Lesson - Statistics: scatter graphs and averages

Year: Year 8

This lesson covers averages and range, frequency tables, grouped data, pie charts, scatter graphs and comparing distributions.

Starter activity

Ask five friends how many hours of sleep they had last night and how many glasses of water they drank today. Write your results as a list. Find the mean and the median of each list. Do the two averages tell the same story? Which would you use if one friend had been awake all night?

Averages and spread

Definition: mean, median and mode

An **average** is a single value that represents a whole set of data.

- The **mean** is the total divided by the number of values. It uses every value but is pulled by very large or small ones.
- The **median** is the middle value when the data are in order. It is not affected by extreme values.
- The **mode** is the most common value. It is the only average for non-numerical data such as favourite colours.

Definition: range

The **range** is the largest value minus the smallest. It measures spread: a small range means the data are consistent.

Worked example 1: finding averages

Find the mean, median, mode and range of 4, 7, 9, 10, 10.

Mean $= \frac{40}{5} = 8$. Median (the middle of five ordered values) $= 9$. Mode $= 10$. Range $= 10 - 4 = 6$. For an even number of values the median is halfway between the two middle ones: the median of 2, 5, 8, 11 is $\frac{5 + 8}{2} = 6.5$.

Frequency tables

Method: mean from a frequency table

1. Multiply each value by its frequency.
2. Add these to get the total of all the values.
3. Divide by the total frequency.

Worked example 2: frequency table

Scores of 1, 2, 3 and 4 occur 3, 5, 4 and 3 times. Find the mean, mode and median.

Total frequency $= 15$. Total of scores $= 3 + 10 + 12 + 12 = 37$. Mean $= \dfrac{37}{15} = 2.47$ (2 d.p.). The mode is 2 (highest frequency). The cumulative frequencies are 3, 8, 12, 15, so the 8th value, the median, is 2.

Grouped data

Definition: midpoint and modal class

When data are in class intervals you cannot find the exact mean. Use the **midpoint** of each class as an estimate. The **modal class** is the class with the highest frequency.

Worked example 3: grouped data

Classes 0-10, 10-20 and 20-30 have frequencies 4, 6 and 11. Estimate the mean and find the modal class.

Midpoints are 5, 15 and 25. Estimated total $= 20 + 90 + 275 = 385$, total frequency 21, so the estimated mean is $\dfrac{385}{21} \approx 18.3$. The modal class is 20-30. The 11th of 21 values also falls in the class 20-30, so the median lies there.

Pie charts

Rule: pie chart angles

The whole circle is 360°. Angle $= \dfrac{\text{frequency}}{\text{total}} \times 360^\circ$.

Worked example 4: drawing and reading a pie chart

In a survey of 40 pupils, 18 chose football, 10 rugby, 8 tennis and 4 other. Find the angles. Then, for a chart of 72 pupils, how many are in a 55° sector?

The angles are 162°, 90°, 72° and 36°, which add to 360°. For 72 pupils, one pupil is 5°, so a 55° sector is 11 pupils.

Scatter graphs

Definition: scatter graph and correlation

A **scatter graph** plots pairs of values to see whether there is a relationship, called **correlation**.

- **Positive correlation:** as one variable increases, the other increases (height and shoe size).
- **Negative correlation:** as one variable increases, the other decreases (hours of television and test score).
- **No correlation:** no pattern (shoe size and maths score).

Key fact: line of best fit

A **line of best fit** is a straight line drawn by eye through the middle of the points, with roughly equal numbers of points on each side. Use it to estimate values within the range of the data; predicting far outside is unreliable. Correlation does not prove that one thing causes the other.

Worked example 5: using a line of best fit

A line of best fit goes through $(2, 10)$ and $(6, 30)$. Estimate y when $x = 4$.

The gradient is $\dfrac{20}{4} = 5$ and the equation is $y = 5x$, so at $x = 4$ the estimate is 20.

Outliers and comparing data

Definition: outlier

An **outlier** is a value that is far from the rest. Use the median for skewed data (such as salaries), the mean when values are fairly symmetrical, and the mode for categories.

Worked example 6: effect of an outlier

In 12, 14, 15, 13, 48, which value is an outlier? Compare the mean and median.

The value 48 is an outlier. The mean is $\dfrac{102}{5} = 20.4$, but without 48 it is 13.5. The median is 14 either way, so it is a better summary here.

Note: comparing distributions

To compare two sets of data, write one comment about an average and one about the spread. Class A has mean 62 and range 20; class B has mean 58 and range 8. Class A did better on average, but class B was more consistent.

Watch out: common mistakes

- Finding the median before putting the data in order.
- Dividing the total by the number of classes instead of the total frequency.
- Giving the frequency instead of the value as the mode.
- Extending a line of best fit far beyond the data.
- Saying that correlation proves cause.

Worked example 7: finding a missing value

The mean of five numbers is 9. Four of the numbers are 4, 7, 11 and 12. Find the fifth number.

The total of five numbers with mean 9 is $5 \times 9 = 45$. The four known numbers add up to $4 + 7 + 11 + 12 = 34$. The fifth number is $45 - 34 = 11$. Check: $(34 + 11) \div 5 = 9$.

Worked example 8: comparing distributions

The boys' median time for a 100 m race is 14.2 s with range 3.1 s. The girls' median is 14.8 s with range 1.9 s. Compare the groups.

The boys' median is lower, so on average the boys were faster. The girls' range is smaller, so the girls' times were more consistent. Always mention both an average and the spread, and write your comparison in the context of the question.

Worked example 9: a pie chart angle

In a survey of 80 people, 18 chose tea. What angle is needed?

Each person is $360 \div 80 = 4.5^\circ$. Tea: $18 \times 4.5 = 81^\circ$. Or as a fraction: $\frac{18}{80} \times 360 = 81^\circ$.

Method: estimating the mean from grouped data

1. Find the midpoint of each class.
2. Multiply each midpoint by its frequency.
3. Add up the products and divide by the total frequency.
4. State that it is an estimate, because the exact values are not known.

Note: choosing an average

The **mean** uses all the data but is pulled by outliers. The **median** is the middle value and is not affected by extreme values. The **mode** is the most common value and is the only average that works for non-numerical data such as favourite colours. Say why you chose your average, using the shape of the data.

Watch out: common mistakes

- Forgetting to order the data before finding the median.
- Finding the mean of a frequency table by adding the frequencies instead of multiplying value × frequency.
- Joining the points on a scatter graph: draw a straight line of best fit instead.
- Using a line of best fit far outside the data (extrapolation).

Did you know?

Florence Nightingale, the famous nurse, was also a pioneer of statistics. In 1858 she used pie-chart-like diagrams, which she called coxcombs, to show that more soldiers in the Crimean War were dying from disease than from wounds. Her charts convinced the government to improve hospital hygiene, and she became the first woman elected to the Royal Statistical Society.

Writing a good comparison

Many statistics questions end with “compare the two sets of data”. A full answer always contains two comparisons, each linked to the context. First compare an average, for example “the median time for Class A is 2 minutes shorter, so Class A were usually faster”. Then compare a measure of spread, for example “Class B has a smaller range, so their times were more consistent”. Saying only that one number is bigger than another does not earn the marks; you must say what it means for the pupils, the plants or the matches in the question.

Did you know?

Florence Nightingale, famous as a nurse, was also a pioneering statistician. In 1858 she drew a new kind of circular chart to show that more soldiers in the Crimean War died from disease than from wounds, and her diagrams persuaded the government to improve hospital hygiene and saved many lives.

Otto's tip

Mean, median, mode, range. Always order the data first and write the average you are finding on each line of your working.