Lesson - Statistics: averages and grouped data

Year: Year 9

This lesson covers averages and range, choosing an average, frequency tables, working backwards with the mean, grouped data and comparing distributions.

Starter activity

Everyone in the class writes down the number of letters in their first name. Make a frequency table. Find the most common number, the middle number and the average number. Do the three values agree? What would happen to each if someone with a 20-letter name joined the class?

Mean, median, mode and range

Definition: averages and range

Averages (measures of central tendency) give a single number that represents a set of data.

- The **mean** is the total of all the values divided by how many values there are.
- The **median** is the middle value when the data is in order. With an even number of values, it is halfway between the two middle ones.
- The **mode** is the most common value. A set can have no mode, one mode or more than one.
- The **range** is the largest value minus the smallest. It is a measure of *spread*.

Worked example 1: five values

Find the mean, median and range of 4, 7, 9, 12, 8.

Mean $= \dfrac{40}{5} = 8$. In order: 4, 7, 8, 9, 12, so the median is 8. Range $= 12 - 4 = 8$.

Worked example 2: an even number of values

Find the median of 3, 9, 4, 12, 7, 5.

In order: 3, 4, 5, 7, 9, 12. The median is halfway between 5 and 7, which is 6.

Which average should I use?

Key fact: choosing an average

The mean uses every value, but it is pulled towards **outliers** (extreme values). The median is not affected by outliers. The mode is the only average that works for non-numerical data. For 2, 3, 3, 4, 48 the mean is $\frac{60}{5} = 12$, which is far bigger than most of the values. The median, 3, is a better description of the typical value.

Frequency tables

Method: mean from a frequency table

1. Add a column for $x \times f$ (the value times its frequency).
2. Add up the frequencies and the $x \times f$ column.
3. Divide $\sum xf$ by $\sum f$.

Worked example 3: goals scored

Find the mean, mode and median of the goals data.

Goals x	Frequency f	$x \times f$
0	2	0
1	5	5
2	8	16
3	3	9
4	2	8
Total	**20**	**38**

Mean $= \frac{38}{20} = 1.9$. The mode is the value with the highest frequency, 2. With 20 values we need the 10th and 11th for the median. The cumulative frequencies are 2, 7, 15, so both are 2 and the median is 2.

Watch out: frequency tables

Do *not* add up the frequencies and divide by the number of rows. The total frequency is the number of data values.

Working backwards with the mean

Rule: total from the mean

Total = mean × number of values.

Worked example 4: a missing value

The mean of five numbers is 8 and four of them are 6, 7, 9 and 10. Find the fifth. The mean of 6 numbers is 7; a seventh number makes the mean 8. Find it.

The total of the five is 40 and the four known add to 32, so the missing number is $40 - 32 = 8$. For the second part, the total was 42 and becomes $7 \times 8 = 56$, so the seventh number is $56 - 42 = 14$.

Grouped data

Definition: class intervals, midpoint and modal class

Continuous data (heights or times) is often put into **class intervals**, for example $140 \le h < 150$. We cannot know the exact values inside each group, so we **estimate** using the **midpoint** of each group. The **modal class** is the class with the highest frequency.

Worked example 5: estimating from grouped data

Estimate the mean height, and find the modal class and the class containing the median.

Height h (cm)	Frequency f	Midpoint	$f\times$ midpoint
$140 \le h < 150$	4	145	580
$150 \le h < 160$	10	155	1550
$160 \le h < 170$	12	165	1980
$170 \le h < 180$	4	175	700
Total	**30**		**4810**

Estimated mean $= \dfrac{4810}{30} = 160.3$ cm (1 d.p.); it is an estimate because we used midpoints. The modal class is $160 \le h < 170$. With 30 values we need the 15th and 16th; the cumulative totals are 4, 14, 26, so the median lies in $160 \le h < 170$. The range can only be estimated: at most $180 - 140 = 40$ cm.

Comparing distributions

Worked example 6: comparing two teams

Team A has mean 12 and range 4; team B has mean 14 and range 10. Compare them.

Always compare **one average and one measure of spread** and refer to the context. Team B scores more on average, but team A is more consistent because its range is smaller.

Watch out: common mistakes

- Not putting the data in order before finding the median.
- Dividing by the number of rows in a frequency table rather than by the total frequency.
- Using the class boundaries rather than the midpoints in a grouped mean.
- Saying the mean of grouped data is exact (it is an estimate).
- Comparing only averages and ignoring the spread.

Worked example 7: finding the median from a frequency table

The table shows pets owned by 25 pupils: 0 pets: 6, 1 pet: 9, 2 pets: 7, 3 pets: 3. Find the median.

The median is the 13th value (since $(25 + 1) \div 2 = 13$). Cumulative frequencies: 6, 15, 22, 25. The 13th value is in the "1 pet" group, so the median is 1 pet.

Worked example 8: how an extra value changes the mean

The mean of 9 numbers is 12. A tenth number, 22, is added. Find the new mean.

Old total $= 9 \times 12 = 108$. New total $= 108 + 22 = 130$. New mean $= 130 \div 10 = 13$. The mean has increased by 1, because 22 is more than the old mean.

Worked example 9: estimated mean from a grouped table

Estimate the mean from this table: 0: 10 values, 20: 15 values, 40: 5 values.

Midpoints 10, 30 and 50. Total $\approx 10 \times 10 + 15 \times 30 + 5 \times 50 = 100 + 450 + 250 = 800$ and there are 30 values, so the mean $\approx 800 \div 30 = 26.7$. It is an estimate because we do not know where in each group the values lie.

Worked example 10: comparing using an average and the range

Mia's five scores: 6, 7, 7, 8, 7. Noah's: 3, 10, 5, 9, 8. Both have a mean of 7. Compare them.

Mia's range is $8 - 6 = 2$ and Noah's is $10 - 3 = 7$. They have the same mean, but Mia is much more consistent. Always give one comment about the average and one about the spread.

Note: which average?

The **mean** uses every value and is best for symmetrical data. The **median** is not affected by extreme values and is best for skewed data, such as salaries or house prices. The **mode** is the only average for non-numerical data, such as favourite colours. The **range** tells you the spread, not the typical value.

Watch out: frequency tables

- In a frequency table the mean is (sum of value $\times$ frequency) divided by the *total frequency*, not divided by the number of rows.
- The mode is the *value* with the highest frequency, not the frequency itself.
- In grouped data you can only estimate the mean, and you can only name the modal *class*, not the mode.

Did you know?

The word statistics comes from the Latin for state, because early statistics were facts collected by governments about their people, such as the Domesday Book of 1086 in England. Florence Nightingale, famous as a nurse, was also a pioneer of statistical charts and used them to persuade the government to improve hospitals.

Using statistics to answer questions

Method: comparing two distributions

1. Calculate an average for each set of data, usually the mean or the median, and say which set has the higher average.
2. Calculate a measure of spread, such as the range, and say which set is more consistent: the smaller the spread, the more consistent the data.
3. Write two separate sentences, one about the average and one about the spread, and always refer to the context, for example "On average Team A scored more goals."
4. Mention anything unusual, such as an extreme value, that might make one of the statistics misleading.

Worked example 11: the effect of an outlier

Five friends earn £8, £9, £9, £10 and £64 for a weekend job. Which average best represents a typical friend?

The mean is $100 \div 5 = £20$, but four of the five friends earn much less than this, so the mean does not describe them well. The value £64 is an outlier: it is far away from the rest. The median is the middle value, £9, and the mode is also £9. Both of these are close to what most of the friends earn, so the median is the best average here. The range, $64 - 8 = £56$, is also made very large by the single outlier.

Note: data in the news

Newspapers often report averages, such as the average house price or the average salary. These are usually medians, because a few very expensive houses or very high

salaries would push the mean up and give a misleading picture.

Otto's tip

For frequency tables, write the column $x \times f$ first, add it up, then divide by the total frequency. It never fails. Hoot hoot!