Data for Statistics

Curriculum Ready

Mathletics

www.mathletics.com

First edition printed 2009 in Australia.

A catalogue record for this book is available from 3P Learning Ltd.

ISBN 978-1-925202-70-0

Published 3P Learning Ltd

For more copies of this book, contact us at: www.3plearning.com/contact/

Designed 3P Learning Ltd

This book is about how we can collect information and interpret what it means.

Collecting and interpreting information can be very helpful for improving many things in life.

- Population information can help countries decide what they need to do to help the country move forward.
- Which drink flavour is preferred by customers to help ensure the right amounts of each are ordered every time.

Think of 2 other examples where collecting information can be helpful and briefly explain why below:

1

2

Q The word '**novemnonagintillion**' can be used to name a number starting with 1, followed by 300 zeros!

- Which letter occurs the most often in this word?
- If the letters of this word were arranged alphabetically, which letter would be in the middle?

Work through the book for a great way to do this

Primary and secondary data

Data is often a list of numbers called 'scores' or 'results'.
The heights in cm of 4 people are: 135, 145, 180, 178. This is data.

Primary data:
When no information or results currently exists on a topic, new (primary) data needs to be collected.
Primary data comes from conducting interviews, questionnaires (surveys), experiments or observations.

⊕ Data is recent, reliable and exactly what you want to know.
⊖ Collecting this data can be expensive requiring a lot of work and time.

Secondary data:
When information or results have already been collected by someone else, it is called secondary data.
It is often taken from trade directories, reports or websites etc.

⊕ Less expensive, instantly available as work has been done by someone else.
⊖ Data could be too old to be relevant and not always be the exact information required.

Examples of primary and secondary data

(i) α) Your teacher asks you to personally collect data from your classmates about how many siblings they each have.

The results from 15 students surveyed are: 2, 0, 1, 3, 2, 1, 2, 1, 2, 4, 2, 2, 3, 3, 2.

Since you collected this data by asking the classmates yourself, then these results represent **primary data**.

β) Your teacher decides that instead of you collecting the data, she would simply read out the numbers on file.

The results read out for 15 students are: 2, 0, 1, 3, 2, 1, 2, 1, 2, 4, 2, 2, 3, 3, 2.

The data may be the same, however you didn't personally collect the results.
So this is **secondary data**.

(ii) The range of temperatures for cities around the globe on a particular day are shown below:

	Low	High	Report
Beijing	19	30	Cloudy
Kuala Lumpur	25	33	Showers
New Delhi	28	37	Thunderstorms
Dublin	11	19	Bright
Johannesburg	5	22	Fine
Melbourne	7	14	Showers
London	13	21	Bright
Chicago	16	27	Fine
Toronto	16	20	Bright
Buenos Aires	4	15	Fine
Auckland	15	24	Cloudy

You did not personally collect/measure the weather data yourself in each city.
So this weather information is secondary data.

Primary and secondary data

1 Tick the type of data being collected in each of these situations:

a Counting the total number of fruit eaten by students during recess each day over 1 week.

☐ Primary ☐ Secondary

b Counting how many steps there are in your house.

☐ Primary ☐ Secondary

c Asking a Vet to send you a list of the totals for all the different animals they helped over a month.

☐ Primary ☐ Secondary

d Recording the temperature of your home from a thermometer each day.

☐ Primary ☐ Secondary

e Writing down the speeds of various wild animals from a documentary on TV.

☐ Primary ☐ Secondary

f Recording the sporting results from a radio broadcast.

☐ Primary ☐ Secondary

g Measuring and recording the lengths of the leaves on a plant using a tape measure.

☐ Primary ☐ Secondary

h Watching the number of times different people in a room blink every minute.

☐ Primary ☐ Secondary

i Listening to your friend tell you how many horses his family counted during holiday road trip.

☐ Primary ☐ Secondary

j Using the odometer in the car to calculate how far you travel each time you are driven somewhere.

☐ Primary ☐ Secondary

k Counting how often other people yawn when you do deliberately.

☐ Primary ☐ Secondary

l Checking to see how many friends your classmates have on a social network site.

☐ Primary ☐ Secondary

m Writing down the 3 most popular types of cars sold over the past month from an automotive magazine.

☐ Primary ☐ Secondary

Primary and secondary data

2 You are tasked to do a report covering all sorts of general information about the student population of your school.
For example preferred food, types of pet, favourite subject, sport, music, colour, book, movie genre etc.
Write down two ways you can collect primary and secondary data to help with this report.

Primary:

a

b

Secondary:

a

b

Range

A set of data scores will always have a lowest and highest value. All other scores will sit somewhere in between. The **range** is the difference between the highest and lowest scores.

$$\text{Range} = \text{Highest score} - \text{Lowest score}$$

Calculate the range for each of these sets of scores.

(i) 4 cm, 3 cm, 4 cm, 6 cm, 5 cm, 9 cm, 8 cm, 7 cm.

Find the highest and lowest score:

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

(3 cm, — lowest score; 9 cm, — highest score)

$$\begin{aligned}\text{Range} &= \text{Highest score} - \text{Lowest score}\\ &= 9\,\text{cm} - 3\,\text{cm}\\ &= 6\,\text{cm}\end{aligned}$$

Make sure to include the units in your answer.

(ii) 4.3, 3.4, 5.7, 4.2, 5.4, 3.6, 2.9, 5.1

(2.9, 3.4, 3.6, 4.2, 4.3, 5.1, 5.4, 5.7)

(2.9 — lowest score; 5.7 — highest score)

Rearranging scores into ascending order first helps.

$$\begin{aligned}\text{Range} &= 5.7 - 2.9\\ &= 2.8\end{aligned}$$

Ascending order = smallest to highest.

If the values are in mixed forms, it can help to change them to the same form first.

(iii) $15,\ 14\frac{1}{2},\ 14.2,\ 15.1,\ 15\frac{1}{4},\ 14$

$15,\ 14.5,\ 14.2,\ 15.1,\ 15.25,\ 14$

All values in the same decimal form.

(14, 14.2, 14.5, 15, 15.1, 15.25)

(14 — lowest score; 15.25 — highest score)

Scores rearranged into ascending order.

$$\begin{aligned}\text{Range} &= 15.25 - 14\\ &= 1.25 \ \left(\text{or } 1\tfrac{1}{4}\right)\end{aligned}$$

Range

1 **a** Circle the highest and lowest scores in each of these sets of data:

(i) 2, 7, 3, 1, 9, 10, 9, 11, 6, 8, 3, 4

(ii) 23.2, 24.1, 22.9, 23.5, 24.8, 25.0

(iii) −3°C, −5°C, −2°C, 0°C, −1°C, −2°C, −4°C, −4°C, −6°C

(iv) 2 cm, 24 mm, 1.8 cm, 2.5 cm, 26 mm

b Calculate the range of the scores for parts (i) – (iv)

(i) Range = ☐ (ii) Range = ☐

(ii) Range = ☐ °C (iv) Range = ☐ mm

2 (i) Rearrange these sets of data into ascending order, then ...

(ii) Calculate the range of the scores for each set.

a −3°C, 9°C, −1°C, 2°C, −4°C, 0°C, 5°C, 3°C, 1°C

(i) ☐, ☐, ☐, ☐, ☐, ☐, ☐, ☐, ☐

(ii) Range = ☐ °C

b 15 min, 0.5 hours, 10 min, 18 min, 20 min, 0.2 hours

(i) ☐, ☐, ☐, ☐, ☐, ☐

(ii) Range = ☐ min

c $2\frac{3}{4}$ m, $1\frac{1}{2}$ m, 100 cm, 2.1 m, 180 cm.

(i) ☐, ☐, ☐, ☐, ☐

(ii) Range = ☐ cm

Range

3 The range of scores can be limited by the tool being used to collect the data.

a (i) What is the lowest and highest possible score on a normal 6-sided die?

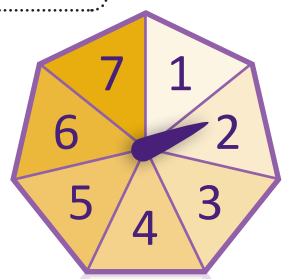
Lowest = _____ Highest = _____

(ii) What is the range of possible scores if this die is rolled many times?

Range = _____

b (i) Write the lowest and highest scores shown on the spinner.

Lowest = _____ Highest = _____



(ii) What is the range of possible scores if this is spun many times?

Range = _____

c (i) Write the lowest and highest scores possible from a bag that has all the multiples of 6 that are between 0 and 30.

Lowest = _____ Highest = _____

(ii) What is the range of possible scores if numbers are drawn and replaced many times over?

Range = _____

Measures of central tendency: Mode

Mathematicians try to find the 'typical' result for data collected using calculations called 'Measures of Central Tendency'.

The first measure of central tendency we will look at is called the mode.

The mode (or modal score) = the score that occurs the most.

So the modal score is the one with the highest frequency!

Mode = Most Frequent Score

Special cases:

- If there are no repeated scores in a data set, then there is no mode.
- If two scores share the same highest frequency, then there are two modes, and the data set is called **bimodal**.

The prefix 'bi' means 2.

- If more than two scores share the same highest frequency, then the data set is called **multimodal**.

Here are some modal score examples

a Find the mode for this data set: 3, 4, 4, 5, 6, 7, 8, 9.

The score '4' has the highest frequency of 2.

∴ Mode = 4 *or* The data has a modal score of 4.

b A class has 15 students and the numbers of brothers or sisters each classmate has is given by the data set:

2, 3, 2, 0, 0, 1, 3, 2, 4, 2, 0, 1, 2, 2, 1.

Find the modal score.

If there are many scores, you can draw a frequency table or reorder the scores into ascending order.

Using a frequency table:

Score	Tally	Frequency
0	|||	3
1	|||	3
2	𝍸 |	6
3	||	2
4	|	1

Frequencies of each score

Rearranging into ascending order:

0, 0, 0, 1, 1, 1, 2, 2, 2, 2, 2, 2, 3, 3, 4.

(3, 3, 6, 2, 1 — frequencies of each score)

The score that occurs the most is 2. It has a frequency of 6.

∴ Mode = 2 *or* The data has a modal score of 2.

c The points scored by each player of a basketball team for one game were: 3, 3, 3, 4, 4, 4, 5, 5, 5, 14 .
Find the mode(s) and comment whether the data is bimodal or multimodal?

3, 3, 3, 4, 4, 4, 5, 5, 5, 8

(3, 3, 3, 1 — frequencies of each score)

There are three scores with the same highest frequency.

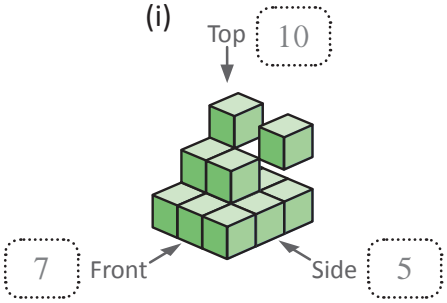
∴ The data has 3 modal scores of 3, 4 and 5.

∴ This data is multimodal as there are more than 2 modes.

Mode

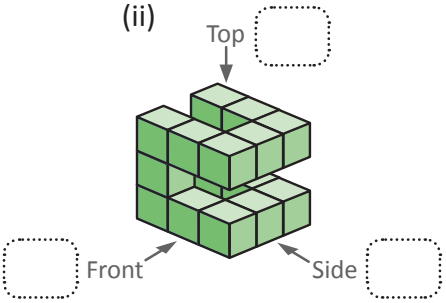
1 **a** Count the number of square faces visible for each view of these images.
Which view has the modal score for square faces visible?

(i)



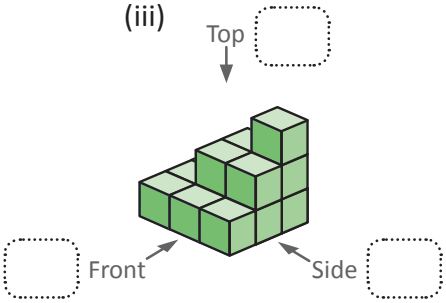
Modal score view = TOP

(ii)



Modal score view =

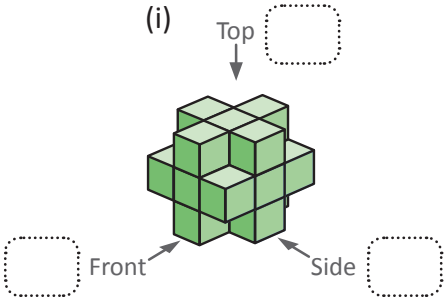
(iii)



Modal score view =

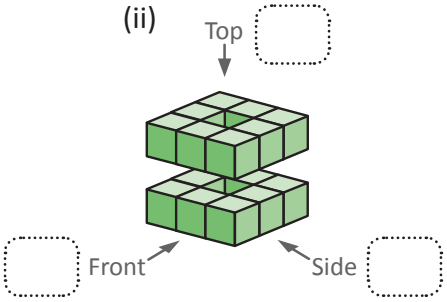
b Count the number of square faces visible for each view of these images.
Tick the correct description for the number of modes of square faces visible.

(i)



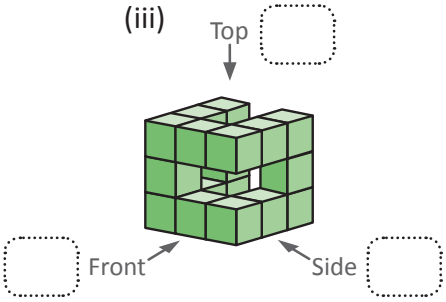
☐ No mode ☐ One mode
☐ Bimodal ☐ Multimodal

(ii)



☐ No mode ☐ One mode
☐ Bimodal ☐ Multimodal

(iii)



☐ No mode ☐ One mode
☐ Bimodal ☐ Multimodal

2 Each student in a year 7 class recorded the number of bicycles their family home has:

0, 1, 4, 3, 4, 2, 3, 0, 0, 1, 1, 2, 1, 1, 0, 2, 3, 1, 0, 2, 3, 2, 1, 1, 2, 3, 3, 3, 3, 2, 1

a Complete the tally table for this set of scores.

Score	Tally	Frequency
0		
1		
2		
3		
4		

Every fifth tally is drawn as a line through the middle
∴ 5 = 𝍸

b Modal score = ☐ ?

Mode

3 Find the mode(s) for these data sets after rearranging them into ascending order.

a 2, 6, 8, 3, 1, 10, 4, 9, 8, 5, 5 — Mode = ☐ ☐ No mode ☐ One mode ☐ Bimodal ☐ Multimodal

b 8, 5, 7, 9, 9, 7, 8, 5, 8, 9, 5, 7 — Mode = ☐ ☐ No mode ☐ One mode ☐ Bimodal ☐ Multimodal

c 15, 30, 7, 9, 12, 15, 5, 29, 7 — Mode = ☐ ☐ No mode ☐ One mode ☐ Bimodal ☐ Multimodal

d 107, 110, 99, 97, 101, 112, 108, 100, 111, 101 — Mode = ☐ ☐ No mode ☐ One mode ☐ Bimodal ☐ Multimodal

e 5.4, 6.6, 1.9, 1.2, 2.4, 2.6, 3.4, 0.1, 2.6, 1.2, 2.4 — Mode = ☐ ☐ No mode ☐ One mode ☐ Bimodal ☐ Multimodal

4 Complete the missing numbers from the following single mode data sets arranged in ascending order:

Data Set	Range	Mode
1, 2, 2, 3, 4	☐	☐
5, 5, 6, 7, ☐	3	☐
4, 6, ☐, 9, ☐	23	6
☐, 2, 2, 3, ☐, ☐, ☐, 9, 12, 12, 13, 24, 30	30	6
1, 1, 3, 3, 7, 9, 9, ☐, 16, ☐	19	☐
12, 13, 17, 17, ☐, ☐, ☐, 19, 22, 22 ,26	☐	19

Mode

5 Change the mode of these data sets as requested:

a What single score must be added to each of these data sets to make them bimodal?

(i) 3, 4, 5, 2, 3, 1, 4, 3, 6, 7, 6, 2, 7, 3, 4

Score needed to make this data set bimodal = _____

(ii) 12.2, 12.1, 12.6, 12.3, 12.5, 12.4, 12.1, 12.6, 12.1, 12.7

Score needed to make this data set bimodal = _____

b What single score must be added to each of these data sets to make multimodal?

(i) 8, 9, 10, 1, 10, 8, 10, 9, 10, 8, 8, 7, 9, 1, 9

Score needed to make this data set multimodal = _____

(ii) 23.5, 23, 22.5, 23.5, 22, 24, 22.5, 22, 23.5, 22, 21

Score needed to make this data set multimodal = _____

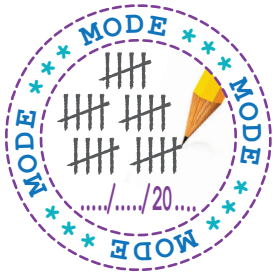
c Write down a score that could be added to each of these data sets to give them a single modal value?

(i) 0.4, 0.3, 0.1, 0.3, 0.5, 0.4, 0.5, 0.2

Score needed to give this data set a single mode = _____

(ii) −8, 6, 0, 2, −1, 7, −3, 3, −8, 2, 0, 3, 8

Score needed to give this data set a single mode = _____



Measures of central tendency: Mean

The **mean** score is the average score in a data set.

Symbol for the **mean score** is $\bar{x}$ (so an x with a bar over the top).

To find the mean score from a set of data, just add them all up and divide by how many there are.

$$\text{Mean (or average)} \quad \bar{x} = \frac{\text{Add (summate) all the scores together}}{\text{Total number of scores}}$$

Here are some examples of finding the mean (or average)

Find the mean (or average) of the given numbers:

a 4 and 8 — Just add them together and divide by 2 — Because you are averaging two scores.

$$\therefore \text{The mean score is: } \frac{2+8}{2} = \frac{12}{2}$$

$$\therefore \bar{x} = 6.$$

b 10, 15, 8 — Find their sum and divide by 3 — Because you are averaging three scores.

$$\therefore \text{The mean score is: } \frac{10+15+8}{3} = \frac{33}{3}$$

$$\therefore \bar{x} = 11.$$

c 2, 4, 6, 8, 10, 12 — Summate the scores and divide by 6 — Because you are averaging six scores.

$$\therefore \text{The mean score is: } \frac{2+4+6+8+10+12}{6} = \frac{42}{6}$$

$$\therefore \bar{x} = 7.$$

Sometimes we need to know what score is required to finish with a particular average score.

d Jack Racer finished 3^{rd}, 1^{st}, 2^{nd}, 1^{st}, 4^{th}, 1^{st}, 3^{rd}, 6^{th}, 2^{nd} and 2^{nd} in the previous 9 races.

What place must Jack finish in the 10^{th} race to have an average finishing position of 3^{rd}?

There are 10 races in total, so the sum of all the positions divided by 10 must $= 3$ (for 3^{rd})

Let the last race position be called 'p'

$$\therefore \bar{x} = \frac{3+1+2+1+4+1+3+6+2+2+p}{10} = 3$$

$$\therefore 3+1+2+1+4+1+3+6+2+2+p = 30. \quad \text{Multiply 3 by 10.}$$

$$\therefore 25 + p = 30.$$

$$\therefore p = 5.$$

So Jack Racer must finish 5^{th} in the last race to have an average finishing position of 3^{rd}.

Mean

1 Find the mean (or average) of these data sets:

a 5, 8, 2

Mean score $\bar{x} =$ ⬚

b 4, 3, 6, 8, 9, 9, 5, 4, 2, 0, 5, 7, 3

Mean score $\bar{x} =$ ⬚

c 2.3 , 4.1 , 8.4 , 9.2 , 3.5

Mean score $\bar{x} =$ ⬚

d 4.4 , 7 , 5.8 , 8

Mean score $\bar{x} =$ ⬚

e $\frac{1}{2}$, $\frac{1}{4}$, $\frac{3}{4}$

Mean score $\bar{x} =$ ⬚

f $\frac{1}{14}$, $\frac{4}{7}$, $\frac{6}{7}$, 1

Mean score $\bar{x} =$ ⬚

g -8 , 4 , 2.5 , -1 , 0 , 6 , 11 , -4.5 , 3 , 10

Mean score $\bar{x} =$ ⬚

2 Calculate the mean value of these data sets containing a mix of integers, decimals, fractions and mixed numerals.

a $\frac{1}{2}$, 4.2 , $1\frac{2}{5}$, 6 , 0.7 , 8 , $\frac{1}{5}$

Mean score $\bar{x} =$ ⬚

b 12 , $-\frac{3}{4}$, -5 , $8\frac{1}{2}$, $7\frac{1}{4}$, -18 , -4.8 , 8.2

Mean score $\bar{x} =$ ⬚

Mean

3 Rock fishermen caught twenty two fish on Monday, twenty five fish on Tuesday, and forty fish on Wednesday. What is the average number of fish they caught over those 3 days?

Average number of fish caught $\bar{x} =$ .....

4 Wendy went bird watching for seven days in a row. The table below shows how many birds she spotted on the successful days only. She spotted no birds on the other days.

Day	Number of birds spotted
MON	18
TUES	15
FRI	9
SUN	7

What was the average number of birds spotted per day in the seven days that she was bird watching?

Average number of birds spotted daily $\bar{x} =$ .....

5 If the mean of twenty scores was 3.5, what is the sum of all the scores for this data?

Sum of 20 scores $=$ .....

Mean

6 The mean score for a set of three numbers is 7 find the new mean when a score of 21 is added.

New mean $\bar{x} =$ .....

7 The average of two numbers is 15 and one of the numbers is 29, what is the other number?

Other number = ...

8 If the mean of three numbers is 7 and two of the numbers are 6 and 8, what is the third number?

Third number = ...

9 The sum of a set of scores is 234 and the mean score is $\bar{x} = 13$. How many scores are there in the set?

Number of scores = ...

Measures of central tendency: Median

The last measure of central tendency we need to learn is the **median** score.

It is the score in the middle of a data set that has been arranged into ascending order.

There are two methods for finding the median. They depend on whether there are an even or odd number of scores.

This example shows how to find the median for an **ODD** number of scores.

Find the median of the data set: $12, 9, 4, 6, 7, 3, 2$.

Arrange the data set into ascending order (i.e. from lowest score to highest):

$$2, \ 3, \ 4, \ 6, \ 7, \ 9, \ 12$$

There are always an equal number of scores on either side of the medium.

There are 7 scores in this data set and the middle score is 6.

$$\cancel{2}, \ \cancel{3}, \ \cancel{4}, \ \mathbf{6}, \ \cancel{7}, \ \cancel{9}, \ \cancel{12}$$

There are 3 scores either side of 6.

$$\therefore \text{ Median score} = 6$$

If there is an even number of scores in the data set, the median is the average of the two middle scores, and so you add them and divide by two.

This example shows how to find the median for an **EVEN** number of scores.

Find the median of the data set: $8, 9, 1, 0, 1, 4, 3, 6, 6, 9$.

Arrange the data set into ascending order (i.e. from lowest score to highest):

$$0, \ 1, \ 1, \ 3, \ 4, \ 6, \ 6, \ 8, \ 9, \ 9$$

There are 10 scores in this data set and this is an even number so there are 2 middle scores.

$$\cancel{0}, \ \cancel{1}, \ \cancel{1}, \ \cancel{3}, \ \mathbf{4}, \ \mathbf{6}, \ \cancel{6}, \ \cancel{8}, \ \cancel{9}, \ \cancel{9}$$

The two middle scores are: 4 and 6.

The median is the sum of these two scores divided by 2 (ie their average).

$$\therefore \text{ Median score} = \frac{4+6}{2} = 5$$

Median

1 Find the median for each of these data sets: [Hint: Write in ascending order first if necessary.]

a 2, 3, 1

Median score = ☐

b 6, 8, 2, 0, 4

Median score = ☐

c 9, 3, 7

Median score = ☐

d 10, 6, 4, 9

Median score = ☐

e 25, 17, 9, 7, 29, 3, 13, 27, 11, 21, 19, 23, 1, 15, 5

Median score = ☐

f 12, 23, 10, 3, 8, 10

Median score = ☐

g 2, 5, 7, 12, 9, 1, 0, 3, 6, 8, 13, 2, 9, 8, 7, 6

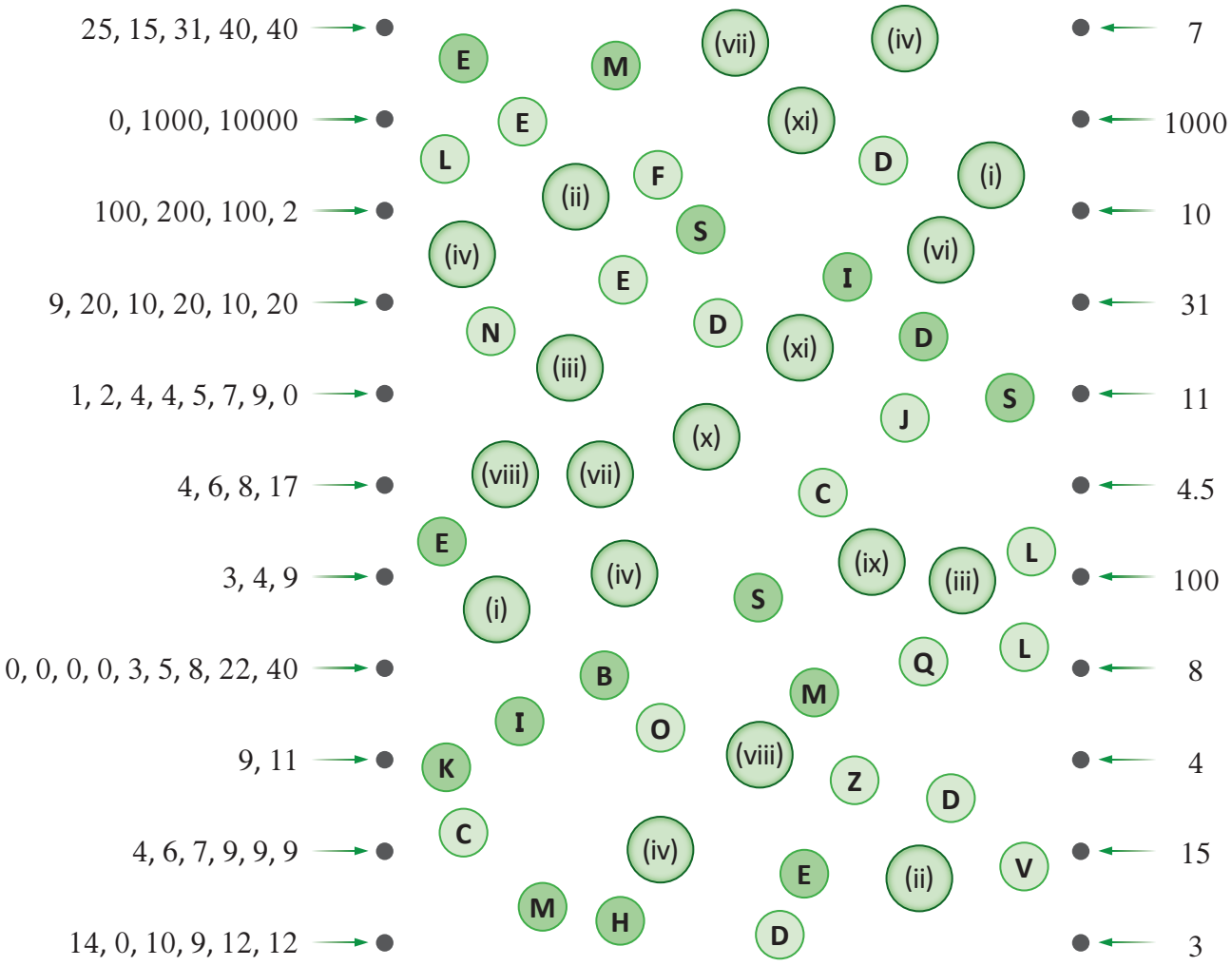
Median score = ☐

2 Find the unknown score, x, if the median of the scores for individual players in a basketball game is 6 and the scores are: 3, 4, 4, 4, 4, x, 8, 9, 10, 11.

x = ☐

Median

3 For each data set on the left, draw a line to the correct median score on the right to discover a unique set of letters.
Fill in the squares below by matching Roman numeral to solve this puzzle.



(i)	(ii)	(iii)	(iv)	(v)	(vi)	(vii)	(viii)	(ix)	(x)	(xi)

What one word starting with 'M' does this relate to? _____

Outliers and measures of central tendency

Outliers are scores which seem to be extremely low or extremely high when compared to the others scores.

$1, 2, 3, 4, 4, 4, 5, 6, \boxed{100}$ ← Outliers → $\boxed{-20}, 1.1, 3.4, 5.6, 7.5, 9.5$

$12, 15, 18, 19, \boxed{97}$ ← Outliers → $\boxed{1, 2}, 31, 34, 40, 35, 38, 35, 36$

Measures of central tendency – Which one?

The mode, mean and median all tell us the typical score for data collected.
Each one is useful for different situations.

- **Mean:** Better one to use when the data has no outliers.

  E.g. For: $2, 2, 4, 4, 4, 5, 5, 6$ $\quad\quad$ For: $2, 2, 4, 4, 4, 5, 5, 6, \boxed{100}$ ← Outlier

  $\bar{x} = 4$ $\quad\quad$ $\bar{x} = 14.\dot{6}$

  One outlier score makes the mean no longer look like a 'typical' score.

- **Median:** Better one to use when the data does have outliers.

  E.g. For: $2, 2, 4, 4, 4, 5, 5, 6$ $\quad\quad$ For: $2, 2, 4, 4, 4, 5, 5, 6, \boxed{100}$ ← Outlier

  median $= 4$ $\quad\quad$ median $= 4$

  The outlier score does not affect the median representing a 'typical' score.

- **Mode:** Better for when you want to know what is the most common score or category.

  E.g. A shop sells these sizes of a popular shirt: $6, 6, 8, 8, 8, 8, 10, 10, 12, 14, 16$.

  What size should the shop order more of? (i.e. what it the typical size being sold?)

  $6, 6, 8, 8, 8, 8, 10, 10, 12, 14, 16.$

  mode $= 8$

  $\bar{x} = 9.6\dot{3}$ and median $= 9$ ← Neither of these scores represent a typical shirt size sold.

  NOTE: Sometimes more than one measure can be good for the same set of data.

Find outliers in these sets of scores and the best measures for a typical score

a $1\ 2\ 2\ 3\ 3\ 3\ 4\ 4\ 4\ 4\ 5\ 5\ 5\ 6\ 6\ 7$

Outlier score/s? = No $\quad$ Mode $= 4$ $\quad$ Mean $(\bar{x}) = 4$ $\quad$ Median $= 4$

This data is considered 'symmetric' as any one of them returns a typical score.

b $1\ 2\ 2\ 3\ 3\ 3\ 4\ 4\ 4\ 5\ 5\ 5\ 6\ 6\ 20$

Outlier score/s? = yes, 20 $\quad$ Mode = multimodal $(3, 4$ & $5)$ $\quad$ Mean $(\bar{x}) = 4.8$ $(1$ d.p.$)$ $\quad$ Median $= 4$

The mean is affected. So the better measures of the data here are the median or the mode.
The median is the best here as the data is multimodal

c $1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 3, 3, 3, 4, 4, 4, 4, 4, 52, 98$

Outlier score/s? = yes, 52 & 98 $\quad$ Mode $= 1$ $\quad$ Mean $(\bar{x}) = 6.1$ $(1$ d.p.$)$ $\quad$ Median $= 1$

The best measure of central tendency here is the mode or the median because there are outliers.
The outliers affect (or skew) the mean $(\bar{x})$ to a value of 6.1 which is clearly not a typical score.

Outliers and measures of central tendency

1 Circle the outliers (if any) in these data sets:

a 1, 4, 8, 3, 9, 10, −2, −4, −6, −700, 12

○ No outlier ○ One outlier ○ Two outliers

b 20, 22, 19, 18, 21, 19, 21, 20

○ No outlier ○ One outlier ○ Two outliers

c 1, 36, 37, 40, 43, 52, 57, 1000

○ No outlier ○ One outlier ○ Two outliers

d 62, −66, 54, 56, 70, 66, −71, 60

○ No outlier ○ One outlier ○ Two outliers

e 1, 3, 4, 6, 8, 8, 9, 12, 14, 16, 19, 21, 23, 24, 25

○ No outlier ○ One outlier ○ Two outliers

f 1.9, 2.1, 2.4, 2.5, 7.7, 2.9, 3.1, 3.3, 3.6, 4.0

○ No outlier ○ One outlier ○ Two outliers

g 2.1, 2.3, 2.2, 2.0, 2.4, 3.5, 2.1, 2.3, 2.1, 2.2

○ No outlier ○ One outlier ○ Two outliers

h $\frac{1}{2}, \frac{3}{4}, \frac{12}{5}, \frac{21}{25}, \frac{2}{5}, \frac{7}{8}, \frac{1}{6}$

○ No outlier ○ One outlier ○ Two outliers

i −1, −0.5, 0.7, 1, 0.2, −0.6, 0.1, 0.9, −0.3

○ No outlier ○ One outlier ○ Two outliers

j 1, 30, 32, 42, 29, 53, 33, 14, 28, 39, 40, 48, 36

○ No outlier ○ One outlier ○ Two outliers

Outliers and measures of central tendency

2 The times taken in minutes to get from Ashville to Blanderton over eleven consecutive days was:

$$23, 25, 28, 29, 27, 125, 32, 24, 28, 24, 20$$

a Which score is an outlier?

Outlier = ⬚

b Calculate the mean score $\bar{x}$.

$\bar{x}$ = ⬚

c Find the median score.

Median = ⬚

d Which is the best representative of the centre of these scores, the mean or the median?

◯ Mean ◯ Median

e Calculate the mean score $\bar{x}$ again, this time ignoring the outlier.

$\bar{x}$ (without outlier) = ⬚

f Is the mean found in part **e** a better representative of the centre of the data?

◯ Yes ◯ No

Outliers and measures of central tendency

3 For each of these data sets:
(i) Find the mode, median and mean scores
(ii) Identify which score/s are the best measure of central tendency for each data set. Explain your reasoning.

a 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 5, 5, 6, 6, 9

Mode =

Median =

$\bar{x}$ =

b 5, 9, 6, 8, 7, 11, 8, 10, 8, 8, 8, 7, 9, 5, 9, 6, 10, 7, 11

Mode =

Median =

$\bar{x}$ =

c 7.0, 1.4, 1.2, 1.4, 1.2, 1.6, 1.5, 1.5, 6.5, 1.5, 1.7, 1.5, 1.6, 1.7, 1.5, 1.0

Mode =

Median =

$\bar{x}$ =

Outliers and measures of central tendency

3 d 16, 16, 9, 19, 16, 15, 7, 16, 4, 16, 16, 16, 16, 16, 16, 16, 9, 16, 16, 16, 16, 16, 12, 16, 16, 16, 16, 16

Mode =

Median =

$\bar{x}$ =

e 0, 7, 1, 6, 2, 7, 9, 3, 8, 4, 5, 5, 1, 6, 0, 2, 8, 4, 9, 3

Mode =

Median =

$\bar{x}$ =

f −5 ,−2, 0, 3, 2, −15, −6, 1, 2, 4, 5, −5, 0, −4, 2, 1, −2, −1, 0, 1

Mode =

Median =

$\bar{x}$ =

Outliers and measures of central tendency

4 The word 'novemnonagintillion' can be used to name a number starting with 1, followed by three hundred zeros!

a Which letter is the modal letter?

Modal letter = ⬚

b If the letters of this word were arranged alphabetically, which letter would be in the middle?

Middle letter = ⬚

c Do both results represent a typical letter used to form the word novemnonagintillion?

◯ Yes ◯ No

d Try steps **a** and **b** again for the word 'circumcentre'.
Use the results to explain why arranging letters into alphabetical order to find the median letter is really not a good way to find the 'typical letter' used in a particular word.

Modal letter = ⬚ Middle letter = ⬚

e The letter 's' in the word 'cheese' could be considered an outlier alphabetically.
Explain why 's' would not affect the measure of central tendency method needed to find the letter typically used in that word.

Outliers and measures of central tendency

5 Earn yourself an awesome passport stamp with this question.

The number of correct answers achieved in 10 games of live Mathletics for one student are:

$$15,\ 24,\ 18,\ 16,\ 20,\ 23,\ 11,\ 18,\ 21,\ 24.$$

a If the student played an 11^{th} game, what exact number of correct answers in that game will ensure the median score remains the same for the new data set?

b What score achieved in the 11^{th} game would increase the mean score of the new data set by 1?

c Is the score required for the 11^{th} game in **b** an outlier? Explain why.

○ Yes ○ No

Comparing two data sets

Comparing data sets using the range, mean, mode and median scores, depends on the nature and context of the data. There are many ways in which two data sets can be different.

Examples of comparing data sets.

(i) This table shows the number of kilometres two cross-country skiers travelled over 6 weeks.

Callum	12	20	25	25	26	31
Omar	10	15	25	25	41	42

α) Find the means and medians for these data sets:

	Mean	Median	Range
Callum:	23.2 km	25 km	19 km
Omar:	26.3 km	25 km	32 km

β) By just comparing the means, who skied further?

Omar's weekend average is higher.

γ) By just comparing the medians, who skied further?

The medians are equal, so neither if just comparing using the median.

δ) Find the total kilometres skied by Callum, and by Omar.

Callum $= 12 + 20 + 25 + 25 + 26 + 31 = 139$ km,
Omar $= 10 + 15 + 25 + 25 + 41 + 42 = 153$ km

ε) Which measure, the mean or the median, best reflects/compares each skiers performance?

Overall, the total kilometres cycled for Omar was greater than that for Callum, and so the mean reflects this as it is higher for Omar than for Callum, whereas the two medians are equal. So the mean was the best method for comparison here.

(ii) Two students have the following marks for their subjects at the end of a school term:

Student A:	5%	52%	65%	70%	74%	79%	85%	90%
Student B:	50%	58%	66%	68%	74%	78%	85%	85%

α) Find the range, median and mean for each students results and summarize in a table (to 1 d.p.).

	Mean	Median	Range
Student A:	85%	72%	65%
Student B:	35%	71%	70.5%

β) Which student has the highest range, and what does this show?

Student A has the highest range and this shows there are at least two marks that differ by a large amount. The outlier (mark of 5% here) has resulted in a much larger range for student A.

γ) Compare and discuss the mean and median scores for the two data sets.

- The means differ by a large amount due to student A's outlier of 5%.
- The medians are similar. So if the outliers are removed, the centre of the data for both student A and B are about the same.

∴ The median score is a better comparison since most of the scores (except for the one low outlier for student A) are similar.

Comparing two data sets

1 The table shows the number of personal basketball scores contributed in each game over a series of games for two players:

Player A	5	13	18	18	19	20
Player B	3	8	18	18	34	35

a Find the means and medians for these data sets and complete the table:

	Mean	Median	Range
Player A:			
Player B:			

b By just comparing the means which player did better?

c By just comparing the medians which player did better?

d Find the total number of scores contributed by each player over the series of games.

e Which measure, the mean or the median, better reflects their overall performance in comparison to each other?

Comparing two data sets

2 The marks for a test in mathematics from two classes are as follows:

Maths A (%):
15, 35, 40, 45, 50, 52, 55, 58, 57, 57, 59, 59, 60, 60, 61, 65, 68, 68, 68, 75, 75, 78, 81, 83, 99

Maths B (%):
15, 30, 35, 48, 53, 54, 55, 55, 60, 57, 58, 60, 65, 65, 66, 66, 67, 67, 67, 70, 73, 74, 81, 83, 99

Fill in the details of each class for these statistical measures:

	Range	Mode	Mean	Median
Maths A:				
Maths B:				

a If the pass mark is 50% or more, how many students were below this in each class?

Maths A = ☐ **Maths B** = ☐

b Are there any outliers?

Maths A = ○ Yes ○ No **Maths B** = ○ Yes ○ No

c How many students achieved 75% or above in each class?

Maths A = ☐ **Maths B** = ☐

d If you just compared the averages of the two classes, which one would you say did better?

e If you just compared the medians of the two classes, which one would you say did better?

f Which of the two classes do you think performed better overall? Explain.

Investigating secondary data

Renewable energy: Energy obtained from resources that do not reduce over time.

Eg: water, wind, the sun, geothermal sources, and biomass sources such as energy crops.

Non-renewable energy: Energy obtained from finite resources that do not renew fast enough for continued use. One day these resources will run out.

Eg: Crude oil/petroleum, coal, natural gas, uranium/nuclear energy, aluminium, copper, silver, gold, and diamonds.

Comment on the data in the table relating to the distribution and use of non-renewable resources

The table shows crude oil production in bbl/day (barrels per day) for several countries.

The web link where this data was collected from is:
http://*www.en.wikipedia.org/wiki/List_of_countries_by_oil_production*

	Country	Population (bbl./day)	Date of Information
	World	84 951 000	2014 est.
	Saudi Arabia	9 693 200	2013 est.
	Russia	10 950 000	2014 est.
	United States	9 020 000	2015 est.
	Iran	3 518 000	2014 est.
	China	4 372 000	2014 est.
	Mexico	2 934 000	2013 est.
	South Africa	191 000	2009

So what can we say or interpret from this data?

- The data compares crude oil production (in barrels per day) for selected countries.
- This table was compiled by someone else and posted onto a website, so the data is secondary.
- Because all the values in the second column end in '000' they are only accurate to the nearest thousand.
- Some years have the label 'est' for the word estimate. So that data is not 100% accurate.
- The data was collected at different times. This makes it harder to compare the countries accurately.

Example comparisons:

- Iran and Chinas estimated values are very close (both around 4 000 000 bbl/day). So we cannot safely say one produces more than the other.
- Saudi Arabia, Russia and the United States all produce more than the combined production of crude oil by Iran and China.

Investigating secondary data

1 This table shows the electricity consumption (in billions of Kilo Watt Hours or kWh) for the world and several countries.

Rank	Country	kWh	Date of Information
1	World	19 540 000 000 000	2009 est.
2	China	4 693 000 000 000	2011
3	United States	3 889 000 000 000	2010 est.
4	European Union	3 037 000 000 000	2009 est.
5	Russia	1 038 000 000 000	2012 est.
6	Japan	859 700 000 000	2012 est.
7	India	637 600 000 000	2009 est.
8	Germany	549 100 000 000	2010 est.
9	Canada	504 800 000 000	2009 est.
10	Korea, South	455 100 000 000	2011 est.
11	France	451 400 000 000	2009 est.
12	Brazil	438 300 000 000	2010 est.
13	United Kingdom	325 800 000 000	2009 est.
14	Italy	313 800 000 000	2011 est.

(i) What nearest kWh have these values been rounded to?
[Hint: use the minimum number of zero's for each value in column 3 to help] _____ kWh

(ii) How does the information in column 4 affect your ability to use the data for comparison?

(iii) Can you confidently compare the United Kingdom and Italy from this table? Explain.

(iv) Can you confidently compare the United Kingdom and the United States from this table? Explain.

(v) Would it be possible to accurately or approximately predict the kWh for Canada in 2012 from this table?

○ Accurately ○ Approximately

Why?

(vi) Which countries in this table can you confidently say consume more electricity than Russia?

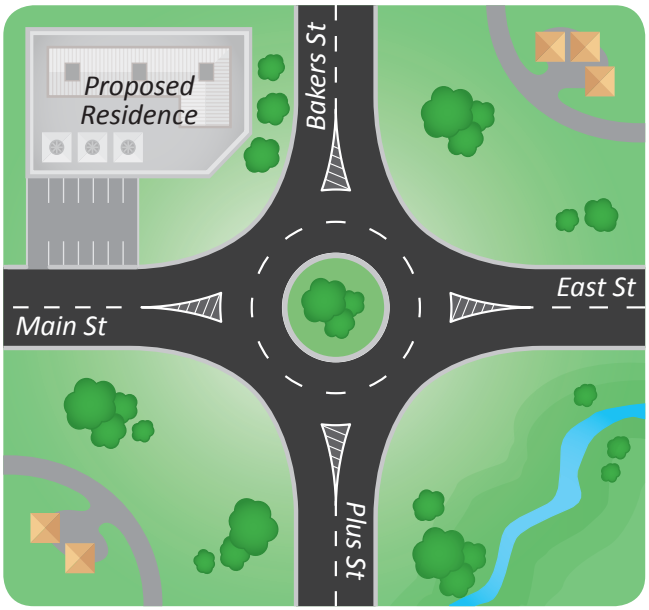
Investigating secondary data

Traffic Volume

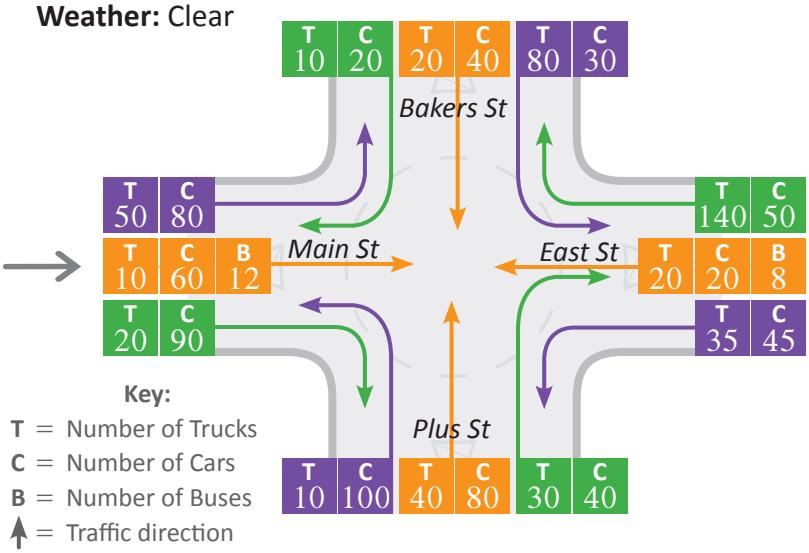
The impact from a lot of people moving into an area on both road and pedestrian traffic is an important planning consideration. It could mean new traffic control measures are required.

Interpreting collected data correctly can help reduce problems occurring later on.

2 Secondary traffic data (for an intersection where a new residential apartment has been proposed), is below.



Traffic survey time: 8:00 am – 10:00 am Monday
Weather: Clear

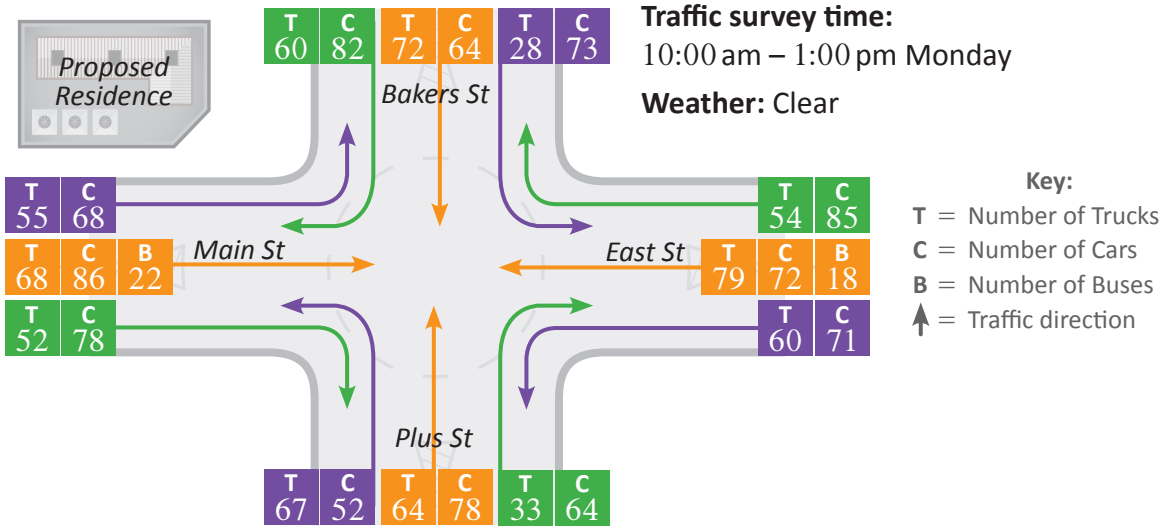


(i) Which street has the modal number of vehicles *entering* the intersection for the time period surveyed?

(ii) Which street has the modal number of vehicles *leaving* the intersection for the time period surveyed?

(iii) Cars *entering/leaving* driveways on busy roads can slow the traffic and cause delays. From the information shown above for this time period, would you recommend the proposed driveway into the residence on Main St be changed to Bakers St? Explain you reasoning.

Investigating secondary data



These survey results for the same intersection were taken over a 3 hour period at a different time on the same day.

(iv) Calculate the hourly average of trucks entering the intersection over the 3 hour period.

(v) If the number of trucks and buses entering an intersection is ever greater than 240 vehicles/hour, the surface of the road needs to be improved.

α) Calculate the hourly average of buses and trucks entering the intersection from each survey.

β) Use your answers to comment on whether or not the road surface needs to be improved.

Investigating secondary data

(vi) If the average number of vehicles entering the intersection between $8{:}00$ am and $1{:}00$ pm is higher than 550 vehicles/hour, the road planners will install traffic lights.

α) Calculate the hourly average of vehicles entering the intersection during this time period.

β) Does this rate exceed the average vehicle rate per hour required for traffic lights?

γ) Use your answers to comment on whether or not traffic lights should be installed and why.

Here is a summary of the important things to remember for data for statistics

Primary data
Data which is collected for the first time and has no previous source available.

The Range
Highest score minus the lowest score.

The Mode
Is the most frequently occurring score.

Bimodal Data
Data set has two modes.

Multimodal
Data set that has more than two modes.

Median Score
- The middle score of a data set where the scores have been arranged in increasing order.
- If there are two middle scores then you add them together and divide by 2.

Mean Score
The average score in a data set. Add up all the scores and divide by the total number of scores.

$$\text{Mean (or average)} \quad \bar{x} = \frac{\text{Add all the scores together}}{\text{Total number of scores}}$$

Measures of central tendency
These are the median, mode, mean.

Outliers
Scores which appear to be outside the main body of scores, so either extremely low or extremely high scores.

Primary and secondary data

1. a Primary b Primary c Secondary
d Primary e Secondary f Secondary
g Primary h Primary i Secondary
j Primary k Primary l Secondary
m Secondary

2. Primary:

a Create a survey sheet for students to complete.

b Walking around the playground during lunch and asking students questions.

Basically anything that involves them gathering the information directly from the students themselves. ie interviews, surveys, observations, etc.

Secondary:

a Reading a survey of the students conducted the previous year.

b Asking parents what they know about their childs preferred items from the list are.

Basically Anything that involves them sourcing information from somewhere else, and not asking the students directly.

Range

1. a (i) (11) (1) (ii) (25.0) (22.9)
(iii) (−6°C) (0°C) (iv) (26 mm) (1.8 cm)

b (i) $11 - 1 = 10$

(ii) $25.0 - 22.9 = 2.1$

(iii) $0°C - (-6°C) = 6\ °C$

(iv) $26 - 18 = 8$ mm

2. a (i) −4°C −3°C −1°C 0°C 1°C
2°C 3°C 5°C 9°C

(ii) $9 - (-4) = 13°C$

Range

2. b (i) 10 min 12 min 15 min 18 min
20 min 30 min

(ii) $30 - 10 = 20$ min

c (i) 100 cm 150 cm 180 cm 210 cm 275 cm

(ii) $275 - 100 = 175$ cm

3. a (i) 1 and 6 (ii) $6 - 1 = 5$

b (i) 1 and 7 (ii) $7 - 1 = 6$

c (i) 6 and 30 (ii) $30 - 6 = 24$

Mode

1. a (i) Top = 10 Front = 7 Side = 5
Modal score view = TOP

(ii) Top = 7 Front = 7 Side = 8
Modal score view = SIDE

(iii) Top = 9 Front = 6 Side = 6
Modal score view = TOP

b (i) Top = 9 Front = 9 Side = 9
No mode

(ii) Top = 7 Front = 7 Side = 8
One mode

(iii) Top = 7 Front = 8 Side = 8
Bimodal

2. a

Score	Tally	Frequency
0	𝍸	5
1	𝍸 ||||	9
2	𝍸 ||	7
3	𝍸 |||	8
4	||	2

b Modal score = 1

Mode

3. (a) Mode = 8 — One mode
(b) Mode = n/a — No mode
(c) Mode = 7, 15 — Bimodal
(d) Mode = 101 — One mode
(e) Mode = 1.2, 2.4, 2.6 — Multimodal

4.

Data Set	Range	Mode
1, 2, 2, 3, 4	3	2
5, 5, 6, 7, **8**	3	5
4, 6, **6**, 9, **27**	23	6
0, 2, 2, 3, **6**, **6**, **6**, 9, 12, 12, 13, 24, 30	30	6
1, 1, 3, 3, 7, 9, 9, **9**, 16, **20**	19	9
12, 13, 17, 17, **18**, **19**, **19**, 19, 22, 22, 26	14	19

5. (a) (i) 4 (ii) 12.6
(b) (i) 9 (ii) 23.5
(c) (i) 0.3, 0.4 or 0.5 (ii) −8, 0, 2 or 3

Mean

1. (a) 5 (b) 5 (c) 5.5
(d) 6.3 (e) $\frac{1}{2}$ (f) $\frac{5}{8}$
(g) 2.3

2. 29 **3.** 7 **4.** 70

5. 10.5 **6.** 1 **7.** 7

8. 18

Median

1. (a) 2 (b) 4 (c) 7 (d) 7.5
(e) 15 (f) 10 (g) 6.5

Median

2. $x = 8$

3.

M	I	D	D	L	E	S	C	O	R	E
(i)	(ii)	(iii)	(iv)	(v)	(vi)	(vii)	(viii)	(ix)	(x)	(xi)

MEDIAN

Outliers and measures of central tendency

1. (a) One outlier (b) No outlier (c) One outlier
(d) Two outliers (e) No outlier (f) One outlier
(g) One outlier (h) One outlier (i) No outlier
(j) Two outliers

2. (a) 125 (b) 35 (c) 27
(d) Median (e) 26 (f) Yes

3. (a) (i) Mode = 2 Median = 2 $\bar{x} = 1.5$
(ii) Mode and median are the best measures because there are outliers in the data.
(b) (i) Mode = 8 Median = 8 $\bar{x} = 8$
(ii) All scores are good measures because the data is evenly spread (or symmetrical).
(c) (i) Mode = 1.5 Median = 1.5 $\bar{x} = 2.1125$
(ii) Mode and median are the best measures because there are two outliers in the data.
(d) (i) Mode = 16 Median = 16 $\bar{x} = 14.68$
(ii) Mode and median are the best measures because there are two outliers in the data.
(e) (i) No Mode Median = 4.5 $\bar{x} = 4.5$
(ii) Mean and median are the best measures because the data is evenly spread and there is no mode.
(f) (i) Multi Model Median = 0 $\bar{x} = -1$
(ii) Mode and median are the best measures because there is an outlier in the data and the median score is equal to one of the modal score.

Outliers and measures of central tendency

4. (a) Modal letter = **n** (b) Middle letter = **n**

(c) Yes

(d) Modal letter = **c**
Middle letter = Between '**e**' and '**i**'

For words with an even number of letters, finding the average between two different middle letters does not make sense since. It implies that it is a letter that is not actually in the word itself.

(e) For problems like this, the mode is the only useful measure of central tendency to find the typical letter used in a word and mode is not affected by outliers.

5. (a) 19 (b) 30

(c) No

The new mean score of 20 is equal to the new median score of 20, so the score has not changed the even spread (or symmetry) of the data. Also the smallest and new highest score are roughly the same distance away from the median/mean scores.

Comparing two data sets

1. (a)

	Mean	Median	Range
Player A:	15.5	18	15
Player B:	$19\frac{1}{3}$	18	33

(b) Player B. (c) Both players are equal.

(d) **Player A**: 93 **Player B**: 116

(e) The mean score reflects their overall performance in comparison as it indicates that **Player B** generally contributed more points over 6 games. The median only shows that each players middle score was the same.

Comparing two data sets

2.

	Range	Mode	Mean	Median
Maths A:	84	68	61	60
Maths B:	84	67	61	65

(a) **Maths A** = 4 **Maths B** = 4

(b) **Maths A** = No **Maths B** = No

(c) **Maths A** = 6 **Maths B** = 3

(d) Both classes have the same average, so by comparing average score, they performed equalily.

(e) The median score for class **B** is higher, so using this measure to compare, you would say they did better.

(e) The classes performed similarly, the averages were the same, the ranges were the same, but class B had a slightly higher median. This indicates that class **B** did a better, since 50% of students got 65% or more whereas in class **A**, 50% of students got 60 or more.

Investigating secondary data

1. (i) 100 million kWh

(ii) The data is from different years, so comparing might usage might not be accurate due to different conditions changing each year.

(iii) Since the data is labelled as an estimate only, the values are close yet collected 2 years apart from each other. It would be difficult to confidently compare usage.

(iv) Despite the data being collected 1 year apart and estimates, the usage in the US is far greater than in the UK, so the usage can still be compared confidently.

(v) Approximately

The data is from 2009, so the usage could have changed a great deal in three years due to population, changing weather, etc.

(vi) China and the United States. The European Union uses more, but this represents a large group of countries all together.

Investigating secondary data

2. (i) Modal street for vehicles *entering* the intersection is Main St.

(ii) Modal street for vehicles *leaving* the intersection is Baker St.

(iii) No, the proposed driveway should remain on Main St.

Keeping the driveway on Main St should have less impact on the traffic.

(iv) Hourly average $= 692 \div 3 = 230.7$ (to 1 d.p.)

(v) α) Hourly average $= 732 \div 3 = 244$

β) The road surface does need to be improved as there are more than 240 trucks + Buses per hour.

(vi) α) Hourly average $= 2745 \div 5 = 549$/hour

β) No this does not currently exceed the average vehicle rate needed for lights.

γ) Yes traffic lights should be installed. With the proposed new development, in increase in traffic will occur and there will be more than 550 vehicles/hour using the intersection.

www.mathletics.com