AQA GCSE Maths: Higher

Your notes

Cumulative Frequency & Box Plots

Contents

Your notes

Cumulative Frequency

What is cumulative frequency?

- **Cumulative** refers to a “running total" or "adding up as you go along”
- So in a table of **grouped data**
  - **cumulative frequency** means all of the frequencies for the different groups totalled up to the **end** of the group in a given row
- When working out cumulative frequencies you may see **tables** presented in two ways
  - A regular grouped data table with an extra column for cumulative frequencies
    - E.g. rows labelled $0 \leq x < 20$, $20 \leq x < 40$, $40 \leq x < 60$, etc

Group	Frequency	Cumulative frequency
$0 \leq x < 20$	14	14
$20 \leq x < 40$	25	39 (because 14+25=39)
$40 \leq x < 60$	29	68 (because 14+25+29=68)
$60 \leq x < 80$	12	80 (because 14+25+29+12=80)

  - or a separate table where every group is relabelled as starting at the beginning (often zero)
    - E.g. rows labelled $0 \leq x < 20$, $0 \leq x < 40$, $0 \leq x < 60$, etc.
    - Or $x < 20$, $x < 40$, $x < 60$, etc.

Group	Cumulative frequency

$0 \leq x < 20$ (or $x < 20$)	14
$0 \leq x < 40$ (or $x < 40$)	39
$0 \leq x < 60$ (or $x < 60$)	68
$0 \leq x < 80$ (or $x < 80$)	80

Your notes

- In the second type of table, you can **subtract** to find the individual frequencies
  - E.g. the frequency of the $20 \leq x < 40$ class interval is $39 - 14 = 25$
  - The frequency of the $40 \leq x < 60$ class interval is $68 - 39 = 29$, etc.

Drawing Cumulative Frequency Diagrams

What is a cumulative frequency diagram?

- A **cumulative frequency diagram** is a way of representing **grouped continuous data**
- A cumulative frequency diagram can be used to **estimate** other **statistical values**
  - For example the **median**, **quartiles** or **percentiles**

How do I draw a cumulative frequency diagram?

- This is best explained with an example
  - The times taken to complete a short general knowledge quiz taken by 50 students are shown in the table below:

Time taken (s seconds)	Frequency
$25 \leq s < 30$	3
$30 \leq s < 35$	8
$35 \leq s < 40$	17
$40 \leq s < 45$	12
$45 \leq s < 50$	7
$50 \leq s < 55$	3
Total	**50**

- Then the **cumulative frequency** is the running total of the frequencies

Time taken (s seconds)	Frequency	Cumulative Frequency

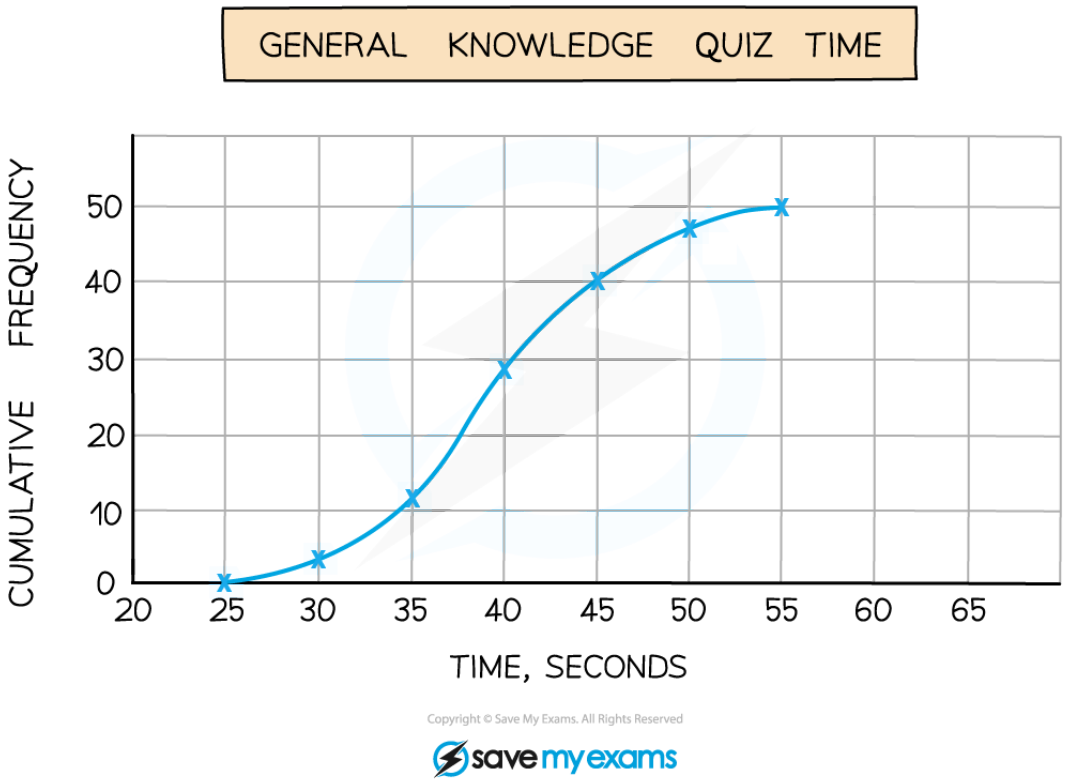
$25 \leq s < 30$	3	3
$30 \leq s < 35$	8	3 + 8 = 11
$35 \leq s < 40$	17	11 + 17 = 28
$40 \leq s < 45$	12	28 + 12 = 40
$45 \leq s < 50$	7	40 + 7 = 47
$50 \leq s < 55$	3	47 + 3 = 50
Total	**50**	

Your notes

- We can now draw the **cumulative frequency diagram**
  - The most important part is that **cumulative frequency** is plotted against the **end** (**upper bound**) of the **class interval**
    - The end of the class interval is the *x*-coordinate
    - The cumulative frequency is the *y*-coordinate
    - For the above example the first two points to plot would be (30, 3) and (35, 11)
  - To explain this, consider the second row ($30 \leq s < 35$)
    - The 8 students in this group could have taken **any** time between 30 and 35 seconds
    - They cannot **all** be guaranteed to have been accounted for until we reach 35 seconds
  - Once all points from the table are plotted, a point for the **start** needs to be added
    - This will be at the lowest time from the table
    - i.e. at 25 seconds with a cumulative frequency of 0
    - Plot the point (25, 0)
  - Join points up with a **smooth curve** (this takes some practice)
    - Make sure it goes through all of the marked points
  - In general a cumulative frequency diagram has a **stretched-S-shape** appearance

- A cumulative frequency diagram will **never** come back towards the x-axis

- Here is the final cumulative frequency diagram for the quiz times



Your notes

Your notes

Interpreting Cumulative Frequency Diagrams

How do I use and interpret a cumulative frequency diagram?

- A cumulative frequency diagram provides a way to **estimate** key facts about the data
  - **median**
  - **lower** and **upper quartiles** (and **interquartile range**)
  - **percentiles**
- These values will be **estimates** as the original raw data is unknown
  - Cumulative frequency diagrams are used with **grouped** data
  - Points are joined by a smooth curve
    - This means the data is assumed to be **smoothly spread out** over each interval

How do I find the median, lower quartile and upper quartile from a cumulative frequency diagram?

- This is all about understanding **how many data values** are represented by the cumulative frequency diagram
  - This may be **stated** in words within the question
  - If not, it will be the **highest value** on the frequency (y-) axis that the curve on the diagram reaches
    - This should be "top right" of the curve on a cumulative frequency diagram
- Finding the **median**:
  - **STEP 1**

    Find the **position** of the median, for n data values, this will be $\frac{n}{2}$
    - This is different from finding the median from a set of data values
    - E.g. for a list of 60 data values the median would be halfway between the 30th and 31st values
    - But for a cumulative frequency diagram it would just be $\frac{60}{2} = 30$

- **STEP 2**

  Draw a **horizontal line** from $\frac{n}{2}$ on the cumulative frequency (y-) axis until it hits the curve

- **STEP 3**
  Draw a **vertical line** from that point on the curve down to the horizontal (x-) axis

  - The value where that line hits the horizontal axis will be the median

- Finding the **lower quartile**:

  - **STEP 1**
    Find the **position** of the lower quartile

    - For n data values this will be $\frac{n}{4}$

  - **STEP 2**

    Draw a **horizontal line** from $\frac{n}{4}$ on the cumulative frequency axis until it hits the curve

  - **STEP 3**
    Draw a **vertical line** from that point on the curve down to the horizontal (x-) axis

    - The value where that line hits the horizontal axis will be the lower quartile

- Finding the **upper quartile**:

  - **STEP 1**
    Find the **position** of the upper quartile

    - For n data values this will be $\frac{3n}{4}$ $\left(\text{i.e. } 3 \times \frac{n}{4}\right)$

  - **STEP 2**

    Draw a **horizontal line** from $\frac{3n}{4}$ on the cumulative frequency axis until it hits the curve

  - **STEP 3**
    Draw a **vertical line** from that point on the curve to the horizontal (x-) axis

    - The value where that line hits the horizontal axis will be the upper quartile

How do I find a percentile from a cumulative frequency diagram?

- **Percentiles** split the data into 100 parts

Your notes

- The **50th percentile** is another way of describing the **median**
- The **25th** and **75th percentiles** are the same as the **lower** and **upper quartiles** (respectively)

- To find the ***p*th percentile**:
  - **STEP 1**
    Find the **position** of the p^{th} percentile
    - For n data values, this will be $\frac{np}{100}$ $\left(\text{i.e. } \frac{n}{100} \times p\right)$
    - So for the 10th percentile ($p = 10$) with 60 data values ($n = 60$)
    - The position is $\frac{10}{100} \times 60 = \frac{1}{10} \times 60 = 6$
  - **STEP 2**
    Draw a **horizontal line** from $\frac{np}{100}$ on the cumulative frequency axis until it hits the curve
  - **STEP 3**
    Draw a **vertical line** from that point on the curve down to the horizontal (x-) axis
    - The value where that line hits the horizontal axis will be the p^{th} percentile

Your notes

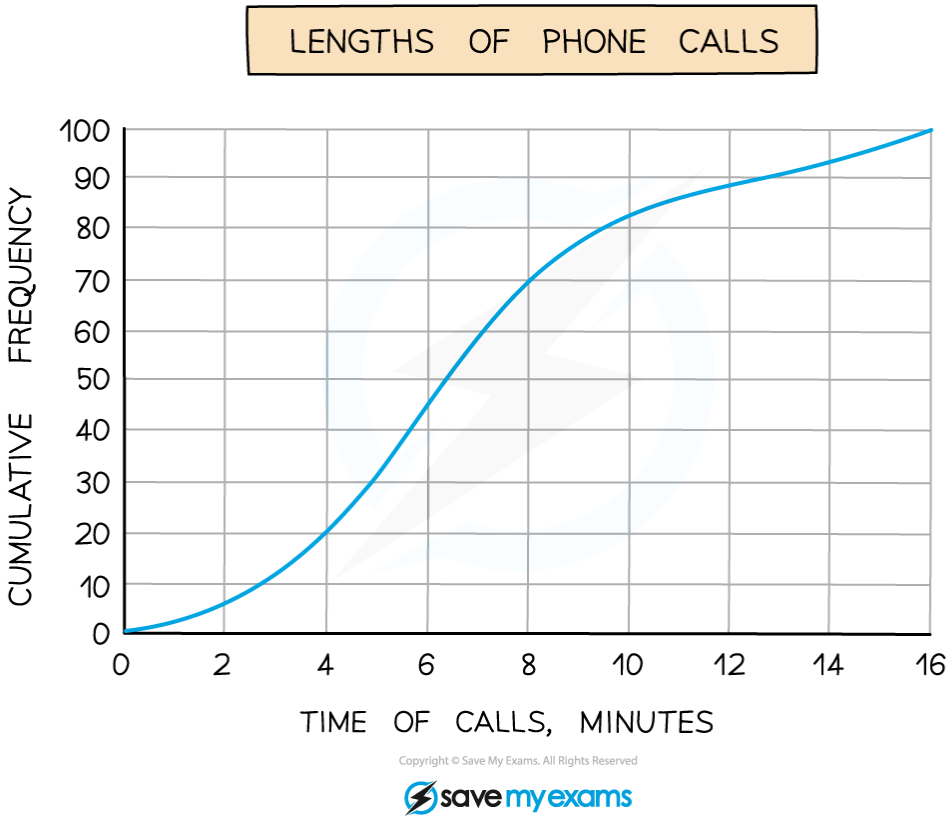
Worked Example

A company is investigating the length of telephone calls customers make to its help centre.

The company randomly selects 100 phone calls from a particular day.

The results are displayed in the cumulative frequency diagram below.

Your notes



a) Estimate the median, the lower quartile and the upper quartile.

There are 100 pieces of data, so $n = 100$

$$\frac{n}{2} = \frac{100}{2} = 50$$

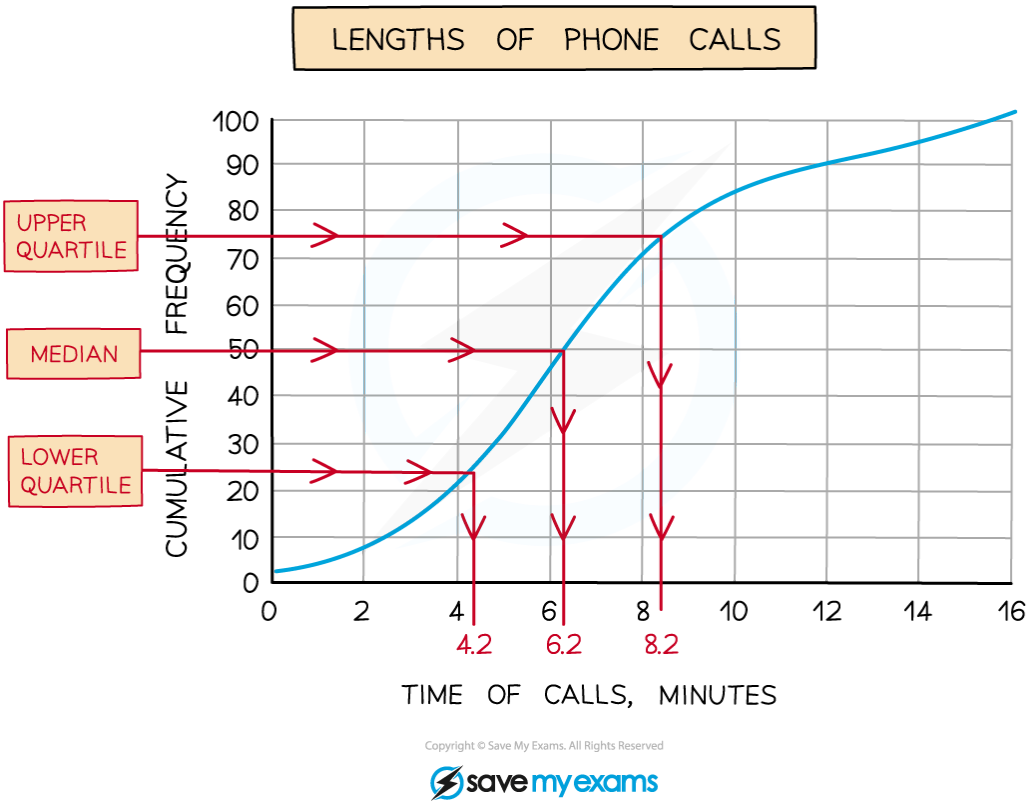
$$\frac{n}{4} = \frac{100}{4} = 25$$

$$\frac{3n}{4} = 3 \times 25 = 75$$

So the median is the 50th value
The lower quartile is the 25th value
The upper quartile is the 75th value

Draw horizontal lines from these on the cumulative frequency axis until they hit the curve
Then draw vertical lines down to the time of calls axis and take readings

Your notes



Median = 6.2 minutes (6 m 12 s)
Lower quartile = 4.2 minutes (4 m 12 s)
Upper quartile = 8.2 minutes (8 m 12 s)

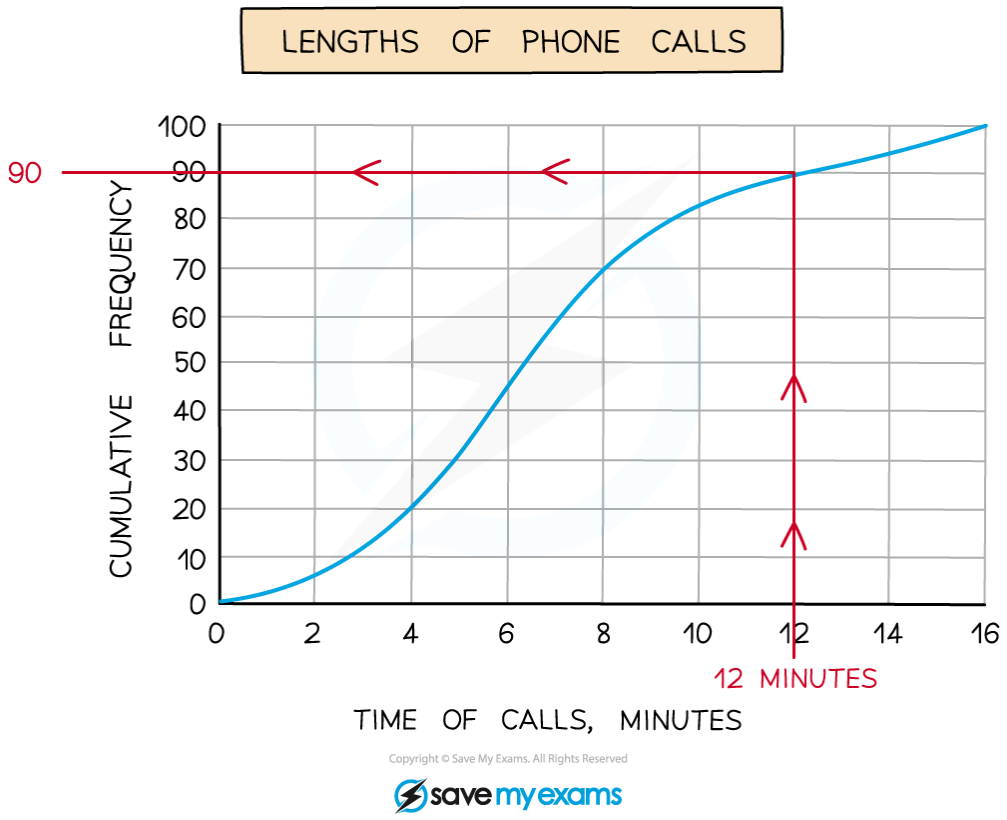
There is no need to convert to minutes and seconds unless the question asks you to
However, writing 6 m 2 s or 6 m 20 s would be incorrect

b) The company is thinking of putting an upper limit of 12 minutes on calls to its help centre.

Estimate the number of these 100 calls that would have been beyond this limit.

Draw a vertical line up from 12 minutes on the time of calls axis until it hits the curve
Then draw a horizontal line across to the cumulative frequency axis and take a reading (in this case, 90)

Your notes



This tells us that up to 12 minutes, 90 of the calls had been accounted for

The question wants the number of calls that were **greater** than 12 minutes so subtract this from the total of 100

$$100 - 90 = 10$$

Approximately 10 (out of 100) calls were beyond the 12 minute limit

Your notes

Box Plots

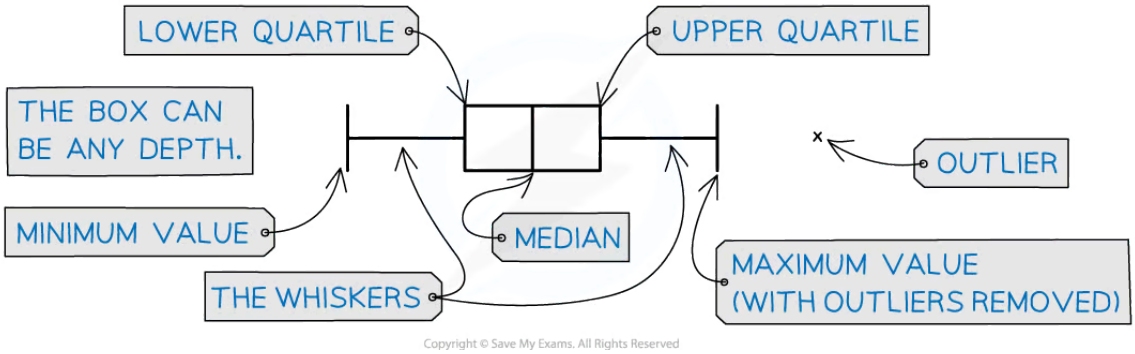
What are box plots and when should they be used?

- **Box plots** are also known as **box-and-whisker diagrams**
- They are used when we are interested in splitting data up into **quartiles**
- Often, data will contain extreme values
  - consider the cost of a car: there are far more family cars around than there are expensive sports cars
  - if you had 50 data values about the prices of cars and 49 of them were family cars but 1 was a sports car
    - the sports car's value would not fit in with the rest of the data
- Using quartiles and drawing a box plot allows us to split the data
  - we can see what is happening at the low, middle and high points
  - and consider any possible extreme values

How do I draw a box plot?

- You need to know five values to draw a box plot
  - Lowest data value
  - Lower quartile
  - Median
  - Upper quartile
  - Highest data value
- Usually on graph paper, box plots are drawn accurately with the five points marked by short vertical lines
  - the middle three values then form a box with the median line inside
    - the median will not necessarily be in the middle of the box!
  - the box represents the interquartile range (middle 50% of the data)

- the lowest data value and highest data value are joined to the box by horizontal lines
  - these are often called whiskers
  - they represent the lowest 25% of the data and the highest 25% of the data
- You may be given a box plot
  - from which you can read off the five values
  - calculate other statistics like the **range** and **interquartile range** (IQR)



How do I compare box plots?

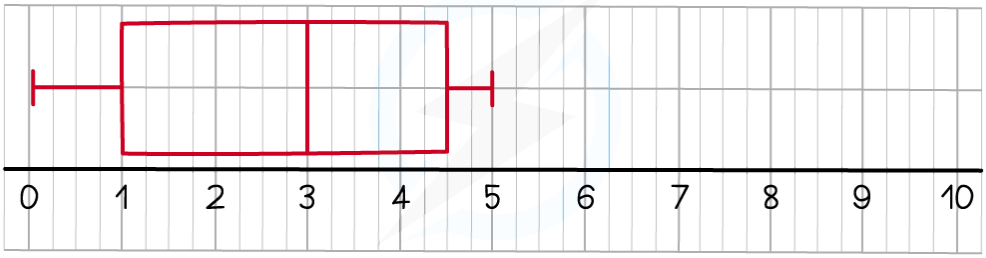
- If you are asked to **compare** box plots aim for two pairs of comments
  - the first pair of comments should mention **average** - i.e. the **median**
    - the first comment should **compare** the **value** of the medians
      e.g. the median for boys (12) is greater than the median for girls (8)
    - the second comment should **explain** it in the **context** of the question
      e.g. the boys were, on average, 4 seconds slower than the girls
  - the second pair of comments should mention **spread** - i.e. the **interquartile range** (or range)
    - the first comment should **compare** the **value** of the IQRs
      e.g. the IQR for boys (6) is lower than the IQR for girls (9)
    - the second comment should **explain** it in the **context** of the question
      e.g. the boys times were less spread out than the girls, the boys were more consistent

Worked Example

Your notes

Your notes

The box plot below shows the number of goals scored per game by Albion Rovers during a football season.

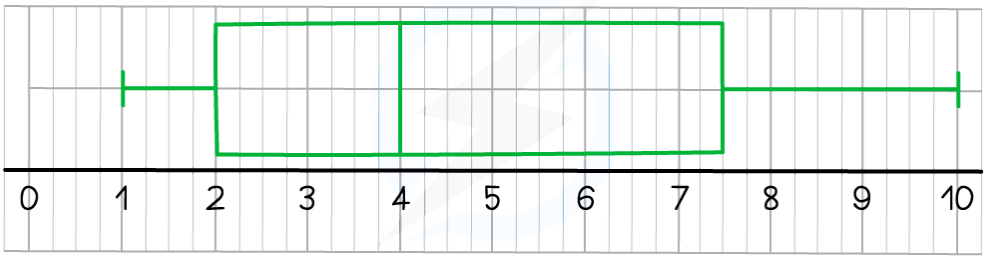


The information below shows the number of goals scored per game by Union Athletic during the same football season.

Median number of goals per game	4
Lower quartile	2
Upper quartile	7.5
Lowest number of goals per game	1
Highest number of goals per game	10

(a) Draw a box plot for the Union Athletic data.

Draw the box plot by first plotting all five points as vertical lines.
Draw a box around the middle three and then draw whiskers out to the outer two.



(b) Compare the number of goals scored per game by the two teams.

Your first comment should be about **averages** - do it in two sentences.
Your first sentence should be just about the maths and numbers involved. The second should be about what it means.

The median number of goals per game is higher for Union Athletic (4 goals) than Albion Rovers (3 goals).
This means that on average, Union Athletic scored more goals per game than Albion Rovers.

Your second comment should be about **spread** - do it in two sentences.
Your first sentence should be just about the maths and numbers involved. The second should be about what it means.

The interquartile range (IQR) is higher for Union Athletic (4) than Albion Rovers (3).
This means that Albion Rovers were more consistent regarding the number of goals they scored per game.

Remember a smaller range/IQR means more consistent which, depending on the situation, may be desirable.


Your notes