AQA GCSE Maths: Higher

Your notes

Averages, Ranges & Data

Contents

Mean, Median & Mode

Your notes

Mean, Median & Mode

What is the mode?

- The **mode** is the value that appears the **most often**
  - The mode of 1, 2, 2, 5, 6 is 2
- There can be **more than one** mode
  - The modes of 1, 2, 2, 5, 5, 6 are 2 and 5
- The mode can also be called the **modal value**

What is the median?

- The **median** is the **middle** value when you put values **in size order**
  - The median of 4, 2, 3 can be found by
    - ordering the numbers: 2, 3, 4
    - and choosing the middle value, 3
- If you have an **even** number of values, find the **midpoint** of the **middle two** values
  - The median of 1, 2, 3, 4 is 2.5
    - 2.5 is the midpoint of 2 and 3
  - The **midpoint** is the **sum** of the **two middle** values **divided by 2**

What is the mean?

- The **mean** is the **sum** of the values **divided** by the **number** of values
  - The mean of 1, 2, 6 is $(1 + 2 + 6) \div 3 = 3$
- The mean can be **fraction** or a **decimal**
  - It may need **rounding**
  - You do **not** need to force it to be a **whole** number
    - You **can** have a mean of 7.5 people, for example!

How do I know when to use the mode, median or mean?

Your notes

- The mode, median and mean are different ways to measure an **average**
- In certain situations it is **better** to use one average over another
- For example:
  - If the data has **extreme values** (outliers) like 1, 1, 4, 50
    The mode is 1
    The median is 2.5
    The mean is 14
    - **Don't use the mean** (it's badly affected by extreme values)
  - If the data has **more** than one mode
    - **Don't use the mode** as it is not clear
  - If the data is **non-numerical**, like dog, cat, cat, fish
    - You can **only** use the **mode**

Worked Example

15 students were timed to see how long it took them to solve a mathematical problem. Their times, in seconds, are given below.

12	10	15	14	17
11	12	13	9	21
14	20	19	16	23

(a) Find the mean time, giving your answer to 3 significant figures.

Add up all the numbers (you can add the rows if it helps)

$$12 + 10 + 15 + 14 + 17 = 68$$
$$11 + 12 + 13 + 9 + 21 = 66$$
$$14 + 20 + 19 + 16 + 23 = 92$$

$$\text{Total} = 68 + 66 + 92 = 226$$

Divide the total by the number of values (there are 15 values)

$$\frac{226}{15} = 15.066\,666\ \ldots$$

Your notes

Write the mean to 3 significant figures
Remember to include the units

The mean time is 15.1 seconds (to 3 s.f.)

(b) Find the median time.

Write the times in order and find the middle value

$$\cancel{9} \quad \cancel{10} \quad \cancel{11} \quad \cancel{12} \quad \cancel{12} \quad \cancel{13} \quad \cancel{14} \quad 14 \quad \cancel{15} \quad \cancel{16} \quad \cancel{17} \quad \cancel{19} \quad \cancel{20} \quad \cancel{21} \quad \cancel{23}$$

The median time is 14 seconds

(c) Explain why the median is a better measure of average time than the mode.

Try to find the mode (the number that occurs the most)

There are two modes: 12 and 14

Explain why the median is better

There is no clear mode (there are two modes, 12 and 14), so the median is better

(d) If a 16th student has a time of 95 seconds, explain why the median of all 16 students would be a better measure of average time than the mean.

The16th value of 95 is extreme (very high) compared to the other values
Means are affected by extreme values

The mean will be affected by the extreme value of 95 whereas the median will not

Your notes

Calculations with the Mean

How do I solve harder problems involving the mean?

- Remember what the mean is
  - **Mean = total of values ÷ number of values**
    - It is a **formula** involving three quantities
    - if you know any two, you can find the other one
- A question may require you to work **backwards** from a **known mean**
  - It helps to rearrange the formula
  - **Total of values = mean × number of values**
- Find the **total** of the values **before and after** to help with question that involve:
  - **missing** values
  - **adding in**, or **taking out**, a value

Examiner Tips and Tricks

- It helps to start thinking of the mean as a formula which you can rearrange
  - Total of values = mean × number of values

Worked Example

A class of 24 students has a mean height of 1.56 metres.
A new student joins the class.
The mean height of the class is now 1.57 metres.

Find the height of the new student.

Rearrange the formula for mean to get 'total of heights = mean height × number of students'
Find the total of heights before

Your notes

$$\text{Total of heights before} = 1.56 \times 24 = 37.44$$

Find the total of heights after
Remember there are now 25 students

$$\text{Total of heights after} = 1.57 \times 25 = 39.25$$

The height of the new student is the difference of the two totals above

$$39.25 - 37.44 = 1.81$$

The height of the new student is 1.81 metres

Averages from Tables

Your notes

Averages from Tables & Charts

What are frequency tables?

- **Frequency tables** are used to **summarise** data in a **neat** format
  - They also put the data in **order**
- For example, the table below shows the number of pets in different houses along a street
  - The number of pets is the **data value**, x
  - The number of houses is the **frequency**, f
    - The **frequency** is **how many times** a data value is **recorded** (or **seen**)
- The **total frequency**, n, can be calculated by **adding** together all the values in the **frequency column**

Number of pets (data *value*, x)	**Number of houses (*frequency*, f)**
0	2
1	7
2	6
3	4
4	1
	Total frequency (n) = 20

How do I find the mode from a frequency table?

- The **mode** is the **data value** with the **highest frequency**
  - The mode for the example above is 1 pet per house
    - The mode is **not** the frequency, 7, this is the number of houses that have exactly 1 pet

Your notes

How do I find the median from a frequency table?

- The **median** is the **data value** in the **middle** of the **frequency**
  - It is the $\left(\frac{n+1}{2}\right)^{th}$ value, where n is the total frequency
- From above, $n = 20$ so the median is the $\left(\frac{20+1}{2}\right)^{th}$ = 10.5^{th} value in the table
  - The first two rows have a **combined** (**cumulative**) frequency of 2 + 7 = 9
  - The first three rows have a combined frequency of 2 + 7 + 6 = 15
  - Therefore the 10th and 11th values are in the third row (x = 2)
    - The median is 2 pets per house

How do I find the mean from a frequency table?

- The **mean** from a **frequency table** has the following **formula**:
  - $\text{mean} = \frac{\text{total of 'data value} \times \text{frequency'}}{\text{total frequency}}$
    - It helps to create a new column of 'data value × frequency'
    - Add up the values in this column
    - Divide by the total frequency
- The mean is $\frac{35}{20}$ = 1.75 pets per house
  - Means do not need to be whole numbers

Number of pets (data *value*, *x*)	Number of houses (frequency, *f*)	data value × frequency (*xf*)
0	2	0 × 2 = 0
1	7	1 × 7 = 7
2	6	2 × 6 = 12

3	4	3 × 4 = 12
4	1	4 × 1 = 4
	Total = 20	Total = 35

Your notes

How do I find the range from frequency tables?

- The **range** is the **difference** of the largest and smallest **data values**
  - The range above is 4 - 0 = 4
    - The range is **not** the difference of the largest and smallest frequencies

What else should I know about frequency tables?

- **Tables** can be **converted** back into a **list of data values** using their **frequencies**
  - From above, 0 pets were recorded twice, 1 pet was recorded 7 times, 2 pets were recorded 6 times, etc
    - The list of pets recorded is 0,0,1,1,1,1,1,1,1,2,2,2,2,2,2,3,3,3,3,4
- You could then find the **mode**, **median** and **mean** from this list of numbers

Worked Example

The table shows data for the shoe sizes of pupils in class 11A.

Shoe size	Frequency
6	1
6.5	1
7	3
7.5	2
8	4
9	6

10	11
11	2
12	1

(a) Find the mean shoe size for the class, giving your answer to 3 significant figures.

It helps to label shoe size (x) and frequency (f)
Add an extra column and calculate the values of 'shoe size × frequency', (xf)
Find the total frequency and total xf value

Shoe size (x)	**Frequency (f)**	xf
6	1	$6 \times 1 = 6$
6.5	1	$6.5 \times 1 = 6.5$
7	3	$7 \times 3 = 21$
7.5	2	$7.5 \times 2 = 15$
8	4	$8 \times 4 = 32$
9	6	$9 \times 6 = 54$
10	11	$10 \times 11 = 110$
11	2	$11 \times 2 = 22$
12	1	$12 \times 1 = 12$
	Total = 31	Total = 278.5

Use the formula that the mean is the total of the xf column divided by the total frequency

$$\text{Mean} = \frac{278.5}{31} = 8.983\ 870\ \ldots$$

Give your final answer to 3 significant figures

The mean shoe size is 8.98 (to 3 s.f.)

Note that the mean does not have to be an actual shoe size

(b) Find the median shoe size.

Your notes

Your notes

The median is the $\left(\frac{n+1}{2}\right)^{th}$ value where n is the total frequency

$$\frac{n+1}{2}=\frac{31+1}{2}=\frac{32}{2}=16$$

The median is the 16$^{\text{th}}$ value
There are 1 + 1 + 3 + 2 + 4 = 11 values in the first five rows of the table
There are 11 + 6 = 17 values in the first six rows of the table
Therefore the 16$^{\text{th}}$ value must be in the sixth row

The median shoe size is 9

(c) Find the range of the shoe sizes.

The range is the highest shoe size subtract the lowest show size

12 - 6

The range of the shoe sizes is 6

Your notes

Averages from Grouped Data

Averages from Grouped Data

What is grouped data?

- Data can be collected into **groups** or **class intervals**
  - It is useful for organising data if you have a lot of individual data points
  - You can present grouped data in a **grouped frequency table**
- Grouped data may be **discrete** or **continuous**
  - Discrete data is numerical data that can only take on specific values, it needs to be counted
    - E.g. Shoe size
  - Continuous data can take **any value** within a range of infinite values, it needs to be **measured**
    - E.g. Length of a foot in cm

Why do I find an estimate for the mean from grouped data?

- It is impossible to find the **mean** for grouped data, because we don't have access to the original data values
  - i.e. there is no way to find the exact sum of all the data values
  - so we can't use the formula, $\text{mean} = \frac{\text{sum of values}}{\text{number of values}}$
- However we can **estimate** the mean for grouped data
  - To do this we use the class **midpoints** as our data values
    - e.g. if a class interval is $150 \leq x < 160$
    - we assume that all the data values are equal to the midpoint, 155

Examiner Tips and Tricks

- When presented with data in a table it may not be obvious whether the data is grouped or not

- When you see the phrase "**estimate** the mean" you know that you are in the world of grouped data!

Your notes

How do I find an estimate for the mean from grouped data?

- To find an estimate for the mean from grouped data, complete the following steps:
- **STEP 1**
  Draw an extra **two** columns on the end of a table of the grouped data
  - In the first new column write down the **midpoint** of each **class interval**
  - If the midpoint isn't obvious, add the endpoints and divide by 2
    - e.g. if a class interval is $150 \le x < 160$
    - the midpoint is $\frac{150+160}{2} = \frac{310}{2} = 155$
- **STEP 2**
  Calculate **"frequency" × "midpoint"** (this is often called *fx*)
  - Write these values in the second column you added to the table
- **STEP 3**
  Find the **total** for the ***fx* column**
  - If the question does not tell you the total number of data values (i.e. the total frequency), find the total of the frequency column also
- **STEP 4**
  Estimate the **mean** by using the **formula**
  - $\text{estimated mean} = \frac{\text{total of (midpoints} \times \text{frequencies)}}{\text{total frequency}}$
  - i.e. **divide** the total of the *fx* column by the total number of data values

How do I find the modal class?

- For grouped data we talk about the **modal class** instead of the mode
  - This is the class with the **highest frequency**
- Find the highest frequency in the table
  - The **corresponding class interval** tells you the modal class

How do I find the class interval that the median lies in?

Your notes

- Find the **position** of the median using $\frac{n+1}{2}$, where n is the number of data values (total of the frequency column)
- Use the table to deduce the **class interval** containing the $\left(\frac{n+1}{2}\right)^{\text{th}}$ value
  - e.g. if the median is the 7$^{\text{th}}$ value and the frequency of the first two class intervals are 4 and 7
    - the median will lie in the second class interval of the table
- Note that rather than 'the median' we refer to the '**class interval containing the median**'

Examiner Tips and Tricks

- Be careful not to confuse the **modal class** with its frequency
  - e.g. if the highest frequency in the table is 34, corresponding to the class interval $40 \le x < 50$
  - then the modal class is $40 \le x < 50$, not '34'!
  - This also applies to the **interval containing the median**

Worked Example

The weights of 20 three-week-old Labrador puppies were recorded at a vet's clinic. The results are shown in the table below.

Weight, w kg	Frequency
$3 \le w < 3.5$	2
$3.5 \le w < 4$	4
$4 \le w < 4.5$	6
$4.5 \le w < 5$	5
$5 \le w < 5.5$	2

$5.5 \le w < 6$	1

(a) Estimate the mean weight of these puppies.

First add two columns to the table
Complete the first new column with the midpoints of the class intervals
Complete the second extra column by calculating "*fx*"
A total row is also useful

Weight, *w* kg	Frequency	Midpoint	"*fx*"
$3 \le w < 3.5$	2	**3.25**	**2 × 3.25 = 6.5**
$3.5 \le w < 4$	4	**3.75**	**4 × 3.75 = 15**
$4 \le w < 4.5$	6	**4.25**	**6 × 4.25 = 25.5**
$4.5 \le w < 5$	5	**4.75**	**5 × 4.75 = 23.75**
$5 \le w < 5.5$	2	**5.25**	**2 × 5.5 = 10.5**
$5.5 \le w < 6$	1	**5.75**	**1 × 5.75 = 5.75**
Total	**20**		**87**

Now we can find the mean using

$$\text{estimated mean} = \frac{\text{total of (midpoints } \times \text{ frequencies)}}{\text{total frequency}}$$

$$\text{estimated mean} = \frac{87}{20} = 4.35$$

4.35 kg

(b) Write down the modal class.

The highest frequency in the table is 6
This corresponds to the interval $4 \le w < 4.5$

$4 \le w < 4.5$

A common error here would be to write down 6 (the frequency) as the modal class

(c) Find the interval that contains the median.

Your notes

There are 20 dogs
The median interval will be the interval containing the 10.5^{th} dog
Keep a running total

Weight, w kg	Frequency	Running Total
$3 \leq w < 3.5$	3	3
$3.5 \leq w < 4$	4	$3 + 4 = 7$
$4 \leq w < 4.5$	6	$7 + 6 = 13$
$4.5 \leq w < 5$	5	$13 + 5 = 18$
$5 \leq w < 6$	2	$18 + 2 = 20$

The 10.5^{th} dog is in the third interval

The median is in the interval $\mathbf{4 \leq w < 4.5}$

Your notes

Range & Interquartile Range

Your notes

Range & IQR

What is the range?

- The **range** is the **difference** between the **highest value** and the **lowest value**
  - range = highest - lowest
    - For example, the range of 1, 2, 5, 8 is 8 - 1 = 7
- It measures how **spread out** the data is
  - Ranges of different data sets can be **compared** to see which is more spread out
  - The range of a data set **can be affected** by very large or small values
- Be careful with **negatives**
  - The range of -2, -1, 0, 4 is 4 - (-2) = 6

How do I know when to use the range?

- The **range** is a simple measure of how **spread out** the data is
  - The range does **not** measure an average value
- It should **not** be used if there are any **extreme values** (outliers)
  - For example, the range of 1, 2, 5, 80 is 80 - 1 = 79
    - This is not a good measure of spread
    - The range is affected by extreme values

What are quartiles?

- The **median** splits the data set into **two parts**
  - Half the data is less than the median
  - Half the data is greater than the median
- **Quartiles** split the data set into **four parts**
  - The **lower quartile (LQ)** lies a **quarter** of the way along the data (when in order)
    - One quarter (25%) of the data is less than the LQ

Your notes

        - Three quarters (75%) of the data is greater than the LQ
    - The **upper quartile (UQ)** lies **three quarters** of the way along the data (when in order)
        - Three quarters (75%) of the data is less than the UQ
        - One quarter (25%) of the data is greater than the UQ
    - You may come across the median being referred to as the second quartile

How do I find the quartiles?

- Make sure the data is written **in numerical order**
- Use the **median** to divide the data set into **lower and upper halves**
    - If there are an **even** number of data values, then
        - the **first half** of those values are the lower half,
        - and the **second half** are the upper half
        - **All** of the data values are included in one or other of the two halves
    - If there are an **odd** number of data values, then
        - all the values **below** the median are the lower half
        - and all the values **above** the median are the upper half
        - The **median** itself is **not included** as a part of either half
- The **lower quartile** is the **median of the lower half** of the data set
    - and the **upper quartile** is the **median of the upper half** of the data set
- Find the quartiles in the **same way** you would usually find the median
    - just **restrict** your attention to the relevant half of the data

What is the interquartile range (IQR)?

- The **interquartile range (IQR)** is the **difference** between the **upper quartile (UQ)** and the **lower quartile (LQ)**
    - Interquartile range (IQR) = upper quartile (UQ) - lower quartile (LQ)
- The IQR measures how **spread out** the **middle 50%** of the data is
    - The IQR is **not affected by extreme values** in the data

Examiner Tips and Tricks

- If asked to find the range in an exam, make sure you show your subtraction clearly (don't just write down the answer)

Your notes

Worked Example

Find the range of the data in the table below.

3.4	4.2	2.8	3.6	9.2	3.1	2.9	3.4	3.2
3.5	3.7	3.6	3.2	3.1	2.9	4.1	3.6	3.8
3.4	3.2	4.0	3.7	3.6	2.8	3.9	3.1	3.0

Range = highest value - lowest value

9.2 - 2.8

The range is 6.4

Worked Example

A naturalist studying crocodiles has recorded the numbers of eggs found in a random selection of 20 crocodile nests

31 32 35 35 36 37 39 40 42 45

46 48 49 50 51 51 53 54 57 60

Find the lower and upper quartiles for this data set.

There are 20 data values (an even number)
So the lower half will be the first 10 values
The lower quartile is the median of that lower half of the data

~~31~~ ~~32~~ ~~35~~ ~~35~~ 36 37 ~~39~~ ~~40~~ ~~42~~ ~~45~~

So the lower quartile is midway between 36 and 37 (i.e. 36.5)

Do the same thing with the upper half of the data to find the upper quartile
The upper quartile is the median of the upper half of the data

~~46~~ ~~48~~ ~~49~~ ~~50~~ 51 51 ~~53~~ ~~54~~ ~~57~~ ~~60~~

So the upper quartile is midway between 51 and 51 (i.e. 51)

Lower quartile = 36.5
Upper quartile = 51

Your notes

Comparing Data Sets

Your notes

Comparing Distributions

How do I compare two data sets?

- You may be given **two** sets of data that relate to a context
- To **compare** data sets, you need to
  - compare their **averages**
    - Mode, median or mean
  - compare their **spreads**
    - Range

How do I write a conclusion when comparing two data sets?

- When comparing **averages** and **spreads**, you need to
  - compare **numbers**
  - **describe** what this means in **real life**
- **Copy** the **exact** wording from the question in your answer
- There should be **four** parts to your conclusion
  - For example:
    - "The **median** score of class A (45) is **higher** than the median score of class B (32)."
    - "This means class A **performed better** than class B in the test."
    - "The **range** of class A (5) is lower than the range of class B (12)."
    - "This means the scores in class A were **less spread out** than scores in class B."
  - Other good phrases for lower ranges include:
    - "scores are **closer together**"
    - "scores are **more consistent**"
    - there is **less variation** in the scores"

What restrictions are there when drawing conclusions?

- The data set may be **too small** to be truly representative
  - Measuring the heights of only 5 pupils in a whole school is not enough to talk about averages and spreads
- The data set may be **biased**
  - Measuring the heights of just the older year groups in a school will make the average appear too high
- The conclusions might be influenced by **who** is presenting them
  - A politician might choose to compare a different type of average if it helps to strengthen their argument!

What else could I be asked?

- You may need to **choose** which, out of mode, median and mean, to compare
  - Check for **extreme values** (outliers) in the data
    - Avoid using the **mean** as it is affected by extreme values
- You may need to think from the **point of view** of another person
  - A teacher might not want a large spread of marks
    - It might show that they haven't taught the topic very well!
  - An examiner might want a large spread of marks
    - It makes it clearer when assigning grade boundaries, A, B, C, D, E, ...

Examiner Tips and Tricks

When comparing data sets in the exam, half the marks are for comparing the numbers and the other half are for saying what this means in real life.

Worked Example

Julie collects data showing the distances travelled by snails and slugs during a ten-minute interval. She records a summary of her findings, as shown in the table below.

Your notes

	Median	Range
Snails	7.1 cm	3.1 cm
Slugs	9.7 cm	4.5 cm

Compare the distances travelled by snails and slugs during the ten-minute interval.

Compare the numerical values of the median (an average)
Describe what this means in real life

Slugs have a higher median than snails (9.7 cm > 7.1 cm)
This suggests that, on average, slugs travel further than snails

Compare the numerical values of the range (the spread)
Describe what this means in real life

Snails have a lower range than slugs (3.1 cm < 4.5 cm)
This suggests that there is less variation in the distances travelled by snails

Your notes

Your notes

Population & Sampling

What are the different types of data?

- **Primary** data is data that has been collected by the person carrying out the research
  - This could be through questionnaires, surveys, experiments etc
- **Secondary** data is data that has been collected previously
  - This could be found on the internet or through other research sources
- **Qualitative** data is data that is usually given in words not numbers to **describe** something
  - For example: the colour of a teacher's car
- **Quantitative** data is data that is given using numbers which **counts or measures** something
  - For example: the number of pets that a student has
- **Discrete** data is quantitative data that needs to be **counted**
  - Discrete data can only take **specific values** from a set of (usually finite) values
  - For example: the number of times a coin is flipped until a 'tails' is obtained
- **Continuous** data is quantitative data that needs to be **measured**
  - Continuous data can take **any value** within a range of infinite values
  - For example: the height of a student
- **Age** can be **discrete or continuous** depending on the context or how it is defined
  - If you mean how many years old a person is then this is discrete
  - If you mean how long a person has been alive then this is continuous

What is a population?

- A **population** refers to the **whole set** of things which you are interested in
  - e.g. if a teacher wanted to know how long pupils in year 11 at their school spent revising each week then the population would be all the year 11 pupils at the school
- Population does **not** necessarily refer to a number of people or animals

    - e.g. if an IT expert wanted to investigate the speed of mobile phones then the population would be all the different makes and models of mobile phones in the world

What is a sample?

- A **sample** refers to a selected **part** (called a subset) **of the population** which is used to collect data from
    - e.g. for the teacher investigating year 11 revision times a sample would be a certain number of pupils from year 11
- A **random sample** is where every item in the population has an **equal chance** of being selected
    - e.g. every pupil in year 11 would have the same chance of being selected for the teacher's sample
- A **biased sample** is where the sample is not random
    - e.g. the teacher asks pupils from just one class

What are the advantages and disadvantages of using a population?

- You may see or hear the word **census** - this is when data is collected from every member of the whole population
- The **advantages** of using a **population**
    - Accurate results - as every member/item of the population is used
        - In reality it would be *close* to every member for practical reasons
    - All options/opinions/responses will be included in the results
- The **disadvantages** of using a **population**
    - Time consuming to collect the data
    - Expensive due to the large numbers involved
    - Large amounts of data to organise and analyse

What are the advantages and disadvantages of using a sample?

- The **advantages** of using a **sample**
    - Quicker to collect the data
    - Cheaper as not so much work involved
    - Less data to organise and analyse

Your notes

Your notes

- The **disadvantages** of using a **sample**
  - A small sample size can lead to unreliable results
    - Sampling methods can usually be improved by taking a larger sample size
  - A sample can introduce **bias**
    - particularly if the sample is **not** random
  - A sample might **not** be **representative** of the **population**
    - Only a selection of options/opinions/responses might be accounted for
    - The members/items used in the sample may all have similar responses
      e.g. even with a random sample it may be possible the teacher happens to select pupils for his sample who all happen to do very little revision
- It is important to recognise that **different samples** (from the same population) may produce **different results**

Worked Example

Mike is a biologist studying mice and has access to 600 mice that live in an enclosure.
Mike wants to sample some of the mice for a study into their response to a new drug.
He decides to sample 10 mice, selecting those nearest to the enclosure's entrance.

a)

State the population in this situation.

The population is the 600 mice living in the enclosure

b)

State two possible issues with the sample method Mike intends using.

The sample size is very small – just 10 mice
The mice are not being selected at random – those nearest the entrance have a greater chance of being selected

c)

Suggest one way in which Mike could improve the reliability of the results from his sample.

Mike should increase the sample size to increase the reliability of the results