

STATISTICAL ANALYSIS

7

STATISTICS

In this chapter, you will study different ways of describing, displaying and summarising statistical data. You will look at measures of central tendency and spread, and use these to interpret and compare data.

CHAPTER OUTLINE

7.01 Types of data
7.02 Displaying numerical and categorical data
7.03 Measures of central tendency
7.04 Quartiles, deciles and percentiles
7.05 Range and interquartile range
7.06 Variance and standard deviation
7.07 Shape and modality of data sets
7.08 Analysing data sets



IN THIS CHAPTER YOU WILL:

- identify different types of data
- display data in tables and graphs
- calculate measures of central tendency: the mean, median and mode
- calculate measures of spread: the range, quantiles, interquartile range, variance and standard deviation
- identify outliers
- recognise different modalities and shapes of data sets
- identify bias in data
- compare 2 sets of data

iStockphoto.com/Sean Pavone

TERMINOLOGY

bar chart: Graph with vertical or horizontal columns, also called a **column graph**.
bimodal: A graph with 2 peaks.
box plot: Graphical display of five-number summary, also called a **box-and-whisker plot**.
categorical data: Data that are named by categories.
continuous data: Numerical data that can take any value that lies within an interval.
decile: One of the values that divide a data set into 10 equal parts.
discrete data: Numerical data that can only take specific distinct values.
dot plot: A column graph of dots.
five-number summary: The lowest and highest values, median, and lower and upper quartiles of a data set.
frequency polygon: Frequency line graph.
histogram: Bar chart of frequencies with no gap between columns.
interquartile range: Measure of spread, the difference between the upper and lower quartiles.
mean: Average score, calculated by dividing the sum of scores by the total number of scores.
median: The middle score when all scores are placed in order.
modality: The number of peaks in a set of data.
mode: The score with the highest frequency.
multimodal: Having many peaks in a set of data.
nominal data: Categorical data that is listed by name with no order.
numerical data: Data whose values are numbers.
ogive: Cumulative frequency polygon.
ordinal data: Categorical data that can be ordered.
outlier: A score that is clearly apart from other scores – it may be much higher or lower than the other scores.
Pareto chart: A chart containing both a bar chart and a line graph where individual values are represented in descending order by the bars and the cumulative total is represented by the line graph.
percentile: One of the values that divide a data set into 100 equal parts.
pie chart: Circular graph showing categories as sectors.
quantile: One of the values that divide a data set into equal parts.
quartile: One of the values that divide a data set into 4 equal parts.
range: Difference between the highest and lowest scores.
skewness: The shape or asymmetry of a graph to one side.
standard deviation: Measure of the spread of data values from the mean. The square root of variance.
stem-and-leaf plot: Graphical display of tens (stem) and units (leaves).
symmetrical distribution: A distribution where the left and right sides are mirror images of each other.
two-way table: A table that combines the effects of 2 separate variables (usually categorical).
variance: Measure of spread, the square of standard deviation.



Statistical data match-up

7.01 Types of data

There are many different types of data, for example, the type of public transport people use to go to work, the heights of basketball players or the marks students gain in an exam.

There are 2 main types of data:

- **Categorical data** uses categories described by words or symbols
- **Numerical data** uses numbers or quantities.

ISBN 9780170413220

These types can be divided further:

Categorical and numerical data

Categorical data:

- **Nominal data**, which cannot be put in order
- **Ordinal data**, which can be ordered

Numerical data:

- **Discrete data**, which can be counted as separate values
- **Continuous data**, which is measured along a smooth scale

For example:

- Public transport – bus, train, tram, ferry – is **categorical nominal** data since it cannot be put into an order.
- Ratings – strongly disagree, disagree, agree, strongly agree – are **categorical ordinal** since they can be put in order.
- Shoe sizes – $6, 6\frac{1}{2}, 7, 7\frac{1}{2}, \ldots$ – are **numerical discrete** since they can be counted.
- Heights of basketball players – 181 cm, 173.64 cm, 192.1 cm … – are **numerical continuous** since they are along a smooth scale.

EXAMPLE 1

Describe each type of data.

a The breeds of dogs

b Exam marks

c The volume of water in a dam

d Audience size for TV programs

e Makes of cars

f Months of the year

Solution

a Categorical nominal

b Numerical discrete

c Numerical continuous

d Numerical discrete

e Categorical nominal

f Categorical ordinal

ISBN 9780170413220

Exercise 7.01 Types of data

1 State whether each type of data is categorical (C) or numerical (N).

a Length of a fence
b Number of koalas in captivity
c Shoe size
d Colour
e Area of land
f Scores on a test
g Number of lollies in a packet
h Gender
i Speed
j Type of swimming strokes
k Attendance at a football match
l Meals on a menu
m Width of a building
n Age
o Weight
p Ranking of quality of a movie
q Surface area of a balloon as it is blown up
r Shirt sizes
s Type of sports offered at a school
t Length of a swimming race

2 State whether each type of data is categorical nominal (N), categorical ordinal (O), numerical discrete (D), numerical continuous (C).

a Survey of radio stations: Excellent, very good, good, poor, very poor
b Weight of truck loads
c Make and model of motorbikes
d Eye colour
e Volume of water in rivers
f Scores on a maths exam
g Number of jellybeans in a packet
h Nationality
i Acceleration
j Olympic sports
k Concert attendance
l Choice of desserts on a menu
m Types of trees in a park

3 Give 3 examples of:

a categorical data
b numerical data
c numerical discrete data
d categorical ordinal data
e numerical continuous data
f categorical nominal data

INVESTIGATION

DATA COLLECTION

Certain organisations are specially set up to collect and analyse data. The Australian Bureau of Statistics (ABS) collects all sorts of data, including the organisation of a regular census.

The census attempts to collect details of every person living in Australia on a particular day. Questions asked include where a person lives, occupation, salary, number of children, religion and marital status. Governments and other organisations use this data to plan future policies in areas such as education, transport, housing. For example, if the number of children in a certain region is increasing, then extra schools could be planned in that area.

ISBN 9780170413220

There is evidence that a census was done back in ancient Roman times. Investigate the methods that the Romans or some other ancient civilisation used for collecting data and writing reports.

What information do you think a census should collect? Is there information that you think that is not useful or invades privacy and therefore should not be collected?

Go to the ABS website and find out more about what this organisation does. Other worldwide organisations such as the World Health Organization (WHO) and the United Nations also collect data. Research these and other organisations that collect data, such as universities and the CSIRO.

7.02 Displaying numerical and categorical data

Data can be displayed in many different ways using tables and graphs.

WS Graphing data

WS Two-way tables and Pareto charts

WS Every picture tells a story

WS Cumulative frequency graphs

Numerical data

Frequency tables, **histograms**, **frequency polygons** and other graphs can be used to display numerical data.

EXAMPLE 2

a For the following Year 12 English essay marks (out of 10):

8, 4, 5, 4, 8, 6, 7, 8, 9, 5, 6, 7, 7, 5, 4, 6, 7, 9, 3, 5, 5

- **i** draw a frequency distribution table
- **ii** draw a histogram for this data
- **iii** draw a frequency polygon on the same set of axes as the histogram
- **iv** how many scores are less than 5?
- **v** what percentage of scores are over 6?

b The assessment scores for a Year 12 mathematics class are below.

75, 53, 58, 71, 68, 51, 60, 87, 62, 62, 89, 65, 69, 47, 70, 72, 75, 68, 76, 83, 62, 88, 94, 53, 85

- **i** Draw a frequency table that shows the results of the class test using groups of 40–49, 50–59 and so on.
- **ii** Add a column for class centre and cumulative frequency.
- **iii** Draw a cumulative frequency histogram and a cumulative frequency polygon (ogive).
- **iv** What percentage of students scored less than 60?

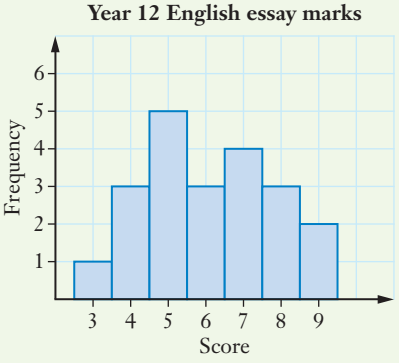
ISBN 9780170413220

Solution

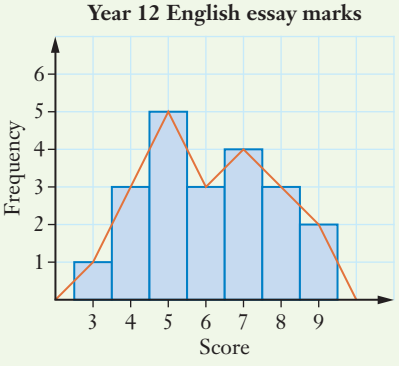
a **i** The scores range from 3 to 9. We arrange them in a table as shown.

Score	Tally	Frequency
3	|	1
4	| | |	3
5	𝍸	5
6	| | |	3
7	| | | |	4
8	| | |	3
9	| |	2

ii The histogram is a bar chart or column graph where the centre of the column is lined up with the score and the columns join together.



iii The frequency polygon is a line graph as shown. It starts and ends on the horizontal axis.



iv Reading from either the table or the graph, scores of 3 and 4 are less than 5. There is one score of 3 and 3 scores of 4.

So there are $1 + 3 = 4$ scores less than 5.

v There are $4 + 3 + 2 = 9$ scores over 6 out of a total of 21 scores.

$$\frac{9}{21} \times 100\% \approx 42.9\%$$

ISBN 9780170413220

b **i**

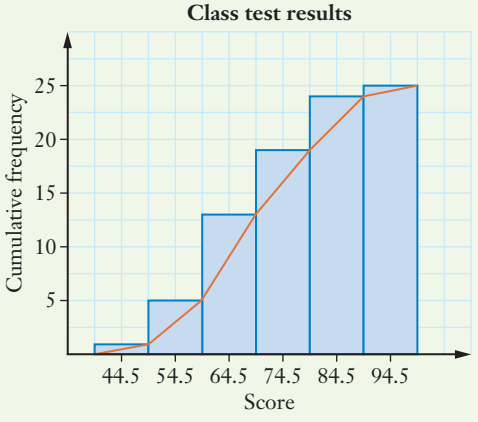
Scores	Tally	Frequency
40–49	|	1
50–59	||||	4
60–69	𝍸 |||	8
70–79	𝍸 |	6
80–89	𝍸	5
90–99	|	1

ii The class centre is the average of the highest and lowest possible score in each group. For example, $\frac{40+49}{2}=44.5$.

Add each score to the previous total for cumulative frequencies.

Scores	Class centre	Frequency	Cumulative frequency
40–49	44.5	1	1
50–59	54.5	4	5
60–69	64.5	8	13
70–79	74.5	6	19
80–89	84.5	5	24
90–99	94.5	1	25

iii Use the class centres for the scores on the graph. The cumulative frequency polygon or **ogive** starts at the bottom left of the first column and ends at the top right corner of the last column.



iv 5 students scored less than 60.

$\frac{5}{25}\times 100\%=20\%$.

ISBN 9780170413220

You can also draw **stem-and-leaf plots** to show discrete data.

EXAMPLE 3

The heartbeat rates in beats per minute of a sample of hospital patients were taken:

75, 53, 58, 71, 68, 51, 60, 87, 62, 62, 89, 65, 69, 47, 70, 72, 75, 68, 76, 83, 62, 88, 94, 53, 85

Draw a stem-and-leaf plot to display these scores.

Solution

On the left of a vertical line, put in the 10s for the scores (the stem). On the right, place the unit for each score in order (the leaf).

For example, for a score of 68 show 6 | 8.

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

Note: The stem-and-leaf plot keeps the actual scores whereas grouping them into a frequency distribution table loses this individual information.

Categorical data

We can display categorical data in different tables and graphs, including **two-way tables**, **bar charts**, **pie charts** and **Pareto charts**.

EXAMPLE 4

In a survey of Year 12 students, it was found that 47 students had a dog but not a cat, 19 had both a dog and a cat, 32 had a cat but not a dog, and 54 had neither a dog nor a cat.

a Draw a two-way table showing this data.

b Find the percentage of students who have:

- **i** both a dog and cat
- **ii** a cat but not a dog
- **iii** neither a cat nor a dog

ISBN 9780170413220

Solution

a A two-way table separates out the students with dogs from those with cats.

	Has a dog	Does not have a dog
Has a cat	19	32
Does not have a cat	47	54

Note that this table could be the other way around, with the cats at the top and the dogs down the side.

b There are $19 + 32 + 47 + 54 = 152$ students altogether.

i 19 students out of 152 have both a dog and a cat: $\frac{19}{152} \times 100\% \approx 12.5\%$

ii 32 students out of 152 have a cat but not a dog: $\frac{32}{152} \times 100\% \approx 21.1\%$

iii 54 students out of 152 have neither a dog nor a cat: $\frac{54}{152} \times 100\% \approx 35.5\%$

EXAMPLE 5

The table shows the eye colour of students.

Represent this data in:

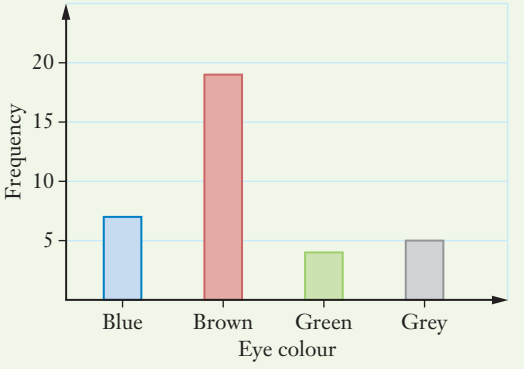
a a bar chart

b a pie chart

Colour	Frequency
Blue	7
Brown	19
Green	4
Grey	5

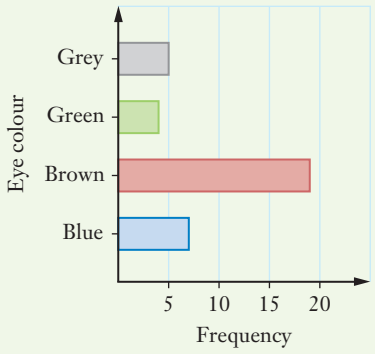
Solution

a Unlike a histogram, in a bar chart the columns do not need to join up.



ISBN 9780170413220

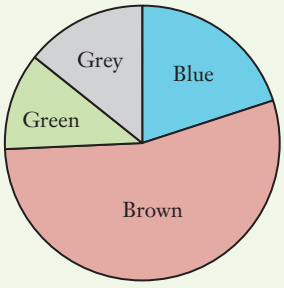
You can also draw the data as a horizontal bar chart like this.



b A pie chart is a circle divided into portions (sectors). Since the angle inside a circle is 360°, each frequency is a proportion of 360°.

There were 35 students surveyed.

Colour	Frequency	Angle
Blue	7	$\frac{7}{35} \times 360° \approx 72°$
Brown	19	$\frac{19}{35} \times 360° \approx 195°$
Green	4	$\frac{4}{35} \times 360° \approx 41°$
Grey	5	$\frac{5}{35} \times 360° \approx 51°$



Note: The number of degrees calculated adds to only 359° because the answers are not exact. This will not greatly affect the pie chart.

WS
Pareto chart template

A **Pareto chart** is useful for displaying categorical data from the most to the least important.

EXAMPLE 6

The table shows a survey group's preferences for types of TV shows.

a Arrange the table in descending order of frequency and add a percentage frequency column and a cumulative percentage frequency column.

b Draw a Pareto chart to show this data.

Type	Frequency
News	68
Drama	78
Comedy	73
Reality	107
Sport	174
	500

ISBN 9780170413220

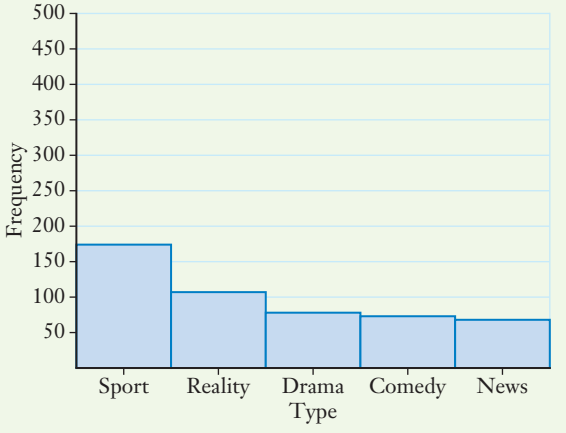
Solution

a

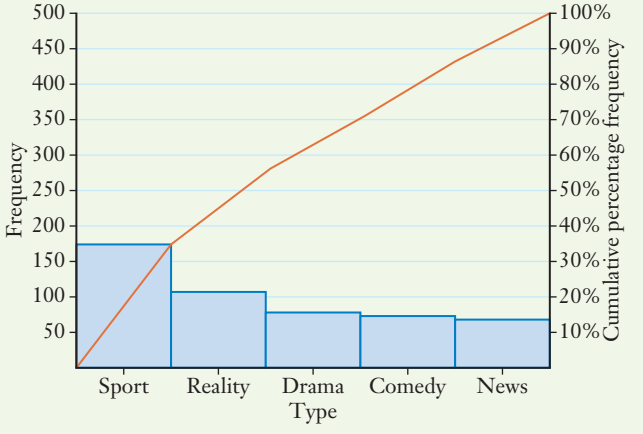
Type	Frequency	Percentage frequency	Cumulative percentage frequency
Sport	174	34.8%	34.8%
Reality	107	21.4%	56.2%
Drama	78	15.6%	71.8%
Comedy	73	14.6%	86.4%
News	68	13.6%	100%
Total	500		

For example, percentage frequency for Sport $= \frac{174}{500} \times 100\% = 34.8\%$

b Step 1: Draw a bar chart of the frequencies using the left axis.



Step 2: Draw a line graph of cumulative percentages using the right axis.



ISBN 9780170413220


Statistical graphs

INVESTIGATION

GRAPHS AND SPREADSHEETS

You can draw different types of graphs, including Pareto charts, using a spreadsheet. Enter the data into the spreadsheet, highlight the table and select the chart you want to use. If you are not sure of how to do this, search for online tutorials.

DID YOU KNOW?

Vilfredo Pareto

The Pareto chart is named after **Vilfredo Pareto** (1848–1923), an economist, sociologist, engineer and philosopher. The chart can be used as a tool for quality control.

Research the Pareto chart, the Pareto principle and the 80/20 rule. Find examples of its uses.

Exercise 7.02 Displaying numerical and categorical data

1 For each set of scores on the next page:

- **i** draw a frequency distribution table
- **ii** draw a histogram and frequency polygon
- **iii** find the highest and lowest scores (groups for parts **d** and **e**)
- **iv** find the most frequent score (group for parts **d** and **e**).

a Results of a class quiz:

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

b The number of people ordering pizzas each night:

15, 12, 17, 18, 18 ,15, 16, 13, 15, 17, 18, 12, 17, 14, 16, 15, 17, 18, 19, 15, 15, 12

c The number of people attending a gym:

110, 112, 114, 109, 112, 113, 108, 110, 113, 112, 113, 110, 109, 110, 110, 112, 114, 114, 112, 114, 113

d The results of an assessment task:

45, 79, 65, 48, 69, 50, 62, 74, 38, 69, 88, 96, 90, 58, 52, 68, 63, 61, 79, 74, 50, 65, 77, 91, 56, 77, 63, 81, 90, 59, 67, 50, 61

(Use groups of 30–39, 40–49 and so on.)

e The heights of students (in cm) in a Year 12 class:

159, 173, 182, 166, 172, 179, 181, 163, 178, 169, 183, 158, 162, 167, 174, 175, 180, 174, 176, 159, 161, 171, 174, 179, 180, 159, 157

(Use groups of 155–159, 160–164, 165–169 and so on.)

ISBN 9780170413220

2 For each data set:

i add a cumulative frequency column and class centre where necessary

ii sketch a cumulative frequency histogram and ogive (cumulative frequency polygon).

a Number of cars in a school car park

Number of cars	Frequency
10	4
11	8
12	11
13	9
14	5

b Results of a science experiment

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

c Number of sales made in a shoe shop

Sales	Class centre	Frequency
0–4		6
5–9		2
10–14		3
15–19		5
20–24		8
25–29		9
30–34		5

d Results of an assessment task

Scores	Class centre	Frequency
0–19		3
20–39		2
40–59		7
60–79		6
80–99		1

3 The table shows the number of daily rescues at a beach over a period of time.

a Draw a histogram showing this data.

b How many times were more than 6 rescues made?

c What was the most common number of daily rescues during the survey?

Rescues	Frequency
4	1
5	3
6	6
7	4
8	5
9	0
10	2

4 The volume of traffic on a stretch of highway was measured and the results are shown in the table.

a Draw a histogram to show this data.

b What was the percentage volume of traffic between 21 and 40 minutes?

Volume/min	Frequency
1–10	7
11–20	10
21–30	8
31–40	5
41–50	4

ISBN 9780170413220

5 **a** Draw a two-way table for the following data:

- 27 people play soccer but not tennis
- 35 people play tennis but not soccer
- 28 play neither sport
- 12 play both sports

b What percentage of people play:

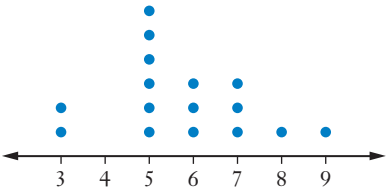
i both sports? **ii** neither sport?

c What percentage of people who play at least one of these sports play only soccer but not tennis?

d What fraction of people who play soccer play tennis as well?

e What percentage of tennis players do not also play soccer?

6 Lauren surveyed her friends and had them rank a film from 1 to 10. The **dot plot** shows the results of her survey.



a How many friends did Lauren survey?

b What was the most common ranking?

c Draw the results in a histogram.

d What percentage of rankings were above 4?

e What fraction of Lauren's friends ranked the film below 4?

7 The stem-and-leaf plot shows the weights (in kg) of a group of people surveyed at a local gym.

Stem	Leaf
5	4 6 8 9
6	1 3 7 8
7	0 3 4 5 5 9
8	1 2 2 4 7 8 9
9	3 5
10	2 6 7

a Arrange these weights in a frequency distribution table using groups of 50–59, 60–69 and so on, and include class centre and cumulative frequency columns.

b Draw an ogive of this data.

c How many people weighed:

i 80 kg or more?

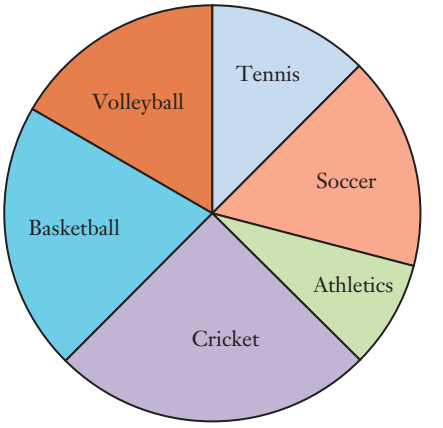
ii less than 60 kg?

d What percentage of people surveyed weighed from 70 kg to 89 kg?

e What fraction of people weighed between 50 kg and 80 kg?

ISBN 9780170413220

8 The pie chart shows the number of students taking different school sports.



a What percentage of students play cricket?

b There are 720 students at the school who play these sports. By measuring the angles in the pie chart, complete a table showing the frequencies for the different sports.

9 The table shows the results of a survey of university students asking what degree they were doing.

Degree	Frequency
Medicine	104
Arts	87
Music	58
Science	93
Economics	79
Law	101

a What percentage of students were studying law?

b What percentage of students were studying medicine or music?

c Draw a pie chart showing this information.

10 The two-way table shows the results of a survey into the protective effect of vaccination on a new virus.

	Vaccinated	Not vaccinated
Infected	11	76
Not infected	159	58

a How many people in the survey were infected with the virus?

b What percentage of people surveyed were vaccinated?

c What percentage of vaccinated people had the virus?

d How many people with the virus were not vaccinated?

11 The two-way table shows the number of people taking part in a trial of a new medication to prevent asthma.

	Taking medication	Control group
Asthmatic	104	105
Not asthmatic	112	109

a How many people took part in the trial?

b What percentage of people were asthmatic?

c What percentage of asthmatic people were in the control group?

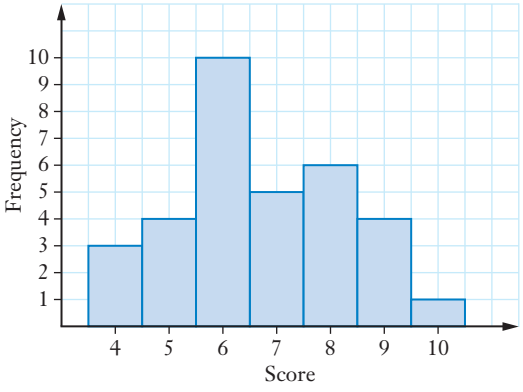
d How many people who were not asthmatic took part in the trial?

e What fraction of the non-asthmatic people took medication?

ISBN 9780170413220

12 The frequency histogram shows the scores on a maths quiz.

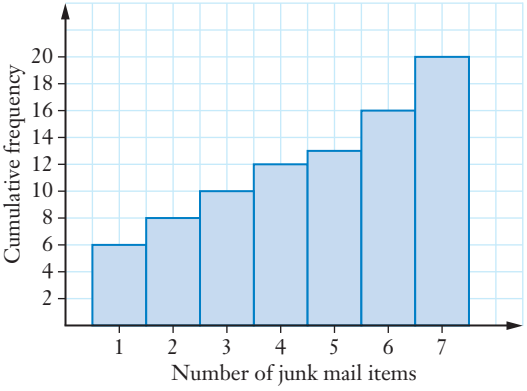
Draw a cumulative frequency histogram and polygon.



13 The cumulative frequency histogram shows data collected in a survey on the number of junk mail items that people received daily in their inbox.

a Draw a frequency distribution table to show the number of junk mail items people receive daily.

b Construct a frequency histogram to show this data.



14 Explain how a stem-and-leaf plot and grouped frequency distribution table can be used for the same data. What are the advantages and disadvantages of each?

15 Draw a Pareto chart for each set of data.

a A survey into why people like a movie:

Reason	Votes
Acting	33
Storyline	29
Music	12
Characters	26

b Customer complaints about an Internet provider:

Complaint	Frequency
Internet speed	34
Cost	61
Data allowance	59
Technical difficulties	46

c Votes for best café in a suburb:

Café	Votes
Coffee Haus	32
Coffee Bean	48
Café Focus	21
Jumping Bean	63
Caffeine Café	36

ISBN 9780170413220

7.03 Measures of central tendency

When we analyse data, we try to find a 'typical', 'normal' or 'average' score. For example we might want to know the average crowd size at football matches through the season. You would usually expect to find this score somewhere in the centre of the data.
There are 3 **measures of central tendency**: the mean, the mode and the median.

The mean

$$\text{Mean} = \frac{\text{Sum of scores}}{\text{Total number of scores}}$$

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

The **mean** has symbol $\bar{x}$, n is the number of scores and Σx is the sum of scores. Note: Σ is the Greek letter 'sigma' and is used in mathematics to stand for a sum.

The symbol $\bar{x}$ usually represents the mean of a **sample**. For the mean of a **population**, the correct symbol is μ, the Greek letter mu.



Shutterstock.com/max blain

WS Measures of central tendency

Statistical measures

Statistics from a frequency table

WS Mean, median and mode

ISBN 9780170413220

EXAMPLE 7

There are 5 children in a family, aged 13, 19, 11, 17 and 10. Find the mean of their ages.

Solution

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

$$= \frac{13+19+11+17+10}{5}$$ ← There are 5 children so $n = 5$.

$$= \frac{70}{5}$$

$$= 14$$

So the mean age of the children is 14.

The mean can also be calculated using a calculator's statistics mode.

	Casio scientific	Sharp scientific
Place your calculator in statistical mode.	MODE STAT 1 VAR	MODE STAT =
Clear the statistical memory.	SHIFT 1 EDIT, DEL -A	2ndF DEL
Enter data.	SHIFT 1 Data to get table. 13 = 19 = etc. to enter in column. AC to leave table.	13 M+ 19 M+ etc.
Calculate mean.	SHIFT 1 VAR $\bar{x}$ =	RCL $\bar{x}$
Check the number of scores.	SHIFT 1 VAR n =	RCL n
Change back to normal mode.	MODE COMP	MODE 0

For the mean of larger data sets, it is easier to sort the data into a frequency distribution table to add up the scores.

The mean of data in a frequency table

$$\bar{x} = \frac{\Sigma fx}{\Sigma f}$$

where Σfx is the sum of each score $\times$ its frequency and Σf is the sum of frequencies

ISBN 9780170413220

EXAMPLE 8

Find the mean number of hours that members of a class practise piano each week. The number of hours for each student is:

1, 4, 6, 1, 3, 6, 2, 1, 1, 3, 2, 5, 6, 6, 1, 2, 6, 2, 5, 6, 6, 2, 3, 6, 2

Solution

First draw up a frequency distribution table for the hours of practice. Include an fx column for multiplying each score by its frequency.

The table gives us a quick way of finding the sum of scores. For example, we know from the table that there are 6 lots of 2, so we can use $2 \times 6 = 12$. The sum of the fx column gives us the sum of all scores.

Hours (x)	Frequency (f)	Score × frequency (fx)
1	5	5
2	6	12
3	3	9
4	1	4
5	2	10
6	8	48
	$\Sigma f = 25$	$\Sigma fx = 88$

$$\bar{x} = \frac{\Sigma fx}{\Sigma f}$$
$$= \frac{88}{25}$$
$$= 3.52$$

So the mean time students practise each week is 3.52 hours.

On a calculator:

Operation	Casio scientific	Sharp scientific
Clear the statistical memory first (see previous example, previous page).		
Enter data.	SHIFT MODE scroll down to STAT, Frequency? ON SHIFT 1 Data to get table. 1 = 2 = etc. to enter in x column. 5 = 6 = etc. to enter in FREQ column. AC to leave table.	1 2ndF STO 5 M+ 2 2ndF STO 6 M+ etc.
Calculate mean.	SHIFT 1 VAR $\bar{x}$ =	RCL $\bar{x}$

ISBN 9780170413220

Although grouped data is not completely accurate because we don't know exactly what scores are in each group, we can still calculate an estimate of the mean.

EXAMPLE 9

From the table, find the mean commuting time that a sample of people take to travel to work.

Minutes	Frequency
0–8	3
9–17	5
18–26	7
27–35	8
36–44	2

Solution

Add **class centre** and fx columns to the table.

Use the class centres as the scores when calculating fx.

Minutes	Class centre (x)	Frequency (f)	Score × frequency (fx)
0–8	4	3	12
9–17	13	5	65
18–26	22	7	154
27–35	31	8	248
36–44	40	2	80
		$\Sigma f = 25$	$\Sigma fx = 559$

$$\begin{aligned}\bar{x} &= \frac{\Sigma fx}{\Sigma f} \\ &= \frac{559}{25} \\ &\approx 22.36\end{aligned}$$

The mean time taken to travel to work is 22.36 minutes.

The mode

The **mode** is the most frequent score.

ISBN 9780170413220

There is no mode if all the scores are different, or there could be several scores with the same frequency.

EXAMPLE 10

a Find the mode of these scores:

7, 4, 3, 5, 7, 1, 2

b Find the mode for these shoes sold at a shoe store.

Shoe size	5	$5\frac{1}{2}$	6	$6\frac{1}{2}$	7	$7\frac{1}{2}$	8	$8\frac{1}{2}$	9	$9\frac{1}{2}$
Frequency	8	9	15	28	53	61	58	29	12	10

Solution

a There are two 7s and only one of the other scores, so the mode is 7.

b The shoe size with the highest frequency is $7\frac{1}{2}$ (there were 61 of them).

So the mode is $7\frac{1}{2}$.

With grouped data, instead of finding the mode we find the **modal class**.

EXAMPLE 11

Find the modal class in this data set showing the ages of people at a caravan park.

Scores	Frequency
10–19	2
20–29	0
30–39	1
40–49	5
50–59	7
60–69	3

Solution

The group or class with the highest frequency is 50–59.

While we do not know the individual score with the highest frequency, we say the modal class is 50–59.

The mode is useful when looking at trends such as the most popular types of clothing. It can also be used for categorical data.

The median

The **median** is the middle score when all scores are in order.

If there are 2 middle scores, the median is the average of those scores.

ISBN 9780170413220

EXAMPLE 12

Find the median age of a group of people in a band:

18, 15, 20, 18, 17, 16, 11, 13

Solution

Put the ages in order.

11, 13, 15, 16, 17, 18, 18, 20

There are 2 middle ages, 16 and 17, so we find their average.

$$\text{Median} = \frac{16+17}{2} = 16.5$$

So the median age of the band members is 16.5.

The median can also be calculated using a calculator's statistics mode.

Operation	Casio scientific
Clear the statistical memory.	
Enter data.	SHIFT 1 Data to get table. 18 = 15 = etc. to enter in column. AC to leave table.
Calculate the median.	SHIFT 1 MinMax med =
Change back to normal mode.	MODE COMP

You can find the median of data in a frequency table. If there is a large number of scores, you can find the position of the middle score using a cumulative frequency column.

EXAMPLE 13

Find the median of this data set.

Score	Frequency
5	3
6	2
7	4
8	7
9	6
10	3

ISBN 9780170413220

Solution

Add a column for cumulative frequencies.

Score	Frequency	Cumulative frequency
5	3	3
6	2	5
7	4	9
8	7	16
9	6	22
10	3	25

There are 25 scores so the position of the middle score is $\frac{25+1}{2} = 13$th.

The 10th to 16th scores are 8, so the 13th score is 8.

You can check this by writing the scores in order.

The median is 8.

The position of the median

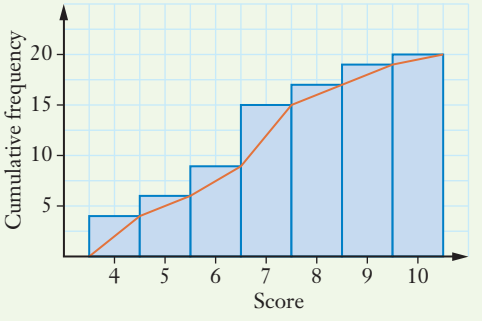
The median of n scores is the $\frac{n+1}{2}$ th score.

If n is even, then the median is the average of the 2 middle scores on both sides of the $\frac{n+1}{2}$ th position.

Another way to find the median is from an ogive (cumulative frequency polygon). We simply use the halfway point on the cumulative frequency axis of the graph.

EXAMPLE 14

Find the median from the cumulative frequency polygon below.

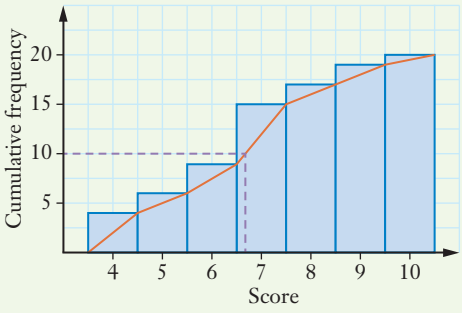


ISBN 9780170413220

Solution

There are 20 scores in the data set, so the halfway point is at the 10th score as shown on the cumulative frequency axis. The dotted line meets the ogive inside the 7 column.

The median is 7.



Outliers

Sometimes a set of data contains a score that is unusual, compared with the other scores. This unusual or extreme value is called an **outlier**.

EXAMPLE 15

The prices of houses sold in the town of Greenfield in a particular week are:

$355 000, $420 000, $320 000, $285 000, $390 000, $1 200 000, $415 000, $320 000, $435 000, $380 000

a Is there an outlier? Why do you think an outlier may be in this data?

b Find the mean house price with and without the outlier.

c Find the median with and without the outlier.

d Find the mode with and without the outlier.

Solution

a The outlier is $1 200 000 as this is much higher than the other prices. It may be that there is one special house in the area that is much larger than the others, or a certain street with huge houses in it that is unusual for the area.

b Using a calculator:

With the outlier, the mean house price is $452 000. ← Notice the big difference.

Without the outlier, the mean house price is $368 888.89.

c With the outlier, the median house price is $385 000. ← Notice the small difference.

Without the outlier, the median house price is $380 000.

d The mode is $320 000 in both cases since this is the most frequent price with or without the outlier.

ISBN 9780170413220

Note: When real estate agents talk about house prices, they usually use the median price since this is not as affected by outliers as the mean.



Photo courtesy Margaret Grove

INVESTIGATION

OUTLIERS

Which measures of central tendency do outliers tend to affect most?

Find other examples of data that contain outliers and find the mean, mode and median. How do they change if the outlier is removed? Should outliers be looked at closely and discarded or is there a place for them?

Exercise 7.03 Measures of central tendency

1 For each data set find:

i the mean **ii** the mode **iii** the median

a Number of people auditioning for parts in a play: 5, 5, 7, 6, 5, 6, 1

b Number of minutes for an ambulance to respond to a call: 1, 4, 6, 8, 7, 4, 6, 4, 5

c Ages of students on a basketball team: 15, 18, 14, 19, 18, 17, 11

d Scores on a class quiz: 4, 6, 5, 4, 7, 8

e Prices of petrol (in dollars): 1.43, 1.66, 1.55, 1.49, 1.27, 1.81, 1.49, 1.38

2 Find the mode of each data set:

a

Hair colour	Frequency
Brown	28
Blond	21
Red	8
Black	12
Grey	17

b

Type of cat	Frequency
Siamese	18
Burmese	12
Russian blue	9
Tabby	32
Ginger	26
Persian	19

ISBN 9780170413220

3 For each data set find:

i the mean **ii** the median **iii** the mode

a Judge's scores on a dance contest

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

b Number of matches in each match box surveyed

Score	Frequency
50	1
51	6
52	5
53	3
54	4
55	2

c Results in a History assignment

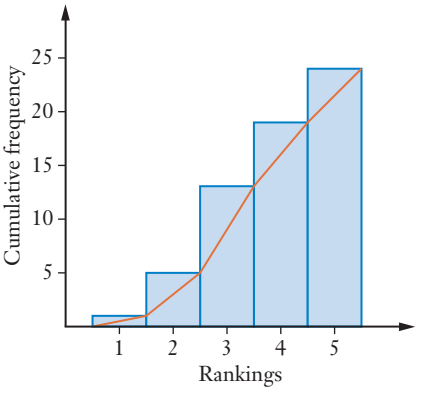
Score	Frequency
14	4
15	2
16	1
17	4
18	3
19	5
20	6

d Attendances at hockey matches

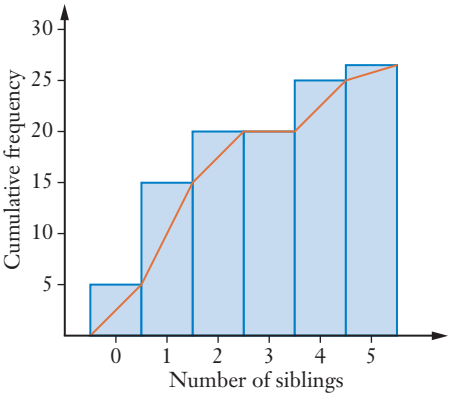
Attendance	Frequency
100	3
101	0
102	2
103	1
104	6
105	5

4 Find the median from each ogive:

a

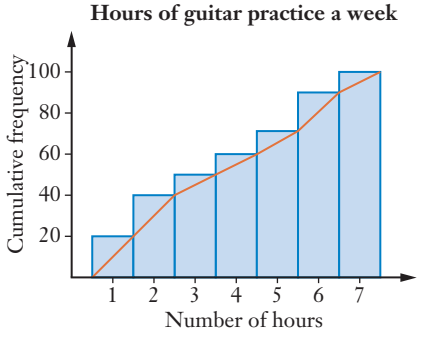


b

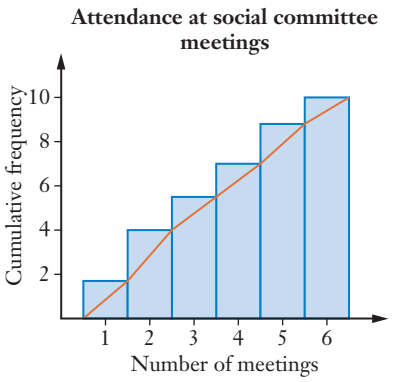


ISBN 9780170413220

c



d



5 For each data set find:

i the mean **ii** the modal class

a Games of chess played each week by members of a chess club:

Score	Frequency
2–4	5
5–7	4
8–10	7
11–13	4
14–16	3
17–19	2

b Hours per week that gymnasts train:

Score	Frequency
0–4	3
5–9	2
10–14	6
15–19	8
20–24	9
25–29	5

c Results in a Legal Studies exam:

Score	Frequency
10–24	4
25–39	0
40–54	1
55–69	5
70–84	9
85–99	8

d Time it takes for computers to boot up:

Time (s)	Frequency
20–24	12
25–29	8
30–34	9
35–39	7
40–44	8
45–49	11
50–54	12
55–59	6

ISBN 9780170413220

6 For each data set:

- **i** add a cumulative frequency column
- **ii** draw a cumulative frequency polygon
- **iii** find the median from the graph (estimate for parts **c** and **d**)

a Number of athletes representing their school over a 30-year period:

Athletes	Frequency
1	5
2	6
3	4
4	8
5	5
6	2

b Number of lollies in a bag:

Number	Frequency
45	3
46	5
47	1
48	7
49	3
50	1

c Hours a week worked by employees in a cafe:

Hours	Frequency
1–5	7
6–10	5
11–15	3
16–20	6
21–25	7
26–30	2

d Time to complete a race:

Time (min)	Frequency
2.5–2.8	3
2.9–3.2	2
3.3–3.6	0
3.7–4.0	6
4.1–4.4	1
4.5–4.8	4
4.9–5.2	4

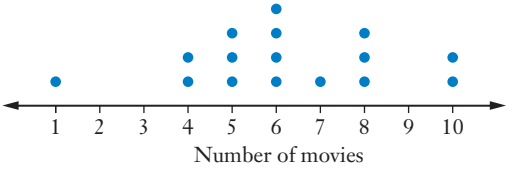
7 For each data set:

- **i** draw a frequency distribution table including cumulative frequency
- **ii** find the mean
- **iii** find the mode or modal class
- **iv** draw an ogive
- **v** find the median from the ogive

a Home runs scored over a baseball season

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

b Number of movies seen in a year:



ISBN 9780170413220

c Ages of people living in a block of units (use classes of 20–29, 30–39 and so on):

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

8 For each set of data, find:

- **i** the outlier
- **ii** the mean, mode and median
- **iii** the mean, mode and median without the outlier

a Weights (in kg) of people in a lift:

69, 75, 58, 77, 32, 68, 60, 64, 59

b Number of questions attempted in an exam:

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

c Ages of people at a family party:

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

d Rating of a venue for a dance party:



9 For the set of times (in minutes) students are recorded as late for school shown below, find the outlier. Which measures of central tendency (mean, median, mode) does it change?

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

ISBN 9780170413220

10 The stem-and-leaf plot below shows the results of a class test:

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

a State which score is an outlier and what this means.

b Draw a frequency table including a cumulative frequency column, using groups 10–19, 20–29, 30–39 and so on.

c Use the table to estimate the mean and find the modal class:

i with the outlier included **ii** without the outlier included.

d Draw an ogive excluding the outlier and find the median.

INVESTIGATION

LIMITATIONS OF CENTRAL TENDENCY

Find the mean, mode and median of each set of data. What do you notice?

Set 1: 5, 6, 7, 7, 8, 9

Set 2: 1, 2, 7, 7, 12, 13

How do the 2 sets of data differ? Can we find out by using the measures of central tendency? How else could we describe how they are different from each other?

WS Quartiles, deciles and percentiles

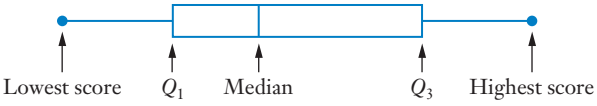
Box-and-whisker plots

7.04 Quartiles, deciles and percentiles

The measures of central tendency give us good information about data sets, but they don't describe the spread of data. As we have seen, the median divides data sets so that half the values lie below the median and half lie above it. A measure that divides a data set into parts of equal size is called a **quantile**. The median gives only a very rough description of the data set, but with more divisions we can describe the data's spread in more detail.

Quartiles

A **quartile** divides a data set into quarters.



ISBN 9780170413220

The 1st quartile (Q_1) is called the **lower quartile** and the 3rd quartile (Q_3) is called the **upper quartile**. The 2nd quartile (Q_2) is the median.

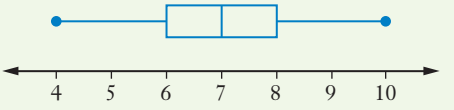
A **box plot** (also called a **box-and-whisker plot**) gives a way of showing a five-number summary: the quartiles and highest and lowest scores.

EXAMPLE 16

a Find Q_1, Q_2 and Q_3 for the number of runs in a softball game:

4, 3, 7, 8, 7, 9, 5, 6, 8, 3, 9

b Find Q_1, Q_3, the median, the highest and lowest score for this data.



c Find the quartiles of the data in this frequency table.

Score	Frequency
1	1
5	3
6	5
7	8
8	11
9	13
10	9

Solution

a Put the 11 scores in order.

3, 3, 4, 5, 6, 7, 7, 8, 8, 9, 9

Find Q_2 (the median) in the usual way: the 6th score is 7.

3, 3, 4, 5, 6, (7), 7, 8, 8, 9, 9

Q_1 is the middle of the scores below the median: the 3rd score, 4.

If a quartile falls between 2 scores, we take their average.

3, 3, (4), 5, 6, (7), 7, 8, (8), 9, 9

Q_3 is the middle of the scores above the median: the 9th score, 8.

So $Q_1 = 4$, $Q_2 = 7$, $Q_3 = 8$

ISBN 9780170413220

On a calculator:

Operation	Casio Scientific
Clear the statistical memory.	
Enter data.	SHIFT 1 Data to get table. 4 = 3 = etc. to enter in column. AC to leave table.
Calculate Q_1.	SHIFT 1 MinMax Q_1 =
Calculate Q_2 (median).	SHIFT 1 MinMax med =
Calculate Q_3.	SHIFT 1 MinMax Q_3 =

b From the box plot:

Lowest score = 4, $Q_1 = 6$, Median = 7, $Q_3 = 8$, Highest score = 10.

c Add a cumulative frequency column to the table.

Score	Frequency	Cumulative frequency
1	1	1
5	3	4
6	5	9
7	8	17
8	11	28
9	13	41
10	9	50

There are 50 scores so the median is the $\frac{50+1}{2}$ or 25.5th score (average of 25th and 26th scores).

So $Q_2 = 8$, reading from the cumulative frequency column.

Q_1 is the $\frac{25+1}{2}$ or 13th score (middle of the 1st 25 scores).

So $Q_1 = 7$, reading from the cumulative frequency column.

Q_3 is the $\frac{50+25+1}{2}$ or 38th score (middle of the last 25 scores, halfway between the 26th to 50th scores).

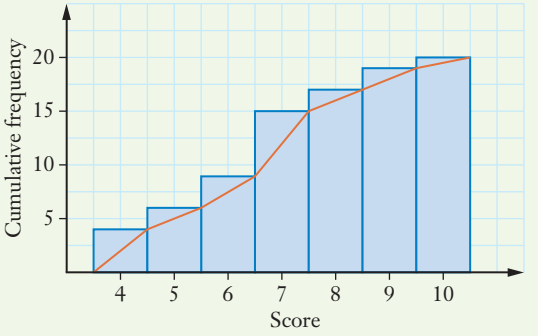
So $Q_3 = 9$, reading from the cumulative frequency column.

So $Q_1 = 7$, $Q_2 = 8$, $Q_3 = 9$.

ISBN 9780170413220

EXAMPLE 17

Find the median and upper and lower quartiles from the ogive.



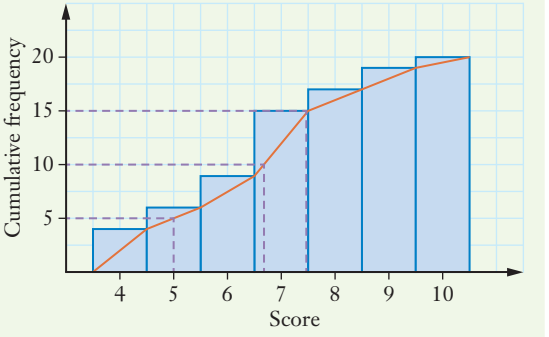
Solution

There are 20 scores.

The median is halfway:

$$\frac{1}{2} \times 20 = 10$$

So the median is 7, reading across from 10 on the cumulative frequency axis.



To find the 1st (lower) quartile:

$$\frac{1}{4} \times 20 = 5$$

So reading across from 5, $Q_1 = 5$.

To find the 3rd (upper) quartile:

$$\frac{3}{4} \times 20 = 15$$

So reading across from 15, $Q_3 = 7.5$.

ISBN 9780170413220

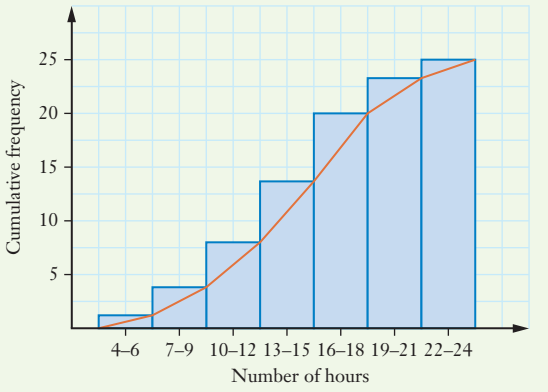
Deciles and percentiles

For a more detailed description of the spread we can divide the data set into smaller parts. **Deciles** divide the data set into 10 parts and **percentiles** divide data sets into 100 parts.


Deciles and percentiles

EXAMPLE 18

The ogive shows the number of hours that a rock group rehearses each week over 25 weeks. The scores have been sorted into groups.



Use the ogive to estimate:

a the 35th percentile

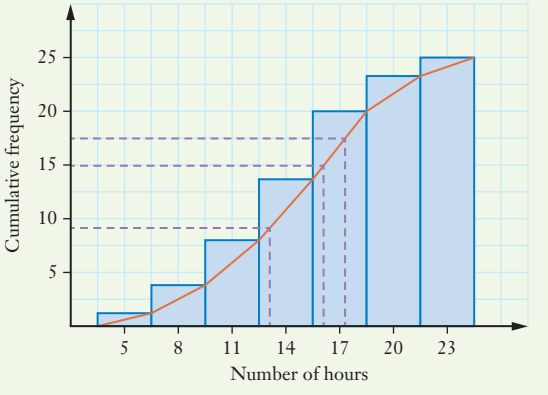
b the 60th percentile

c the 7th decile

We can only **estimate** because the scores have been grouped into classes.

Solution

Redraw the ogive using **class centres** for number of hours, and use the cumulative frequency axis to find answers.



a The 35th percentile is the score that separates the lower 35% of the scores.

35% of 25 scores = 8.75th score.

The 35th percentile ≈ 13.

b The 60th percentile separates the lower 60% of the scores.

60% of 25 scores = 15th score.

The 60th percentile ≈ 16.

c The 7th decile separates the lower $\frac{7}{10}$ of the scores.

$\frac{7}{10} \times 25$ scores = 17.5th score.

The 7th decile ≈ 17.

You can use a graphics calculator or software to draw graphs and find quartiles, deciles and percentiles more accurately.

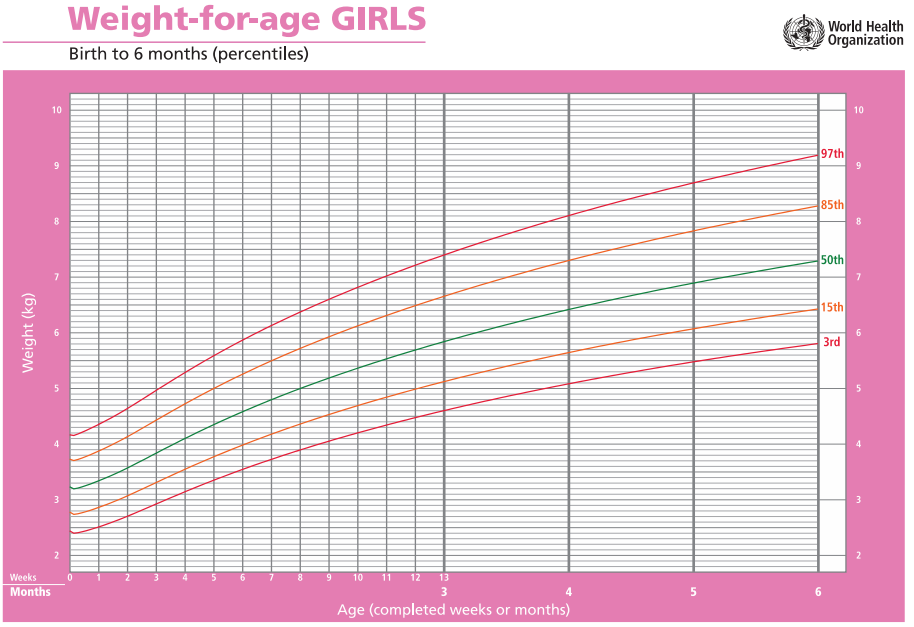
ISBN 9780170413220

INVESTIGATION

RESEARCHING QUANTILES

Research the words *quartile*, *decile* and *percentile*. When were they first used? Where are they used now? There are other measures such as tercile and quintile. What are they?

Percentiles are used in many applications including graphs of infants' and children's growth rates.



http://www.who.int/childgrowth/standards/cht_wfa_girls_p_0_6.pdf?ua=1

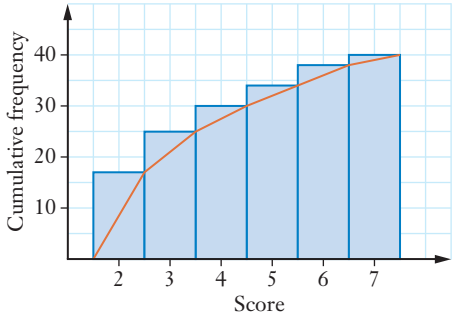
Research and compare different types of growth charts used in Australia and those of the World Health Organization (WHO).

Exercise 7.04 Quartiles, deciles and percentiles

1 For each set of data, find:

i the 1st quartile **ii** the 2nd quartile **iii** the 3rd quartile

a

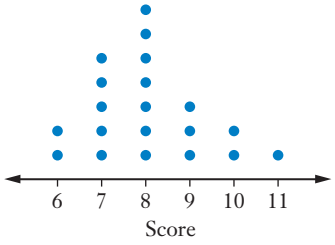


b

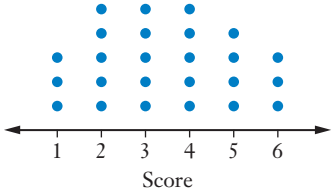
Score	Frequency
10	3
11	5
12	4
13	6
14	0
15	2

ISBN 9780170413220

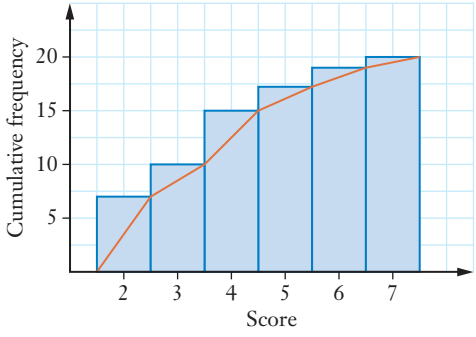
c



d



2 Find the 1st and 3rd quartiles for the following data.



3 For the following data find:

a the 23rd, 55th and 91st percentiles

b the 2nd and 8th decile.

Score	Frequency
23	13
24	19
25	23
26	21
27	9
28	15

4 For the dot plot:

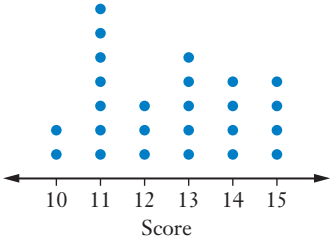
a draw a cumulative frequency polygon

b find:

i the 1st quartile **ii** the 3rd quartile

iii the 35th percentile **iv** the 7th decile

v the 1st decile



5 John measured the weights of children in a particular year at school and organised his findings in a table. Estimate the weight that is:

a the median

b the 1st quartile

c the 3rd quartile

d the 60th percentile

Weight (kg)	Frequency
30–34	1
35–39	9
40–44	8
45–49	5
50–54	2

ISBN 9780170413220

6 The number of different dress sizes in Huang's Sportwear shop was counted in a stocktake, and the results set out in a table.

Size	Frequency
8	12
10	23
12	20
14	21
16	13
18	11

a How many dresses were counted?

b What percentage of dresses in the store were size 12?

c Find the median dress size.

d Find the 3rd quartile.

e What percentile is size 14?

7 Antonietta surveyed a number of people to find out how many pets they have. Her results are shown.

Number of pets	Frequency
0	7
1	11
2	3
3	2
4	1

a What percentage of the people surveyed had 2 pets?

b Draw a cumulative frequency polygon for this data.

c From the graph, find:

i the median

ii the 1st quartile

iii the 3rd quartile

8 Abdul measured the reaction times of a group of drivers and placed his results in a table.

Time (s)	Frequency
0.65–0.69	2
0.70–0.74	14
0.75–0.79	19
0.80–0.84	8
0.85–0.89	7

a What was the mean reaction time?

b What percentage of people surveyed reacted within 0.75 and 0.79 seconds?

c Draw an ogive to show this data.

d Use the graph to estimate:

i the 30th percentile

ii the median reaction time

iii reaction times between the 1st and 3rd quartiles

9 For each data set, find:

i the median **ii** the lower quartile **iii** the upper quartile

a The number of dogs at a pound over several days:
36, 79, 38, 29, 45, 83, 85, 47, 51, 72, 64

b The number of flying hours that Alexis had during a helicopter flying course:
3, 4, 9, 8, 14, 17, 15, 11, 12

c The distance (in km) travelled by a taxi during several shifts:
128.3, 143.2, 103.7, 99.5, 137.5, 203.4, 154.6, 115.3, 192.3, 125.4

d The number of people attending a choir rehearsal over several weeks:
15, 14, 12, 16, 15, 19, 17, 18

ISBN 9780170413220

Range and interquartile range

Interquartile range

Interquartile range

Statistical match-up

Cumulative frequency graphs

7.05 Range and interquartile range

The **range** and **interquartile range** measure the spread of data.

Range and interquartile range

Range = highest score – lowest score

Interquartile range = $Q_3 - Q_1$

EXAMPLE 19

Find the range and interquartile range of these scores: 8, 13, 5, 15, 20, 21, 17, 16, 9

Solution

Range = highest score – lowest score

$= 21 - 5$

$= 16$

Put the 9 scores in order to find the quartiles.

5, 8, 9, 13, 15, 16, 17, 20, 21

$Q_2 = 15$

5, 8, | 9, 13, (15), 16, 17, | 20, 21

$Q_1 = \frac{8+9}{2} = 8.5$ $\quad$ $Q_3 = \frac{17+20}{2} = 18.5$

Interquartile range = $Q_3 - Q_1$

$= 18.5 - 8.5$

$= 10$



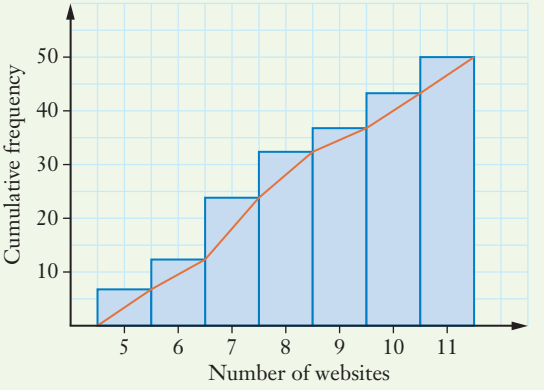
age-fotostock/Martin Barraud

ISBN 9780170413220

EXAMPLE 20

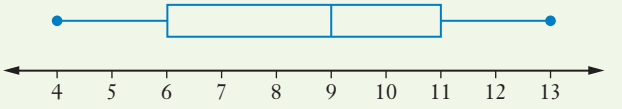
a This set of data shows the results of a survey into the number of travel websites people visit regularly. Find:

i the range **ii** the interquartile range



b From this box plot, find:

i the median **ii** the range **iii** the interquartile range



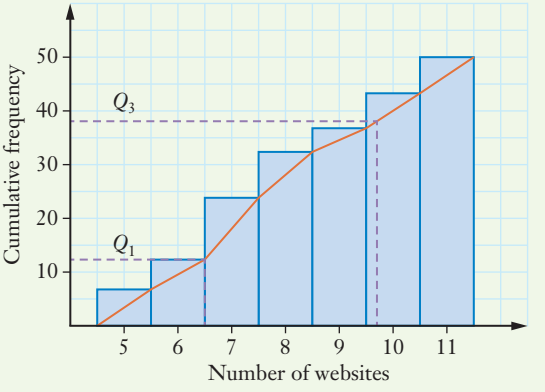
Solution

a **i** Range = 11 – 5 = 6 highest – lowest

ii There are 50 scores.

$\frac{1}{4} \times 50 = 12.5$ so Q_1 will be the 12.5th score.

$\frac{3}{4} \times 50 = 37.5$ so Q_3 will be the 37.5th score.



$Q_1 = 6.5$ and $Q_3 = 10$

Interquartile range $= Q_3 - Q_1$

$= 10 - 6.5$

$= 3.5$

b **i** Median = 9 **ii** Range = 13 – 4 = 9 **iii** Interquartile range = 11 – 6 = 5

ISBN 9780170413220


Outliers

Outliers

Outliers are extreme scores. They affect the range because an outlier will be the highest or lowest score. However, outliers do not affect the interquartile range because the interquartile range does not depend on the highest or lowest scores.

Some outliers are more obvious than others. There is a formal definition of outlier that allows us to test if it's an outlier rather than just deciding by inspection.

Outlier

A score is an outlier if it is more than 1.5 times the interquartile range (IQR) below Q_1 or above Q_3.

An outlier is below $Q_1 - 1.5 \times \text{IQR}$ or above $Q_3 + 1.5 \times \text{IQR}$.

EXAMPLE 21

a For the scores 5, 2, 9, 10, 6, 7, 6, 5, 10, 9, 5, 7, 8, 7, 6, determine if 2 is an outlier.

b For this table of data, find a score that looks like an outlier and use the definition to determine if it is an outlier.

Score	Frequency
1	1
2	0
3	0
4	0
5	3
6	5
7	8
8	11
9	13
10	9

Solution

a $Q_1 = 5$ and $Q_3 = 9$

$\text{IQR} = Q_3 - Q_1$
$= 9 - 5$
$= 4$

$1.5 \times \text{IQR} = 1.5 \times 4$
$= 6$

$Q_1 - 1.5 \times \text{IQR} = 5 - 6$
$= -1$

$Q_1 + 1.5 \times \text{IQR} = 9 + 6$
$= 15$

Any outlier would have to be less than –1 or greater than 15. So 2 is not an outlier.

ISBN 9780170413220

b A score of 1 looks like an outlier.

From Example 16c,

$Q_1 = 7$ and $Q_3 = 9$

$\text{IQR} = Q_3 - Q_1$

$= 9 - 7$

$= 2$

$1.5 \times \text{IQR} = 1.5 \times 2$

$= 3$

$Q_1 - 1.5 \times \text{IQR} = 7 - 3$

$= 4$

$Q_3 + 1.5 \times \text{IQR} = 9 + 3$

$= 12$

Any outlier would have to be less than 4 or greater than 12. So 1 is an outlier.

Exercise 7.05 Range and interquartile range

1 Find the range of each data set.

a 7, 4, 9, 8, 11, 4, 3, 19, 7, 16

b 56, 89, 43, 99, 45, 28, 37, 78

c 103, 108, 99, 112, 126, 87, 101, 123

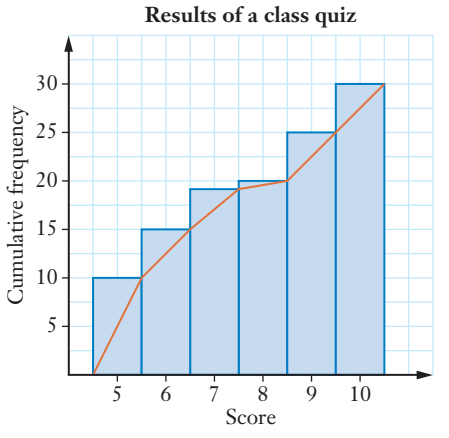
d

Score	Frequency
8	5
9	3
10	7
11	0
12	8
13	7

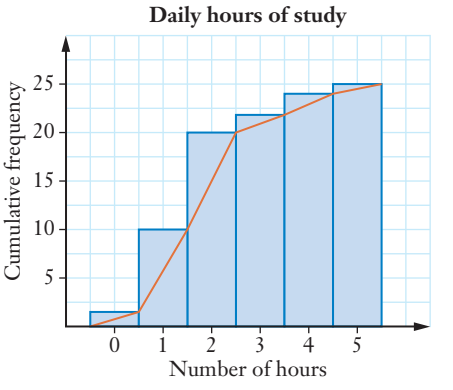
2 For each set of data, find:

i the median **ii** the range **iii** the interquartile range

a

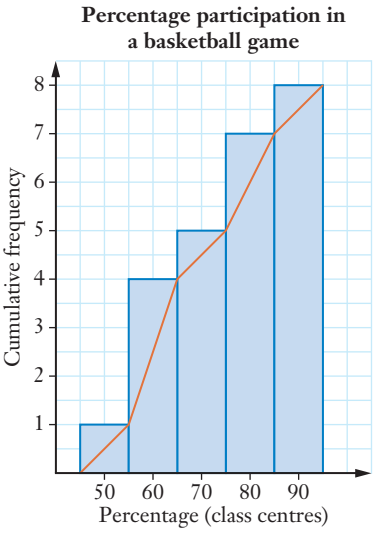


b

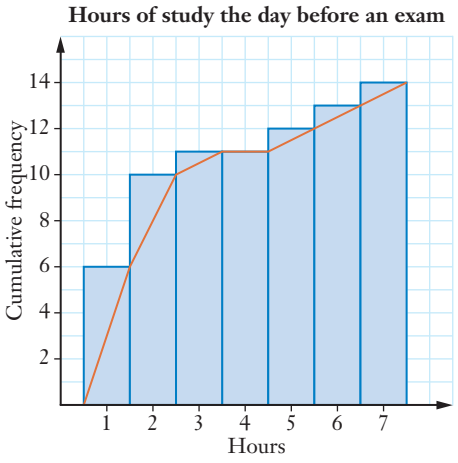


ISBN 9780170413220

c



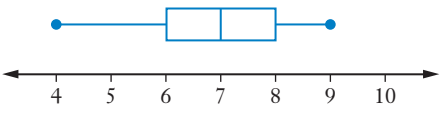
d



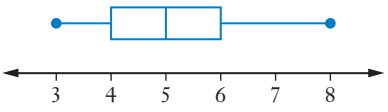
3 For each set of data, find:

i the median **ii** the range **iii** the interquartile range

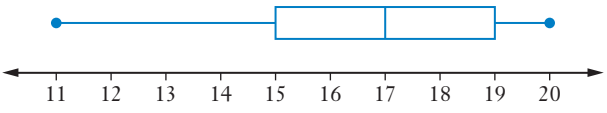
a



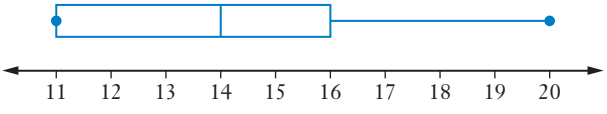
b



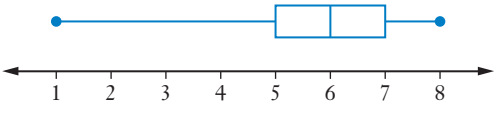
c



d



e



4 For this set of data, find:

a the mean

b the mode

c the median

d the range

e the interquartile range

Rain (mm)	Frequency
5	4
6	7
7	8
8	3
9	3

ISBN 9780170413220

5 Find a potential outlier in each set of data and use the definition to see if it really is an outlier.

a 100, 76. 93, 54, 32, 66, 53, 97, 51, 80

b 11, 5, 7, 19, 5, 3, 7, 5, 6, 10, 11. 2, 5, 7, 4, 6, 1

c

Score	Frequency
1	1
2	0
3	0
4	0
5	5
6	4

7.06 Variance and standard deviation



Variance and standard deviation



Statistical calculations


Statistical measures puzzle

Variance is another measure of spread. It measures how far the scores in a data set are from the mean of the data. You studied variance and standard deviation when studying discrete probability distributions in Year 11, Chapter 10, *Discrete probability distributions*.

The formula for variance, σ^2, is:

$$\sigma^2 = \frac{\Sigma(x-\bar{x})^2}{n}$$

However, you do not have to use it as the calculator's statistical mode can calculate it more easily. The following example will show you what the above formula means, but you don't have to learn it.

EXAMPLE 22

The data below shows the times (in minutes) taken for a fire engine to reach the site of a fire. Find the variance for this data.

4, 7, 3, 9, 4, 5, 1, 3, 5, 9

Solution

First we need to find the mean.

$$\bar{x} = \frac{\Sigma x}{n} = \frac{50}{10} = 5$$

Now we find the difference between each score and the mean. Then we square each difference because we only want positive values. This is shown in the table next page.

ISBN 9780170413220

x	$x - \bar{x}$	$(x - \bar{x})^2$
1	$1 - 5 = -4$	16
3	$3 - 5 = -2$	4
3	$3 - 5 = -2$	4
4	$4 - 5 = -1$	1
4	$4 - 5 = -1$	1
5	$5 - 5 = 0$	0
5	$5 - 5 = 0$	0
7	$7 - 5 = 2$	4
9	$9 - 5 = 4$	16
9	$9 - 5 = 4$	16
		$\Sigma(x - \bar{x})^2 = 62$

Variance is the mean of these squared differences.

$$\sigma^2 = \frac{\Sigma(x - \bar{x})^2}{n}$$
$$= \frac{62}{10}$$
$$= 6.2$$

Standard deviation is another measure of spread, and it is simply the square root of variance. For the data in the previous example, the standard deviation is $\sqrt{6.2} \approx 2.49$.

We use s for standard deviation of a sample and σ (the lowercase Greek sigma) for the standard deviation of a **population**. In this course we will use s most of the time.

The formula for standard deviation, σ, is:

$$\sigma = \sqrt{\frac{\Sigma(x - \bar{x})^2}{n}}$$

This example shows how to calculate standard deviation and variance using the calculator's statistical mode.

EXAMPLE 23

The table shows the number of hours of karate practice that students at a karate club do each week.
Find correct to one decimal place:

a the standard deviation

b the variance

Practice times (h)	Frequency
2	2
3	4
4	3
5	4
6	1
7	2
8	4

ISBN 9780170413220

Solution

a

Operation	Casio scientific	Sharp scientific
Clear the statistical memory.	SHIFT 1 Edit, Del-A	2ndF DEL
Enter data.	SHIFT 1 Data to get table. 2 = 3 = etc. to enter in x column. 2 = 4 = etc. to enter in FREQ column. AC to leave table.	2 2ndF STO 2 M+ 3 2ndF STO 4 M+ etc.
Calculate standard deviation.	SHIFT 1 VAR sx =	RCL sx

$s = 2.0774 \ldots \approx 2.1$

b Variance $= s^2$

$= 2.0774 \ldots^2$

$= 4.3157 \ldots$

≈ 4.3

Exercise 7.06 Variance and standard deviation

1 For each set of data find:

i the mean **ii** the standard deviation

a Number of minutes kept on hold on the telephone: 7, 4, 9, 8, 4, 6, 2, 4, 5

b Travel time (in minutes) to get into the city: 23, 45, 67, 54, 69, 38, 59, 70, 59, 41

c Height of children (in cm): 101, 112, 131, 122, 130, 143, 152, 107, 112

d Number of repetitions on gym equipment: 8, 6, 9, 5, 5, 7, 6, 4, 8, 9, 6, 3, 6

e Age of performers in a play: 18, 19, 17, 16, 20, 18, 15, 19, 14, 20

2 For each data set, find:

i the standard deviation **ii** the variance

a Weights (in kg): 51, 67, 64, 53, 60, 48, 58, 49, 61, 71, 67, 58

b Class quiz results: 4, 6, 5, 3, 7, 9, 8, 10, 4, 6, 7, 6, 5, 8, 6, 7, 9, 10, 5, 4, 8

c Time spent waiting in a queue (in mins): 11, 14, 15, 25, 31, 54, 36, 39, 31, 41, 44, 50

d Weight of crates (in kg): 87, 88, 56, 91, 68, 73, 55

e Response time (in mins) for helicopter rescue: 1, 5, 7, 3, 8, 6, 5, 5, 4, 8, 9, 3

ISBN 9780170413220

3 For each data set find:

i the mean **ii** the standard deviation **iii** the variance

a Number of books read:

Books	Frequency
5	3
6	5
7	6
8	2
9	1
10	3
11	5
12	4

b Piano practice time per week:

Practice (h)	Frequency
1	3
2	0
3	2
4	5
5	7
6	3
7	2

c Weight of luggage:

Weight (kg)	Frequency
31	3
32	0
33	2
34	5
35	7
36	3

d Results of a half-yearly exam:

Score	Frequency
10–19	1
20–29	4
30–39	8
40–49	12
50–59	15
60–69	11
70–79	7

4 In a taste test, the people surveyed had to rank a new biscuit on a scale from 1 to 5. The table shows the results of the survey.

Rank	Frequency
1	5
2	11
3	18
4	21
5	9

a What is the range?

b Find the interquartile range.

c Find the standard deviation.

5 A Year 12 Art class received these results for their major work.

Class	Frequency
20–29	1
30–39	0
40–49	2
50–59	4
60–69	7
70–79	9
80–89	8
90–99	7

a Calculate:

i the mean **ii** the standard deviation

b Remove the outlier and calculate:

i the mean **ii** the standard deviation

6 The data shows the ages of students in an MBA course:

20, 25, 31, 34, 17, 27, 29, 53, 20, 31, 19, 23, 30, 29, 18, 25

a Show that there is an outlier.

b Calculate the standard deviation:

i with the outlier **ii** without the outlier

ISBN 9780170413220

7 Jane recorded the number of crocodile sightings each week in a region of the Northern Territory over several weeks.

5, 8, 9, 4, 11, 7, 9, 15, 17, 10, 8, 5, 9, 12

a What was the mean number of crocodiles sighted per week?

b Find the standard deviation.

c Calculate the variance.

7.07 Shape and modality of data sets



Shape of a distribution


The shape of a frequency distribution

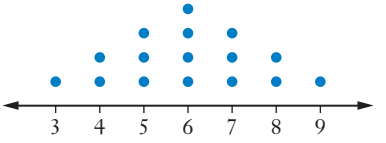


Shapes of distributions

The measures of central tendency and spread are called **summary statistics**, and they help us make decisions about data. Other features of data can help with these decisions, and one of these features is the **shape** of the data. The shape often gives us an idea of the centre and spread even before we measure them, while the **modality** describes the number of peaks in the distribution of data.

A data set where the mean, mode and median are equal has a **symmetrical distribution**.

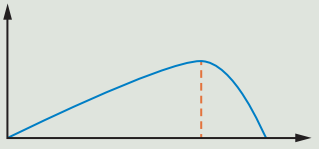
Notice how symmetrical this dot plot is.



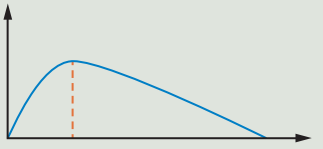
Other graphs that are not so symmetrical can be described by their **skewness**.

The shape of a statistical distribution

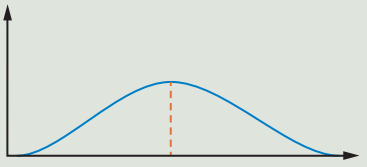
This distribution is **negatively skewed** as most of the area is to the left (or negative direction) of the centre. We can say that the 'tail' points to the low scores in the negative direction.



This distribution is **positively skewed** as most of the area is to the right (or positive direction) of the centre. We can say that the 'tail' points to the high scores in the positive direction.



This distribution is **symmetrical**.

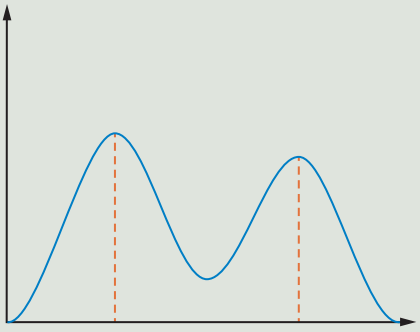


All these graphs are also called **unimodal** since they only have one peak.

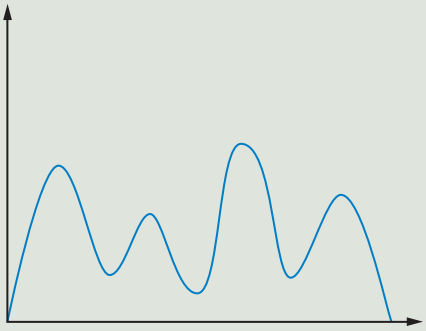
ISBN 9780170413220

The modality of a statistical distribution

This graph is described as **bimodal** because it has 2 peaks.



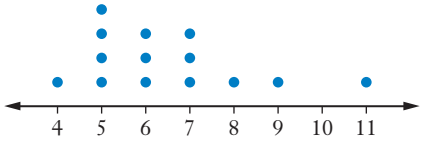
This graph is described as **multimodal** because it has many peaks.



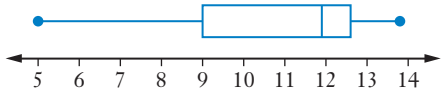
Exercise 7.07 Shape and modality of data sets

1 Describe the shape and modality of each graph.

a



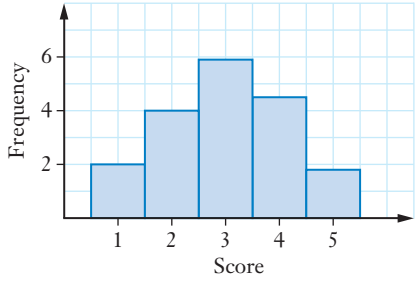
b



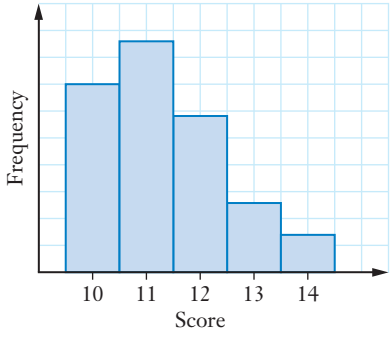
c

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

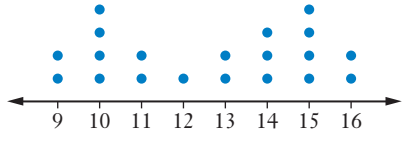
d



e

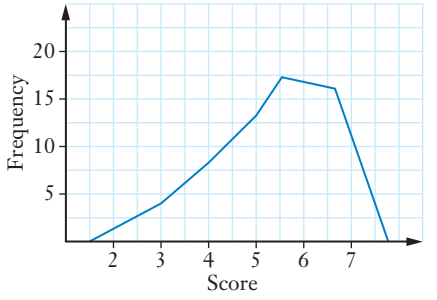


f

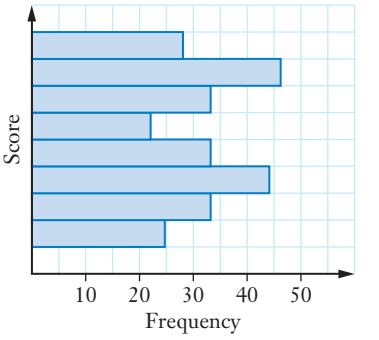


ISBN 9780170413220

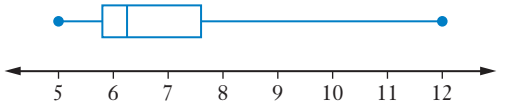
g



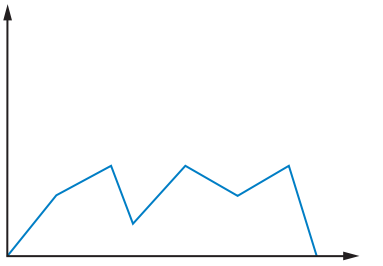
h



i



j



2 **a** Draw a dot plot for this data.

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

b Describe the shape of the dot plot.

3 Describe the shape and modality of each data set.

a

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

b

Score	Frequency
12	3
13	7
14	11
15	14
16	9
17	2

c

Score	Frequency
10–14	3
15–19	6
20–24	7
25–29	4
30–34	7
35–39	5

d

Score	Frequency
6	1
7	2
8	5
9	8
10	7
11	4

e

Score	Frequency
50– 59	2
60– 69	4
70– 79	6
80– 89	3
90– 99	1

4 Draw graphs with the following shapes.

a bimodal **b** skewed negatively

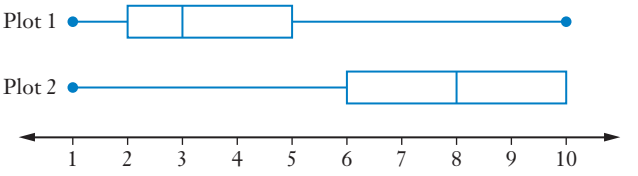
c symmetrical **d** skewed positively

ISBN 9780170413220

5 Describe the shape and modality of each set of data in the back-to-back stem-and-leaf plot.

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

6 **a** Describe the shape of the distributions summarised by the parallel box plots.



b What is the difference between their medians?

c Find the difference in their interquartile ranges.

7 The heights of a number of students were measured and the results are below.

159, 175, 181, 153, 177, 168, 175, 163, 155, 184, 167, 179, 157, 149, 160, 171, 180, 160, 162, 169, 163, 179, 145, 187, 161, 148, 182, 151, 150, 178

a Draw a frequency distribution table for the heights, using groups of 145–149, 150–154, 155–159 and so on.

b Describe the type of distribution for this data.

8 Draw a box plot that describes a symmetrical distribution.

9 This table shows the results of an assessment task.

Score	Frequency
4	3
5	5
6	6
7	9
8	6
9	5
10	3

a Find:

i the mean

ii the median

iii the mode

b Describe the shape of the distribution.

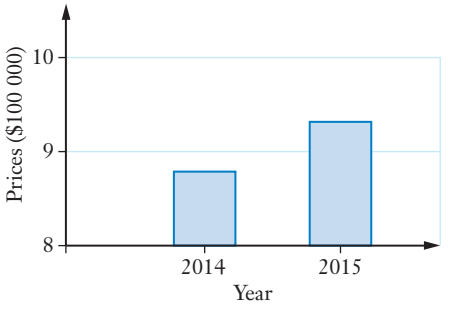
10 Choose a random sample of about 50 people and collect data on the number of siblings (brothers and sisters) each one has. Graph the data and describe the shape and modality of the graph.

ISBN 9780170413220

INVESTIGATION

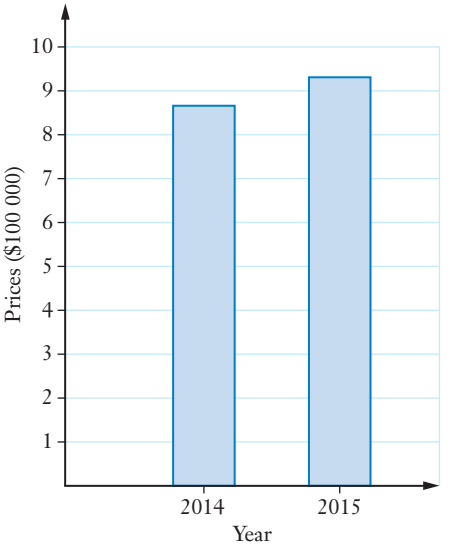
MISLEADING GRAPHS

Sydney's median house price increased from $886 408 in 2014 to $929 842 in 2015.



The column graph shows this information. Looking at the graph, you would think that this was a huge price rise because the second column is almost twice as tall as the first column.

Now look at this graph. What is the difference between the 2 graphs? Which one do you think shows the information better? Is one of the graphs misleading? Why?



Search online for other misleading graphs. Collect them into a portfolio and share with the class. Write an account of why each one is misleading and how you could change it to give a better reading of the information.

Statistics can be misleading in different ways. In the investigation this was caused by the scale on the graph. Sometimes the measures of central tendency or spread can be misleading as well.

ISBN 9780170413220

Double box plots

WS Comparing city temperatures

WS Comparing word lengths

WS Comparing sports scores

7.08 Analysing data sets

EXAMPLE 24

The table shows the heights of students in a Year 12 class.

Height (cm)	Class centre	Frequency
150–154	152	3
155–159	157	18
160–164	162	27
165–169	167	31
170–174	172	12
175–179	177	15
180–184	182	25
185–189	187	11
190–194	192	3

a Find the mean and standard deviation of the heights.

b Are the mean and standard deviation misleading for this data? Why?

Solution

a Using a calculator:

$\overline{x} = 170.6$

$s = 10.2$

b Looking at the table, the data looks to be bimodal. This might be because the survey is for both males and females.

If this is the case, the mean of 170.6 is misleading because it doesn't tell us about differences in male and female heights. The spread may be less than 10.2 if we split the data into male and female data.

ISBN 9780170413220

Comparing two or more sets of data

Sometimes we need to compare different data sets to see how similar or different they are.

EXAMPLE 25

a Two surveys were made into the number of people attending an outdoor cinema: one in 2019 and one in 2020.

For the 2019 survey, the mean was 112 and the standard deviation was 6.7.

For the 2020 survey, the mean was 95 and the standard deviation was 11.9.

Describe how these results differ.

b This back-to-back stem-and-leaf plot shows the results of tests of the life of 2 brands of batteries batteries (measured in hours).

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

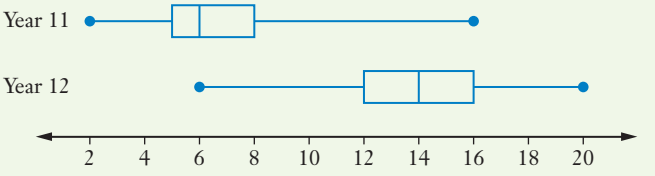
i Describe the shape of the distribution for each brand.

ii Find the mean result for each brand.

iii Find the standard deviation for each brand.

iv Compare the results for the 2 brands of batteries.

c The parallel box plots below show the results of 2 surveys into the number of hours 2 groups of students study each week.



i What is the median number of hours studied for each Year group?

ii Calculate the range for each group.

iii What is the interquartile range for each group?

iv What is the highest number of hours studied in each group surveyed?

v What is the main difference between the 2 groups?

ISBN 9780170413220

Solution

a The mean was lower in the second survey, so it looks as if, on average, fewer people were going to the movies in 2020 than in 2019.

The standard deviation was higher in the second survey, so there was a greater variation in the number of people going to the movies.

b **i** Buzz is slightly positively skewed and Eternity is approximately bimodal.

ii Using a calculator:

The mean for Buzz is 60.4.

The mean for Eternity is 69.4.

iii The standard deviation for Buzz is 12.3 and the standard deviation for Eternity is 14.0.

iv The mean was higher on Eternity so these batteries had longer lives overall.

The standard deviation was slightly higher on Eternity, showing slightly more variability in the life of these batteries. That is, the battery lives were more spread out than for Buzz, but there wasn't a big difference between the 2 brands.

c **i** Year 11: median is 6.

Year 12: median is 14.

ii Year 11: range = 16 – 2 = 14

Year 12: range = 20 – 6 = 14

iii Year 11: interquartile range = 8 – 5 = 3

Year 12: interquartile range = 16 – 12 = 4

iv Year 11: highest hours = 16

Year 12: highest hours = 20

v Year 12 students generally study for more hours.

CLASS DISCUSSION

ANALYSING DATA SETS

Why do you think the surveys in the example give different results? Are they taken from the same population? How could you tell? What other information could help you decide?

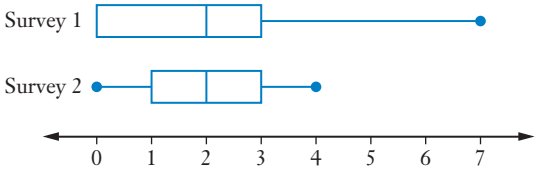
Find other examples online, in newspapers or in magazines that compare 2 or more sets of data. Is the information taken from the same or different populations? Can you tell?

Put these examples in a portfolio and present a report to the class.

ISBN 9780170413220

Exercise 7.08 Analysing data sets

1 The parallel box plots show the results of 2 surveys into the number of children in families.



a What was the largest number of children in a family in:

i survey 1? **ii** survey 2?

b What is the median number of children in the families in:

i survey 1? **ii** survey 2?

c Find the interquartile range for:

i survey 1 **ii** survey 2

d Do you think the surveys were taken from the same population or from different populations?

2 The heights (in centimetres) of people were taken from 2 samples.

Sample 1: 158, 169, 177, 147, 160, 153, 167, 164, 171, 159, 162, 166, 172, 148, 157, 176, 155, 170, 149, 150

Sample 2: 181, 166, 179, 181, 164, 173, 168, 165, 176, 181, 157, 162, 178, 182, 157, 172, 180, 173, 159, 151

a Plot this information in a back-to-back stem-and-leaf plot.

b Find the median of each sample.

c Find the range of each sample.

d Do you think the samples were taken from the same population?

3 Surveys were taken at 3 different banks to measure the length of time customers waited before being served. The results are shown in the table.

Time (min)	Bank 1	Bank 2	Bank 3
0–2	29	59	2
3–5	38	26	8
6–8	15	12	11
9–11	9	3	21
12–14	5	0	28
15–17	3	0	20
18–20	1	0	10

a Find **i** the mean and **ii** the median waiting times for customers at each bank.

b Find the standard deviation for each bank.

c Do you think there is a significant difference in waiting times at the banks?

ISBN 9780170413220

4 Mrs Spell's piano students earned the following marks in their piano examinations:

Class 1: 91, 86, 74, 92, 85, 89, 63, 71, 80, 91, 85, 72, 54, 78

Class 2: 97, 87, 69, 91, 88, 89, 93, 94, 71, 79, 84, 85, 88

a Sketch parallel box plots showing this information.

b For each class, find:

i the median **ii** the interquartile range

iii the mean **iv** the range

5 Two speed cameras at different locations recorded speeds (in km h^{-1}) of vehicles travelling over the speed limit:

Camera 1: 85, 66, 75, 69, 72, 83, 80, 69, 74, 77, 73, 74, 90, 84, 65, 73, 69, 89, 76, 103

Camera 2: 122, 142, 120, 118, 116, 135, 140, 123, 135, 124, 120, 119, 138, 131, 122, 119, 125, 130, 130, 113

a Draw a back-to-back stem-and-leaf plot to show this data.

b Find the mean speeds recorded by each camera.

c What do you think was the speed limit at the site where each camera was placed?

6 Jon sat for the HSC in one year and scored 56, 48, 61, 53, 41 and 35 for his maths assessments. He resat his HSC the next year and his maths assessment scores were 73, 58, 67, 74, 59 and 68.

a By how much did his mean scores increase the second year?

b What was the difference in the median scores?

c Calculate the difference in the range of scores for each year.

d By how much does the standard deviation differ over the 2 years?

7 These 2 sets of scores have the same median.

Test 1: 1, 2, 4, 6, 7, 9, 10 **Test 2:** 4, 5, 5, 6, 6, 7, 8

a What is the mean of:

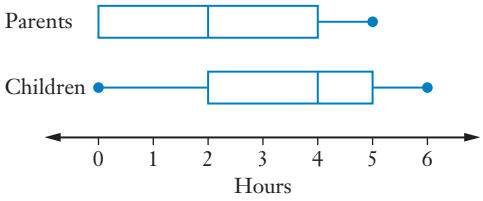
i Test 1? **ii** Test 2?

b Calculate the standard deviation of:

i Test 1 **ii** Test 2

c Describe how the 2 sets of scores differ.

8 The parallel box plots show the results of 2 surveys into the number of hours people spend watching TV each day.



a What is the median for each group?

b For each group, find:

i the interquartile range **ii** the range

c What is the highest number of hours of TV watched?

ISBN 9780170413220

9 For the back-to-back stem-and-leaf plot, find:

a the median of each test

b the mean of each test

c the standard deviation of each test

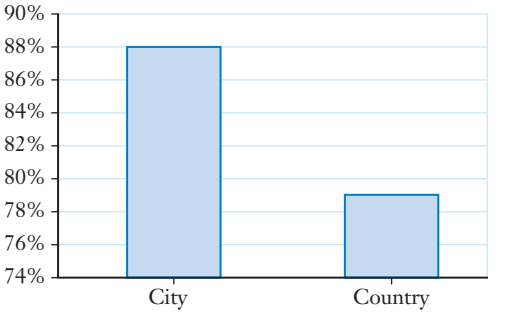
d the difference in range between the 2 tests.

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

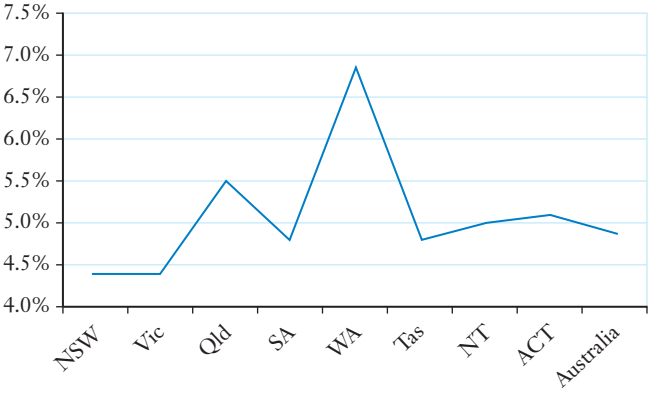
10 The graph compares access to the Internet for city and country households in 2015.

a Describe how this graph is misleading.

b Redraw the graph so it is not misleading.



11 The graph shows the average annual growth in incomes from 2006 to 2011.



Redraw the graph so it is not misleading.

INVESTIGATION

COMPARING SURVEY RESULTS

Conduct a survey among your friends. Make up your own topic, such as what sports they play, subjects they study or their heights. Alternatively, you could carry out an experiment such as counting numbers of people travelling in cars or measuring the time taken for the same journey to school on different days.

To check your results, take another sample and do the same survey or experiment. Are the new results the same? Can you explain why?

ISBN 9780170413220

7. TEST YOURSELF

Practice quiz

Statistics review

Statistics crossword

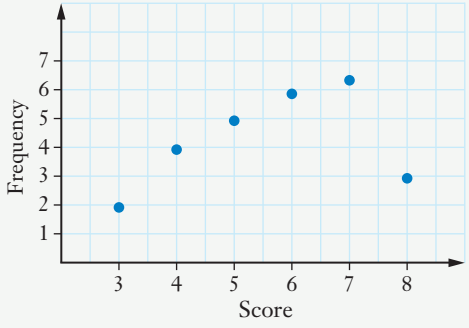
For Questions **1** to **6**, select the correct answer **A**, **B**, **C** or **D**.

1 What type of data is a person's body temperature?

A categorical ordinal
B numerical continuous
C numerical discrete
D categorical nominal

2 This frequency distribution is:

A symmetrical
B positively skewed
C negatively skewed
D multimodal



3 What type of data is a flower's colour?

A categorical ordinal
B numerical continuous
C numerical discrete
D categorical nominal

4 The median of scores 7, 3, 2, 6, 4, 3, 8, 9 is:

A 6 **B** 4 **C** 3 **D** 5

5 Which measure is most affected by an outlier?

A median
B interquartile range
C mode and interquartile range
D mean and range

6 Test 1 has a mean of 4 and standard deviation of 1.2. Test 2 has a mean of 4 and a standard deviation of 2.5. Which statement below is true when comparing the centres and spreads for both tests?

A The centres are the same and test 2 has a larger spread than test 1.
B The spreads are the same and test 2 has a higher centre than test 1.
C The spreads are the same and test 2 has a lower centre than test 1.
D The centres are the same and test 2 has a smaller spread than test 1.

7 Find the mode, median and range of this data set:

8, 6, 8, 4, 5, 6, 8, 5, 7, 4, 7, 8, 6, 8, 9

ISBN 9780170413220

8 The table shows the results of a survey to find the distance people must travel to work.

a Find the modal class.

b Find the mean.

c What is the median?

d Find the standard deviation and variance.

Distance (km)	Frequency
0–4	6
5–9	8
10–14	4
15–19	7
20–24	3

9 Find the mean and standard deviation of this data set:

16, 34, 29, 80, 65, 77, 91, 58, 67, 40

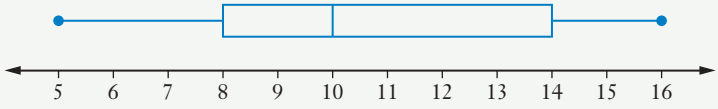
10 From this box plot, find:

a the median

b the range

c the 3rd quartile

d the interquartile range.



11 The table shows the results of a survey into the number of years that people keep a car before selling it.

a Draw a cumulative frequency polygon to show this data.

b Estimate the median age of the cars.

c Estimate:

i the 20th percentile **ii** the 3rd quartile

iii the 91st percentile **iv** the 3rd decile **v** the 9th decile

Years	Frequency
0–2	15
3–5	37
6–8	13
9–11	18
12–14	17

12 The table records the number of times people visited a doctor in the past 6 months.

a Find the mean number of visits.

b Find the median number of visits.

c What is the range?

d Find the interquartile range.

Visits	Frequency
0	9
1	8
2	5
3	1
4	3
5	2

13 'Most families have 2 children.' Is this statement about a mean, median or mode?

14 Nikola conducted a survey to find the number of people in each car that travelled across the Anzac Bridge one morning. These were her results:

a Draw a cumulative frequency polygon to illustrate the results.

b From the polygon, find:

i the median number of people in a car

ii the interquartile range

Occupants	Frequency
1	43
2	32
3	12
4	8
5	5

ISBN 9780170413220

15 The table shows the results of Mr Cheung's history class.

Score	Frequency
6	7
7	6
8	3
9	6
10	2

a Find the mean score.

b Find the standard deviation.

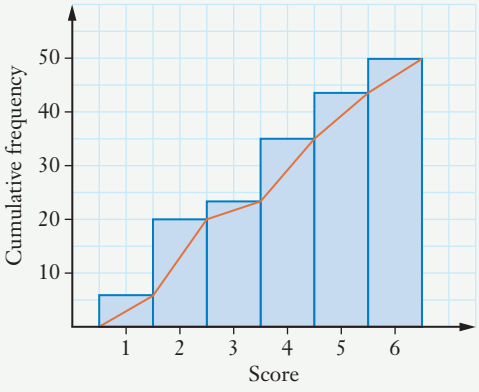
c What is the mode?

16 Find the mean, mode and median of this data set:
45, 49, 49, 48, 43, 45, 41, 40, 49, 48, 44, 40, 42

17 From the graph, find:

a the interquartile range

b the median



18 A class test gave the following scores:
15, 19, 12, 2, 19, 16, 13, 18, 11, 15, 17, 11, 18, 14, 14, 16, 18, 14, 12

a Find:

i the range **ii** the mean

iii the mode **iv** the median

b Show that one score is an outlier. Which is it?

c Find, without this outlier:

i the range **ii** the mean

iii the mode **iv** the median

d Does the outlier have much effect on all these measures?

19 The two-way table shows the results of a survey into the number of pets microchipped at a veterinary surgery.

	Male	Female	Total
Cats	369	473	842
Dogs	578	664	1242
Total	947	1137	2084

a What percentage of microchipped animals were:

i female? **ii** female cats? **iii** male dogs? **iv** dogs?

b Draw this information in a Pareto chart.

ISBN 9780170413220

20 The table shows the reasons employees gave for leaving their jobs in a large organisation. Draw a Pareto chart of this data.

Work too difficult	185
Boring work	139
Not paid enough	104
Not getting on with co-workers	56
Unsuitable hours	116

21 A back-to-back stem-and-leaf plot shows the number of mushrooms found in 2 regions of a forest in New Zealand.

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

a Calculate the median number of mushrooms recorded in each region.

b Find the mean and standard deviation of North region.

c Find the mean and standard deviation of West region.

d Compare and contrast the 2 regions.

22 Below are the results of 2 English assessments.

Term 1: 8, 7, 9, 8, 6, 5, 8, 7, 7, 5, 9, 9

Term 2: 5, 7, 8, 4, 6, 6, 5, 5, 5, 4, 7

a Draw a box plot for each set of data.

b What is the median of each assessment?

c Find the interquartile range of each set.

d Find the mean and standard deviation for each assessment.

e Compare and contrast the 2 assessments.

ISBN 9780170413220

7. CHALLENGE EXERCISE

1 The table shows the heights of students in Year 12.

a Describe the modality of the distribution. Can you explain this?

b Find the mean height and the variance.

Height (cm)	Frequency
150–154	7
155–159	5
160–164	15
165–169	9
170–174	8
175–179	15
180–184	6
185–189	2

2 The mean of a data set is 4.5. The table containing the data has a frequency missing. What is it?

Score	Frequency
2	2
3	5
4	
5	3
6	2
7	4

3 **a** Sketch a box plot for the following scores.
7, 9, 5, 6, 8, 11, 25, 14, 16

b What does this plot show? What is its shape? Can you make a more symmetrical box plot by taking out one of the scores?

4 For the scores 7, 11, 15, 19, 23

a Find:

i the mean **ii** the standard deviation

b Copy and complete this table:

Score x	$x - \bar{x}$	$(x - \bar{x})^2$
7		
11		
15		
19		
23		
	$\Sigma(x - \bar{x}) =$	$\Sigma(x - \bar{x})^2 =$

c Find the standard deviation by using the formula $\sigma = \sqrt{\frac{\Sigma(x - \bar{x})^2}{n}}$.

ISBN 9780170413220

5 The mean of 7 scores is 25. If an extra score of 28 is also included, what will be the new mean?

6 Compare and contrast each pair of data sets.

a Set A has a mean of 54 and a standard deviation of 5.6
Set B has a mean of 76 and a standard deviation of 2.1

b Set A has a mean of 11.6 and a standard deviation of 2.7
Set B has a mean of 21.3 and a standard deviation of 9.2

7 A Year 12 class received the following scores out of 10 on their maths quiz.

a Find the mean and standard deviation of the scores.

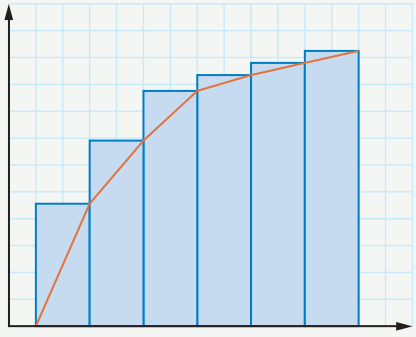
b Draw a box plot for this set of scores.

c Describe the shape of the distribution.

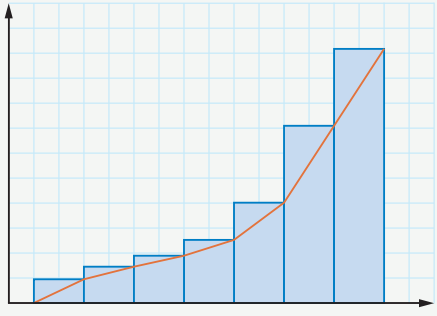
Score	Frequency
4	7
5	11
6	15
7	9
8	4
9	2
10	1

8 Describe the shape of each distribution, given the ogive.

a



b



ISBN 9780170413220