Lesson - Charts, averages and cumulative frequency

Year: Year 11

This lesson covers averages and spread, frequency tables, grouped data, charts, and the Higher tier topics of histograms, cumulative frequency and box plots (marked Higher). Statistics questions ask you to draw, read and compare.

Starter activity

Look at the results of a school survey on how long Year 11 pupils sleep: 6, 7, 7, 8, 8, 8, 9, 10 and 5 hours. Without calculating, guess the typical amount of sleep. Now calculate the mean, median and mode. Are they the same? Would the answer change if one pupil slept for only 1 hour? Talk about which average describes the group most fairly.

Averages and spread

Definition: mean, median, mode and range

- **Mean**: add all the values and divide by how many there are.
- **Median**: the middle value when the data are in order. With an even number of values take the mean of the middle two.
- **Mode**: the most common value.
- **Range**: largest minus smallest, a measure of spread.

Worked example 1: averages

Find the mean, median, mode and range of $4, 7, 7, 9, 12, 15, 16$.

Mean $= \dfrac{70}{7} = 10$, median $= 9$, mode $= 7$, range $= 16 - 4 = 12$.

Key fact: which average to use

The mean uses every value but is affected by extreme values. For $5, 6, 6, 7, 8, 38$ the mean is 11.7 but the median is 6.5, which is more typical. The mode is the only average for non-numerical data (favourite colour).

Frequency tables

Rule: mean and median from a frequency table

The total frequency is the sum of the frequencies, and the mean is $\frac{\sum fx}{\sum f}$: multiply each value by its frequency, add, then divide by the total frequency. The median is in position $\frac{n+1}{2}$.

Worked example 2: a frequency table

Scores 1 to 5 have frequencies 3, 4, 6, 5, 2. Find the mean, mode and median.

Total $= 20$; $\sum fx = 3 + 8 + 18 + 20 + 10 = 59$; mean $= 2.95$. The mode is 3 and the median (10th and 11th values) is 3.

Grouped data

Definition: midpoint and modal class

With classes such as $0 < h \leq 10$ you cannot find exact values, so you **estimate** the mean using the **midpoint** of each class. The **modal class** is the one with the highest frequency, and the median is in the class containing the $\frac{n+1}{2}$th value.

Worked example 3: estimated mean

Frequencies 4, 6, 10, 5 in classes of width 10 starting at 0. Estimate the mean and give the modal class.

Midpoints are 5, 15, 25, 35. $\sum fx = 20 + 90 + 250 + 175 = 535$ and $\sum f = 25$, so the estimated mean is 21.4. The modal class is $20 < h \leq 30$.

Method: estimating the mean from grouped data

1. Write the midpoint of each class.
2. Multiply each midpoint by its frequency and add up the results to get $\sum fx$.
3. Divide by the total frequency $\sum f$ and say that the answer is an estimate.

Charts

Rule: pie charts, polygons and stem and leaf

Pie charts: the angle for a category is $\frac{\text{frequency}}{\text{total}} \times 360°$. With 90 students, 18 of them are $\frac{18}{90} \times 360 = 72°$. **Frequency polygons** plot the frequency against the class midpoint and join with straight lines. **Stem and leaf diagrams** need a key and

ordered leaves.

(Higher) Histograms

Definition (Higher): histogram and frequency density

Histograms are used for continuous data with classes of *different widths*. The height of a bar is the **frequency density** and the **area** of a bar is the frequency:

$$\text{frequency density} = \frac{\text{frequency}}{\text{class width}}$$

The vertical axis is labelled "frequency density", not "frequency".

Worked example 4 (Higher): frequency density

Class $10 < t \leq 30$ has frequency 50. A bar $30 < t \leq 50$ has height 1.5. Find the frequency density of the first class and the frequency of the second.

The class width is 20, so the frequency density is $50 \div 20 = 2.5$. The second bar has frequency $1.5 \times 20 = 30$.

(Higher) Cumulative frequency

Definition (Higher): cumulative frequency and quartiles

Cumulative frequency is a running total. Plot each cumulative frequency against the **upper class boundary** and join with a smooth curve. For n values, the median is at $\frac{n}{2}$, the lower quartile Q_1 at $\frac{n}{4}$ and the upper quartile Q_3 at $\frac{3n}{4}$. The interquartile range is $IQR = Q_3 - Q_1$, a measure of spread that ignores extreme values.

Worked example 5 (Higher): reading quartiles

Ages have class frequencies 5, 15, 30, 35, 15 in classes of width 10 from 0. Find the median and the interquartile range.

The cumulative frequencies are 5, 20, 50, 85, 100. The median is at the 50th value, 30. Q_1 at the 25th value is $20 + \frac{5}{30} \times 10 \approx 21.7$ and Q_3 at the 75th value is $30 + \frac{25}{35} \times 10 \approx 37.1$. So $IQR \approx 15.5$.

(Higher) Box plots

Key fact (Higher): box plots

A box plot shows five numbers: minimum, Q_1, median, Q_3, maximum. If the five numbers are 12, 20, 28, 35 and 52, the range is 40 and the IQR is 15.

Comparing data sets

Rule: comparing distributions

A good comparison has two sentences, in context: one comparing a typical value (mean or median) and one comparing the spread (range or IQR).

Worked example 6: comparing two teams

Team A has mean 52 and range 10; Team B has mean 50 and range 30. Compare them. Then compare two box plots.

Team A scores higher on average and is more consistent. For box plots, for example: "Group B has a higher median, so they are generally faster, and a smaller IQR, so their times are more consistent."

Watch out: common mistakes

- Not ordering the data before finding the median.
- Using the class boundary instead of the midpoint when estimating a mean.
- Plotting a cumulative frequency graph against the midpoint rather than the upper boundary.
- Drawing histograms with frequency on the vertical axis.
- Comparing without using the context.

Worked example 7: Choosing the best average

The times (in seconds) of eight pupils: 12, 15, 15, 18, 21, 24, 24, 35. Find the mean, median and modes, and explain which average best represents the group.

Total $= 12 + 15 + 15 + 18 + 21 + 24 + 24 + 35 = 164$, so the mean is $164 \div 8 = 20.5$. The median is the mean of the 4th and 5th values: $(18 + 21) \div 2 = 19.5$. There are two modes, 15 and 24. The value 35 is much larger than the others, which pulls the mean up, so the median (19.5) is the better description of a typical time.

Worked example 8: Comparing two data sets

Team A has mean 14 and range 6. Team B has mean 12 and range 15. Compare the two teams.

Compare an average and a spread. Team A has the higher mean (14 compared with 12), so they score more on average. Team A also has the smaller range (6 compared with 15), so they are more consistent. Always write the comparison in the context of

the question and use the word "consistent" for spread.

Worked example 9: Frequency density without a table

A histogram bar for $20 < x \leq 50$ has height 1.6 on the frequency density axis. How many items does it represent?

The class width is $50 - 20 = 30$. Frequency = frequency density $\times$ class width $= 1.6 \times 30 = 48$. So the bar represents 48 items. The area of a bar in a histogram is the frequency, which is why the bars can have different widths.

Worked example 10: Using a cumulative frequency graph (Higher)

From a cumulative frequency graph for 200 pupils, the reading at 60 marks is 150. How many pupils scored more than 60? Give the percentage.

150 pupils scored 60 or fewer, so $200 - 150 = 50$ pupils scored more than 60. As a percentage: $50 \div 200 \times 100 = 25\%$. If the question asks for the median, read across from 100, which is half of 200. For the lower quartile read across from 50 and for the upper quartile from 150.

Worked example 11: Drawing and reading a box plot (Higher)

A data set has minimum 4, lower quartile 9, median 13, upper quartile 20 and maximum 31. Work out the range and the interquartile range, and say what the box contains.

Range $= 31 - 4 = 27$. Interquartile range $= 20 - 9 = 11$. The box runs from 9 to 20 and so contains the middle half of the data. The median line is closer to the lower quartile ($13 - 9 = 4$ against $20 - 13 = 7$), which tells us the data is positively skewed: the upper half is more spread out.

Note: choosing and describing a chart

Choose a **bar chart** for categories, a **pie chart** to show proportions of a whole, a **frequency polygon** (plot at the midpoints) to compare distributions, a **histogram** when class widths are unequal and a **scatter graph** for two variables. Every chart needs a title, labelled axes with units and an equal-width scale. When you are asked to *compare*, always write one sentence about an average and one about the spread, and refer to the context in the question.

Watch out: grouped data traps

- In a grouped frequency table the mean is only an **estimate**. Use the midpoint, not the class boundary.
- The median of grouped data lies in a **class**; it is not the middle class.
- For a histogram, plot **frequency density**, not frequency, when the class widths are unequal.

- For a cumulative frequency graph, plot against the **upper** class boundary.

Did you know?

The first statistical graph was drawn by a Scottish engineer called William Playfair in 1786. He invented the bar chart and the line graph to show how much Britain was exporting and importing. The pie chart followed in 1801. Florence Nightingale used a coloured version called a polar area diagram to convince the government to improve hospitals during the Crimean War.

Otto's tip

Write the formula before you use it, and put a total row in your table. Check that your cumulative frequency finishes at the total number of data values. When you read a value from a graph, draw your construction lines with a ruler so the examiner can see which point you used.