TOPIC 1
Investigating data distributions

1.1 Overview

1.1.1 Introduction



Due to advances in technology, the collection, analysis and interpretation of data is at an unprecedented level in the world today. These advances have ensured that after the collection and analysis of large datasets, the results can be distributed very quickly. It is very important that future planning and decisions are based on current empirical data as change is fast-paced and costly in the modern world. In 1880 a census was conducted in the USA and it took 7 years for the information collected by the census to be published. The very detailed dataset collected by the Australian census in August 2016 was released in June 2017.

Data has the power to initiate change and business and governments ignore it at their peril. Meteorologists, urban planners, health scientists, traffic engineers, pharmaceutical companies, environmental scientists, educators and agriculturalists to name a few, use data to monitor change and make future predictions. These changes may be as small as choosing to block off a suburban street to stop too many people using it as a short cut and creating traffic problems, or as big as ensuring that there is enough food production to feed an ever-increasing world population. Decisions that require major societal change, practice and investment need to be justified. Empirical evidence can provide this justification and therefore it is important that everyone has some knowledge and understanding of the collection and interpretation of data.

COPYRIGHTED MATERIAL

LEARNING SEQUENCE

1.1 Overview
1.2 Types of data
1.3 Stem plots
1.4 Dot plots, frequency tables and histograms, and bar charts
1.5 Describing the shape of stem plots and histograms
1.6 The median, the interquartile range, the range and the mode
1.7 Boxplots
1.8 The mean of a sample
1.9 Standard deviation of a sample
1.10 Populations and simple random samples
1.11 The 68–95–99.7% rule and *z*-scores
1.12 Review: exam practice

Fully worked solutions for this topic are available in the Resources section of your eBookPLUS at www.jacplus.com.au.

1.1.2 Kick off with CAS

Exploring histograms with CAS

Histograms can be used to display numerical data sets. They show the shape and distribution of a data set, and can be used to gather information about the data set, such as the range of the data set and the value of the median (middle) data point.

1. The following data set details the age of all of the guests at a wedding.

Age	Frequency
0–	4
10–	6
20–	18
30–	27
40–	14
50–	12
60–	6
70–	3
80–	2

Use CAS to draw a histogram displaying this data set.



2. Comment on the shape of the histogram. What does this tell you about the distribution of this data set?

3. a. Use CAS to display the mean of the data set on your histogram.

b. What is the value of the mean?

4. a. Use CAS to display the median of the data set on your histogram.

b. What is the value of the median?

on Resources

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

1.2 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.

Data can be classified as either numerical or categorical. The methods we use to display data depend on the type of information we are dealing with.

1.2.1 Types of data

Numerical data is data that has been assigned a numeric value. Numerical data can be:

- **discrete** — data that can be counted but that can have only a particular value, for example the number of pieces of fruit in a bowl
- **continuous** — data that is not restricted to any particular value, for example the temperature outside, which is measured on a continuous scale.



Categorical data is data split into two or more categories. Categorical data can be:

- **nominal** — data that can be arranged into categories but not ordered, for example arranging shoes by colour or athletes by gender
- **ordinal** — data that can be arranged into categories that have an order, for example levels of education from high school to post-graduate degrees.

1.2.2 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. 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. If variables can be measured, then the data are continuous.

WORKED EXAMPLE 1

MC Which of the following is *not* numerical data?

A. Maths test results **B.** Ages **C.** Names of AFL football teams
D. Heights of students in a class **E.** Lengths of bacteria

THINK	WRITE
1. To be numerical data, it is to be measurable or countable.	
2. Look for the data that does not fit the criteria.	A: Measurable B: Countable C: Names, so not measurable or countable D: Measurable E: Measurable

3. Answer the question.	The data that is not numerical is names of AFL football teams. The correct option is C.

WORKED EXAMPLE 2

MC Which of the following is *not* discrete data?
A. Number of students older than 17.5 years old
B. Number of girls in a class
C. Number of questions correct in a multiple choice test
D. Number of students above 180 cm in a class
E. Height of the tallest student in a class

THINK	WRITE
1. To be discrete data, it is to be a whole number (countable).	
2. Look for the data that does not fit the criteria.	A: Whole number (countable) B: Whole number (countable) C: Whole number (countable) D: Whole number (countable) E: May not be a whole number (measurable)
3. Answer the question.	The data that is not discrete is the height of the tallest student in a class. The correct option is E.

studyon

Units 3 & 4 > AOS DA > Topic 1 > Concepts 1 & 2

Classification of data Summary screen and practice questions
Defining univariate data Summary screen and practice questions

Exercise 1.2 Types of data

1. MC WE1 Which of the following is *not* numerical data?
   - **A.** Number of students in a class
   - **B.** The number of supporters at an AFL match
   - **C.** The amount of rainfall in a day
   - **D.** Finishing positions in the Melbourne Cup
   - **E.** The number of coconuts on a palm tree

2. MC Which of the following is *not* categorical data?
- **A.** Preferred political party
- **B.** Gender
- **C.** Hair colour
- **D.** Salaries
- **E.** Religion

3. 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 exhibit 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

4. MC Which of the following represent categorical data?
- **A.** The numbers of subjects offered to VCE students at various schools
- **B.** Life expectancies
- **C.** Species of fish
- **D.** Blood groups
- **E.** Years of birth
- **F.** Countries of birth
- **G.** Tax brackets

5. MC WE2 Which of the following is *not* discrete data?
- **A.** Number of players in a netball team
- **B.** Number of goals scored in a football match
- **C.** The average temperature in March
- **D.** The number of Melbourne Storm members
- **E.** The number of twins in Year 12

6. MC Which of the following is *not* continuous data?
- **A.** The weight of a person
- **B.** The number of shots missed in a basketball game
- **C.** The height of a sunflower in a garden
- **D.** The length of a cricket pitch
- **E.** The time taken to run 100 m

7. For each set of numerical data identified in question **3** above, state whether the data are discrete or continuous.

8. 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



9. 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

10. When reading the menu at the local Chinese restaurant, you notice that the dishes are divided into sections. The sections are labelled chicken, beef, duck, vegetarian and seafood. What type of data is this?
11. MC NASA collects data on the distance to other stars in the universe. The distance is measured in light years. What type of data is being collected?
    A. Discrete **B.** Continuous **C.** Nominal **D.** Ordinal **E.** Bivariate
12. MC The number of blue, red, yellow and purple flowers in an award winning display is counted. What type of data is being collected?
    A. Nominal **B.** Ordinal **C.** Discrete **D.** Continuous **E.** Bivariate
13. Students in a performing arts class watch a piece of modern dance and are then asked to rate the quality of the dance as poor, average, above average or excellent. What type of data is being collected?
14. Given the set of data: 12, 6, 21, 15, 8, 2, describe what type of numerical data this data set is.
15. If a tennis tournament seeds the players to organise the draw, what type of categorical data is this?
16. MC The height of the players in the basketball team is measured. The data collected is best described as:
    A. categorical and nominal **B.** categorical and numerical
    C. numerical and discrete **D.** numerical and continuous
    E. none of the above

1.3 Stem plots

A *stem-and-leaf plot*, or **stem plot** for short, is a way of ordering and displaying a set of data, with the advantage that all of the raw data is kept. Since all the individual values of the data are being listed, it is only suitable for smaller data sets (up to about 50 observations).

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

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

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.

With your stem-and-leaf plot it is important to include a key so it is clear what the data values represent.

In cases where there are numerous leaves attached to one stem (meaning that the data is heavily concentrated in one area), the stem can be subdivided. Stems are commonly subdivided into halves or fifths. By splitting the stems, we get a clearer picture about the data variation.

WORKED EXAMPLE 3

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

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.
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.
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.

WRITE

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

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

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

WORKED EXAMPLE 4

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

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.
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.
3. Construct an ordered stem plot. Provide a key.

WRITE

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

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 7
73	
74	1
75	3 6

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

WORKED EXAMPLE 5

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

on Resources

- **Interactivity:** Stem plots (int-6242)
- **Interactivity:** Create stem plots (int-6495)

studyon

Units 3 & 4 > AOS DA > Topic 2 > Concept 4 **Stem plots** Summary screen and practice questions

Exercise 1.3 Stem plots

1. **WE3** The number of iPads sold in a month from a department store over 16 weeks is shown.

28, 31, 18, 48, 38, 25, 21, 16, 33, 42, 35, 39, 49, 30, 29, 28

Construct a stem plot to display the number of iPads sold over the 16 weeks.

2. The money (correct 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. **WE4** The test scores (as percentages) of a student in a Year 12 Further Maths class are shown.

88.0 86.8 92.1 89.8 92.6 90.4 98.3 94.3 87.7

94.9 98.9 92.0 90.2 97.0 90.9 98.5 92.2 90.8

Display the data in a stem plot.

4. 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

5. 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

6. **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

7. 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 young' (less than 30 years old).

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



8. 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

9. The 2015 weekly median rental price for a 2-bedroom unit in a number of Melbourne suburbs is given in the following table. Construct a stem plot for these data and comment on it.

Suburb	Weekly rental ($)
Alphington	400
Box Hill	365
Brunswick	410
Burwood	390
Clayton	350
Essendon	350
Hampton	430
Ivanhoe	395
Kensington	406
Malvern	415

Suburb	Weekly rental ($)
Moonee Ponds	373
Newport	380
North Melbourne	421
Northcote	430
Preston	351
St Kilda	450
Surrey Hills	380
Williamstown	330
Windsor	423
Yarraville	390

10. **WE5** LeBron James's scores for his last 16 games are shown in the following stem plot.

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

Key: 2|0 = 20



Produce another stem plot for these data by splitting the stems into:

a. halves

b. fifths.

11. The data below give the head circumference (correct 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.

12. For each of the following, write down all the pieces of data shown on the stem plot. The key used for each stem plot is 3|2 = 32.

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

13. 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.

14. The first 20 scores that came into the clubhouse in a local golf tournament are shown. Construct a stem plot for these data.

102 98 83 92 85 99 104 112 88 91

78 87 90 94 83 93 72 100 92 88

15. Golf handicaps are designed to even up golfers on their abilities. Their handicap is subtracted from their score to create a net score. Construct a stem plot on the golfer's net scores below and comment on how well the golfers are handicapped.



76, 76, 73, 74, 69, 72, 73, 86, 73, 72, 75, 74, 77, 73, 75, 75, 71, 71, 68, 67

16. The following data represents percentages for a recent Further Maths test.

63 71 70 89 88 69 76 83 93 80 73

77 91 75 81 84 87 78 97 89 98 60

Construct a stem plot for the test percentages, using:

a. the stems 6, 7, 8 and 9

b. the stems 6, 7, 8 and 9 split into halves.

17. The following data was collected from a company that compared the battery life (measured in minutes) of two different Ultrabook computers. To complete the test they ran a series of programs on the two computers and measured how long it took for the batteries to go from 100% to 0%.

Computer 1	358	376	392	345	381	405	363	380	352	391	410	366
Computer 2	348	355	361	342	355	362	353	358	340	346	357	352

a. Draw a back-to-back stem plot (using the same stem) of the battery life of the two Ultrabook computers.

b. Use the stem plot to compare and comment on the battery life of the two Ultrabook computers.

18. The heights of 20 Year 8 and Year 10 students (to the nearest centimetre), chosen at random, are measured. The data collected is shown in the table below.

Year 8	151	162	148	153	165	157	172	168	155	164	175	161	155	160	149	155	163	171	166	150
Year 10	167	164	172	158	169	159	174	177	165	156	154	160	178	176	182	152	167	185	173	178

a. Draw a back-to-back stem plot of the data.

b. Comment on what the stem plot tells you about the heights of Year 8 and Year 10 students.

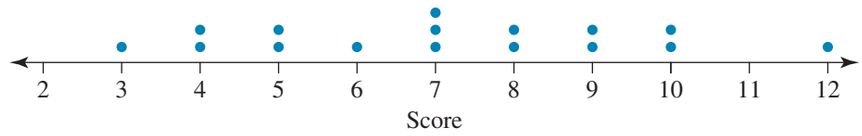
1.4 Dot plots, frequency tables and histograms, and bar charts

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

1.4.1 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.

When representing discrete 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 6

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

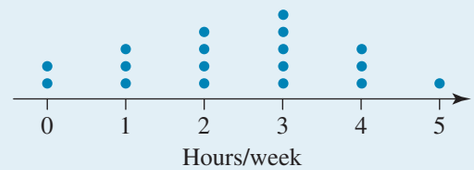
4	**0**	**3**	**1**	**3**	**4**	**2**	**2**	**3**
4	**1**	**3**	**2**	**5**	**3**	**2**	**1**	**0**

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



1.4.2 Frequency tables and 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).

The vertical bars in a histogram are adjacent with no gaps between them, as we generally consider the numerical data scale along the horizontal axis as continuous. Note, however, that histograms can also represent discrete data. It is common practice to leave a small gap before the first bar of a histogram.

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** before constructing a histogram.

WORKED EXAMPLE 7

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.

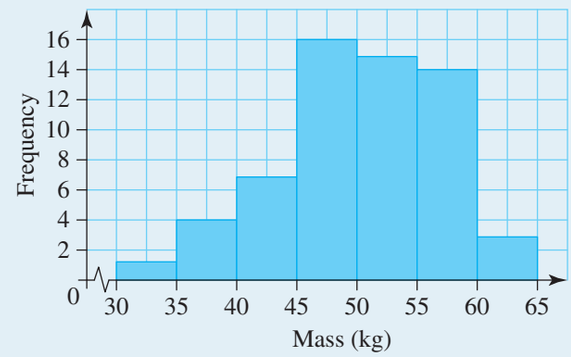
45.7	**45.8**	**45.9**	**48.2**	**48.3**	**48.4**	**34.2**	**52.4**	**52.3**	**51.8**	**45.7**	**56.8**	**56.3**	**60.2**	**44.2**
53.8	**43.5**	**57.2**	**38.7**	**48.5**	**49.6**	**56.9**	**43.8**	**58.3**	**52.4**	**54.3**	**48.6**	**53.7**	**58.7**	**57.6**
45.7	**39.8**	**42.5**	**42.9**	**59.2**	**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. 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–	|	1
35–	||||	4
40–	𝍸 ||	7
45–	𝍸 𝍸 𝍸 |	16
50–	𝍸 𝍸 𝍸	15
55–	𝍸 𝍸 ||||	14
60–	|||	3
	Total	60



When constructing a histogram to represent continuous data, as in Worked example 7, the bars will sit between two values on the horizontal axis which represent the class intervals. When dealing with discrete data the bars should appear above the middle of the value they're representing.

WORKED EXAMPLE 8

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

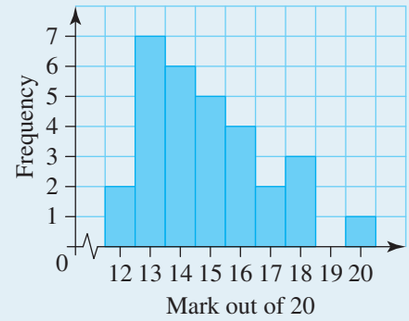
Mark	12	13	14	15	16	17	18	19	20
Frequency	2	7	6	5	4	2	3	0	1

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



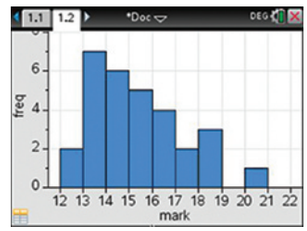
TI | THINK

1. On a Lists & Spreadsheet page, enter the data into the lists, named Mark and Freq.

WRITE



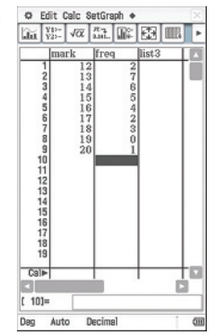
2. Press CTRL + I and select:
   5: Add Data & Statistics
   Use the tab key to move to the horizontal axis and press the click key or ENTER to place *mark* on the horizontal axis.
   To place *freq* on the vertical axis, move the cursor to the vertical axis and press MENU
   Select:
   1: Plot Type
   3: Histogram
   Press MENU:
   2: Plot Properties
   9: Add Y Summary List
   Select freq.



CASIO | THINK

1. On a Statistics screen, enter the data into the lists, named Mark and Freq.

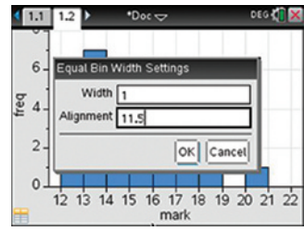
WRITE



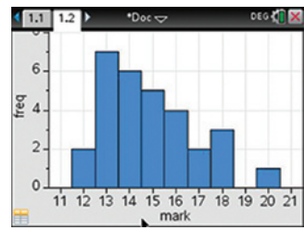
2. Tap:
   SetGraph
   Setting
   Complete the fields under Set StatGraphs as:
   Draw: On
   Type: Histogram
   XList: main\mark
   Freq: main\freq
   Tap Set to lock your selection.
   Tap graph icon.



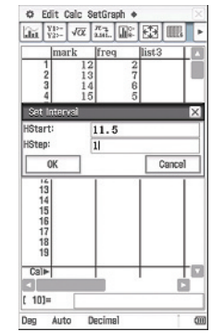
3. As the data in this example is discrete data, to change the column width, press MENU and select:
   2: Plot Properties
   2: Histogram Properties
   2: Bin Settings
   1: Equal Bin Width
   Complete the fields as shown.
   Press ENTER.
   To Zoom out, press MENU and select:
   5: Window/Zoom
   2: Zoom-Data



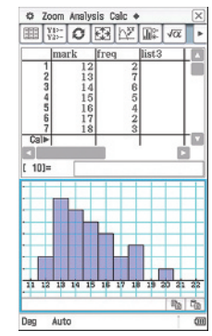
4. The histogram is shown on the screen.



3. Complete the Set Interval fields as:
   HStart: 11.5
   HStep: 1
   Tap OK.



4. The histogram is shown on the screen.

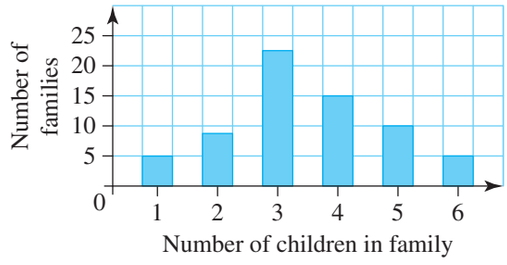


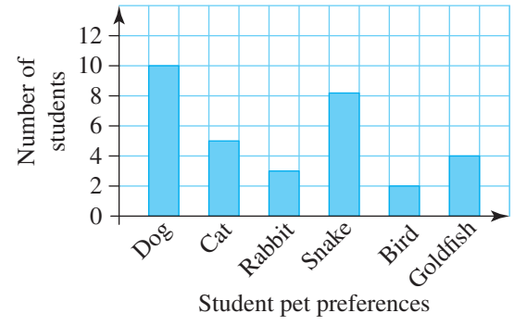
1.4.3 Bar charts

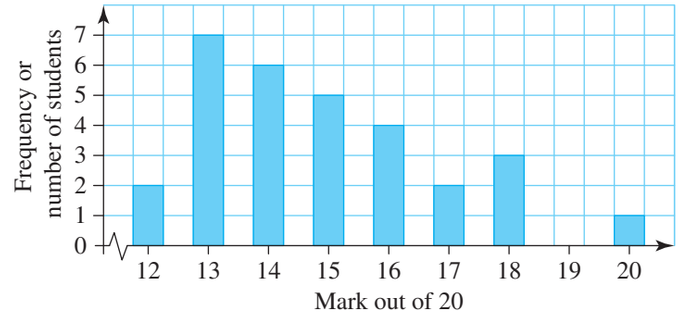
A *bar chart* consists of bars of equal width *separated* by small, equal spaces and may be arranged either *horizontally* or *vertically*. Bar charts are often used to display categorical data.

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

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 bar chart shown represents the data presented in Worked example 8. It could also have been drawn with horizontal bars (rows).

1.4.4 Segmented bar charts

A **segmented 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.

WORKED EXAMPLE 9

The table shown represents fatal road accidents for one year in Australia. Construct a segmented bar chart to represent this data.

	Accidents involving fatalities								
Year	NSW	Vic.	Qld	SA	WA	Tas.	NT	ACT	Aust.
2008	376	278	293	87	189	38	67	12	**1340**

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

THINK

1. To draw a segmented bar chart the data needs to be converted to percentages.

WRITE/DRAW

State	Number of accidents	Percentage
NSW	376	$376 \div 1340 \times 100\% = 28.1\%$
Vic.	278	$278 \div 1340 \times 100\% = 20.7\%$
Qld	293	$293 \div 1340 \times 100\% = 21.9\%$
SA	87	$87 \div 1340 \times 100\% = 6.5\%$
WA	189	$189 \div 1340 \times 100\% = 14.1\%$
Tas.	38	$38 \div 1340 \times 100\% = 2.8\%$
NT	67	$67 \div 1340 \times 100\% = 5.0\%$
ACT	14	$12 \div 1340 \times 100\% = 0.9\%$

2. To draw the segmented bar chart to scale decide on its overall length, let's say 100 mm.

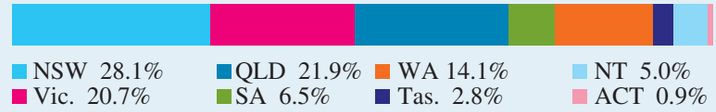
Measure a line 100 mm in length.

3. Therefore NSW = 28.1%, represented by 28.1 mm.
Vic = 20.7%, represented by 20.7 mm and so on.

Measure off each segment and check it adds to the set 100 mm.

4. Draw the answer and colour code it to represent each of the states and territories.

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 28.1 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.

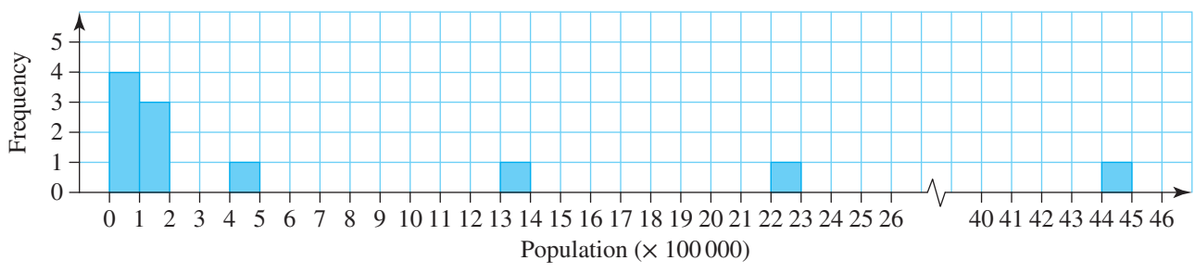


1.4.5 Using a log (base 10) scale

Sometimes a data set will contain data points that vary so much in size that plotting them using a traditional scale becomes very difficult. For example, if we are studying the population of different cities in Australia we might end up with the following data points:

City	Population
Adelaide	1 304 631
Ballarat	98 543
Brisbane	2 274 460
Cairns	146 778
Darwin	140 400
Geelong	184 182
Launceston	86 393
Melbourne	4 440 328
Newcastle	430 755
Shepparton	49 079
Wagga Wagga	55 364

A histogram splitting the data into class intervals of 100 000 would then appear as follows:



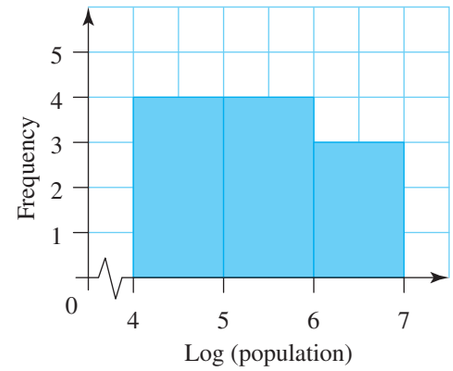
A way to overcome this is to write the numbers in **logarithmic (log) form**. The log of a number is the power of 10 which creates this number.

$$\log(10) = \log\left(10^1\right) = 1$$
$$\log(100) = \log\left(10^2\right) = 2$$
$$\log(1000) = \log\left(10^3\right) = 3$$
$$\vdots$$
$$\log(10^n) = n$$

Not all logarithmic values are integers, so use the log key on CAS to determine exact logarithmic values.
For example, from our previous example showing the population of different Australian cities:
log(4 440 328) = 6.65 (correct to 2 decimal places)
log(184 182) = 5.27 (correct to 2 decimal places)
log(49 079) = 4.69 (correct to 2 decimal places)
and so on …

We can then group our data using class intervals based on log values (from 4 to 7) to come up with the following frequency table and histogram.

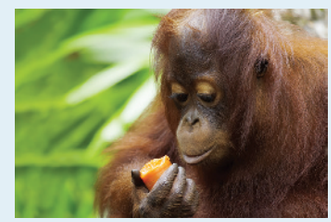
Population	Log (population)	Frequency
10 000–	4–5	4
100 000–	5–6	4
1 000 000–	6–7	3



WORKED EXAMPLE 10

The following table shows the average weights of 10 different adult mammals.

Mammal	Weight (kg)
African elephant	4800
Black rhinoceros	1100
Blue whale	136 000
Giraffe	800
Gorilla	140
Humpback whale	30 000
Lynx	23
Orang-utan	64
Polar bear	475
Tasmanian devil	7



Display the data in a histogram using a log base 10 scale.

THINK

WRITE/DRAW

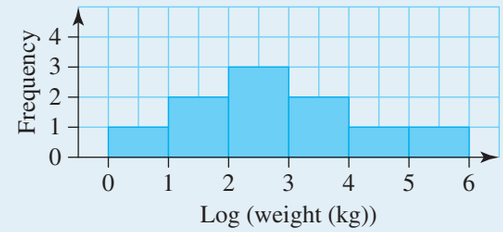
1. Using CAS, calculate the logarithmic values of all of the weights, e.g.:
   log(4800) = 3.68 (correct to 2 decimal places)

Weight	Log (weight (kg))
4800	3.68
1100	3.04
136 000	5.13
800	2.90
140	2.15
30 000	4.48
23	1.36
64	1.81
475	2.68
7	0.85

2. Group the logarithmic weights into class intervals and create a frequency table for the groupings.

Log (weight (kg))	Frequency
0–1	1
1–2	2
2–3	3
3–4	2
4–5	1
5–6	1

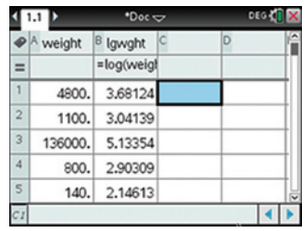
3. Construct a histogram of the data set.



TI | THINK

1. On a Lists & Spreadsheet page, enter the data into the list named weight.
   Label Column B as lgwght and in the formula cell for Column B, complete the entry line as:
   = log(*weight*)
   Press ENTER.

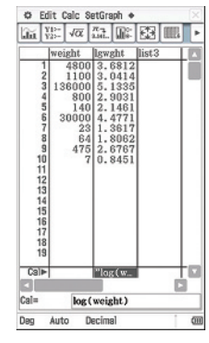
WRITE



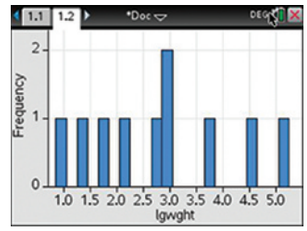
CASIO | THINK

1. On a Statistics screen, enter the data into List1 and rename the list weight.
   Label List2 as lgwght.
   Place the cursor in the Cal cell at the base of the List2 column and complete the entry line as:
   = log(*weight*)
   Tap EXE.

WRITE



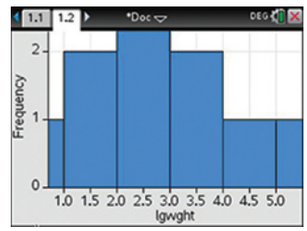
2. Press CTRL + I and select:
5: Add Data & Statistics
Use the tab key to move to the horizontal axis and press the click key or ENTER to place *lgwght* on the horizontal axis.
Press MENU and select:
1: Plot Type
3: Histogram
Press ENTER.



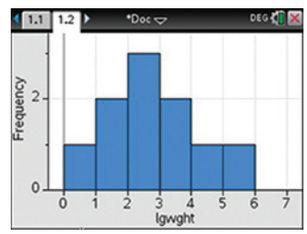
3. To change the column width, press MENU and select:
2: Plot Properties
2: Histogram Properties
2: Bin Settings
1: Equal Bin Width
Complete the fields as shown.
Press ENTER.



4. To Zoom out, press MENU and select:
5: Window/Zoom
2: Zoom-Data



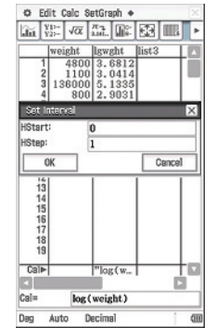
5. The histogram is shown on the screen.



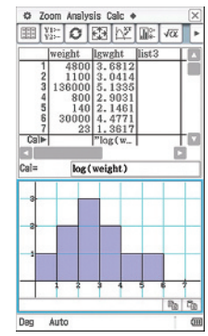
2. Tap:
SetGraph
Setting
Complete the fields under Set StatGraphs as:
Draw: On
Type: Histogram
XList: main\lgwght
Freq: main\1
Tap Set to lock your selection.
Tap graph icon.



3. Complete the Set Interval fields as:
HStart: 11.5
HStep: 1
Tap OK.



4. The histogram is shown on the screen.



Interpreting log (base 10) values

If we are given values in logarithmic form, by raising 10 to the power of the logarithmic number we can determine the conventional number.

For example, the number 3467 in log (base 10) form is 3.54, and $10^{3.54} = 3467$.

We can use this fact to compare values in log (base 10) form, as shown in the following worked example.

WORKED EXAMPLE 11

The Richter Scale measures the magnitude of earthquakes using a log (base 10) scale. How many times stronger is an earthquake of magnitude 7.4 than one of magnitude 5.2? Give your answer correct to the nearest whole number.

THINK	WRITE
1. Calculate the difference between the magnitude of the two earthquakes.	$7.4 - 5.2 = 2.2$
2. Raise 10 to the power of the difference in magnitudes.	$10^{2.2} = 158.49$ (correct to 2 decimal places)
3. Express the answer in words.	The earthquake of magnitude 7.4 is 158 times stronger than the earthquake of magnitude 5.2.

on Resources

- **Interactivity:** Create histograms (int-6494)
- **Interactivity:** Dot plots, frequency tables and histograms, and bar charts (int-6243)
- **Interactivity:** Create bar charts (int-6493)

studyon

Units 3 & 4 > AOS DA > Topic 2 > Concepts 1, 2, 3, 5, 7

Dot plots Summary screen and practice questions
Frequency histograms Summary screen and practice questions
Bar charts Summary screen and practice questions
Segmented graphs Summary screen and practice questions
Log base 10 scale Summary screen and practice questions

Exercise 1.4 Dot plots, frequency tables and histograms, and bar charts

1. **WE6** The number of questions completed for maths homework each night by 16 students is shown below.

   5 6 5 9 10 10 6 8

   10 9 8 5 7 8 7 9

   Display the data as a dot plot.

2. 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

3. The number of hours spent on homework for a group of 20 Year 12 students each week is shown.

15 15 18 21 20 21 24 24 20 18

18 21 22 20 24 20 24 21 18 15

Display the data as a dot plot.

4. **WE7** The data shows the distribution of heights (in cm) of 40 students in Year 12. Construct a frequency histogram to display the data more clearly.

167 172 184 180 178 166 154 150 164 161

187 159 182 177 172 163 179 181 170 176

177 162 172 184 188 179 189 192 164 160

166 169 163 185 178 183 190 170 168 159

5. 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

Using the frequency tables above, construct a histogram by hand for each set of data.

6. The number of dogs at a RSPCA kennel each week is as shown. Construct a frequency table for these data.

7, 6, 2, 12, 7, 9, 12, 10, 5, 7, 9, 4, 5, 9, 3, 2, 10, 8, 9, 7, 9, 10, 9, 4, 3, 8, 9, 3, 7, 9

7. **WE8** The number of fish caught by 30 anglers in a fishing competition are given in the frequency table below.



Fish	0	1	2	3	4	5	7
Frequency	4	7	4	6	5	3	1

Display these data on a histogram.

8. The number of fatal car accidents in Victoria each week is given in the frequency table for a year.

Fatalities	0	1	2	4	6	7	9	10	13
Frequency	12	3	6	10	7	6	5	2	1

Display these data on a histogram.

9. Using the frequency table in question **6**, construct a histogram by hand.

10. Using CAS construct a histogram from the data in question **6**, and compare it to the histogram in question **9**.

11. WE9 The following table shows how many goals each of the 18 AFL team's leading goal kickers scored in the 2014 regular season. Construct a segmented bar chart to represent this data.

Club	Goals
Adelaide Crows	51
Brisbane Lions	33
Carlton	29
Collingwood	39
Essendon	27
Fremantle	49
Geelong Cats	62
Gold Coast Suns	46
GWS Giants	29

Club	Goals
Hawthorn	62
Melbourne	20
North Melbourne	41
Port Adelaide	62
Richmond	58
St Kilda	49
Sydney Swans	67
West Coast Eagles	61
Western Bulldogs	37

12. Information about adult participation in sport and physical activities in 2017–18 is shown in the following table. 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 — 2017–18

	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**

13. A group of students was surveyed, asking how many children were in their family. The data is shown in the table.

Number of children	1	2	3	4	5	6	9
Number of families	12	18	24	10	8	3	1

Construct a bar chart that displays the data.

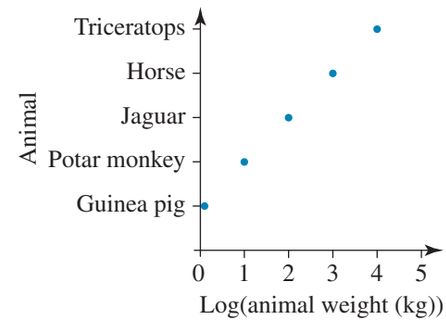
14. **WE10** The following table shows the average weights of 10 different adult mammals.

Mammal	**Weight (kg)**
Black wallaroo	18
Capybara	55
Cougar	63
Fin whale	70 000
Lion	175
Ocelot	9
Pygmy rabbit	0.4
Red deer	200
Quokka	4
Water buffalo	725

Display the data in a histogram using a log base 10 scale, using class intervals of width 1.

15. **MC** The following graph shows the weights of animals. If a gorilla has a weight of 207 kilograms then its weight is between that of:

A. Potar monkey and jaguar.
B. horse and triceratops.
C. guinea pig and Potar monkey.
D. jaguar and horse.
E. none of the above.



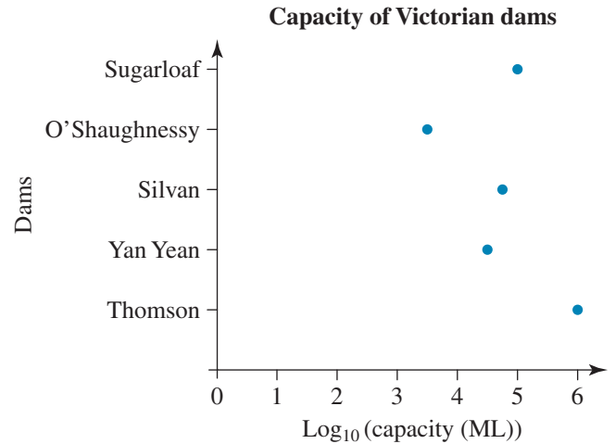
16. **MC** The following table shows a variety of top speeds.

F1 racing car	370 000 m/h
V8 supercar	300 000 m/h
Cheetah	64 000 m/h
Space shuttle	28 000 000 m/h
Usain Bolt	34 000 m/h

The correct value, to 2 decimal places, for a cheetah's top speed using a log (base 10) scale would be:

A. 5.57 **B.** 5.47 **C.** 7.45
D. 4.81 **E.** 11.07

17. MC The following graph represents the capacity of five Victorian dams.



a. The capacity of Thomson Dam is closest to:
A. 100 000 ML **B.** 1 000 000 ML **C.** 500 000 ML
D. 5000 ML **E.** 40 000 ML

b. The capacity of Sugarloaf Dam is closest to:
A. 100 000 ML **B.** 1 000 000 ML **C.** 500 000 ML
D. 5000 ML **E.** 40 000 ML

c. The capacity of Silvan Dam is between which range?
A. 1 and 10 ML **B.** 10 and 100 ML **C.** 1000 and 10 000 ML
D. 10 000 and 100 000 ML **E.** 100 000 and 1 000 000 ML

The table shows the flow rate of world famous waterfalls. Use this table to complete questions **18** and **19**.

Waterfalls	Flow rate
Victoria Falls	1 088 m^3/s
Niagara Falls	2 407 m^3/s
Celilo Falls	5 415 m^3/s
Khane Phapheng Falls	11 610 m^3/s
Boyoma Falls	17 000 m^3/s

18. MC The correct value, correct to 2 decimal places, to be plotted for Niagara Falls' flow rate using a log(base 10) scale would be:
A. 3.38 **B.** 3.04 **C.** 4.06
D. 4.23 **E.** 5.10

19. MC The correct value, correct to 2 decimal places, to be plotted for Victoria Falls' flow rate using a log (base 10) scale would be:
A. 3.38 **B.** 3.04 **C.** 4.06
D. 4.23 **E.** 5.10

20. WE11 The Richter scale measures the magnitude of earthquakes using a log (base 10) scale. How many times stronger is an earthquake of magnitude 8.1 than one of magnitude 6.9? Give your answer correct to the nearest whole number.

21. The pH scale measures acidity using a log (base 10) scale. For each decrease in pH of 1, the acidity of a substance increases by a factor of 10.



If a liquid's pH value decreases by 0.7, by how much has the acidity of the liquid increased?

22. Using the information provided in the table below:

a. calculate the proportion of residents who travelled in 2017 to each of the countries listed

b. draw a segmented bar graph showing the major destinations of Australians travelling abroad in 2014.

Short-term resident departures by major destinations					
	2013 (×1000)	**2014 (×1000)**	**2015 (×1000)**	**2016 (×1000)**	**2017 (×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
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

23. The number of people who attended the Melbourne Grand Prix are shown.

Year	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
Attendance 000s	360	359	301	301	303	287	305	298	313	323

a. Construct a bar chart by hand to display the data.

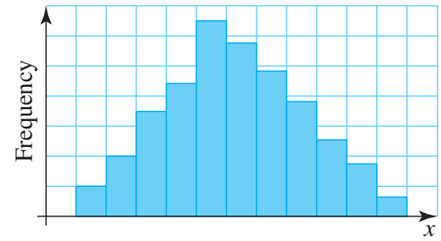
b. Use CAS to construct a bar chart and compare the two charts.

1.5 Describing the shape of stem plots and histograms

1.5.1 Symmetric distributions

The data shown in the histogram below 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	7 8 9 9
5	2 7 8
6	1 3

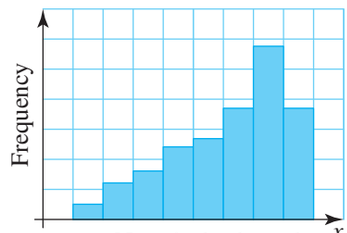
Key: 0|7 = 7

1.5.2 Skewed distributions

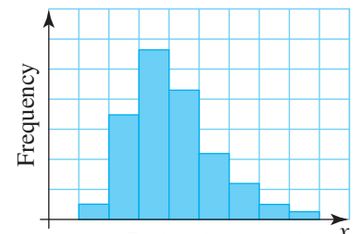
Each of the histograms shown on next page are examples of skewed distributions.

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

The figure on the 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 12

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	1 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.

studyon

Units 3 & 4 > AOS DA > Topic 2 > Concept 6 **Shape** Summary screen and practice questions

Exercise 1.5 Describing the shape of stem plots and histograms

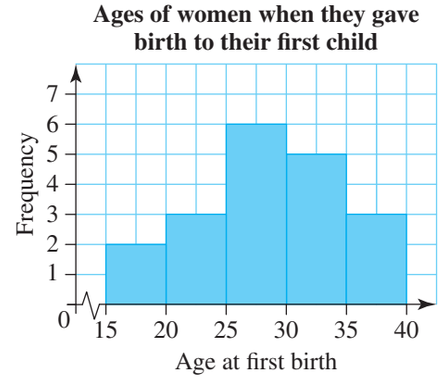
1. WE12 The ages of a group of people when they bought their first car are shown.

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

Key: 1|7 = 17 years old

Describe the shape of the distribution of these data.

2. The ages of women when they gave birth to their first child is shown. Describe the shape of the distribution of the data.



3. 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 = 04

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

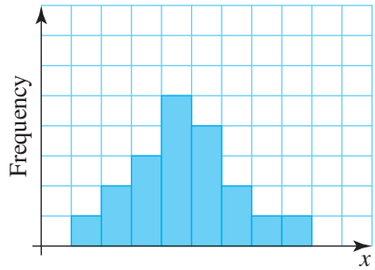
f.

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

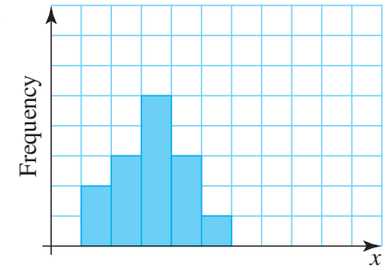
Key: 62|3 = 623

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

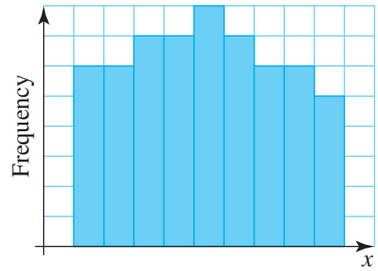
a.



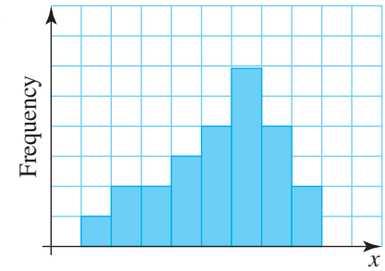
b.



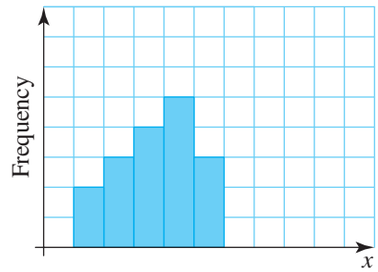
c.



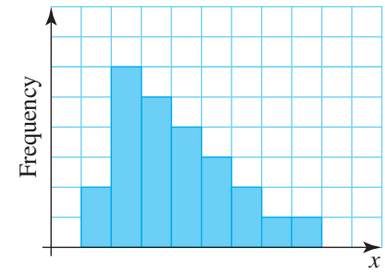
d.



e.



f.



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

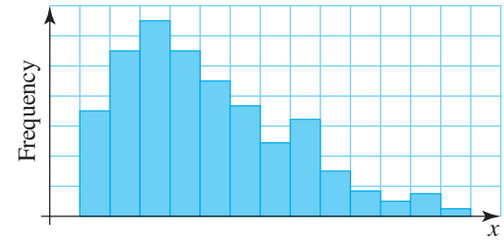
A. negatively skewed
B. negatively skewed and symmetric
C. positively skewed
D. positively skewed and symmetric
E. symmetric

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

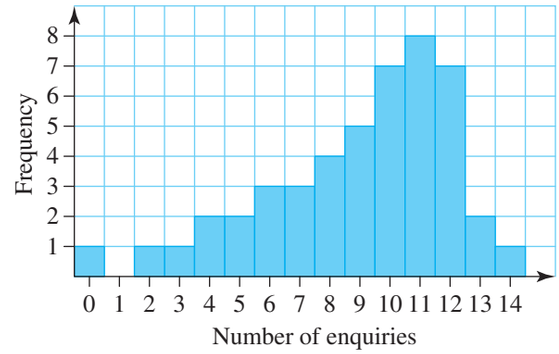
Key: 1|8 = 18

6. MC The distribution of the data shown in the histogram could be described as:

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



7. 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 below. Describe the shape of the distribution of these data.



8. The number of nights per month spent interstate by a group of flight attendants is shown in the stem plot. 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

9. The mass (correct to the nearest kilogram) of each dog at a dog obedience school is shown in the stem plot.

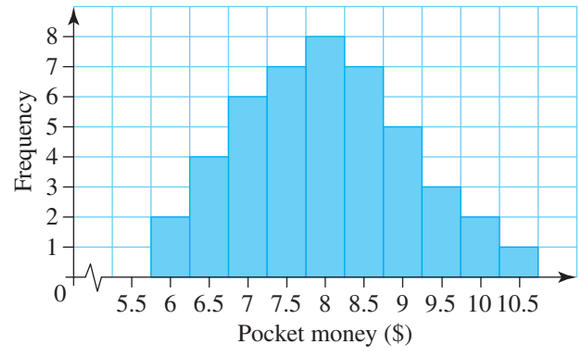
a. Describe the shape of the distribution of these data.
b. What does this information tell us about this group of dogs?



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 = 4 kg

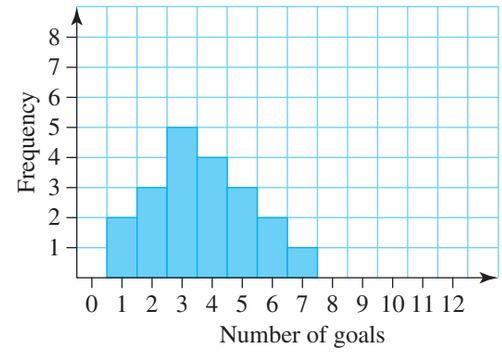
10. The amount of pocket money (correct to the nearest 50 cents) received each week by students in a Grade 6 class is illustrated in the histogram.



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?

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



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 inclusive over the 3 weeks.

12. The number of hours a group of students exercise each week is shown in the stem plot.

a. Describe the shape of the distribution of these data.

b. What does this sample data tell us about this group of students?

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

Key: 0|1 = 1

13. The stem plot shows the age of players in two bowling teams.

a. Describe the shape of the distribution of Club A and Club B.

b. What comments can you make about the make-up of Club A compared to Club B?

c. How many players are over the age of 70 from:

i. Club A

ii. Club B?

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

Key: 5|5 = 55

14. The following table shows the number of cars sold at a dealership over eight months.

Month	April	May	June	July	August	September	October	November
Cars sold	9	14	27	21	12	14	10	18

a. Display the data on a bar chart.
b. Describe the shape of the distribution of these data.
c. What does this sample data tell us about car sales over these months?
d. Explain why the most cars were sold in the month of June.

1.6 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.

1.6.1 The median

The *median* is the midpoint of an ordered set of data. Half the data are less than or equal to the median and half the data are greater 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. $\left(\text{Alternatively, median} = \frac{7+8}{2} = 7.5.\right)$

When there are n records in a set of ordered data, the median is 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 13

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

1.6.2 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, 1, 0. 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 $Q_3 = 15$.
So the IQR is $15 - 6 = 9$.

Case 2

Consider a set of data containing the 7 observations: 11, 13, 30, 4, 17, 9, 23.
Order the data. 4, 9, 11, 13, 17, 23, 30
Locate the median. 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 $Q_3 = 23$.
So the IQR is $23 - 9 = 14$.

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 $Q_3 = 19$.
So the IQR is $19 - 6 = 13$.

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 $Q_3 = 23$.
So the IQR is $23 - 10 = 13$.

WORKED EXAMPLE 14

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.

$$\begin{aligned}\text{IQR} &= Q_3 - Q_1 \\ &= 59.5 - 17 \\ &= 42.5\end{aligned}$$

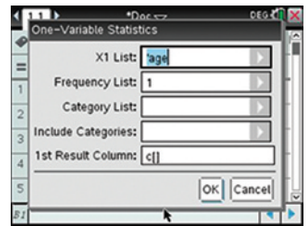
TI | THINK

1. On a Lists & Spreadsheet page, enter the data into Column A and rename the column *age*.

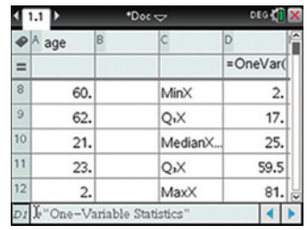
WRITE



2. Press MENU and select:
4: Statistics
1: Stat Calculations
1: One-Variable Statistics
Complete the field:
Num of Lists: 1
Press ENTER.
Complete the fields under One-Variable Statistics as shown.
Press ENTER or Click OK.



3. The One-Variable Statistics are displayed in Column D.



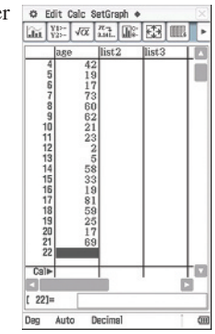
4. The answer is displayed on the screen.

$Q_1 = 17$
$Q_3 = 59.5$
$\text{IQR} = 59.5 - 17$
$= 42.5$

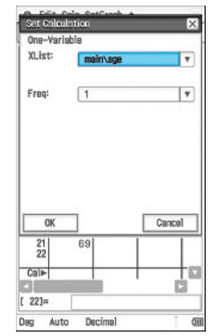
CASIO | THINK

1. On a Statistics screen, enter the data into List 1 and rename the list *age*.

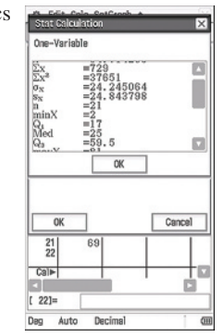
WRITE



2. Tap:
Calc
One-variable
Complete the fields under Set Calculation as:
XList: main\age
Freq: 1
Tap OK.



3. The One-Variable Statistics are displayed.



4. The answer is displayed on the screen.

$Q_1 = 17$
$Q_3 = 59.5$
$\text{IQR} = 59.5 - 17$
$= 42.5$

WORKED EXAMPLE 15

Parents are often shocked at the amount of money their children spend. The data below give the amount spent (correct 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 CAS 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, IQR $= Q_3 - Q_1$ $= 23 - 17$ $= 6$

1.6.3 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 Worked example 15, the minimum and maximum values were 11 and 37, respectively. The range, therefore, can be calculated as follows.

$$\begin{aligned}\text{Range} &= \max_x - \min_x \\ &= 37 - 11 \\ &= 26\end{aligned}$$

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.

1.6.4 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 13, the mode is 48 since it occurs three times and hence is the score with the highest frequency. In Worked example 14 there are two modes, 17 and 19, because they equally occur most frequently.

on Resources

Interactivity: The median, the interquartile range, the range and the mode (int-6244)
Interactivity: Mean, median, mode and quartiles (int-6496)

studyon

Units 3 & 4 > AOS DA > Topic 3 > Concepts 2 & 3

Measures of centre—median and mode Summary screen and practice questions

Measures of spread—range and interquartile range Summary screen and practice questions

Exercise 1.6 The median, the interquartile range, the range and the mode

1. **WE13** The stem plot shows 30 observations. What is the median value?

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

Key: 2|1 = 21

2. The following data represents the number of goals scored by a netball team over the course of a 16 game season. What was the team's median number of goals for the season?

28, 36, 24, 46, 37, 21, 49, 32, 33, 41, 47, 29, 45, 52, 37, 24

3. **WE14** From the following data find the interquartile range.

33 21 39 45 31 28 15 13 16

21 49 26 29 30 21 37 27 19

12 15 24 33 37 10 23 28 39

4. The ages of a sample of people surveyed at a concert are shown.

21 25 24 18 19 16 19 27 32 24

15 20 31 24 29 33 27 18 19 21

Find the interquartile range of these data.

5. **WE15** The data shows the amount of money spent (to the nearest dollar) at the school canteen by a group of students in a week.

3 5 7 12 15 10 8 9 21 5

7 9 13 15 7 3 4 2 11 8

Calculate the interquartile range for the data set.

6. The amount of money, in millions, changing hands through a large stocks company, in one-minute intervals, was recorded as follows.

45.8 48.9 46.4 45.7 43.8 49.1 42.7 43.1 45.3 48.6

41.9 40.0 45.9 44.7 43.9 45.1 47.1 49.7 42.9 45.1

Calculate the interquartile range for these data.

7. 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

8. For each of the following sets of data, determine 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



9. a. 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?

10. 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?

11. Give an example of a data set where:

a. the lower quartile equals the lowest score

b. the IQR is zero.

12. 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.

13. 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

14. For each set of data shown in 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

15. For the data in the stem plots shown, find the range, median, mode and interquartile range.

a.

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

Key: 0|1 = 1

b.

Stem	Leaf
40	3 5 7
41	1 1 1 3 4 6 7 8 9
42	0 2 3 6 7 8 9
43	2 3 3 6 8
44	1 2
45	0

Key: 40|3 = 403

16. From the following set of data, find the median and mode.

4, 7, 9, 12, 15, 2, 3, 7, 9, 4, 7, 9, 2, 8, 13, 5, 3, 7, 5, 9, 7, 10

17. The following data represents distances (in metres) thrown during a javelin-throwing competition. Use CAS to calculate the interquartile range and median.



40.3	42.8	41.0	50.3	52.2	46.1	44.5
41.6	44.3	47.4	45.1	48.8	46.1	44.5
45.3	42.9	41.1	49.0	47.5	40.8	51.1

18. The data below shows the distribution of golf scores for one day of an amateur tournament.

a. Use CAS to calculate the median, interquartile range, mode and range.

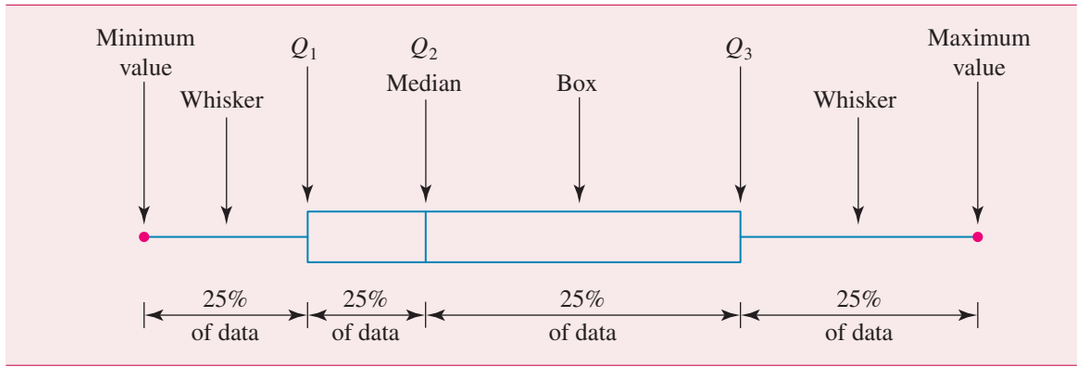
b. Comment on what you suggest the average handicap of the players listed should be if par for the course is 72.

111	93	103	85	81	90	101	95	84
93	101	85	87	85	93	100	86	91
93	95	93	99	95	93	92	96	93
97	93	93	97	96	92	100	95	104

1.7 Boxplots

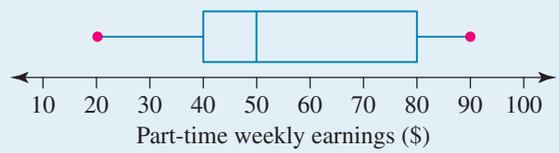
The five number summary statistics that we looked at in the previous section ($\min_x$, Q_1, Q_2, Q_3, $\max_x$) can be illustrated very neatly in a special diagram known as a boxplot (or box-and-whisker diagram). The diagram is made up of a box with straight lines (whiskers) extending from opposite 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.



WORKED EXAMPLE 16

The boxplot below 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

1. Range = maximum value − minimum value
   The minimum value is 20 and the maximum value is 90.

WRITE

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 $Q_3 = 40$
IQR = 80 − 40
= 40

Boxplots can be drawn horizontally or vertically.

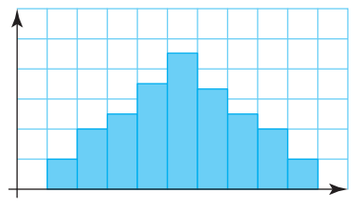


Horizontal boxplot



Vertical boxplot

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.

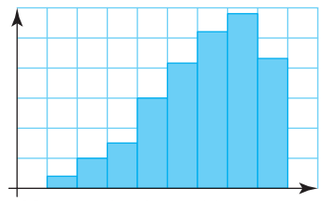


Symmetric histogram



Symmetric boxplot

The median is in the centre of the boxplot.

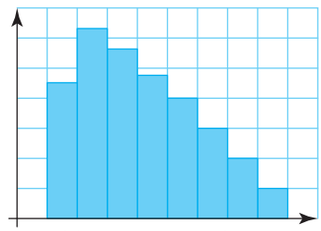


Negatively skewed histogram



Negatively skewed boxplot

The median is to the right of the centre of the boxplot.



Positively skewed histogram

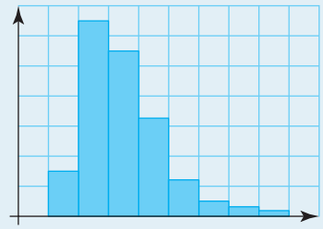


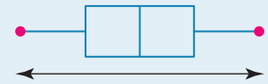
Positively skewed boxplot

The median is to the left of the centre of the boxplot.

WORKED EXAMPLE 17

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





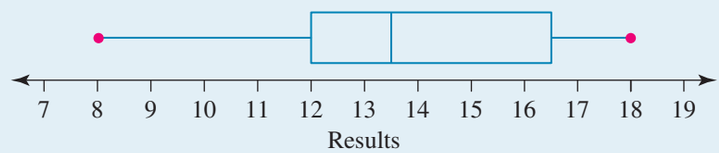
THINK	WRITE
The histogram shows a distribution which is positively skewed. The boxplot shows a distribution which is approximately symmetric.	The histogram and the boxplot could not represent the same data since the histogram shows a distribution that is positively skewed and the boxplot shows a distribution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

WORKED EXAMPLE 18

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

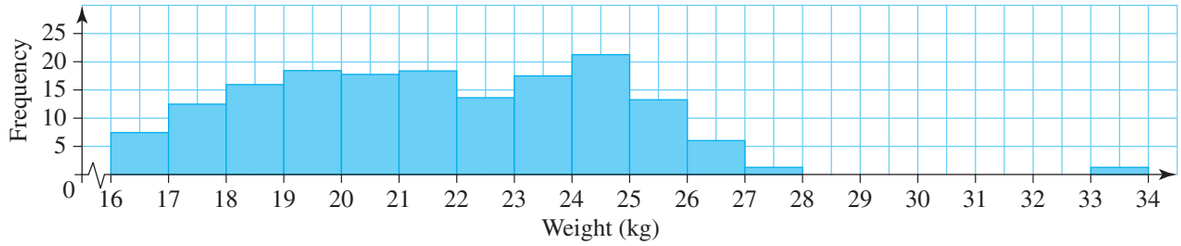
15, 12, 17, 8, 13, 18, 14, 16, 17, 13, 11, 12

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.

1.7.1 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, here 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.

1.7.2 The lower and upper fences

We can identify outliers by calculating the values of the lower and upper fences in a data set. Values which lie either below the lower fence or above the upper fence are outliers.

$$\textbf{The lower fence} = Q_1 - 1.5 \times \text{IQR}$$

$$\textbf{The upper fence} = Q_3 + 1.5 \times \text{IQR}$$

An outlier is not included in the boxplot, but should instead be plotted as a point beyond the end of the whisker.

WORKED EXAMPLE 19

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

1. Determine the five number summary statistics by first ordering the data and obtaining the interquartile range.

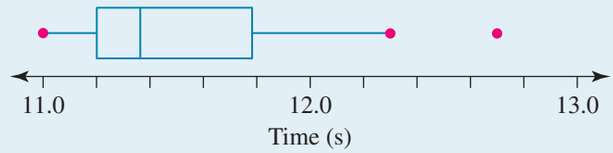
WRITE/DRAW

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 calculating the values of the lower and upper fences.

$Q_1 - 1.5 \times \text{IQR} = 11.2 - 1.5 \times 0.55$
$= 10.375$
The lowest score lies above the lower fence of 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 the upper fence of 12.575, so it is an outlier and 12.3 becomes the end of the upper whisker.

3. Draw the boxplot with the outlier.



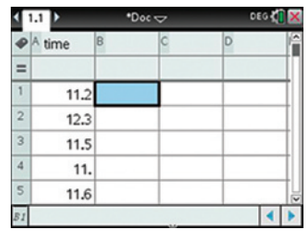
4. Describe the shape of the distribution. Data peak to the left and trail off to the right with one outlier.

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.

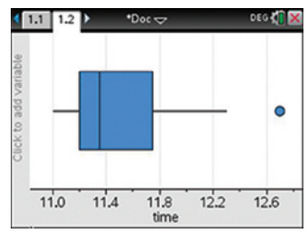
TI | THINK

1. On a Lists & Spreadsheet page, enter the data into Column A and rename the column *time*.

WRITE



2. Press Ctrl + I and select:
5: Add Data and Statistics
Move the cursor to select *time* on the horizontal axis.
Press MENU and select:
1: Plot Type
2: Box Plot



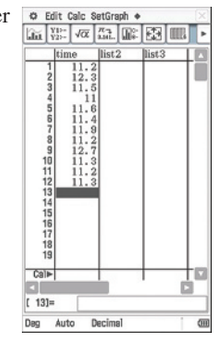
3. The answer is displayed in the screen.

The data is positively skewed with 12.7 seconds being an outlier.

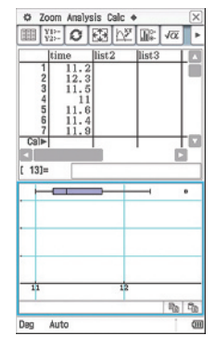
CASIO | THINK

1. On a Statistics screen, enter the data into List1 and rename the list *time*.

WRITE



2. Tap:
SetGraph
Settings
Complete the fields under Set StatGraphs as:
Draw: On
Type: MedBox
XList: main\time
Freq: 1
Tick Show Outliers.
Tap Set.



3. The answer is displayed in the screen.

The data is positively skewed with 12.7 seconds being an outlier.

on Resources

Interactivity: Boxplots (int-6245)

studyon

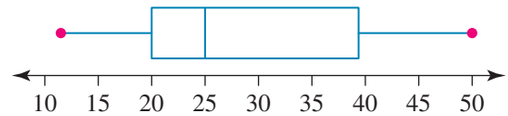
Units 3 & 4 > AOS DA > Topic 3 > Concepts 5 & 6

Boxplots Summary screen and practice questions

Boxplots with outlier(s) Summary screen and practice questions

Exercise 1.7 Boxplots

1. **WE16** Write down the range, median and interquartile range for the data in the boxplot shown.

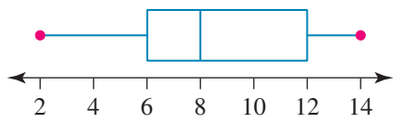


2. Find the median, range and interquartile range of the data displayed in the boxplot shown.

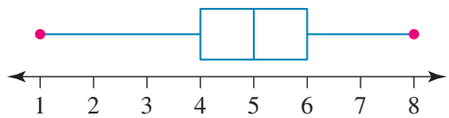


3. For the boxplots shown, write down the range, the interquartile range and the median of the distributions which each one represents.

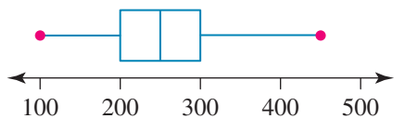
a.



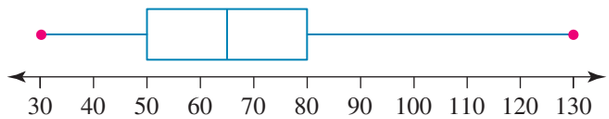
b.



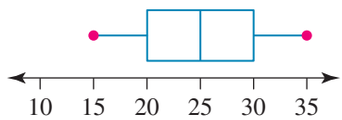
c.



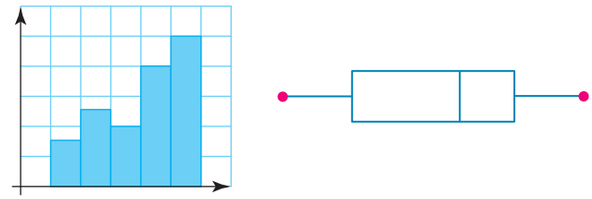
d.



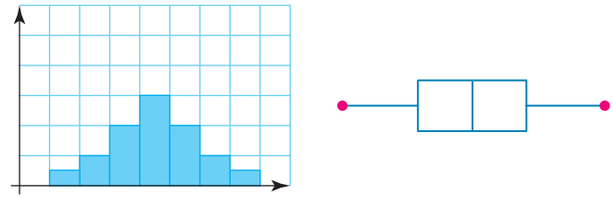
e.



4. **WE17** Explain whether or not the histogram and the boxplot shown could represent the same data.

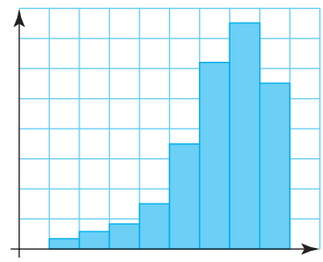


5. Do the histogram and boxplot represent the same data? Explain.

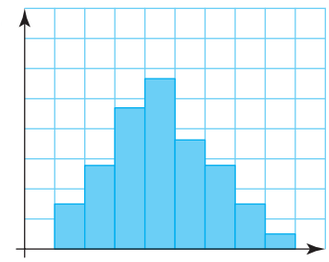


6. Match each histogram below with the boxplot which could show the same distribution.

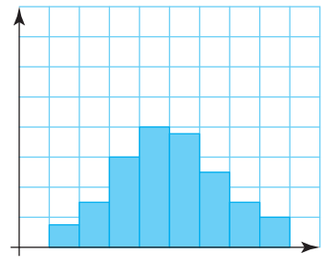
a.



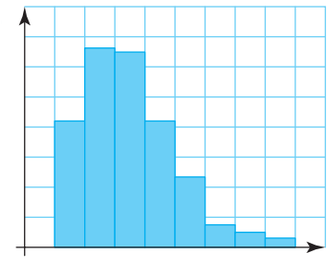
b.



c.



d.



i.



ii.



iii.



iv.



7. **WE18** The results for a Physics test (out of 50) are shown.

32, 38, 42, 40, 37, 26, 46, 36, 50, 41, 48, 50, 40, 38, 32, 35, 28, 30

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

8. The numbers of hours Year 12 students spend at their part-time job per week are shown.

4, 8, 6, 5, 12, 8, 16, 4, 7, 10, 8, 20, 12, 7, 6, 4, 8

Display the data using a boxplot and discuss what the boxplot shows in relation to part-time work of Year 12 students.

9. For each of the following sets of data, construct a boxplot.

a. 3, 5, 6, 8, 8, 9, 12, 14, 17, 18

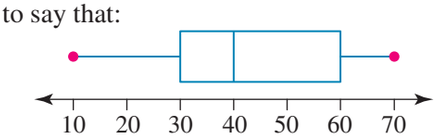
b. 3, 4, 4, 5, 5, 6, 7, 7, 7, 8, 8, 9, 9, 10, 10, 12

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

10. 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

11. 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.

12. 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.

13. WE19 The heights jumped (in metres) at a school high jump competition are listed:

1.35, 1.30, 1.40, 1.38, 1.45, 1.48, 1.30, 1.36, 1.45, 1.75, 1.46, 1.40

a. Draw a boxplot to represent the data.

b. Describe the shape of the distribution and comment on the existence of any outliers by finding the lower and upper fences.

14. The amount of fuel (in litres) used at a petrol pump for 16 cars is listed:

48.5 55.1 61.2 58.5 46.9 49.2 57.3 49.9

51.6 30.3 45.9 50.2 52.6 47.0 55.5 60.3

Draw a boxplot to represent the data and label any outliers.

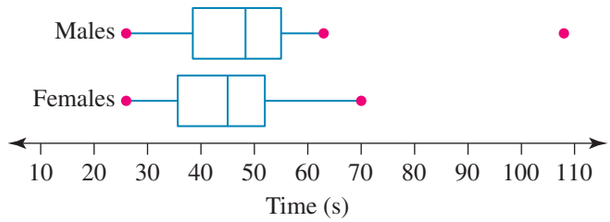
15. The number of rides that 16 children had at the annual show are listed below.

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 CAS to draw a boxplot for these data.

16. 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.





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 by finding the lower and upper fences.

17. You work in the marketing department of a perfume company. You completed a survey of people who purchased your perfume, asking them how many times a week they used it. Analyse the data by drawing a boxplot and comment on the existence of any outliers by finding the lower and upper fences.

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

7 7 9 8 5 6 5 3 15 8 7 5

18. For the data set shown:

11, 11, 14, 16, 19, 22, 24, 25, 25, 27, 28, 28, 36, 38, 38, 39

a. construct a boxplot by hand
b. comment on the presence of outliers by finding the lower and upper fences
c. construct a boxplot using CAS and compare the two boxplots.

19. From the stem plot shown construct a boxplot using CAS and comment on any outliers if they exist in the data.

Stem	Leaf
2	3
3	
4	
5	
6	1 5 8
7	2 5 5 7
8	1 2 3 3 6 7 7
9	1 1 3 7 7
10	6 7 9
11	1 3

Key: 2|3 = 23

1.8 The mean of a sample

When dealing with sets of data, we are always working with either the **population**, or a **sample** from the population. The means of a data set representing either the population or a sample are calculated in the same way; however, they are represented by different symbols. The **mean** of a population is represented by the Greek letter μ (mu) and the mean of a sample is represented by $\bar{x}$ (x-bar), a lower case x with a bar on top. In this section we will only use $\bar{x}$ to represent 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}:

$$\bar{x} = \frac{4 + 7 + 9 + 12 + 18}{5}$$
$$= 10$$

The formal definition of the mean is:

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

where $\sum x$ 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, $\sum$ 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.



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.

1.8.1 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.

1.8.2 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 20

Calculate the mean of the set of data shown.
10, 12, 15, 16, 18, 19, 22, 25, 27, 29

THINK

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.
2. Substitute the values into the formula and evaluate.

WRITE

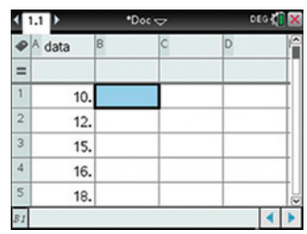
$$\bar{x} = \frac{\sum x}{n}$$
$$= \frac{10 + 12 + 15 + 16 + 18 + 19 + 22 + 25 + 27 + 29}{10}$$
$$\bar{x} = 19.3$$

The mean, $\bar{x}$, is 19.3.

TI | THINK

1. On a Lists & Spreadsheet page, enter the data into Column A and rename the column *data*.

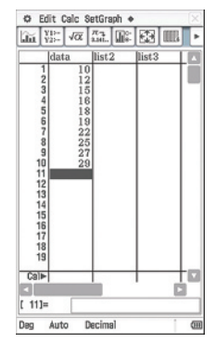
WRITE



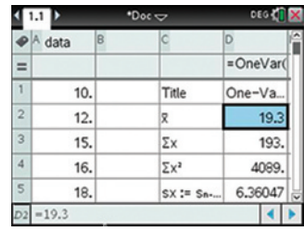
CASIO | THINK

1. On a Statistics screen, enter the data into List1 and rename the list *data*.

WRITE



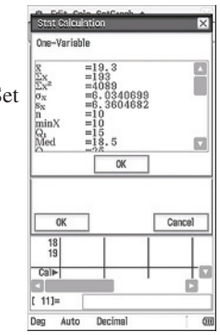
2. Press MENU and select:
4: Statistics
1: Stat Calculations
1: One-Variable Statistics
Complete the field:
Num of Lists: 1
Press ENTER.
Complete the fields under One-Variable Statistics as:
X1List: data
Frequency List: 1
Leave other fields blank.
Press ENTER or Click OK.
The One-Variable Statistics are displayed in Column D.



4. The answer is displayed on the screen.

The mean, $\bar{x} = 19.3$

2. Tap:
Calc
One-variable
Complete the fields under Set Calculation as:
XList: main\data
Freq: 1
Tap OK.



4. The answer is displayed on the screen.

The mean, $\bar{x} = 19.3$

When calculating the mean of a data set, sometimes the answer you calculate will contain a long stream of digits after the decimal point.

For example, if we are calculating the mean of the data set

$$44,\ 38,\ 55,\ 61,\ 48,\ 32,\ 49$$

Then the mean would be:

$$\begin{aligned}\bar{x} &= \frac{\sum x}{n} \\ &= \frac{327}{7} \\ &= 46.714\,285\,71\ldots\end{aligned}$$

In this case it makes sense to round the answer to either a given number of decimal places, or a given number of **significant figures**.

1.8.3 Rounding to a given number of significant figures

When rounding to a given number of signified figures, we are rounding to the digits in a number which are regarded as 'significant'.

To determine which digits are significant, we can observe the following rules:

- All digits greater than zero are significant
- Leading zeros can be ignored (they are placeholders and are not significant)
- Zeros included between other digits are significant
- Zeros included after decimal digits are significant
- Trailing zeros for integers are not significant (unless specified otherwise)

The following examples show how these rules work:

0.003 561 — leading digits are ignored, so this has 4 significant figures

70.036 — zeros between other digits are significant, so this has 5 significant figures

5.320 — zeros included after decimal digits are significant, so this has 4 significant figures

450 000 — trailing zeros are not significant, so this has 2 significant figures

78 000.0 — the zero after the decimal point is considered significant, so the zeros between other numbers are also significant; this has 6 significant figures

As when rounding to a given number of decimal places, when rounding to a given number of significant figures consider the digit after the specified number of figures. If it is 5 or above, round the final digit up; if it is 4 or below, keep the final digit as is.

5067.37 — rounded to 2 significant figures is 5100
3199.01 — rounded to 4 significant figures is 3199
0.004 931 — rounded to 3 significant figures is 0.004 93
1 020 004 — rounded to 2 significant figures is 1 000 000

studyon

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

Measures of centre—mean Summary screen and practice questions

Exercise 1.8 The mean of a sample

1. WE20 Calculate the mean of the data set shown.
   9, 12, 14, 16, 18, 19, 20, 25, 29, 33, 35, 36, 39
2. Calculate the mean of the data set shown.
   5.5, 6.3, 7.7, 8.3, 9.7, 6.7, 12.9, 10.5, 9.9, 5.1
3. Find the mean of each of the following sets of data.
   a. 5, 6, 8, 8, 9 (correct to 2 significant figures)
   b. 3, 4, 4, 5, 5, 6, 7, 7, 7, 8, 8, 9, 9, 10, 10, 12 (correct to 4 significant figures)
   c. 4.3, 4.5, 4.7, 4.9, 5.1, 5.3, 5.5, 5.6 (correct to 5 significant figures)
   d. 11, 13, 15, 15, 16, 18, 20, 21, 22 (correct to 1 decimal place)
   e. 0.4, 0.5, 0.7, 0.8, 0.8, 0.9, 1.0, 1.1, 1.2, 1.0, 1.3 (correct to 4 decimal places)
4. 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
5. The number of people attending sculpture classes at the local TAFE college for each week during the first semester is given.
   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 correct to 2 significant figures.)
6. MC The ages of a group of junior pilots joining an international airline are indicated in the stem plot below.

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

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

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

7. 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

8. 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

1.9 Standard deviation of a sample

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 ($x - x$)
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:

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

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

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.

Note: This is the formula for the standard deviation of a sample. The standard deviation for a population is given by σ (sigma) and is calculated using a slightly different formula, which is outside the scope of this course.

While some of the theory or formulas associated with standard deviation may look complex, the calculation of this measure of spread is straightforward using CAS. Manual computation of standard deviation is therefore rarely necessary, however application of the formula is required knowledge.

WORKED EXAMPLE 21

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

1. Enter the data into CAS to determine the sample statistics.
2. Round the value correct to 2 decimal places.

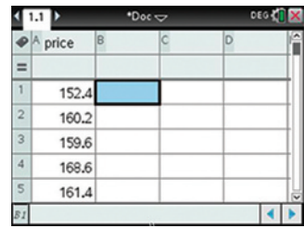
WRITE

$s = 4.515\,92$
$= 4.52$ cents/L

TI | THINK

1. On a Lists & Spreadsheet page, enter the data into Column A and rename the column *price*.

WRITE



2. Press MENU and select:
   4: Statistics
   1: Stat Calculations
   1: One-Variable Statistics
   Complete the field:
   Num of Lists: 1
   Press ENTER.
   Complete the fields under One-Variable Statistics as:
   X1List: price
   Frequency List: 1
   Leave other fields blank.
   Press ENTER or Click OK.
   The One-Variable Statistics are displayed in Column D.



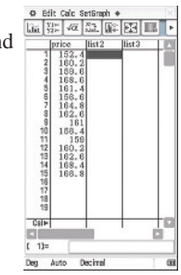
4. The answer is displayed on the screen.

The standard deviation, correct to 2 decimal places is, $s = 4.52$ cents/L.

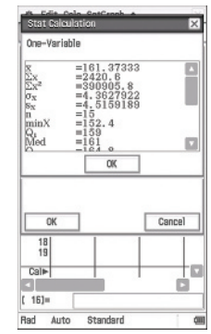
CASIO | THINK

1. On a Statistics screen, enter the data into List1 and rename the list *price*.

WRITE



2. Tap:
   Calc
   One-variable
   Complete the fields under Set Calculation as:
   XList: main\price
   Freq: 1
   Tap OK.



4. The answer is displayed on the screen.

The standard deviation, correct to 2 decimal places, is $s = 4.52$ cents/L.

WORKED EXAMPLE 22

The number of students attending SRC meetings during the term is given in the stem plot shown. Calculate the standard deviation for this set of data, correct to 4 significant figures, by using the formula $s = \sqrt{\dfrac{\sum(x - \bar{x})^2}{n - 1}}$.

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

Key: 1|4 = 14 students

THINK

1. Calculate the value of the mean ($\bar{x}$).

WRITE

$$\bar{x} = \frac{\sum x}{n}$$
$$= \frac{4 + 8 + 8 + 11 + 13 + 14 + 15 + 18 + 23 + 25}{10}$$
$$= 13.9$$

2. Set up a table to calculate the values of $(x - \bar{x})^2$.

x	$x - \bar{x}$	$(x - \bar{x})^2$
4	−9.9	98.01
8	−5.9	34.81
8	−5.9	34.81
11	−2.9	8.41
13	−0.9	0.81
14	0.1	0.01
15	1.1	1.21
18	4.1	16.81
23	9.1	82.81
25	11.1	123.21
$\sum(x - \bar{x})^2 = 400.9$		

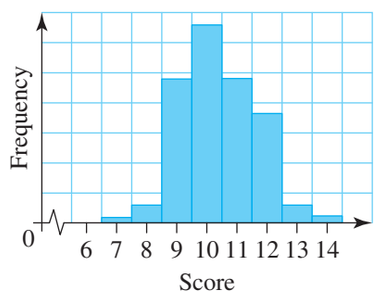
3. Enter the values into the formula to calculate the standard deviation (s).

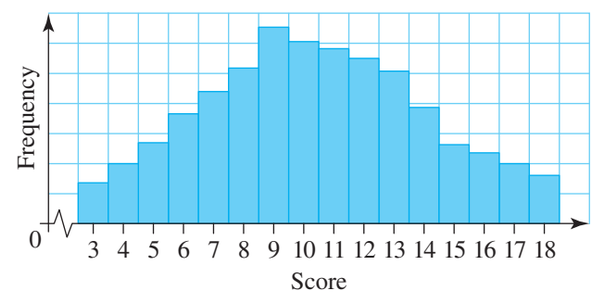
$$s = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}}$$
$$= \sqrt{\frac{400.9}{9}}$$
$$= 6.6741 \ldots$$

4. Round the value correct to 4 significant figures.

$s = 6.674$ (correct to 4 significant figures)

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





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

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

on Resources

Interactivity: The mean and the standard deviation (int-6246)

studyon

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

Measures of spread—standard deviation Summary screen and practice questions

Exercise 1.9 Standard deviation of a sample

1. **WE21** The Australian dollar is often compared to the US dollar. The value of the Australian dollar compared to the US dollar each week over a six-month period is shown.

97.2	96.8	98.0	98.3	97.5	95.9	96.1	95.6	95.0	94.8	94.6	94.9
93.9	93.3	92.9	90.6	90.4	89.6	88.1	88.3	89.5	88.8	88.5	88.1

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

2. The test results for a maths class are shown.

67	98	75	81	70	64	55	52	78	90	76
92	78	80	83	59	67	45	78	48	82	62

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

3. **WE22** Calculate the standard deviation of the data set shown, correct to 4 significant places, by using the formula $s = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}}$.

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

Key: 2|4 = 24

4. The number of students attending the service learning meetings in preparation for the year's fundraising activities is shown by the stem plot. Calculate the standard deviation for this set of data, correct to 3 decimal places, by using the formula $s = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$.

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

Key: 1|3 = 13

5. 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

6. 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

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

x	$x-\bar{x}$	$(x-\bar{x})^2$
2.3	0.6625	□
0.8	□	0.7014
1.6	−.0.0375	□
2.1	□	0.2139
1.7	0.0625	□
1.3	−0.3375	0.1139
1.4	−0.2375	0.05 64
1.9	0.2625	1.5987

$$\frac{\sum(x-\bar{x})^2}{n-1} = \frac{1.5987}{\square}$$

$$s = \square$$

7. The number of outgoing phone calls from an office each day over a 4-week period is shown in 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 4 significant figures.

8. 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

9. The speed of 20 cars (in km/h) is monitored along a stretch of road that is a designated 80 km/h zone. Calculate the standard deviation of the data, correct to 2 decimal places.



80, 82, 77, 75, 80, 80, 81, 78, 79, 78, 80, 80, 85, 70, 79, 81, 81, 80, 80, 80

10. Thirty pens are randomly selected off the conveyor belt at the factory and are tested to see how long they will last, in hours. Calculate the standard deviation of the data shown, correct to 3 decimal places.

20, 32, 38, 22, 25, 34, 47, 31, 26, 29, 30, 36, 28, 40, 31,
26, 37, 38, 32, 36, 35, 25, 29, 30, 40, 35, 38, 39, 37, 30

11. Calculate the standard deviation of the data shown, correct to 2 decimal places, representing the temperature of the soil around 25 germinating seedlings:

28.9	27.4	23.6	25.6	21.1	22.9	29.6	25.7	27.4
23.6	22.4	24.6	21.8	26.4	24.9	25.0	23.5	26.1
23.7	25.3	29.3	23.5	22.0	27.9	23.6		

12. Aptitude tests are often used by companies to help decide who to employ. An employer gave 30 potential employees an aptitude test with a total of 90 marks. The scores achieved are shown.

67, 67, 68, 68, 68, 69, 69, 72, 72, 73, 73, 74, 74, 75, 75,
77, 78, 78, 78, 79, 79, 79, 81, 81, 81, 82, 83, 83, 83, 86
Calculate the mean and standard deviation of the data, correct to 1 decimal place.

13. The number of players attending basketball try-out sessions is shown by the stem plot. Calculate the standard deviation for this set of data, correct to 4 significant figures.

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

Key: 2|1 = 21

14. The scores obtained out of 20 at a dancing competition are shown. Calculate the standard deviation of the scores, correct to 3 decimal places.

18.5	16.5	18.0	12.5	13.0	18.0	15.5	17.5	18.5	19.0
17.0	12.5	16.5	13.5	19.0	20.0	17.5	19.5	16.0	15.5

1.10 Populations and simple random samples

1.10.1 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.



1.10.2 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 CAS technology.

WORKED EXAMPLE 23

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

THINK	WRITE
1. Find the appropriate menu in CAS to generate random integers.	
2. Generate 5 random numbers between 1 and 50.	
3. An example of a set of numbers is displayed.	{48, 46, 8, 26, 21}.

1.10.3 Displaying the data

The raw data for a sample of 12 students from our population of 95 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

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

$\bar{x} = 32.2$
$s = 6$
$Q_1 = 27.5$
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.

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.

The summary statistics for these data are as follows:
$\bar{x} = 47.5$, $s = 16.3$, $Q_1 = 35$, median $= 50$, $Q_3 = 60$.



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

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 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.

studyon

Units 3 & 4 › AOS DA › Topic 5 › Concepts 1& 2

Population Summary screen and practice questions

Sample and simple random sample Summary screen and practice questions

Exercise 1.10 Populations and simple random samples

1. WE23 Generate 5 random numbers (integers) between 1 and 100.
2. Generate 10 random numbers (integers) between 1 and 250.
3. 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
4. Generate 10 random numbers (integers) between 1 and 100.
5. Generate 20 random numbers (integers) between 1 and 500.
6. Which is larger: A population or a sample? Explain why.

7. MC When selecting students for a simple random sample of a year level, the students selected should be:
A. of similar age **B.** a group of mates **C.** independent
D. female **E.** the tallest students

8. MC The students selected for a simple random sample of a year level should be selected by:
A. a group of mates
B. a group who all dance
C. a selection of males
D. the students with the best test results
E. using random numbers

9. Would the mean be a good measure of the centre of the distribution shown? Explain.

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

Key: 1|2 = 12

10. MC The mean is a good measure of the centre of a distribution if the data is:
A. skewed left
B. symmetric
C. skewed right
D. has outliers
E. bimodal

11. MC The interquartile range is 12, since $Q_1 = 24$ and $Q_3 = 36$. The percentage of data that fit between 24 and 36 is:
A. 12% **B.** 30% **C.** 50% **D.** 68% **E.** 95%

12. 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.

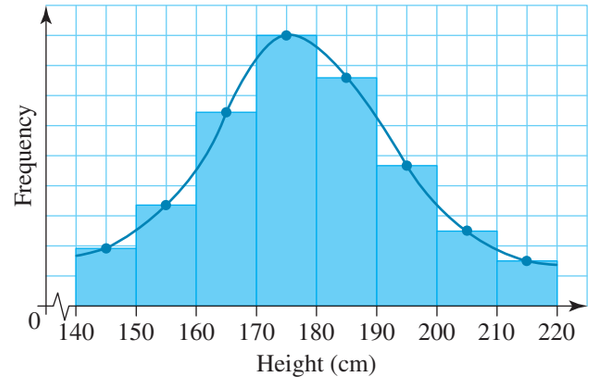
13. Repeat question **12**, but this time investigate the following for students in your year level:
a. the number of hours spent on homework each week
b. the number of hours spent working in part-time jobs.

14. Conduct a similar investigation to that which you completed in questions **12** and **13**; however, this time sample students in another year group. Compare these data with those obtained for your year level.

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

1.11.1 The 68–95–99.7% rule

The heights 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, $\bar{x}$.

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

Note: For the purpose of this course, we will use s, to denote standard deviations, but for large populations the symbol for standard deviation is σ.

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.**

Figure 1

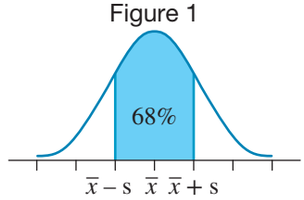


Figure 2

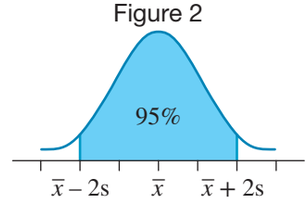
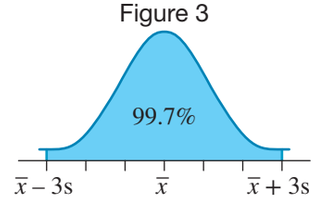


Figure 3



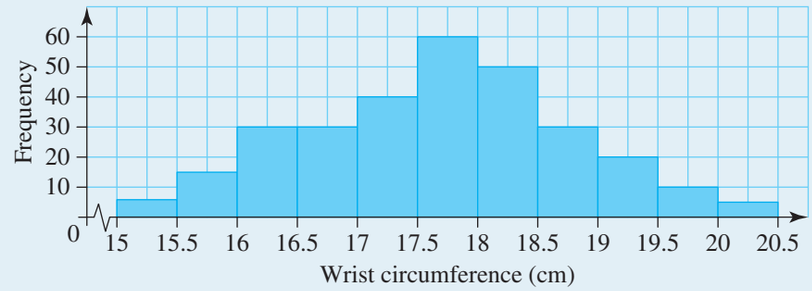
In Figure 1, 68% of the data shown lie between the value which is 1 standard deviation below the mean, that is $\bar{x} - s$, and the value which is 1 standard deviation above the mean, that is, $\bar{x} + s$.

In Figure 2, 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, 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 24

The wrist circumferences of a group of people were recorded and the results are shown in the histogram below. 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. 98.7% 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).

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

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

WRITE

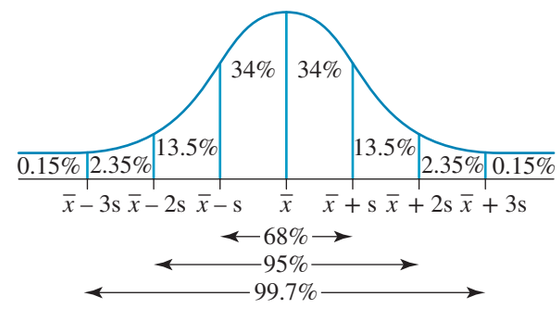
a. $\bar{x} - s = 17.7 - 0.9 = 16.8$
$\bar{x} + s = 17.7 + 0.9 = 18.6$
So approximately 68% of the people have a wrist size between 16.8 and 18.6 cm.

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

c. $\bar{x} - 3s = 17.7 - 2.7 = 15.0$
$\bar{x} + 3s = 17.7 + 2.7 = 20.4$
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.

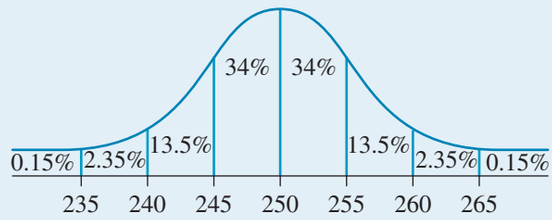
Note that 50% of the data lie below the mean and 50% lie above the mean due to the symmetry of the distribution about the mean.



WORKED EXAMPLE 25

The distribution of the masses of a number 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 26

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 is: $34\% + 50\% = 84\%$
2. Find 84% of the total sample.	Number of boxes = 84% of 200 = 168 boxes

1.11.2 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 27

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 Worked example 28, 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.

1.11.3 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

$$\begin{aligned} z &= \frac{x - \bar{x}}{s} \\ &= \frac{74 - 60}{12} \\ &= 1.17 \end{aligned}$$

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

$$\begin{aligned} z &= \frac{x - \bar{x}}{s} \\ &= \frac{63 - 50}{8} \\ &= 1.625 \end{aligned}$$

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

WORKED EXAMPLE 28

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

a. 1. Write the formula for each subject.

2. Substitute for x, $\bar{x}$ and s.

3. Calculate each z-score.

b. Explain that the subject with the highest z-score is the better result.

WRITE

a. Physics: $z = \frac{x - \bar{x}}{s} = \frac{82 - 62}{10} = 2$

Chemistry: $z = \frac{x - \bar{x}}{s} = \frac{78 - 66}{5} = 2.4$

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.

on Resources

Interactivity: 68–95–99.7% rule in a normal distribution (int-6247)

studyon

Units 3 & 4 > AOS DA > Topic 4 > Concepts 1, 2, 3

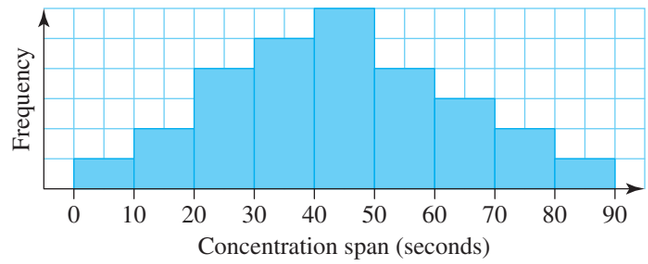
Normal distributions—the 68—95—99.7% rule Summary screen and practice questions

Standard *z*-scores Summary screen and practice questions

Comparing data values Summary screen and practice questions

Exercise 1.11 The 68–95–99.7% rule and *z*-scores

1. **WE24** The concentration ability of a 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.



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:

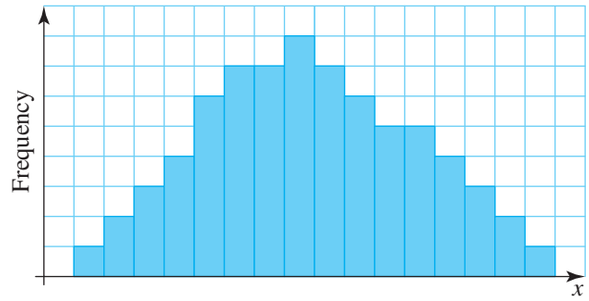
   a. 68% of the group's concentration spans to fall
   b. 95% of the group's concentration spans to fall
   c. 99.7% of the group's concentration spans to fall.

2. The monthly rainfall in Mathmania Island was found to follow a bell-shaped curve with a mean of 45 mm and a standard deviation of 1.7 mm. Write down the rainfall range which we would expect approximately:

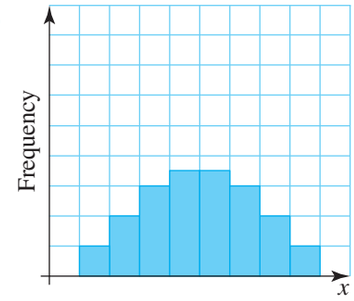
   a. 68% of the group to lie
   b. 95% of the group to lie
   c. 99.7% of the group to lie.

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

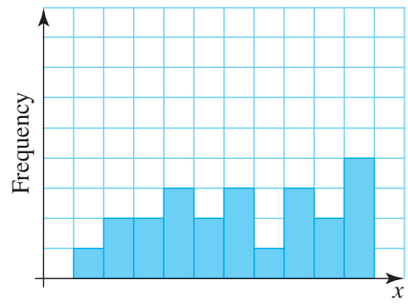
a.



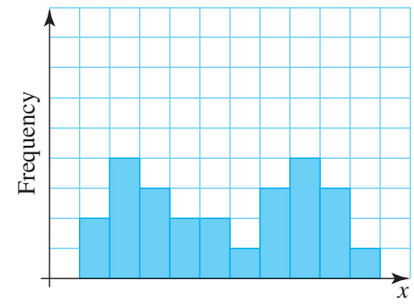
b.



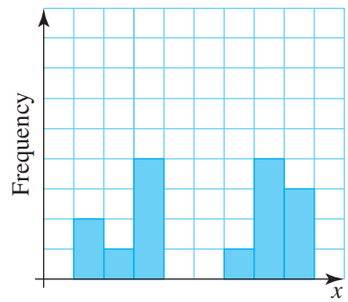
c.



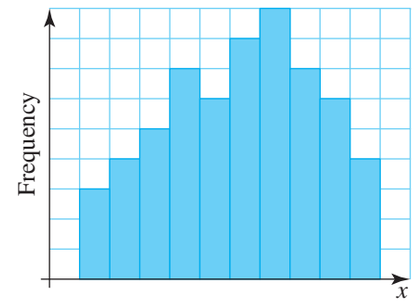
d.



e.

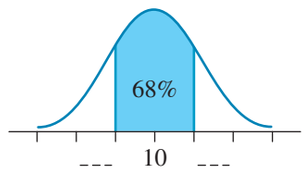


f.

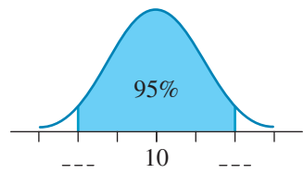


4. Copy and complete the entries on the horizontal scale of the following distributions, given that $\bar{x} = 10$ and $s = 2$.

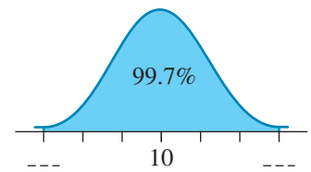
a.



b.

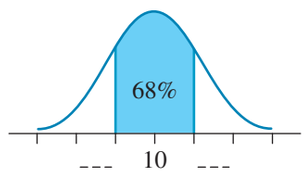


c.

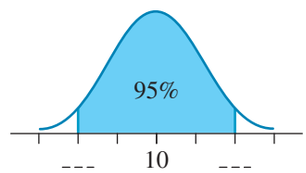


5. Copy and complete the entries on the horizontal scale of the following distributions, given that $\bar{x} = 5$ and $s = 1.3$.

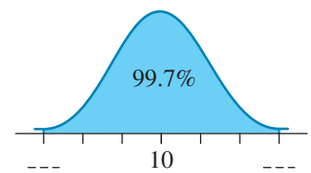
a.



b.



c.



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

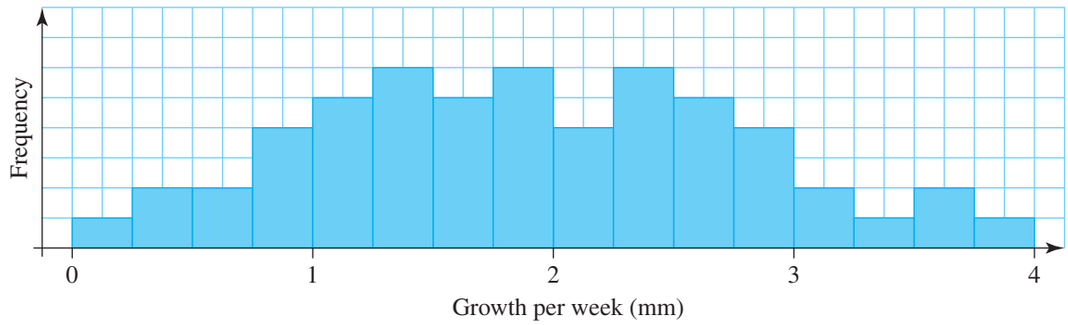
The mean growth per week was 1.9 mm and the standard deviation was 0.6 mm. Write down the hair growth rates between which approximately:

a. 68% of the values fall

b. 95% of the values fall

c. 99.7% of the values fall.





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 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

9. WE25 The distribution of masses of potato chips is known to follow a bell-shaped curve with a mean of 200 g and a standard deviation of 7 g. Find the percentage of the potato chips with a mass which is:

a. more than 214 g

b. more than 200 g

c. less than 193 g

d. between 193 g and 214 g.

10. 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. **WE26** The number of marbles in a bag is not always the same. From a sample of boxes it was found that the number of marbles in a bag approximated a normal distribution with a mean number of 20 and a standard deviation of 1. In a sample of 500 bags, how many would be expected to have more than 19 marbles?



13. 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. **WE27** In a Physics test on electric power, the mean result for the class was 76% and a standard deviation of 9%. Drew's result was 97%. Calculate his mark as a z-score.

16. 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.

17. **WE28** Bella's Specialist Maths mark was 83 and her English mark was 88. In Specialist Maths the mean was 67 with a standard deviation of 9, while in English the mean was 58 with a standard deviation of 14.
a. Convert the marks in each subject to a z-score.
b. Which subject is the better result for Bella? Explain.

18. 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.

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

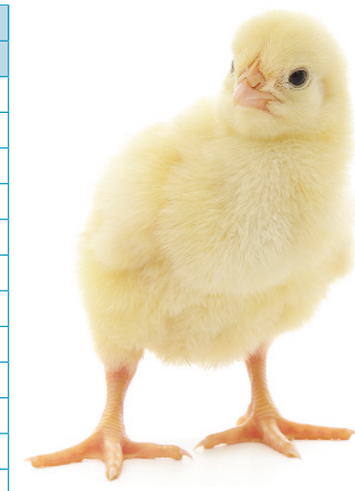
20. MC In a normal distribution the mean is 58. A score of 70 corresponds to a standardised score of 1.5. The standard deviation of the distribution is:

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

21. 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.

22. 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 correct to 3 significant figures.
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, prepare five-number summaries 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?

1.12 Review: exam practice

A summary of this topic is available in the resources section of your eBookPLUS at www.jacplus.com.au.

Multiple choice

1. **MC** 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. **MC** The observations shown on the stem-and-leaf plot are:

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

Key: 2 | 1 = 21

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

3. **MC** 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. Which of the following statements is untrue about the data?

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

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.

4. **MC** Which one of the frequency tables on the next page 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. MC The distribution of data shown in the stem plot could best be described as:

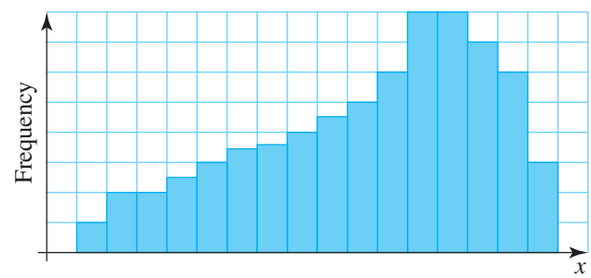
Stem	Leaf
2	3 4
2∗	5 6 8
3	0 1 2 3 4 4
3∗	5 5 7 9 9
4	0 1 3 3
4∗	6 8 8
5	0 1
5∗	6
6	
6∗	9

Key: 3|1 = 31

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

6. MC The distribution of the data shown in the histogram 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. MC 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

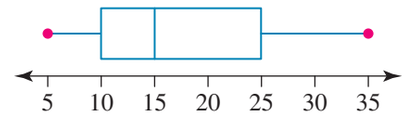
8. MC The median of the set of data shown in the stem plot below is:

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



A. the range is 35

B. the interquartile range is 10

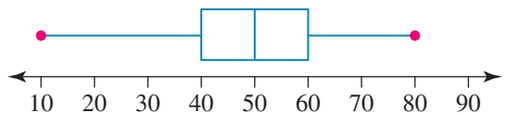
C. the median is 20

D. the interquartile range is 25

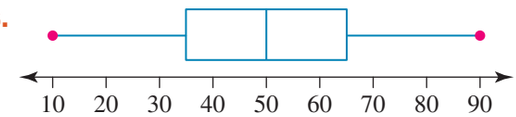
E. the median is equal to the interquartile range

10. MC 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?

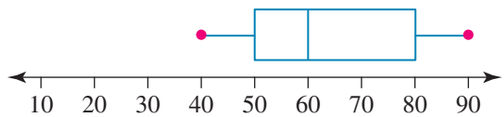
A.



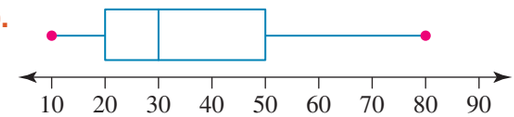
B.



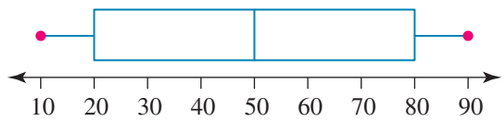
C.



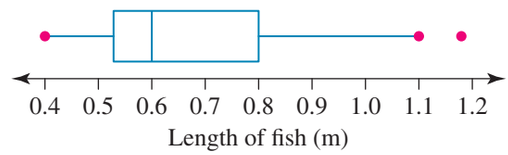
D.



E.



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

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 ∗ 15 = 15 years

A. 18

B. 18.9

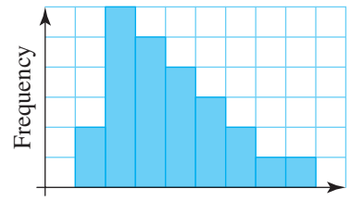
C. 19

D. 20.9

E. 21

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

A.

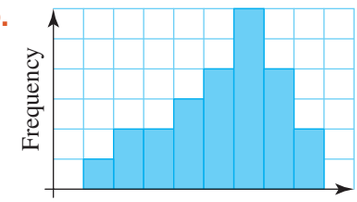


B.

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 Key: 2|4 = 24
	Key: 2|4 = 24
4	3 4
4*	

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

D.



E.



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

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

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

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

18. MC 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. MC 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. MC 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. MC 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.



22. MC When selecting simple random sample of 20 teachers from all teachers in a particular high school, the teachers should be selected by:

- **A.** Choosing teachers from the maths department
- **B.** Choosing teachers of similar age
- **C.** Assigning numbers to a list of teachers' names and using random number generator to select random numbers
- **D.** Choosing from Year 12 teachers
- **E.** Choosing first 20 teachers from the list of teachers' names

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 (correct to the nearest whole dollar) by each student in a Year 3 class on the school walkathon is shown in the stem plot.

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

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.

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. A local council monitored the use of the public swimming pool over a 12-day period. The attendance is recorded in the table below.

Attendance	275	180	201	345	480	150	500	155	302	650	362	410

a. What is the range?

b. For what percentage of days was the attendance greater than 300? Give your answer correct to one decimal place.

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.
Calculate, correct 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.

11. Experimental physicist recorded two measurements using log (base 10) scale. The measurements were 8.2 and 6.6. How many times greater the first measurement compared to the second? Give your answer correct to the nearest whole number.

12. The quarterly incomes (in dollars) of the sample of 12 local businesses was recorded as follows:
$45 000, $20 000, $360 000, $750 000, $3 200, $1 048 500,
$34 590, $37 250, $65 710, $290 060, $59 070, $8 450
Complete the frequency table below for log (base 10) of incomes using class intervals of width 1.

$\log_{10}$(income)	Frequency

Extended response

1. Mr Fahey gives the same test to the two Year 10 classes that he teaches, 10C and 10E. This 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 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 (correct to 2 decimal places) for this sample and use it, along with the mean, to estimate the number of fish (out of 10 000) that would have 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 of the sample correct 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 given data, comment on this statement.
d. Using the mean and standard deviation (correct to 2 decimal places) for this sample of 50 birth weights, estimate 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.

studyon

Units 3 & 4 Sit exam

Answers

Topic 1 Investigating Data Distributions

Exercise 1.2 Types of data

1. D
2. D
3. Numerical: **a**, **b**, **c**
4. Categorical: **c**, **d**, **e**, **f**, **g**
5. C
6. B
7. Discrete: **c**
   Continuous: **a**, **b**
8. C
9. C
10. Categorical
11. B
12. A
13. Categorical and ordinal
14. Discrete
15. Ordinal
16. D

Exercise 1.3 Stem plots

1.

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

Key: 1|6 = 16

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 = $5

The busker's earnings are inconsistent.

3.

Stem	Leaf
86	8
87	7
88	0
89	8
90	2 4 8 9
91	
92	0 1 2 6
93	
94	3 9
95	
96	
97	0
98	3 5 9

Key: 86|8 = 86.8%

4.

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

Key: 18|5 = 1.85 cm

5.

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

Key: 3|7 = 37 years

It seems to be an activity for older people.

6. C

7.

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|5 = 25 years

More than half of the parents are 30 or older with a considerable spread of ages, so this statement is not very accurate.

8.

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|1 = 21 hit outs

Bulldogs, Melbourne, St Kilda

9.

Stem	Leaf
33	0
34	
35	0 0 1
36	5
37	3
38	0 0
39	0 0 5
40	0 6
41	0 5
42	1 3
43	0 0
44	
45	0

Key: 33|0 = $330

The stem plot shows a fairly even spread of rental prices with no obvious outliers.

10. a.

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

Key: 2|0 = 20 points

b.

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

Key: 2|0 = 20 cm

11. a.

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

Key: 4|3 = 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|3 = 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|3 = 43 cm

12. 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

13. 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|5 = 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|5 = 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|5 = 15 mm

Values are bunched together; they vary little.

14.

Stem	Leaf
7	2 8
8	3 3 5 7 8 8
9	0 1 2 2 3 4 8 9
10	0 2 4
11	2

Key: 7|2 = 72 shots

15.

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

Key: 7|1 = 71 net score

The handicapper has done a good job as most of the net scores are around the same scores; that is, in the 70s.

16. a.

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

Key: 6|0 = 60%

b.

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

Key: 6|0 = 60%

17. a.

Computer 1	Stem	Computer 2
5	34	0 2 6 8
8 2	35	2 3 5 5 7 8
6 3	36	1 2
6	37	
1 0	38	
2 1	39	
5	40	
0	41	

Key: 34|0 = 340 minutes

b. Computer 1 lasts longer but is not as consistent.
Computer 2 is more consistent but doesn't last as long.

18. a.

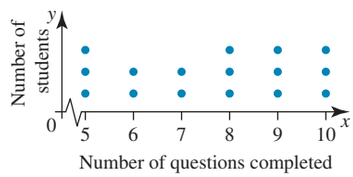
Year 8	Stem	Year 10
9 8	14	
7 5 5 5 3 1 0	15	2 4 6 8 9
8 6 5 4 3 2 1 0	16	0 4 5 7 7 9
5 2 1	17	2 3 4 6 7 8 8
	18	2 5

Key: 14|8 = 148 cm

b. As you would expect the Year 10 students are generally taller than the Year 8 students; however, there is a large overlap in the heights.

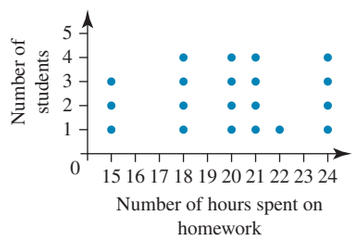
Exercise 1.4 Dot plots, frequency tables and histograms, and bar charts

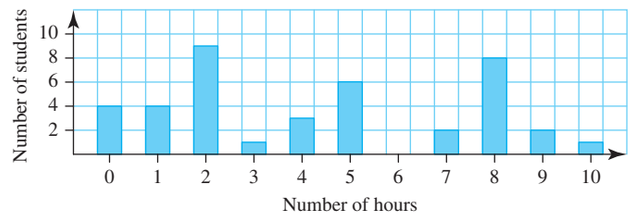
1.



2. See graph at the bottom of the page

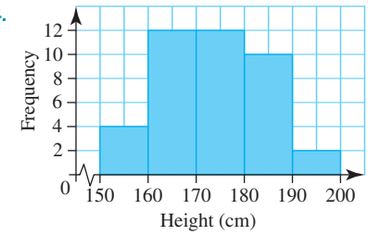
3.





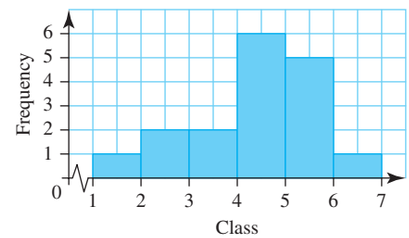


4.



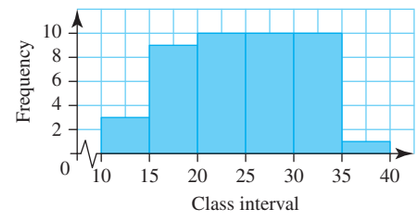
5. a.

Class	Frequency
1–	1
2–	2
3–	2
4–	6
5–	5
6–	1



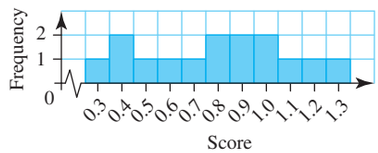
b.

Class interval	Frequency
10–	3
15–	9
20–	10
25–	10
30–	10
35–	1



c.

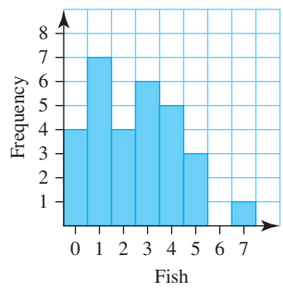
Score	Frequency
0.3	1
0.4	2
0.5	1
0.6	1
0.7	1
0.8	2
0.9	2
1.0	2
1.1	1
1.2	1
1.3	1



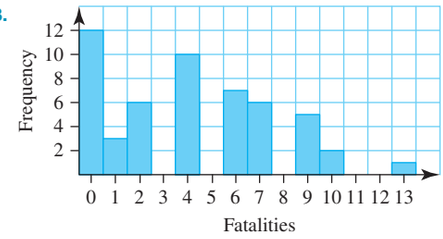
6.

Number of dogs	Tally	Frequency
2	II	2
3	III	3
4	II	2
5	II	2
6	I	1
7	IIII	5
8	II	2
9	IIII III	8
10	III	3
11		0
12	II	2
Total		30

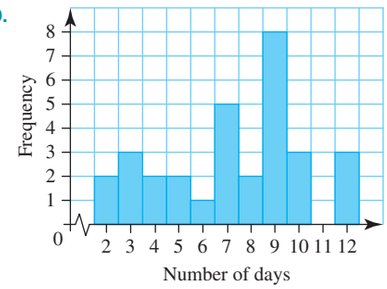
7.



8.

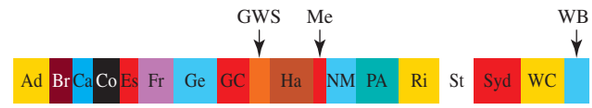


9.

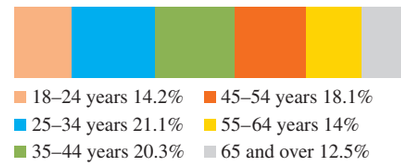


10. Check your histogram against that shown in the answer to question **9**.

11.

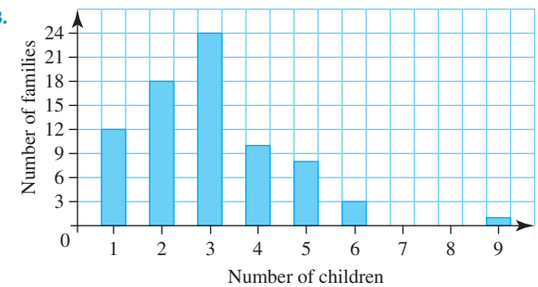


12. Participation in activities

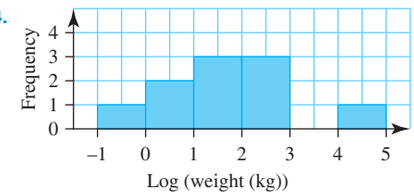


The statement seems untrue as there are similar participation rates for all ages. However, the data don't indicate types of activities.

13.



14.



15. D

16. D

17. **a.** B **b.** A **c.** D

18. A

19. B

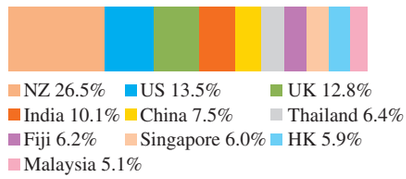
20. 16

21. 5 times

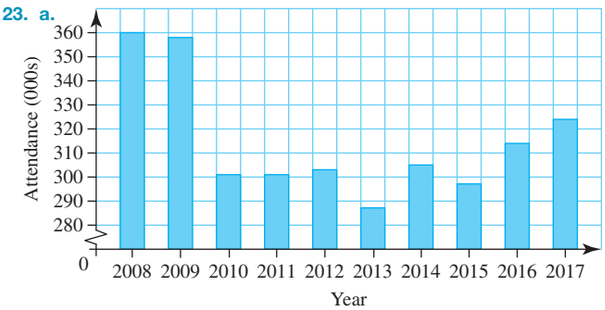
22. **a.**

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.



23. **a.**



b. Check your bar chart against that shown in the answer to part **a**.

Exercise 1.5 Describing the shape of stem plots and histograms

1. Positively skewed

2. Negatively skewed

3. **a.** Symmetric **b.** Negatively skewed
c. Positively skewed **d.** Symmetric
e. Symmetric **f.** Positively skewed

4. **a.** Symmetric, no outliers
b. Symmetric, no outliers
c. Symmetric, no outliers
d. Negatively skewed, no outliers
e. Negatively skewed, no outliers
f. Positively skewed, no outliers

5. E

6. C

7. Negatively skewed

8. 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.

9. **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.

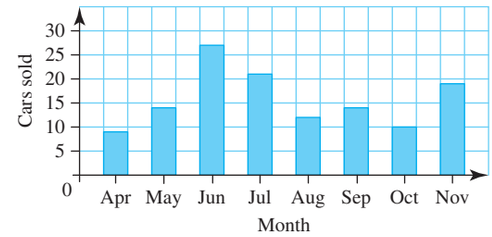
10. **a.** Symmetric
b. Most students receive about $8 (give or take $2).

11. **a.** Positively skewed
b. **i.** 15 **ii.** 85%

12. **a.** Positively skewed
b. Since most of the data is linked to the lower stems, this suggests that some students do little exercise, but those students who exercise, do quite a bit each week. This could represent the students in teams or in training squads.

13. **a.** Club A: negatively skewed
Club B: positively skewed
b. Since Club A has more members of its bowling team at the higher stems as compared to Club B; you could say Club A has the older team as compared to Club B.
c. **i.** Club A: 11 members over 70 years of age
ii. Club B: 4 members over 70 years of age.

14. **a.**



b. Positively skewed
c. June, July and November represent the months with the highest number of sales.
d. This is when the end of financial year sales occur.

Exercise 1.6 The median, the interquartile range, the range and the mode

1. Median = 33

2. Median = 36.5 goals

3. IQR = 14

4. IQR = 8

5. IQR = 6.5

6. IQR = 3.3

7.

	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

8.

	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

9. a. 10
b. 8
c. The IQRs (middle 50%) are similar for the two restaurants, but they don't give any indication about the number of cars in each data set.

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

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

12. C

13.

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

14.

	Median	IQR	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).

15. a. Range = 72, Median = 37.5, Mode = 46, IQR = 22
b. Range = 47
Median = 422
Mode = 411
Mode = 20

16. Median = 7, Mode = 7

17. $Q_1 = 42.2$, $Q_3 = 48.15$, IQR = 5.95, Median = 45.1

18. a. Median = 93, $Q_1 = 91.5$, $Q_3 = 97$, IQR = 5.5, Range = 30, Mode = 93
b. The average handicap of the golfers should be around 21.

Exercise 1.7 Boxplots

1. Range = 39
Median = 25
IQR = 19

2. Range = 3
Median = 7.5
IQR = 1.4

3.

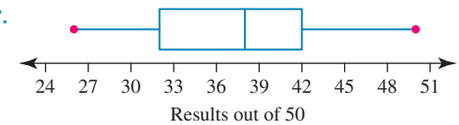
	Range	IQR	Median
a	12	6	8
b	7	2	5
c	350	100	250
d	100	30	65
e	20	10	25

4. They could represent the same data.

5. They could represent the same data.

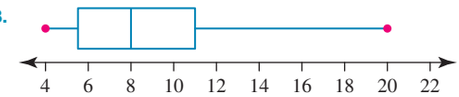
6. a. iii **b.** iv **c.** i **d.** ii

7.



Negatively skewed; 50% of results are between 32 and 42.

8.



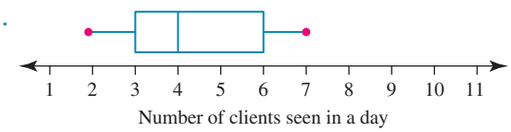
Fairly symmetrical.

9. The boxplots should show the following:

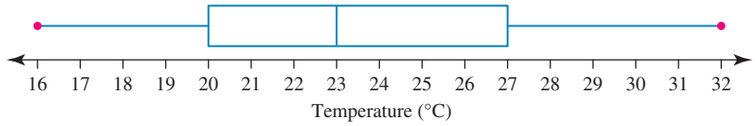
	Min	Q_1	Med	Q_3	Max
a	3	6	8.5	14	18
b	3	5	7	9	12
c	4.3	4.6	5	5.4	5.6
d	11	15.5	18	20	22
e	0.4	0.7	0.9	1.1	1.3

10. D

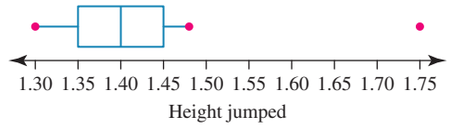
11.



12.

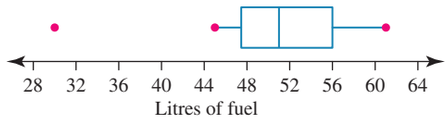


13. a.



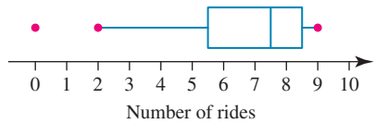
b. The data is symmetrical and 1.75 is an outlier.

14.



30.3 is an outlier.

15. a.



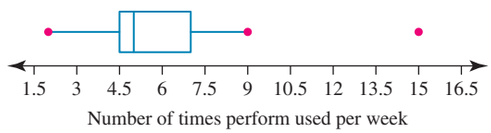
The data are negatively skewed with an outlier on the lower end. The reason for the outlier may be that the person wasn't at the show for long or possibly didn't like the rides.

16. a. Two similar properties: both sets of data have the same minimum value and similar IQR value.

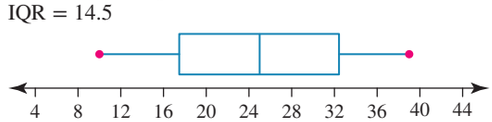
b. Boys IQR = 16
Girls IQR = 16.5

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.

17. Median = 5
$Q_1 = 4.5$
$Q_3 = 7$
$\text{Min}_x = 2$, $\text{Max}_x = 15$
IQR = 2.5
15 is an outlier.



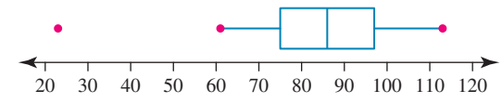
18. a. Median = 25
$Q_1 = 17.5$
$Q_3 = 32$
$\text{Min}_x = 11$, $\text{Max}_x = 39$
IQR = 14.5



b. No outliers

c. Check your boxplot against that shown in the answer to part **a**.

19. Median = 86
$Q_1 = 75$
$Q_3 = 97$
$\text{Min}_x = 23$, $\text{Max}_x = 113$
IQR = 22
23 is an outlier.



Exercise 1.8 The mean of a sample

1. 23.46

2. 8.26

3. a. 7.2 **b.** 7.125 **c.** 4.9875
d. 16.7 **e.** 0.8818

4. 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.

5. 12

6. D

7. A

8. a. Median **b.** Mean **c.** Median **d.** Median

Exercise 1.9 Standard deviation of a sample

1. 3.54 cents

2. 14.27%

3. 9.489

4. 7.306

5. a. 1.21 **b.** 2.36 **c.** 6.01 **d.** 2.45 **e.** 0.06

6. 0.48%

7. 15.49

8. C

9. 2.96 km/h

10. 6.067 pens

11. 2.39 °C

12. $\bar{x} = 75.7$, $s = 5.6$

13. 3.786 players

14. 2.331

Exercise 1.10 Populations and simple random samples

1. Answers will vary.

2. Answers will vary.

3. B

4. Answers will vary.

5. Answers will vary.

6. Population is larger, since a sample is taken from the population.
7. C
8. E
9. Yes, because the distribution is reasonably symmetric with no outliers.
10. B
11. C
12. Answers will vary.
13. Answers will vary.
14. Answers will vary.

Exercise 1.11 The 68–95–99.7% rule and *z*-scores

1. a. 68% of group's concentration span falls between 35 secs and 63 secs.
   b. 95% of group's concentration span falls between 21 secs and 77 secs.
   c. 99.7% of group's concentration span falls between 7 secs and 91 secs.
2. a. 68% of the group to lie between 43.3 mm and 46.7 mm
   b. 95% of the group to lie between 41.6 mm and 48.4 mm
   c. 99.7% of the group to lie between 39.9 mm and 50.1 mm
3. a. Yes b. Yes c. No
   d. No e. No f. Yes
4. a. 8 and 12
   b. 6 and 14
   c. 4 and 16
5. a. 3.7 and 6.3
   b. 2.4 and 7.6
   c. 1.1 and 8.9
6. a. 1.3 mm and 2.5 mm
   b. 0.7 mm and 3.1 mm
   c. 0.1 mm and 3.7 mm
7. a. 5 and 9
   b. 3 and 11
   c. 1 and 13
8. C
9. a. 2.50% b. 50% c. 16% d. 81.5%
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. 420 bags
13. a. 336 containers
    b. 10 containers
    c. 380 containers
14. a. i. 1360 ii. 1950 iii. 317
    b. 100
15. 2.33
16. 3
17. a. Specialist: $\bar{x} = 67, s = 9$
    English: $\bar{x} = 58, s = 14$
    $z_s = 1.78, z_e = 2.14$
    b. English has the higher result as it has the higher *z*-score.
18. a. English 1.25, Maths 1.33
    b. Maths mark is better as it has a higher *z*-score.
19. a. −0.48 b. 1.44 c. 0.08 d. −2.24 e. 2.8
20. B
21. Second test, Barbara's *z*-score was −0.33 compared to −0.5 in the first test.
22. a. Barn: $\bar{x} = 4.4$ $s = 0.3$
    FR: $\bar{x} = 4.1$ $s = 0.2$
    b. 1.18
    c. 84%
    d.

	Cage	Barn	Free range
Min_x	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
Max_x	5.8	4.9	4.4

    i. Cage: 5.15
    Barn: 4.35
    Free: 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.

1.12 Review: exam practice

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	22. C		

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. See first graph at the bottom of the page

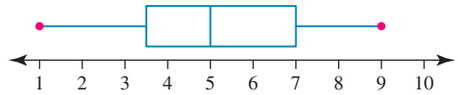
c. The data are approximately symmetrical. More than half the drivers exceeded the limit. The fastest drivers were about 15 km/h over the limit. Many other conclusions are possible.

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. a. 500 **b.** 58.3%

8. 0.05 mL

9. a. i. 10.8 and 13.2 years

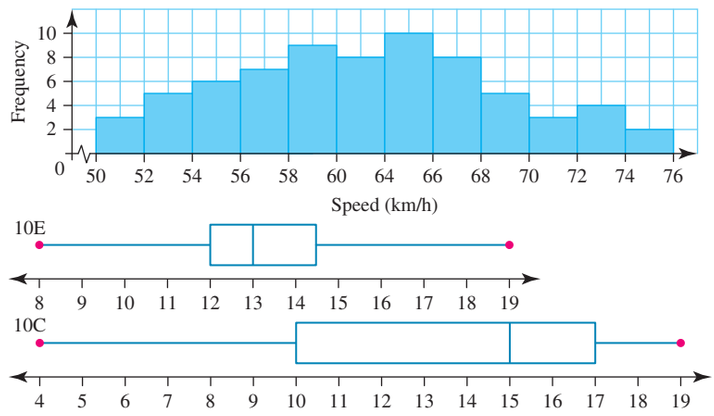
ii. 9.6 and 14.4 years

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).

11. 40 times



12.

$\log_{10}$(income)	Frequency
3 – <4	2
4 – <5	6
5 – <6	3
6 – <7	1

Extended response

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

	10C		10E
Stem	Leaf	Stem	Leaf
0		0	
0		0	
0	4	0	
0	7 7	0	
0	9 9	0	8 9
1	0 0 1	1	0 1 1
1	2 3	1	2 2 2 3 3 3 3 3
1	4 4 5 5 5	1	4 4 4 4 4 4 5 5 5
1	6 7 7 7 7	1	6 6
1	8 8 8 9 9	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	IQR	Range	Mode
10C	15	7	15	17
10E	13	2.5	11	14

iii. See second graph at the bottom of the page

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.

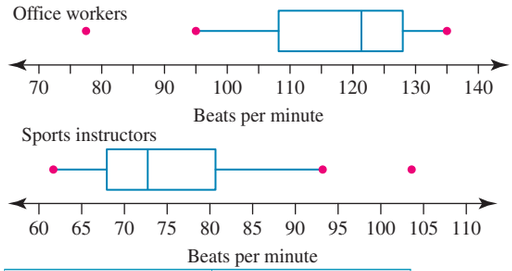
b. We use the median of the 10C scores to give us an indication of the centre of the distribution. The median of the 10C scores is 15. For 10E, the distribution is symmetric and hence we use the mean to give us an indication of the centre of the distribution. The mean of 10E scores is 13.2. The range 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

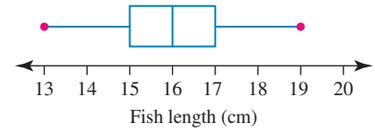
b.

Stem	Leaf
1	
1	3 3
1	4 4 5 5 5
1	6 6 6 6 6 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



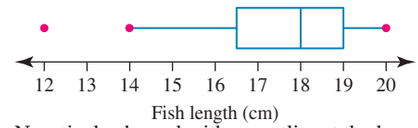
d. Symmetric

e. Given the data itself, the boxplot does.

f. Mean = 16 cm = median

g. $s = 1.75$ cm, 1600

h.



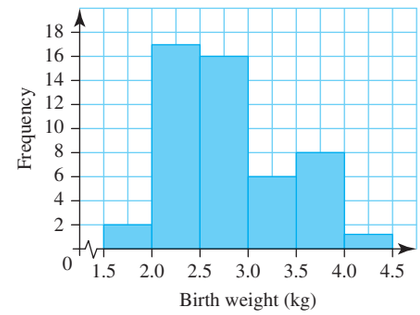
Negatively skewed with an outlier at the lower end

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	Midpoint
1.5 – 1.9	II	2	1.7
2.0 – 2.4	𝍸 III	8	2.2
2.5 – 2.9	𝍸 𝍸 𝍸 II	17	2.7
3.0 – 3.4	𝍸 𝍸 𝍸 I	16	3.2
3.5 – 3.9	𝍸 I	6	3.7
4.0 – 4.4	I	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