on whether the particular observation is lower or higher in value than the mean. If we were to add all the deviations from the mean we would obtain zero.

If we square the deviations from the mean we will overcome the problem of positive and negative deviations cancelling each other out. With this in mind, a quantity known as *sample variance* (s^2) is defined:

$$s^2 = \frac{\Sigma(x - \bar{x})^2}{n - 1}.$$

Technically, this formula for variance is used when the data set is a sub-set of a larger population.

Variance gives the average of the squared deviations and is also a measure of spread. A far more useful measure of spread, however, is the standard deviation, which is the square root of variance (s). One reason for it being more useful is that it takes the same unit as the observations (for example, cm or number of people). Variance would square the units, for example, cm^2 or number of people squared, which is not very practical.

Other advantages of the standard deviation will be dealt with later in the chapter.

In summary,

$$s = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$$

where s represents sample standard deviation
$\sum$ represents 'the sum of'
x represents an observation
$\bar{x}$ represents the mean
n represents the number of observations.

While some of the theory or formulas associated with standard deviation may look complex, the calculation of this measure of spread is straightforward using a statistical, graphics or CAS calculator. Manual computation of standard deviation is therefore rarely necessary.

WORKED EXAMPLE 17

The price (in cents) per litre of petrol at a service station was recorded each Friday over a 15-week period. The data are given below.

152.4	**160.2**	**159.6**	**168.6**	**161.4**	**156.6**	**164.8**	**162.6**
161.0	**156.4**	**159.0**	**160.2**	**162.6**	**168.4**	**166.8**	

Calculate the standard deviation for this set of data, correct to 2 decimal places.

THINK	WRITE
1 On a CAS calculator, enter the data into the first list in the spreadsheet and label it *price*. Select 'One-Variable Statistics' option and choose 'price' for the X1 List. Press OK to see all statistics.	
2 The entry, $SX := S_{n-1}$, gives us the standard deviation. Round the value correct to 2 decimal places.	$S_x = 4.515\ 92$ $s = 4.52$ cents/L

WORKED EXAMPLE 18

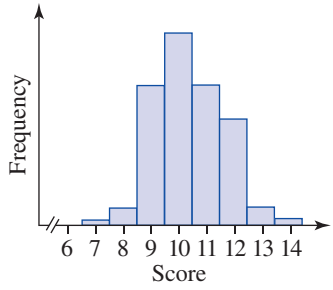
The number of students attending SRC meetings during the term is given in the stem plot at right. Calculate the standard deviation for this set of data, correct to 3 decimal places.

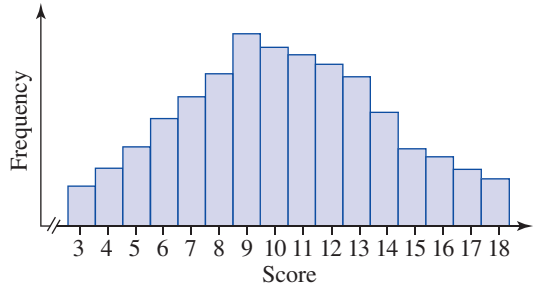
Stem	Leaf
0	4
0*	8 8
1	1 3 4
1*	5 6 8
2	3
2*	5

Key: 1|4 = 14 students

THINK	WRITE
1 On your calculator, enter the data from the stem plot into a spreadsheet. Using the given key, the scores are: 4, 8, 8, 11, 13, 14, 15, 16, 18, 23 and 25. To calculate the summary statistics, repeat the instructions for Worked example 17.	
2 The entry, *SX*, gives us the standard deviation. Round the value correct to 3 decimal places.	*SX* =6.363 25 *s* = 6.363 students

The standard deviation is a measure of the spread of data from the mean. Consider the two sets of data shown below.





Each set of data has a mean of 10. The set of data above left has a standard deviation of 1 and the set of data above right has a standard deviation of 3.

As we can see, the larger the standard deviation, the more spread are the data from the mean.

Exercise 1H Standard deviation

1 **WE17** For each of the following sets of data, calculate the standard deviation correct to 2 decimal places.
   - **a** 3 4 4.7 5.1 6 6.2
   - **b** 7 9 10 10 11 13 13 14
   - **c** 12.9 17.2 17.9 20.2 26.4 28.9
   - **d** 41 43 44 45 45 46 47 49
   - **e** 0.30 0.32 0.37 0.39 0.41 0.43 0.45

2 First-quarter profit increases for 8 leading companies are given below as percentages.
   2.3 0.8 1.6 2.1 1.7 1.3 1.4 1.9
   Calculate the standard deviation for this set of data and express your answer correct to 2 decimal places.

3 The heights in metres of a group of army recruits are given below.
   1.8 1.95 1.87 1.77 1.75 1.79 1.81 1.83 1.76 1.80 1.92 1.87 1.85 1.83
   Calculate the standard deviation for this set of data and express your answer correct to 2 decimal places.

4 **WE18** Times (to the nearest tenth of a second) for the heats in the 100 m sprint at the school sports carnival are given at right.
   Calculate the standard deviation for this set of data and express your answer correct to 2 decimal places.

Stem	Leaf
11	0
11	2 3
11	4 4 5
11	6 6
11	8 8 9
12	0 1
12	2 2 3
12	4 4
12	6
12	9

Key: 11|0 = 11.0 s

5 The number of outgoing phone calls from an office each day over a 4-week period is shown on the stem plot below.

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

Key: 2|1 = 21 calls

Calculate the standard deviation for this set of data and express your answer correct to 2 decimal places.

6 **MC** A new legal aid service has been operational for only 5 weeks. The number of people who have made use of the service each day during this period is set out at right.

The standard deviation (to 2 decimal places) of these data is:

A 6.00 **B** 6.34 **C** 6.47
D 15.44 **E** 16.00

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

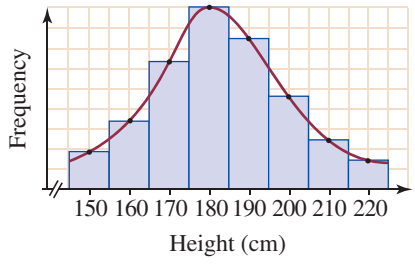
Key: 1|0 = 10 people
1*|6 = 16 people

11 The 68–95–99.7% rule and *z*-scores

The 68–95–99.7% rule

The heights of a large number of students at a graduation ceremony were recorded and are shown in the histogram at right.

This set of data is approximately symmetric and has what is termed a *bell* shape. Many sets of data fall into this category and are often referred to as normal distributions. Examples are birth weights and people's heights. Data which are normally distributed have their symmetrical, bell-shaped distribution centred on the mean value, $\bar{x}$.



An astounding feature of this type of distribution is that we can predict what percentage of the data lie 1, 2 or 3 standard deviations either side of the mean using what is termed the *68–95–99.7% rule*.

The 68–95–99.7% rule for a bell-shaped curve states that approximately:

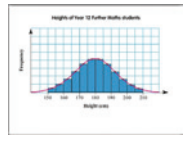
1. 68% of data lie within *1 standard deviation* either side of the mean
2. 95% of data lie within *2 standard deviations* either side of the mean
3. 99.7% of data lie within *3 standard deviations* either side of the mean.

studyon

Units:	3 & 4
AOS:	DA
Topic:	4
Concept:	1

Concept summary
Read a summary of this concept.

See more
Watch a video about normal distributions.



eBook*plus*

INTERACTIVITY
int-0182
The 68–95–99.7% rule and *z*-scores

Figure **1**

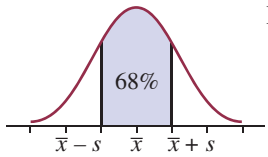


Figure **2**

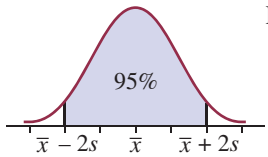
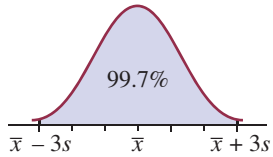


Figure **3**



In figure **1** above, 68% of the data shown lie between the value which is 1 standard deviation below the mean, that is $\bar{x} - s$, and the value which is 1 standard deviation above the mean, that is, $\bar{x} + s$.

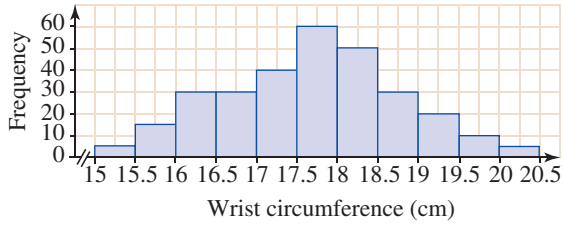
In figure **2** above, 95% of the data shown lie between the value which is 2 standard deviations below the mean, that is, $\bar{x} - 2s$, and the value which is 2 standard deviations above the mean, that is $\bar{x} + 2s$.

In figure **3** above, 99.7% of the data shown lie between the value which is 3 standard deviations below the mean, that is, $\bar{x} - 3s$, and the value which is 3 standard deviations above the mean, that is, $\bar{x} + 3s$.

WORKED EXAMPLE 19

The wrist circumferences of a group of people were recorded and the results are shown in the histogram at right. The mean of the set of data is 17.7 and the standard deviation is 0.9. Write down the wrist circumferences between which we would expect approximately:

a 68% of the group to lie
b 95% of the group to lie
c 99.7% of the group to lie.



THINK

a The distribution can be described as approximately bell-shaped and therefore the 68–95–99.7% rule can be applied. Approximately 68% of the people have a wrist circumference between $\bar{x} - s$ and $\bar{x} + s$ (or one standard deviation either side of the mean).

WRITE

a $\bar{x} - s = 17.7 - 0.9 = 16.8$
$\bar{x} + s = 17.7 + 0.9 = 18.6$
So approximately 68% of the people have a wrist size between 16.8 and 18.6 cm.

b Similarly, approximately 95% of the people have a wrist size between $\bar{x} - 2s$ and $\bar{x} + 2s$.

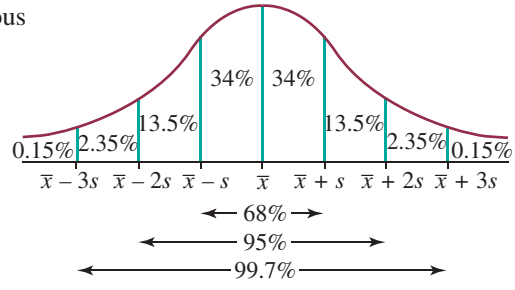
b $\bar{x} - 2s = 17.7 - 1.8 = 15.9$
$\bar{x} + 2s = 17.7 + 1.8 = 19.5$
Approximately 95% of people have a wrist size between 15.9 cm and 19.5 cm.

c Similarly, approximately 99.7% of the people have a wrist size between $\bar{x} - 3s$ and $\bar{x} + 3s$.

c $\bar{x} - 3s = 17.7 - 2.7 = 15.0$
$\bar{x} + 3s = 17.7 + 2.7 = 20.4$
Approximately 99.7% of people have a wrist size between 15.0 cm and 20.4 cm.

Using the 68–95–99.7% rule, we can work out the various percentages of the distribution which lie between the mean and 1 standard deviation from the mean and between the mean and 2 standard deviations from the mean and so on. The diagram at right summarises this.

Note that 50% of the data lie below the mean and 50% lie above the mean due to the symmetry of the distribution about the mean.



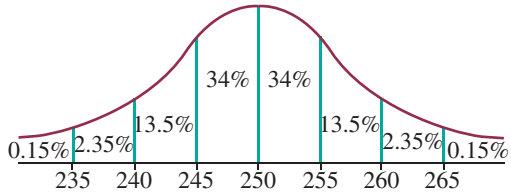
WORKED EXAMPLE 20

eBook*plus*

TUTORIAL
eles-1258
Worked example 20

The distribution of the masses of packets of 'Fibre-fill' breakfast cereal is known to be bell-shaped with a mean of 250 g and a standard deviation of 5 g. Find the percentage of Fibre-fill packets with a mass which is:

a less than 260 g **b** less than 245 g
c more than 240 g **d** between 240 g and 255 g.

THINK	WRITE/DRAW
1 Draw the bell-shaped curve. Label the axis. $\bar{x} = 250$, $\bar{x} + s = 255$, $\bar{x} + 2s = 260$ etc.	
a 260 g is 2 standard deviations above the mean. Using the summary diagram, we can find the percentage of data which is less than 260 g.	**a** Mass of 260 g is 2 standard deviations above the mean. Percentage of distribution less than 260 g is 13.5% + 34% + 34% + 13.5% + 2.35% + 0.15% = 97.5% or 13.5% + 34% + 50% = 97.5%
b 245 g is 1 standard deviation below the mean.	**b** Mass of 245 g is 1 standard deviation below the mean. Percentage of distribution less than 245 g is 13.5% + 2.35% + 0.15% = 16% or 50% – 34% = 16%
c 240 g is 2 standard deviations below the mean.	**c** Mass of 240 g is 2 standard deviations below the mean. Percentage of distribution more than 240 g is 13.5% + 34% + 34% + 13.5% + 2.35% + 0.15% = 97.5% or 13.5% + 34% + 50% = 97.5%
d Now, 240 g is 2 standard deviations below the mean while 255 g is 1 standard deviation above the mean.	**d** Mass of 240 g is 2 standard deviations below the mean. Mass of 255 g is 1 standard deviation above the mean. Percentage of distribution between 240 g and 255 g is 13.5% + 34% + 34% = 81.5%

WORKED EXAMPLE 21

The number of matches in a box is not always the same. When a sample of boxes was studied it was found that the number of matches in a box approximated a normal (bell-shaped) distribution with a mean number of matches of 50 and a standard deviation of 2. In a sample of 200 boxes, how many would be expected to have more than 48 matches?

THINK	WRITE
1 Find the percentage of boxes with more than 48 matches. Since 48 = 50 – 2, the score of 48 is 1 standard deviation below the mean.	48 matches is 1 standard deviation below the mean. Percentage of boxes with more than 48 matches = 34% + 50% = 84%
2 Find 84% of the total sample.	Number of boxes = 84% of 200 = 168 boxes

study on

Units:	3 & 4
AOS:	DA
Topic:	4
Concept:	2

Concept summary
Read a summary of this concept.

Standard *z*-scores

To find a comparison between scores in a particular distribution or in different distributions, we use the *z-score*. The *z*-score (also called the *standardised score*) indicates the position of a certain score in relation to the mean.

A *z*-score of 0 indicates that the score obtained is equal to the mean, a negative *z*-score indicates that the score is below the mean and a positive *z*-score indicates a score above the mean.

The *z*-score measures the distance from the mean in terms of the standard deviation. A score that is exactly one standard deviation above the mean has a *z*-score of 1. A score that is exactly one standard deviation below the mean has a *z*-score of $^{-}1$.

To calculate a *z*-score we use the formula:

$$z = \frac{x - \bar{x}}{s}$$

where x = the score, $\bar{x}$ = the mean and s = the standard deviation.

WORKED EXAMPLE 22

In an IQ test, the mean IQ is 100 and the standard deviation is 15. Dale's test results give an IQ of 130. Calculate this as a *z*-score.

THINK	WRITE
1 Write the formula.	$z = \frac{x - \bar{x}}{s}$
2 Substitute for x, $\bar{x}$ and s.	$= \frac{130 - 100}{15}$
3 Calculate the *z*-score.	$= 2$

Dale's *z*-score is 2, meaning that his IQ is exactly two standard deviations above the mean.

Not all *z*-scores will be whole numbers; in fact most will not be. A whole number indicates only that the score is an exact number of standard deviations above or below the mean.

Using the previous example, an IQ of 88 would be represented by a *z*-score of $^{-}0.8$, as shown below.

$$\begin{aligned} z &= \frac{x - \bar{x}}{s} \\ &= \frac{88 - 100}{15} \\ &= {}^{-}0.8 \end{aligned}$$

The negative value indicates that the IQ of 88 is below the mean but by less than one standard deviation.

WORKED EXAMPLE 23

To obtain the average number of hours of study done by Year 12 students per week, Kate surveys 20 students and obtains the following results.

12	18	15	14	9	10	13	12	18	25
15	10	3	21	11	12	14	16	17	20

a Calculate the mean and standard deviation (correct to 2 decimal places).

b Robert studies for 16 hours each week. Express this as a *z*-score based on the above results. (Give your answer correct to 2 decimal places.)

THINK	WRITE
a 1 Enter the data into your calculator.	a
2 Obtain the mean from your calculator.	$\bar{x} = 14.25$
3 Obtain the standard deviation from your calculator.	$s = 4.88$

b 1 Write the formula for z-score.

b $z = \dfrac{x - \bar{x}}{s}$

2 Substitute for x, $\bar{x}$ and s into the formula and evaluate.

$= \dfrac{16 - 14.25}{4.88}$

$= 0.36$

Comparing data

An important use of z-scores is to compare scores from different data sets. Suppose that in your maths exam your result was 74 and in English your result was 63. In which subject did you achieve the better result?

At first glance, it may appear that the maths result is better, but this does not take into account the difficulty of the test. A mark of 63 on a difficult English test may in fact be a better result than 74 if it was an easy maths test.

The only way that we can fairly compare the results is by comparing each result with its mean and standard deviation. This is done by converting each result to a z-score.

If, for maths, $\bar{x} = 60$ and $s = 12$, then

$$z = \frac{x - \bar{x}}{s} = \frac{74 - 60}{12} = 1.17$$

And if, for English, $\bar{x} = 50$ and $s = 8$, then

$$z = \frac{x - \bar{x}}{s} = \frac{63 - 50}{8} = 1.625$$

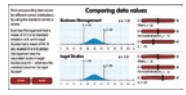
The English result is better because the higher z-score shows that the 63 is higher in comparison to the mean of each subject.

studyon

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

Concept summary
Read a summary of this concept.

Do more
Interact with comparisons of data values.



WORKED EXAMPLE 24

Janine scored 82 in her physics exam and 78 in her chemistry exam. In physics, $\bar{x} = 62$ and $s = 10$, while in chemistry, $\bar{x} = 66$ and $s = 5$.

a Write both results as a standardised score.

b Which is the better result? Explain your answer.

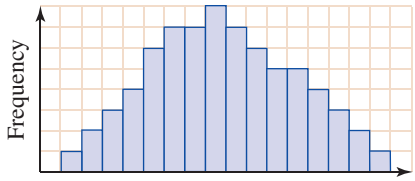
THINK	WRITE
a 1 Write the formula for each subject.	**a** Physics: $z = \dfrac{x - \bar{x}}{s}$ Chemistry: $z = \dfrac{x - \bar{x}}{s}$
2 Substitute for x, $\bar{x}$ and s.	$= \dfrac{82 - 62}{10}$ $= \dfrac{78 - 66}{5}$
3 Calculate each z-score.	$= 2$ $= 2.4$
b Explain that the subject with the highest z-score is the better result.	**b** The chemistry result is better because of the higher z-score.

In each example the circumstances must be analysed carefully to see whether a higher or lower z-score is better. For example, if we were comparing times for runners over different distances, the lower z-score would be the better one.

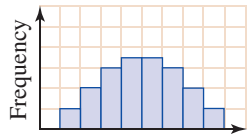
Exercise 1I The 68–95–99.7% rule and *z*-scores

1 In each of the following, decide whether or not the distribution is approximately bell-shaped.

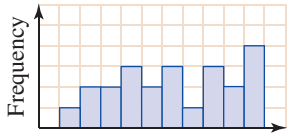
a



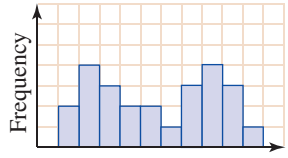
b



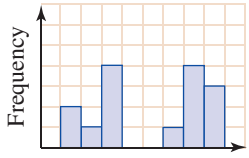
c



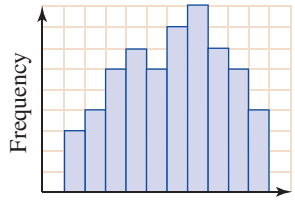
d



e

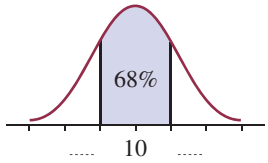


f

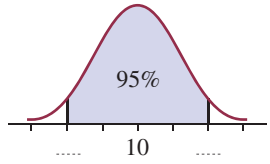


2 Copy and complete the entries on the horizontal scale of the following distributions, given that $\bar{x} = 10$ and $s = 2$.

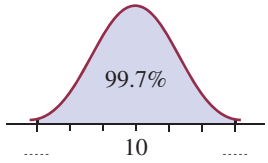
a



b

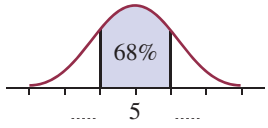


c

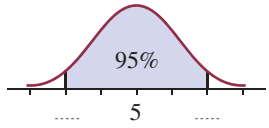


3 Copy and complete the entries on the horizontal scale of the following distributions, given that $\bar{x} = 5$ and $s = 1.3$.

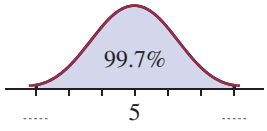
a



b



c



4 **WE19** The concentration ability of a randomly selected group of adults is tested during a short task which they are asked to complete. The length of the concentration span of those involved during the task is shown at right.

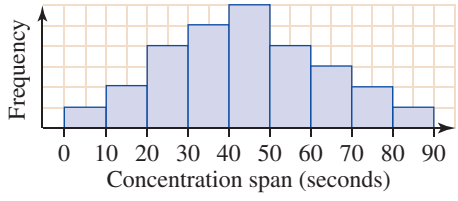
The mean, $\bar{x}$, is 49 seconds and the standard deviation, s, is 14 seconds.

Write down the values between which we would expect approximately:

a 68% of the group's concentration spans to fall

b 95% of the group's concentration spans to fall

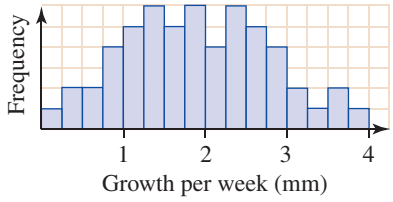
c 99.7% of the group's concentration spans to fall.



5 A research scientist measured the rate of hair growth in a group of hamsters. The findings are shown in the histogram below.

The mean growth per week was 1.9 mm and the standard deviation was 0.6 mm. Write down the hair growth rates between which approximately:

a 68% of the values fall
b 95% of the values fall
c 99.7% of the values fall.



6 The force required to break metal fasteners has a distribution which is bell-shaped. A large sample of metal fasteners was tested and the mean breaking force required was 12 newtons with a standard deviation of 0.3 newtons.

Write down the values between which approximately:

a 68% of the breaking forces would lie
b 95% of the breaking forces would lie
c 99.7% of the breaking forces would lie.

7 The heights of the seedlings sold in a nursery have a bell-shaped distribution. The mean height is 7 cm and the standard deviation is 2.

Write down the values between which approximately:

a 68% of seedling heights will lie
b 95% of seedling heights will lie
c 99.7% of seedling heights will lie.

8 **MC** A set of scores in a competition has a mean of 15 and a standard deviation of 3. The distribution of the scores is known to be bell-shaped. Which one of the following could be true?

A 68% of the scores lie between 3 and 15.
B 68% of the scores lie between 15 and 18.
C 68% of the scores lie between 12 and 15.
D 68% of the scores lie between 13.5 and 16.5.
E 68% of the scores lie between 12 and 18.

9 **MC** A distribution of scores is bell-shaped and the mean score is 26. It is known that 95% of scores lie between 21 and 31.

It is true to say that:

A 68% of the scores lie between 23 and 28.
B 97.5% of the scores lie between 23.5 and 28.5.
C The standard deviation is 2.5.
D 99.7% of the scores lie between 16 and 36.
E The standard deviation is 5.

10 **WE20** The distribution of heights of a group of Melbourne-based employees who work for a large international company is bell-shaped. The data have a mean of 160 cm and a standard deviation of 10 cm.

Find the percentage of this group of employees who are:

a less than 170 cm tall
b less than 140 cm tall
c greater than 150 cm tall
d between 130 cm and 180 cm in height.

11 The number of days taken off in a year by employees of a large company has a distribution which is approximately bell-shaped. The mean and standard deviation of this data are shown below.

Mean = 9 days Standard deviation = 2 days

Find the percentage of employees of this company who, in a year, take off:

a more than 15 days
b fewer than 5 days
c more than 7 days
d between 3 and 11 days
e between 7 and 13 days.

12 **MC** The mean number of Drool-mints in a packet is 48. The data have a standard deviation of 2. If the number of mints in a packet can be approximated by normal distribution. The percentage of packets which contain more than 50 Drool-mints is:

A 0.15% **B** 2.5% **C** 16%
D 50% **E** 84%

eBook*plus*

DIGITAL DOC
doc-9407
SkillSHEET 1.2
Percentages

13 **WE21** The volume of fruit juice in a certain type of container is not always the same. When a sample of these containers was studied it was found that the volume of juice they contained approximated a normal distribution with a mean of 250 mL and a standard deviation of 5 mL. In a sample of 400 containers, how many would be expected to have a volume of:

a more than 245 mL?
b less than 240 mL?
c between 240 and 260 mL?

14 A particular bolt is manufactured such that the length is not always the same. The distribution of the lengths of the bolts is approximately bell-shaped with a mean length of 2.5 cm and a standard deviation of 1 mm.

a In a sample of 2000 bolts, how many would be expected to have a length:
  i between 2.4 cm and 2.6 cm?
  ii less than 2.7 cm?
  iii between 2.6 cm and 2.8 cm?

b The manufacturer rejects bolts which have a length of less than 2.3 cm or a length of greater than 2.7 cm. In a sample of 2000 bolts, how many would the manufacturer expect to reject?

15 **WE22** In a maths exam, the mean score is 60 and the standard deviation is 12. Chifune's mark is 96. Calculate her mark as a z-score.

16 In an English test, the mean score was 55 with a standard deviation of 5. Adrian scored 45 on the English test. Calculate Adrian's mark as a z-score.

17 IQ tests have a mean of 100 and a standard deviation of 15. Calculate the z-score for a person with an IQ of 96. (Give your answer correct to 2 decimal places.)

18 The mean time taken for a racehorse to run 1 km is 57.69 s, with a standard deviation of 0.36 s. Calculate the z-score of a racehorse that runs 1 km in 58.23 s.

19 In a major exam, every subject has a mean score of 60 and a standard deviation of 12.5. Clarissa obtains the following marks on her exams. Express each as a z-score.

a English 54 **b** Maths 78 **c** Biology 61
d Geography 32 **e** Art 95

20 **WE23** The length of bolts being produced by a machine needs to be measured. To do this, a sample of 20 bolts are taken and measured. The results (in mm) are given below.

20	19	18	21	20	17	19	21	22	21
17	17	21	20	17	19	18	22	22	20

a Calculate the mean and standard deviation of the distribution.
b A bolt produced by the machine is 22.5 mm long. Express this result as a z-score. (Give your answer correct to 2 decimal places.)

21 **MC** In a normal distribution, the mean is 21.7 and the standard deviation is 1.9. A score of 20.75 corresponds to a z-score of:

A $^{-}1$ **B** $^{-}0.5$ **C** 0.5 **D** 1 **E** 0.75

22 **MC** In a normal distribution the mean is 58. A score of 70 corresponds to a standardised score of 1.5. The standard deviation of the distribution is:

A 6 **B** 8 **C** 10 **D** 12 **E** 9

23 WE24 Ken's English mark was 75 and his maths mark was 72. In English, the mean was 65 with a standard deviation of 8, while in maths the mean mark was 56 with a standard deviation of 12.

a Convert the mark in each subject to a z-score.

b In which subject did Ken perform better? Explain your answer.

24 In the first maths test of the year, the mean mark was 60 and the standard deviation was 12. In the second test, the mean was 55 and the standard deviation was 15. Barbara scored 54 in the first test and 50 in the second test. In which test did Barbara do better? Explain your answer.

25 The table below shows the average number of eggs laid per week by a random sample of chickens with 3 different types of living conditions.

Number of eggs per week		
Cage chickens	**Barn chickens**	**Free range chickens**
5.0	4.8	4.2
4.9	4.6	3.8
5.5	4.3	4.1
5.4	4.7	4.0
5.1	4.2	4.1
5.8	3.9	4.4
5.6	4.9	4.3
5.2	4.1	4.2
4.7	4.0	4.3
4.9	4.4	3.9
5.0	4.5	3.9
5.1	4.6	4.0
5.4	4.1	4.1
5.5	4.2	4.1

a Copy and complete the following table by calculating the mean and standard deviation of barn chickens and free range chickens correct to 1 decimal place.

Living conditions	Cage	Barn	Free range
Mean	5.2		
Standard deviation	0.3		

b A particular free range chicken lays an average of 4.3 eggs per week. Calculate the z-score relative to this sample to 2 decimal places.
The number of eggs laid by free range chickens is normally distributed. A free range chicken has a z-score of 1.

c Approximately what percentage of chickens lay fewer eggs than this chicken?

d Referring to the table showing the number of eggs per week, construct boxplots for each set of data.

i State the median of each set of data.

ii What could be concluded about the egg-producing capabilities of chickens in different living conditions?

1J Populations and simple random samples

Populations

A group of Year 12 students decide to base their statistical investigation for a maths project on what their contemporaries — that is, other Year 12 students — spend per year on Christmas and birthday presents for their family members. One of their early decisions is to decide what the population is going to be for their investigation. That is, are they looking at Year 12 students in Australia or in Victoria or in metropolitan Melbourne or in their suburb or just in their school? In practice, it is difficult to look at a large population unless, of course, you have a lot of resources available to you! The students decide that their population will be the Year 12 students at their school. This means that any conclusions they draw as a result of their investigation can be generalised to Year 12 students at their school but not beyond that.

studyon

Units:	3 & 4
AOS:	DA
Topic:	5
Concept:	1

Concept summary
Read a summary of this concept.

Samples

Given that there are 95 students in Year 12 at the school, it would be too time-consuming to interview all of them. A smaller group known as a *sample* is therefore taken from the population. The way in which this smaller group is chosen is of paramount importance. For the investigation to have credibility, the sample should be a random selection from the population and every member of that population should have an equal chance of being chosen in the sample. Also, the selection of one person from the population should not affect whether or not another person is chosen; that is, the selections should be independent. A *simple random sample* provides such a sample.

The students conducting the investigation decide to choose a sample of 12 fellow students. While it would be simplest to choose 12 of their mates as the sample, this would introduce bias since they would not be representative of the population as a whole.

The students obtain a list of names of the 95 students in Year 12. They then write next to the name of each student a number from 1 to 95. Using a calculator, the students generate 12 random numbers between 1 and 95. Alternatively, the students could have used a table of random numbers. Any point on the table can be taken as the starting point. The students decide which direction to move through the table; for example, across the table to the right or to the left or down. Once a direction is chosen, they must stay with that movement and write down the 2-digit numbers as they go along.

The numbers chosen by the students are then matched to the numbers on the name list and the students in their sample can be identified.

These 12 students are then asked what they spent in the last year on family presents.

The students conducting the investigation can then record the data.

Random numbers can also be generated with the aid of a CAS calculator.

studyon

Units:	3 & 4
AOS:	DA
Topic:	5
Concept:	2

Concept summary
Read a summary of this concept.

See more
Watch a video about simple random samples.

WORKED EXAMPLE 25

Generate 5 random numbers (integers) between 1 and 50.

THINK	WRITE
1 From the calculator menu choose probability and random integer.	
2 Since the numbers be generated are to be between 1 and 50 and there are to be 5 numbers enter 1, 50 and 5 in that order.	rand Int (1, 50, 5)
3 An example of a set of numbers is displayed.	{48, 46, 8, 26, 21}.

Displaying the data

The raw data are given below in dollars.

25 30 35 38 34 22 30 40 35 25 32 40

Since there are not many responses, a stem plot is an appropriate way of displaying the data.

Stem	Leaf
2	2
2*	5 5
3	0 0 2 4
3*	5 5 8
4	0 0

Key: 2|2 = 22 dollars
2*|5 = 25 dollars

To summarise and comment further on the sample, it is useful to use some of the summary statistics covered earlier in this chapter. The most efficient way to calculate these is to use a CAS calculator. Using the steps outlined in the previous sections, we obtain a list of summary statistics for these data.

$$\bar{x} = 32.2$$
$$s = 6$$
$$Q_1 = 27.5$$
$$\text{median} = 33$$
$$Q_3 = 36.5$$

To measure the centre of the distribution, the median and the mean are used. Since there are no outliers and the distribution is approximately symmetric, the mean is quite a good measure of the centre of the distribution. Also, the mean and the median are quite close in value.

To measure the spread of the distribution, the standard deviation and the interquartile range are used. Since $s = 6$, and since the distribution is approximately bell-shaped, we would expect that approximately 95% of the data lie between $32.2 + 12 = 44.2$ and $32.2 - 12 = 20.2$. It is perhaps a little surprising to think that 95% of students spend between \$20.20 and \$44.20 on family presents. One might have expected there to be greater variation on what students spend. The data, in that sense, are quite bunched.

The interquartile range is equal to $36.5 - 27.5 = 9$. This means that 50% of those in the sample spent within \$9 of each other on family presents. Again, one might have expected a greater variation in what students spent. It would be interesting to know whether students confer about what they spend and therefore whether they tended to allocate about the same amount of money to spend.

At another school, the same investigation was undertaken and the results are shown in the following stem plot.

Stem	Leaf
2	0
2*	5 5
3	
3*	5
4	
4*	5 5
5	0 0
5*	5 5
6	0 0
6*	5
7	
7*	5

Key: 2|2 = 22 dollars
2*|5 = 25 dollars

The summary statistics for these data are as follows:

$\bar{x} = 47.5$, $s = 16.3$, $Q_1 = 35$, median = 50, $Q_3 = 60$.

The distribution is approximately symmetric, albeit very spread out. The mean and the median are therefore reasonably close and give us an indication of the centre of the distribution. The mean value for this set of data is higher than for the data obtained at the other school. This indicates that students at this school in this year level, in general, spend more than their counterparts at the other school. Reasons for

this might be that this school is in a higher socio-economic area and students receive greater allowances, or perhaps it is that at this school there is a higher proportion of students from cultures where spending more money on family presents is usual.

The range of money spent on family presents at this school and at this particular year level is $55. This is certainly much higher than at the other school. The interquartile range at this school is $25. That is, the middle 50% of students spend within $25 of each other which is greater than the students at the other school.

Exercise 1J Populations and simple random samples

1 MC Students are selecting a sample of students at their school to complete an investigation. Which of the following are examples of choosing this sample randomly?

- **A** Choosing students queuing at the tuckshop.
- **B** Assigning numbers to a list of student names and using a random number table to select random numbers.
- **C** Calling for volunteers.
- **D** Choosing the girls in an all-girls science class.
- **E** Choosing students in a bus on the way home.

2 Conduct an investigation into how much money students in your year level earn per week (this might be an allowance or a wage). Write a report on your findings, ensuring you include:

- **a** an explanation of the population for your investigation
- **b** the manner in which your sample was selected
- **c** the number in your sample
- **d** your results as raw data
- **e** your results in a stem plot or histogram
- **f** the summary statistics for your data.

Comment on your results based on the summary statistics.

3 Repeat question **2**, but this time investigate the following for students in your year level:

- **a** the number of hours spent on homework each week
- **b** the number of hours spent working in part-time jobs.

4 Conduct a similar investigation to that which you completed in questions **2** and **3**; however, this time sample students in another year group. Compare these data with those obtained for your year level.

Summary

Types of data

- Univariate data are data with one variable. Sets of data that contain two variables are called *bivariate* data and those that contain more than two variables are called *multivariate* data.
- Numerical data involve quantities that are measurable or countable.
- Categorical data, as the name suggests, are data that are divided into categories or groups.
- Discrete data are produced when a variable can take only certain fixed values.
- Continuous data are produced when a variable can take any value between two values.

Stem plots

- A stem-and-leaf plot (or stem plot) is a useful way of displaying data containing up to about 50 observations.
- A stem plot is constructed by breaking the numerals of a record into two parts: a 'stem' and a 'leaf'. The last digit is always the 'leaf' and any preceding digits form the stem.
- When asked to represent data using a stem-and-leaf plot, it is always assumed that the stem-and-leaf plot will be ordered.
- If data are bunched then it may be useful to break the stems into halves or even fifths.

Dot plots, frequency histograms and bar charts

- On a frequency histogram, the vertical axis displays the frequency and the horizontal axis displays the class intervals.
- Data given in raw form should be summarised first in a frequency table.

Describing the shape of stem plots and histograms

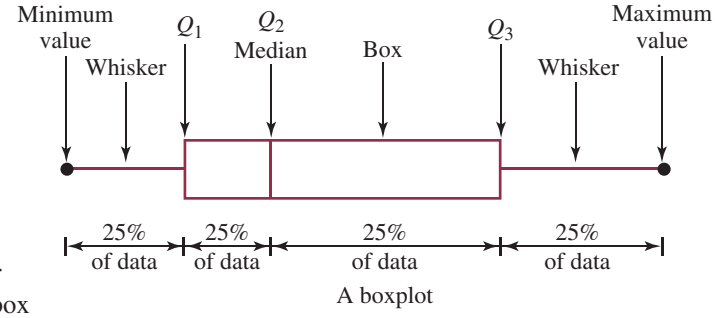
- When data are displayed in a histogram or a stem plot, we say that the distribution of those data is:
  1. *symmetric* if there is a single peak and the data trail off on either side of this peak in roughly the same fashion
  2. *negatively* skewed if the data peak to the right and trail off to the left
  3. *positively skewed* if the data peak to the left and trail off to the right.

The median, the interquartile range, the range and the mode

- The *median* is the midpoint of a set of data. Half the data are less than or equal to the median. When there are n observations in a set of ordered data, the median can be located at the $\left(\frac{n+1}{2}\right)$th position.
- The *interquartile range* IQR $= Q_3 - Q_1$.
- The interquartile range gives us the range of the middle 50% of values in our set of data.
- There are four steps to locating Q_1 and Q_3.
  Step 1: Write down the set of data in ordered form from lowest to highest.
  Step 2: Locate the median, that is, locate Q_2.
  Step 3: Now consider just the lower half of the set of data. Find the middle score. This score is Q_1.
  Step 4: Now consider just the upper half of the set of data. Find the middle score. This score is Q_3.
- The *range* of a set of data is the difference between the highest and lowest values in that set.
- The mode is the score that occurs most often. If there is more than one score with the highest frequency, then all scores with that frequency are the modes.

Boxplots

- When data are displayed in a boxplot we say that the distribution of the data is:
  1. *symmetric* if the whiskers are about the same length and the median is about halfway along the box
  2. *negatively skewed* if the left-hand whisker is longer than the right-hand whisker and the median occurs closer to the right-hand end of the box



A boxplot

3. *positively skewed* if the left-hand whisker is shorter than the right-hand whisker and the median occurs closer to the left-hand end of the box.
4. An outlier is a score, x, which lies *outside* the interval:

$$Q_1 - 1.5 \times \text{IQR} \le x \le Q_3 + 1.5 \times \text{IQR}$$

The mean

- The mean is given by $\bar{x} = \dfrac{\sum x}{n}$ where $\sum x$ represents the *sum* of all the observations in the data set and n represents the *number* of observations in the data set.
- The mean is calculated by using the values of the observations and because of this it becomes a less reliable measure of the centre of the distribution when the distribution is skewed or contains an outlier.
- To find the mean for grouped data, $\bar{x} = \dfrac{\sum(f \times m)}{\sum f}$ where f represents the frequency of the data and m represents the midpoint of the class interval of the grouped data.
- The more symmetrical the distribution, the closer the value of the mean is to the median.

Standard deviation

- The standard deviation is a measure of the spread of data from the mean. The symbol for standard deviation is s.
- $s = \sqrt{\dfrac{\sum(x - \bar{x})^2}{n - 1}}$

  where: $\sum$ represents 'the sum of'
  x represents an observation
  $\bar{x}$ represents the mean
  n represents the number of observations
- The larger the standard deviation, the more spread are the data from the mean.

The 68–95–99.7% rule and *z*-scores

- The 68–95–99.7% rule for a bell-shaped curve states that:
  1. approx. 68% of data lie within 1 standard deviation either side of the mean
  2. approx. 95% of data lie within 2 standard deviations either side of the mean
  3. approx. 99.7% of data lie within 3 standard deviations either side of the mean.
- The z-score is used to measure the position of a score in a data set relative to the mean.
- The formula used to calculate the z-score is $z = \dfrac{x - \bar{x}}{s}$, where x is the score, $\bar{x}$ is the mean and s is the standard deviation.
- Scores can be compared more accurately by their z-scores, which consider the mean and the standard deviation of the data set in their calculations.
- When comparing scores, read the question carefully to see if a higher or lower z-score is a better outcome.

Populations and simple random samples

- A *population*, in statistics, is a group of people (or objects) to whom you can apply any conclusions or generalisations that you reach in your investigation.
- A *sample*, in statistics, is a smaller group of people (or objects) who have been chosen from the population and are involved in the investigation.
- A *simple random sample* is a random selection from the population such that every member of that population has an equal chance of being chosen in the sample and the choice of one member does not affect the choice of another member.

Chapter review

MULTIPLE CHOICE

1 The best distances that a group of twenty 16-year-old competitors achieved in the long jump event at an athletics meeting are recorded.
This is an example of:

A discrete, numerical data
B continuous, numerical data
C categorical data
D discrete, categorical data
E continuous, categorical data

2 The observations shown on the stem-and-leaf plot below right are:

A 20 21 26 27 28 29 30 31 35
B 20 20 21 26 27 28 29 30 31 31 35
C 20 21 26 27 28 29 30 31 35
D 20 20 21 26 27 28 29 29 30 31 31 35
E 20 21 26 27 28 29 29 30 31 35

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

Key: 2|1 = 21

3 The number of people attending 25 of the sessions at an outside court, which has a seating capacity of 150, during the Australian Open Tennis Tournament are displayed in the stem plot at right. Which of the following statements is untrue about the data?

A The smallest number of people attending was 85.
B Only during six sessions did attendance fall below 100.
C The largest number of people attending was 140.
D On six occasions the number of people attending was more than 130.
E On one occasion the number of people attending was only eight less than the seating capacity.

Stem	Leaf
8	5 9
9	2 3 4 9
10	5 5 8
11	0 1 6 6 7
12	4 7 7 8 8
13	5 7 9 9
14	0 2

Key: 9|2 = 92

4 Which one of the following frequency tables accurately summarises the scores shown below?

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

A

Score	Frequency
1	2
2	3
3	2
4	5
5	2
6	4
7	6
8	3
9	2

B

Score	Frequency
1	3
2	2
3	3
4	1
5	5
6	2
7	3
8	4
9	3

C

Score	Frequency
1	3
2	2
3	1
4	3
5	1
6	5
7	3
8	2
9	4

D

Score	Frequency
1	2
2	3
3	3
4	4
5	1
6	3
7	4
8	3
9	4

E

Score	Frequency
1	3
2	4
3	3
4	2
5	2
6	1
7	3
8	2
9	1

5 The distribution of data shown in the stem plot at right could best be described as:

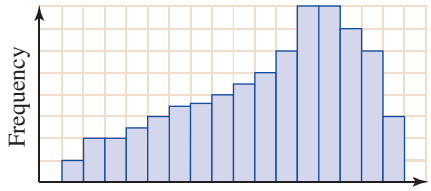
A negatively skewed
B negatively skewed with one outlier
C positively skewed
D positively skewed with one outlier
E symmetric

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

Key: 3|1 = 31

6 The distribution of the data shown in the histogram below could best be described as:

A negatively skewed
B negatively skewed with one outlier
C positively skewed
D positively skewed with one outlier
E symmetric



7 A set of data contains 7 observations and has a median of 5 and a range of 3. The set of data could be:

A 4 4 5 6 7
B 1 1 2 3 4 5 6
C 4 5 5 5 6 7 7
D 1 3 5 5 5 6 7
E 3 5 7

8 The median of the set of data shown in the stem plot below is:

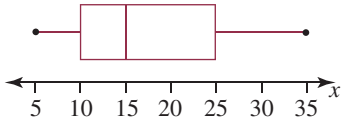
Stem	Leaf
1	2 3
2	0 4 5 7
3	1 2 5 9
4	1 3 6 7
5	2 9 9
6	3

Key: 2|4 = 24

A 5
B 7
C 9
D 9.5
E 37

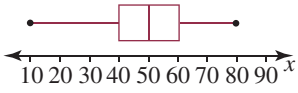
9 For the distribution shown in this boxplot, it is true to say that:

A the range is 35
B the interquartile range is 10
C the median is 20
D the interquartile range is 25
E the median is equal to the interquartile range

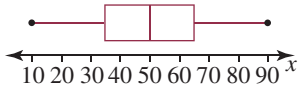


10 A distribution has a range of 80, an interquartile range of 30 and a median of 50. Which one of the following boxplots could represent this distribution?

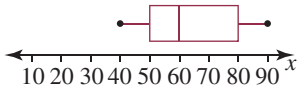
A



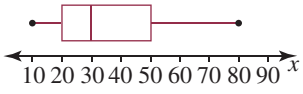
B



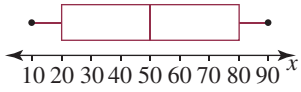
C



D

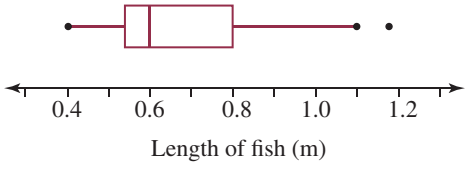


E



11 The boxplot at right represents the lengths of barracuda caught by fishing boats during one day. Which one of the following statements is not true about these data?

A The data contain an outlier.
B The shortest length is 0.4 m.
C The median is 60 cm.
D The interquartile range is 0.2 m.
E The distribution is positively skewed.



12 For the following set of data, 14 18 20 21 23 23 24 25 29 30, the mean is:

A 10 **B** 22.666 666 **C** 22.7 **D** 23 **E** 24.222 222

13 The ages of a group of students entering university for the first time is shown on the stem plot below. What is the mean age?

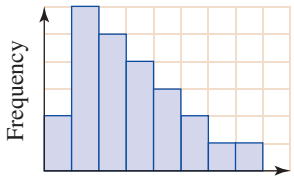
A 18
B 18.9
C 19
D 20.9
E 21

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

Key: 1*|5 = 15 years

14 In which case below would you expect the mean to be greater than the median?

A



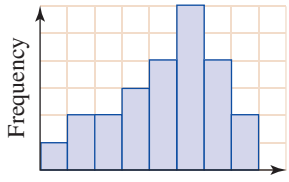
B

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

Key: 2|4 = 24

C The data:
11 13 16 17 18 18 19 20

D



E

15 In which case in question **14** is the median not necessarily the better measure of the centre of the data?

16 The Millers obtained a number of quotes on the price of having their home painted. The quotes, to the nearest hundred dollars, were:

4200	4800
5100	5000
4700	4700
4600	4900

The standard deviation for this set of data, to the nearest whole dollar, is:

A 277 **B** 278 **C** 324 **D** 325 **E** 4750

17 The number of Year 12 students who spent their spare periods studying in the resource centre during each week of terms 3 and 4 is shown on the stem plot at right.

The standard deviation for this set of data, to the nearest whole number is:

A 10 **B** 12 **C** 14 **D** 17 **E** 35

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

Key: 2|5 = 25 students

18 The lifetime (in hours) of a particular type of battery is known to have a distribution which is bell-shaped. A large number of batteries of this type are sampled and are found to have a mean lifetime of 1200 hours and a standard deviation of 10 hours. We would expect that approximately 95% of the batteries in the sample would have a lifetime (in hours) between:

A 10 and 1200 **B** 1170 and 1230 **C** 1200 and 1210
D 1180 and 1220 **E** 1190 and 1210

19 A set of marks from a maths test has a mean of 45 and a standard deviation of 5. The distribution of the marks is known to be bell-shaped. Which of the following statements is false?

A Approximately 68% of the marks lie between 40 and 50.
B The distribution is approximately symmetric.
C Approximately 95% of the marks lie between 35 and 50.
D A mark in the twenties would be most unusual.
E Approximately 99.7% of the marks lie between 30 and 60.

20 The mean length of a large batch of broom handles is 120 cm. The data have a standard deviation of 3 cm. The percentage of broom handles, in this batch, which are shorter than 114 cm is:

A 0.15% **B** 2.5% **C** 13.5%
D 16% **E** 34%

21 The mean birth weight of babies at a hospital is 2.8 kg with a standard deviation of 0.4 kg. The standardised score for a weight of 3.3 kg would be:

A 0.73
B $^{-}1.25$
C 1.04
D 1.25
E $^{-}1.04$

SHORT ANSWER

1 Write an example of a variable which produces:

a categorical data
b numerical data that are:
i discrete **ii** continuous.

2 The money (rounded to the nearest whole dollar) raised by fifteen Year 12 students is shown below.

78 84 61 73 71 83 87 65 60 67 71 82 84 79 78

Construct a stem plot for the amount raised using:

a the stems 6, 7 and 8 **b** the stems 6, 7 and 8 split into halves
c the stems 6, 7 and 8 split into fifths.

Discuss your results.

3 **a** The following frequency table shows the speeds of cars recorded by police. The cars were travelling through a 60 km/h zone. Construct a histogram to display the data.

Class interval	Frequency
50–51.9	3
52–53.9	5
54–55.9	6
56–57.9	7
58–59.9	9
60–61.9	10
62–63.9	9
64–65.9	10
66–67.9	8
68–69.9	5
70–71.9	3
72–73.9	4
74–75.9	2

b Check your work using a CAS calculator.
c Draw two conclusions about these data.

4 The money raised (to the nearest whole dollar) by each student in a Year 3 class on the school walkathon is shown in the stem plot at right.

a Describe the shape of the distribution of these data.

b Describe how this distribution would need to change for it to become a symmetric one.

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

Key: 0|8 = $8

5 Find the range, the median, the mode and the interquartile range of this set of data.

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

Key: 1|4 = 14

6 **a** For the set of data below, construct a boxplot to display the distribution.

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

b Describe the shape of the distribution.

7 The ages of a group of people attending a classical music recital are shown in the frequency table below.

Age (class interval)	Frequency
40–44	4
45–49	6
50–54	11
55–59	16
60–64	23
65–69	18
70–74	10
75–79	2

Calculate the mean age of those attending the recital.

8 A chemical component is added to a filtering system on a weekly basis. The amount of chemical component required each week varies. The amounts required (in mL) over the past 20 weeks are shown in the stem plot below.

Calculate to 2 decimal places the standard deviation of the amounts used.

Stem	Leaf
2	1
2	2 2
2	4 4 4 5
2	6 6
2	8 8 9 9
3	0
3	2 2
3	4 5
3	6
3	8

Key: 3|8 = 0.38 mL

9 The life spans of dogs of a particular breed follow a bell-shaped distribution. A group of this particular breed at a dog club was found to have a mean life span of 12 years with a standard deviation of 1.2 years.

a For this group, write down the expected values between which the life spans of approximately:

- **i** 68% of the dogs would lie
- **ii** 95% of the dogs would lie
- **iii** 99.7% of the dogs would lie.

b What does this information suggest about this breed?

10 Ricardo scored 85 on an entrance test for a job. The test has a mean score of 78 and a standard deviation of 8. Kory sits a similar test and scores 27. In this test, the mean is 18 and the standard deviation is 6. Based on this test, who is the better candidate for the job? Explain your answer.

EXTENDED RESPONSE

1 Mr Fahey gives the same test to the two Year 10 classes that he teaches, 10C and 10E. The test is out of 20. The results in 10C are:

4	7	7	9	9	10	10	11	12
13	14	14	15	15	15	16	17	17
17	17	18	18	18	19	19		

The results in 10E are:

8	9	10	11	11	12	12	12	13
13	13	13	13	14	14	14	14	14
14	15	15	15	16	16	19		

a For each of these sets of data:

- **i** display the data using a histogram or a stem plot. Give reasons for your choices. Also describe the shapes of the distribution
- **ii** calculate the median, the interquartile range, the range and the mode
- **iii** represent each set of data using a boxplot
- **iv** calculate the mean and the standard deviation
- **v** state whether the mean or the median is a better measure of the centre of each distribution
- **vi** comment as to whether the 68–95–99.7% rule can be applied to either of the distributions.

b Using the summary statistics you have calculated, comment on and compare the performances of 10C and 10E on the test.

2 A group of office workers and a group of sports instructors were asked to complete 5 minutes of exercise as part of a study of heart rates. Following the exercise, participants rested for 2 minutes before their pulse rates were measured. The results are set out below in the stem plots.

Pulse rates for office workers (beats/min)

Stem	Leaf
7	6
8	
9	5
10	6 7
11	0 2
12	0 1 2 4 6 7 9
13	0 0 4

Pulse rates for sports instructors (beats/min)

Stem	Leaf
6	2 4 8 8 9
7	2 2 3 5 7 9
8	2 8
9	6
10	8

Key: 12|4 = 124 beats/min

a Describe the shape of each distribution.

b Calculate the median, the interquartile range, the mode and the range for both.

c Represent each set of data using a boxplot.

d Calculate the mean and the standard deviation for both sets of data.

e Comment as to whether the 68–95–99.7% rule can be applied to either of the distributions.

f Use the summary statistics that you have calculated to comment on the pulse rates of each group, noting any differences between the two.

3 A hatch of Atlantic salmon has been reared in a coastal environment over a period of 12 months. The lengths (to the nearest cm) of a sample of 20, out of the total number of 10 000 fish, are shown below.

13	16	17	14	16	19	15	17	16	15
16	18	16	13	17	14	18	15	19	16

a Describe the type of data that the variable produces.
b Construct an appropriate stem plot from these data and use it to describe the shape of the distribution.
c Using your stem plot, calculate the five number summary statistics and then draw a boxplot.
d Describe the shape of the distribution from the boxplot.
e Does the stem plot or boxplot give a better indication of the distribution's shape?
f For a symmetric distribution the mean is the same as the median. Is that the case here?
g Given that the distribution is symmetric, the whole population of these salmon would form a normal or bell-shaped distribution. Find the standard deviation (to 2 decimal places) for this sample and use it, along with the mean, to find the number of fish with lengths greater than 17.75 cm.

The same number of salmon was reared in a river environment over the same period of time. The lengths of 20 fish in a sample are shown below.

18	20	17	19	16	19	19	17	16	18
19	18	12	18	17	14	18	15	19	17

h Use an appropriate method to help you describe the shape of this distribution.
i Determine how many of this population of 10 000 salmon would have a length greater than 19.25 cm (calculate the standard deviation to 2 decimal places).
j Comment on the growth of each hatch of salmon over the 12 months.

4 The birth weights (in kg) of 50 of the 220 babies that were born at a hospital during a one-month period are listed below.

2.9	2.7	3.1	2.5	2.4	2.6	2.9	2.6	3.2	4.1
2.3	2.8	2.4	3.2	2.7	2.5	2.6	2.9	3.0	2.2
3.4	3.1	3.3	2.9	3.2	2.9	3.4	3.1	2.3	3.5
3.1	3.0	2.9	3.6	3.1	2.7	2.6	1.8	1.9	3.6
2.0	3.4	3.5	2.4	3.5	3.0	2.2	2.8	3.5	3.1

a Construct a frequency histogram for the data using class intervals of 1.5–1.9, 2.0–2.4, 2.5–2.9 and so on.
b Comment on the shape of the distribution.
c It has been said that the mean birth weight of babies is 3 kg. Using the data given, comment on this statement.
d Using the mean and standard deviation (to 2 decimal places) for this sample of 50 birth weights, determine how many of the 220 babies born at the hospital had weights:
i between 2.35 kg and 3.43 kg **ii** between 3.43 kg and 3.97 kg **iii** greater than 3.97 kg.

eBook*plus*

DIGITAL DOC
doc-9408
Test Yourself
Chapter 1

study on

Units: 3 & 4
AOS: DA

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

eBookplus ICT activities

Chapter opener

DIGITAL DOC

- 10 Quick Questions doc-9399: Warm up with a quick quiz on univariate data. *(page 1)*

1A Types of data

DIGITAL DOC

- WorkSHEET 1.1 doc-9400: Apply your knowledge of univariate data to construct and analyse stem plots. *(page 2)*

1C Dot plots, frequency histograms and bar charts

DIGITAL DOCS

- Spreadsheet doc-9401: Create a segmented bar chart. *(page 9)*
- SkillSHEET 1.1 doc-9402: Practise converting a fraction into a percentage. *(page 10)*
- Spreadsheet doc-9403: Conduct a survey and plot your results on a histogram. *(page 10)*

1D Describing the shape of stem plots and histograms

DIGITAL DOC

- WorkSHEET 1.2 doc-9404: Apply your knowledge of univariate data to construct and analyse stem plots and boxplots and to construct histograms. *(page 14)*

TUTORIAL

- WE7 eles-1254: Learn how to describe the shape of stem plots and bar charts. *(page 12)*

1E The median, the interquartile range, the range and the mode

DIGITAL DOC

- Spreadsheet doc-9405: Conduct a survey and find the median of a set of data. *(page 18)*

TUTORIAL

- WE9 eles-1255: Discover how to calculate the interquartile range of a set of univariate data. *(page 16)*

INTERACTIVITY

- Measures of centre int-0084: Use the interactivity to calculate the mean, median and mode of a set of univariate data. *(page 15)*

1F Boxplots

DIGITAL DOC

- Spreadsheet doc-9406: Conduct a survey and use a spreadsheet to study the effect of uniformly spread data. *(page 23)*

TUTORIAL

- WE13 eles-1256: Learn how to construct a boxplot using a CAS calculator. *(page 21)*

1G The mean

TUTORIAL

- WE16 eles-1257: Watch a tutorial on calculating the mean using data in a frequency table. *(page 26)*

1I The 68–95–99.7% rule and *z*-scores

DIGITAL DOC

- SkillSHEET 1.2 doc-9407: Refine your knowledge of percentages. *(page 38)*

TUTORIAL

- WE20 eles-1258: See how normal distributions can be used to determine percentages above or below a certain mass. *(page 33)*

INTERACTIVITY

- The 68–95–99.7% rule and *z*-scores int-0182: Use the interactivity to consolidate your understanding of the normal distribution and confidence intervals. *(page 32)*

Chapter review

DIGITAL DOC

- Test Yourself doc-9408: Take the end-of-chapter test to test your progress. *(page 51)*

To access eBookPLUS activities, log on to www.jacplus.com.au

Answers CHAPTER 1

UNIVARIATE DATA

Exercise 1A Types of data

1 Numerical **a**, **b**, **c**, **g**, **h**
Categorical **d**, **e**, **f**, **i**, **j**, **k**, **l**, **m**

2 Discrete **c**, **g**, **m**
Continuous **a**, **b**, **h**

3 C

4 C

Exercise 1B Stem plots

1 **a** 1 2 5 8 12 13 13 16 16 17 21 23 24 25 25 26 27 30 32

b 10 11 23 23 30 35 39 41 42 47 55 62

c 101 102 115 118 122 123 123 136 136 137 141 143 144 155 155 156 157

d 50 51 53 53 54 55 55 56 56 57 59

e 1 4 5 8 10 12 16 19 19 21 21 25 29

2

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

Key: |0 = $5

Busker's earnings are inconsistent.

3

Stem	Leaf
3	7 9
4	2 9 9
5	1 1 2 3 7 8 9
6	1 3 3 8

Key: |7 = 37 years

It seems to be an activity for older people.

4 C

5

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

Key: |2 = 25 years

Ages are spread considerably; not all parents are young.

6

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

Key: |2 = 21 hit outs

Bulldogs, Melbourne, St Kilda

7

Stem	Leaf
18	5 7 9
19	1 5 6 6 7 9
20	1 3 3 5 9
21	7
22	1

Key: |8 = 1.85 m

8

Stem	Leaf
51	5 7
52	5 6
53	6
54	8
55	
56	
57	6
58	3
59	
60	0 0 8
61	
62	0 8
63	3 8
64	
65	2 3
66	
67	0 0 0

Key: |5 = $515 000

There are two groups, one with house prices between $510 000 and $550 000 and the other with prices between $580 000 and$670 000.

9 **a**

Stem	Leaf
4	3 7 7 8 8 9 9 9
5	0 0 0 0 1 2 2 3

Key: |4 = 43 cm

b

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

Key: |4 = 43 cm

c

Stem	Leaf
4	
4	3
4	
4	7 7
4	8 8 9 9 9
5	0 0 0 0 1
5	2 2 3
5	
5	
5	

Key: |4 = 43 cm

10 **a**

Stem	Leaf
1	5 6 7 7 7 8 9 9 9 9
2	0 0 0 1 1 1 2 3 3 3

Key: |1 = 15 mm

b

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

Key: |1 = 15 mm

c

Stem	Leaf
1	
1	
1	5
1	6 7 7 7
1	8 9 9 9 9
2	0 0 0 1 1 1
2	2 3 3 3
2	
2	
2	

Key: |1 = 15 mm

Values are bunched together; they vary little.

Exercise 1C Dot plots, frequency histograms and bar charts

1 **a**

Class	Frequency
1–1.9	1
2–2.9	2
3–3.9	2
4–4.9	6
5–5.9	5
6–6.9	1

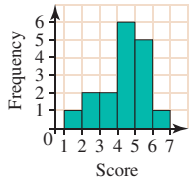
b

Class interval	Frequency
10–14	3
15–19	9
20–24	10
25–29	10
30–34	10
35–39	1

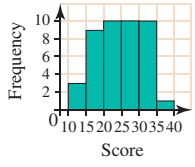
c

Score	Frequency
0.3	1
0.4	2
0.5	1
0.6	1
0.7	1
0.8	2
0.9	2
1.0	2
1.1	1
1.2	1
1.3	1

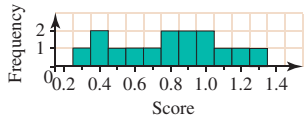
2 **a**



b

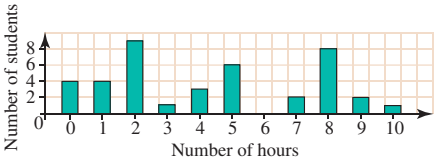


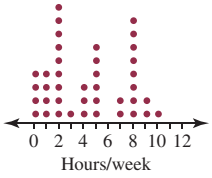
c



3 Check your histograms against those shown in question **2** answers.

4

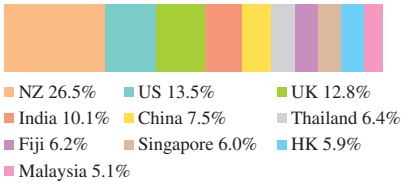




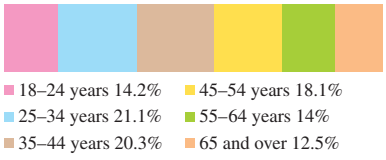
5 a

NZ	26.5%
US	13.5%
UK	12.8%
India	10.1%
China	7.5%
Thailand	6.4%
Fiji	6.2%
Singapore	6.0%
HK	5.9%
Malaysia	5.1%

b



6 Participation in activities



The statement seems untrue as there are similar participation rates for all ages. However, the data don't indicate types of activities.

Exercise 1D Describing the shape of stem plots and histograms

1 a Symmetric
b Negatively skewed
c Positively skewed
d Symmetric
e Symmetric
f Positively skewed

2 a Symmetric, no artliers
b Symmetric, no artliers
c Symmetric, no artliers
d Negatively skewed, no artliers
e Negatively skewed, no artliers
f Positively skewed, no artliers

3 E

4 C

5 Negatively skewed

6 Positively skewed. This tells us that most of the flight attendants in this group spend a similar number of nights (between 2 and 5) interstate per month. A few stay away more than this and a very few stay away a lot more.

7 a Symmetric
b This tells us that there are few low-weight dogs and few heavy dogs but most dogs have a weight in the range of 10 to 19 kg.

8 a Symmetric
b Most students receive about \$8 (give or take \$2).

9 a Positively skewed **b i** 15 **ii** 85%

Exercise 1E The median, the interquartile range, the range and the mode

1

	Median	Range	Mode
a	37	56	38, 49
b	5	17	5
c	11	18	8, 11
d	42.5	18	43
e	628	72	613, 628, 632

2

	Median	Range
a	6	7
b	17	9
c	6	6
d	10	13
e	18.5	14
f	4	7
g	19	17
h	4.5	9
i	23	21

3 a 10 **b** 8
c The IQRs (middle 50%) are similar for the two restaurants but there is no indication of the actual numbers of cars.

4 An example is 2 3 6 8 9. There are many others.

5 a Yes. The lowest score occurs several times. An example is 2 2 2 3 5 6.
b Yes. There are several data points that have the median value. An example is 3 5 5 5 5 5 7.

6 C

7

	Median	Interquartile range	Range	Mode
a	21	18	45	15, 23, 32
b	27.5	8	20	29
c	3.7	3	5.9	3.7

8

	Median	Interquartile range	Range	Mode
a	42	21	91	46
b	32	7	30	34

The data in set **a** have a greater spread than in set **b**, although the medians are similar. The spread of the middle 50% (IQR) of data for set **a** is bigger than for set **b** but the difference is not as great as the spread for all the data (range).

Exercise 1F Boxplots

1

	Range	Interquartile range	Median
a	12	6	8
b	7	2	5
c	350	100	250
d	100	30	65
e	20	10	25

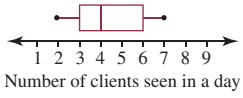
2 a iii **b iv** **c i** **d ii**

3 The boxplots should show the following:

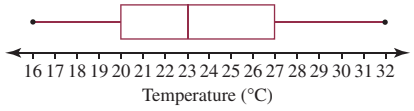
	Minimum value	Q_1	Median	Q_3	Maximum value
a	3	6	8.5	14	18
b	3	5	7	9	12
c	4.3	4.6	5	5.4	5.6
d	11	15.5	18	20	22
e	0.4	0.7	0.9	1.1	1.3

4 D

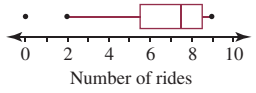
5



6



7 a



The data are negatively skewed with an outlier on the lower end. The reason for the outlier may be that the person wasn't at the show for long or possibly didn't like the rides.

8 a Two similar properties: both sets of data have the same minimum value and similar IQR value.
b Boys IQR = 16
Girls IQR = 16.5
c The reason for an outlier in the boys' data may be that the student did not understand how to do the test, or he stopped during the test rather than working continuously.

Exercise 1G The mean

1 a 7.2 **b** 7.125 **c** 4.9875
d $16.\dot{7}$ **e** $0.88\dot{1}\dot{8}$

2 a $1.078\dot{3}$ No, because of the outlier.
b 17 Yes
c 30.875 Yes
d 15.57 No, because of the outlier.

3 12 **4** D **5** A

6 a Median **b** Mean
c Median **d** Median

7 a 36.09 **b** 16.63
c 168.25 **d** 18.55

8 a $\bar{x} = 24.4$ **b** median = 22
The distribution is positively skewed — confirmed by the table and the boxplot.

Exercise 1H **Standard deviation**

1 a 1.21 b 2.36 c 6.01
d 2.45 e 0.06
2 0.48% 3 0.06 m
4 0.51 seconds
5 15.49 6 C

Exercise 1I **The 68–95–99.7% rule and *z*-scores**

1 a Yes b Yes c No
d No e No f Yes
2 a 8 and 12
b 6 and 14
c 4 and 16
3 a 3.7 and 6.3
b 2.4 and 7.6
c 1.1 and 8.9
4 a 35 s and 63 s
b 21 s and 77 s
c 7 s and 91 s
5 a 1.3 mma nd2.5 mm
b 0.7 mma nd3. 1 mm
c 0.1 mma nd3. 7 mm
6 a 11.7 N and 12.3 N
b 11.4 N and 12.6 N
c 11.1 N and 12.9 N
7 a 5 and 9 b 3 and 11
c 1 and 13
8 E
9 C
10 a 84% b 2.5%
c 84% d 97.35%
11 a 0.15% b 2.5%
c 84% d 83.85%
e 81.5%
12 C
13 a 336 b 10 c 380
14 a i 1360 ii 1950 iii 317
b 100
15 3
16 $^{-}2$
17 $^{-}0.27$
18 1.5
19 a $^{-}0.48$ b 1.44 c 0.08
d $^{-}2.24$ e 2.8
20 a $\bar{x} = 19.55$, $s = 1.76$ b 1.68
21 B
22 B
23 a English 1.25, Maths 1.33
b Maths mark is better as it has a higher *z*-score.
24 Second test, Barbara's *z*-score was $^{-}0.33$ compared to $^{-}0.5$ in the first test.
25 a Barn $\bar{x} = 4.4$ $s = 0.3$
FR $\bar{x} = 4.1$ $s = 0.2$
b $z = 1.18$
c 84%
d

	Cage	Barn	Free range
Xmin	4.7	3.9	3.8
Q_1	5	4.1	4
med	5.15	4.35	4.1
Q_3	5.5	4.6	4.2
Xmax	5.8	4.9	4.4

i boxplots will show:

Cage	5.15
Barn	4.35
FR	4.1

ii It could be concluded that the more space a chicken has, the fewer eggs it lays because the median is greatest for cage eggs.

Exercise 1J **Populations and simple random samples**

1 B
2 Answers will vary.
3 Answers will vary.
4 Answers will vary.

CHAPTER REVIEW

MULTIPLE CHOICE

1 B	2 D	3 C	4 D	5 C
6 A	7 C	8 E	9 E	10 B
11 D	12 C	13 D	14 A	15 B
16 B	17 C	18 D	19 C	20 B
21 D				

SHORT ANSWER

1 Many answers possible.
2 a

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

Key: 6|0 = \$6

b

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

Key: 6|0 = \$6

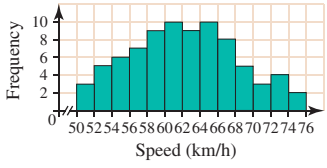
c

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

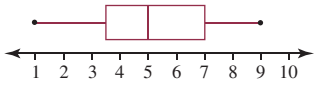
Key: 6|0 = \$6

Stem plot **b** is probably an appropriate display. No real need for stems in fifths.

3 a



c The data are approximately symmetrical. More than half the drivers exceeded the limit. The fastest drivers were about 15 km/h over the limit. Many other conclusions are possible.
4 a Positively skewed
b There would need to be a shift of some of the amounts in the twenties to the thirties and forties.
5 Range = 24, median = 14, mode = 14, interquartile range = 9.5
6 a



b Approximately symmetric
7 60.4 years 8 0.05 mL
9 a i 10.8 and 13.2 years
ii 9.6 and 14.4 years
iii 8.4 and 15.6 years
b There is a large range of life spans for these dogs. The oldest dog is almost twice as old as the youngest.
10 Kory is the better candidate as he has a greater *z*-score (1.5 compared with 0.875).

EXTENDED RESPONSE

1 a i A stem plot is more appropriate since there are only 25 observations in each set.

10C

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

10E

Stem	Leaf
0	
0	
0	
0	
0	8 9
1	0 1 1
1	2 2 2 3 3 3 3 3
1	4 4 4 4 4 4 5 5 5
1	6 6
1	9

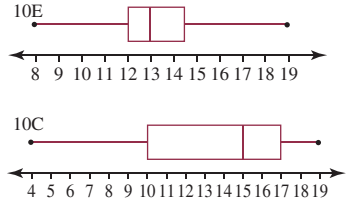
Key: 1|3 = 13

The distribution of 10C is negatively skewed with no outliers.
The distribution of 10E is symmetric with no outliers.

ii

Class	Median	Interquartile range	Range	Mode
10C	15	7	15	17
10E	13	2.5	11	14

iii



iv

Class	Mean	Standard deviation
10C	13.64	4.24
10E	13.2	2.35

v For 10C the median provides a better indication, whereas for 10E the mean and the median are close anyway.

vi The 68–95–99.7% rule can be applied to the distribution of 10E's data since it is approximately bell-shaped.

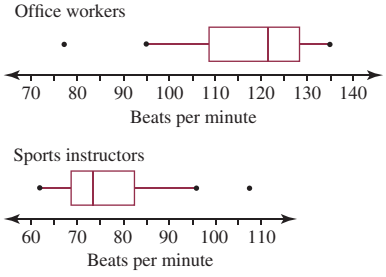
b We use the median of the 10C scores to give us an indication of the centre of the distribution. The median of the 10C scores is 15. For 10E, the distribution is symmetric and hence we use the mean to give us an indication of the centre of the distribution. The mean of 10E scores is 13.2. The range of the 10C scores is 15 whereas for 10E the range is 11. Also, the standard deviation for 10C is 4.24 and for 10E it is 2.35. This means that the scores in 10C are more spread out than those in 10E, which are relatively bunched. So, while in 10C there are more students with higher marks than in 10E, the range of marks in 10C is greater and this would make it a more challenging class to teach.

2 a Office workers: negatively skewed with outlier. Sports instructors: positively skewed with outlier.

b

	Office workers	Sports instructors
Median	121.5 beats/min	73 beats/min
IQR	19.5 beats/min	14 beats/min
Range	58 beats/min	46 beats/min
Mode	130 beats/min	68, 72 beats/min

c



d

Office workers	Sports instructors
$\bar{x} = 116.8$ beats/min	$\bar{x} = 76.9$ beats/min
$s = 15.3$ beats/min	$s = 12.4$ beats/min

e Not used since the distributions are not bell-shaped.

f Office workers: Pulse rates are generally very high, clustered around 120–130 beats/min. Also, there is one person whose rate was much lower than the rest. This outlier (76) produces a large range and makes the mean slightly lower than the median. As a result the median is a more appropriate measure of the centre of the data rather than the mean. Sports instructors: Pulse rates are generally low, clustered around 60–70 beats/min, although there are a few people with rates much higher, which makes the mean slightly higher than the median and also produces quite a large range. As a result of the skewed distribution the median is the more appropriate measure of the centre of the data rather than the mean, although there is little difference between these values.

3 a Discrete, numerical data

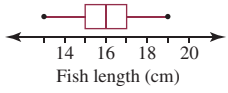
b

Stem	Leaf
1	
1	3 3
1	4 4 5 5 5
1	6 6 6 6 6 6 7 7 7
1	8 8 9 9

Key: 1|3 = 13 cm

Looks to be slightly negatively skewed.

c $Q_2 = 16$ cm
$Q_1 = 15$ cm
$Q_3 = 17$ cm
Lowest score = 13 cm
Highest score = 19 cm



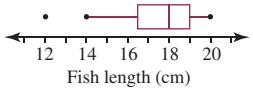
d Symmetric

e Given the data itself, the boxplot does.

f Mean = 16 cm = median

g $s = 1.75$ cm, 1600

h



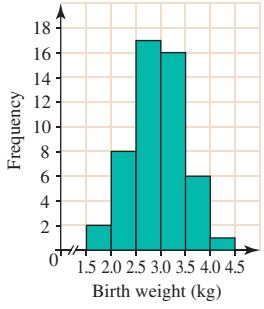
Negatively skewed with an outlier at the lower end

i 1600

j The river fish seem to be larger overall. Only 1600 of the coastal fish lie above 17.75 cm, whereas 1600 of the river fish lie above 19.25 cm. All the quartiles for the river fish are higher than those for the coastal ones. It would seem that the river fish have grown more than the coastal fish.

4 a

Class interval	Tally	Frequency	Mid-point
1.5–1.9	||	2	1.7
2.0–2.4	卌 |||	8	2.2
2.5–2.9	卌 卌 卌 ||	17	2.7
3.0–3.4	卌 卌 卌 |	16	3.2
3.5–3.9	卌 |	6	3.7
4.0–4.4	|	1	4.2



b Somewhat symmetrical or slightly positively skewed

c Mean = 2.89 kg, $s = 0.54$ kg
The mean of the sample is only just less than the suggested mean, in fact it is only about 0.2 of 1 standard deviation away from it. So the suggestion is probably right.

d i 149
ii 29
iii 5