

Statistics – native teaching supplement

AQA · GCSE · Mathematics

Supported teaching and tier boundary

8300: Statistics. Version: Version 1.0, 12 September 2014; first examination 2017.

Foundation teaching and Higher additions are labelled below. This reference packages the existing native-lesson crosswalk. It does not certify unreviewed specification rows or a whole qualification. Original diagnostics are separate and are not reproduced.

■ Centre, spread and data displays · Foundation

The mean is total divided by count. The median is the central value after sorting. The range is maximum minus minimum. Use frequency tables, bar charts and suitable comparisons.

$$\bar{x} = \frac{\sum x_i}{n}$$

For 2,4,4,6,9, total=25 and count=5, so mean=5. The central value is 4, so median=4. The range is 9-2=7. Explain both a typical value and the spread.

The tallest histogram bar need not contain the most observations. A grouped mean is an estimate. Correlation does not prove causation, and extrapolation extends beyond the observed range.

A bar chart uses separate bars for categories. Unequal-class-width histograms and formal density calculations are outside this Foundation/Core lesson.

■ Data summaries, histograms and interpretation · Higher

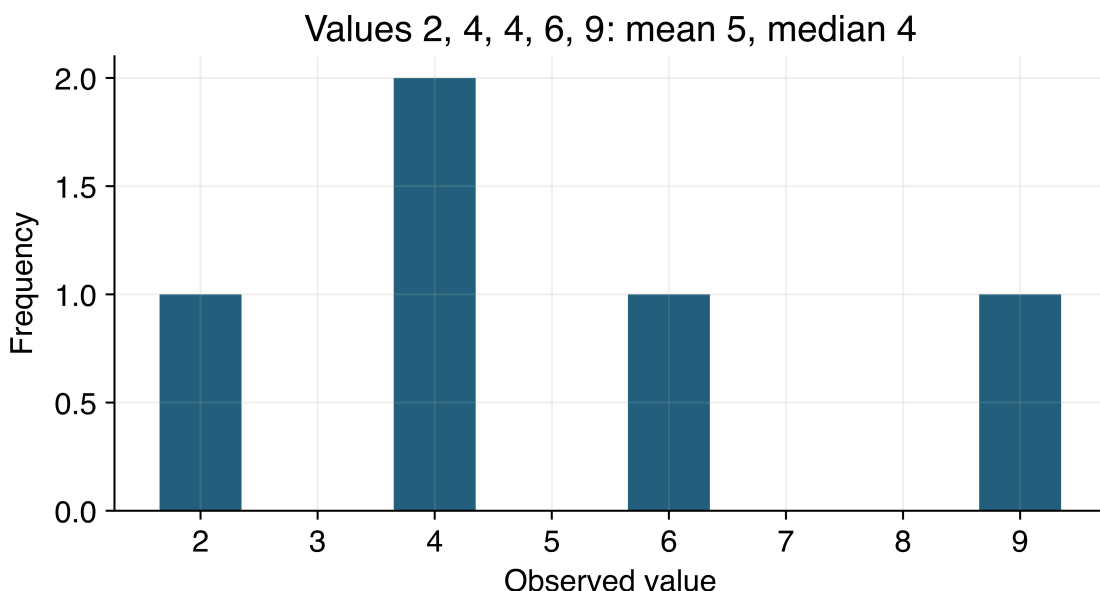
Compare an appropriate average and spread in context. A histogram uses area for frequency, so height=frequency/class width. Grouped estimates assume representative values within intervals.

$$\bar{x} = \frac{\sum x_i}{n}, \quad \text{density} = \frac{\text{frequency}}{\text{class width}}$$

A class from 10 to 20 with frequency 30 has density 30/10=3. A class from 20 to 40 with frequency 20 has density 20/20=1. Its wider bar must not be mistaken for a larger density. For values 2,4,4,6,9, the median is 4 and mean is 5.

The tallest histogram bar need not contain the most observations. A grouped mean is an estimate. Correlation does not prove causation, and extrapolation extends beyond the observed range.

Choose a display that fits the data type. Give both a numerical comparison and what it means for the population; do not infer more precision than the sample supports.



Original native-lesson illustration; labels belong to its worked example.

■ Cumulative frequency and box plots · Higher

A cumulative frequency counts observations below successive class boundaries. Read quartiles at one quarter, one half and three quarters of the total frequency. A box plot represents minimum, lower quartile, median, upper quartile and maximum.

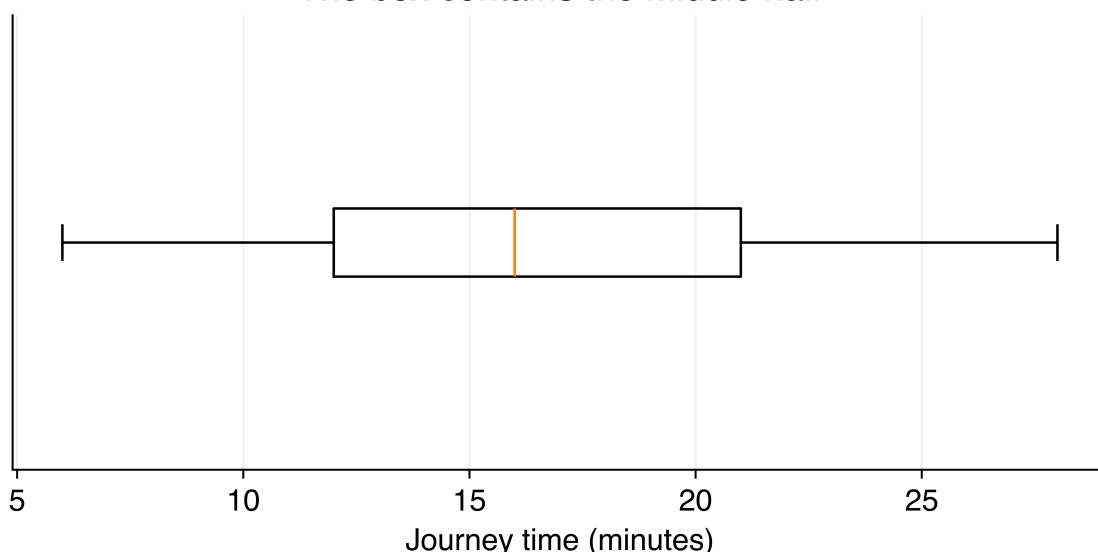
$$\text{IQR} = Q_3 - Q_1$$

For 80 observations, read Q_1 at cumulative frequency 20, median at 40 and Q_3 at 60. If $Q_1=12$, $Q_3=21$, then $\text{IQR}=9$. Compare the medians for typical journey time and the IQRs for consistency.

Plot against class boundaries rather than midpoints. Grouped quartiles are estimates. The range is sensitive to extremes; the IQR describes only the middle half.

Explain a comparison in the context of the measured quantity. An outlier rule may use $Q_1 - 1.5\text{IQR}$ and $Q_3 + 1.5\text{IQR}$; use the rule specified in the task rather than assuming every graph follows it.

The box contains the middle half



Original native-lesson illustration; labels belong to its worked example.

■ Samples, populations and justified comparisons · Foundation

Define the target population and variables before sampling. A sample is a subset; a census includes the whole population. Random selection reduces systematic selection bias but does not eliminate sampling variability or non-response. Primary data is collected for the present investigation; secondary data was collected by another source or purpose. Discrete data is counted; continuous data is measured.

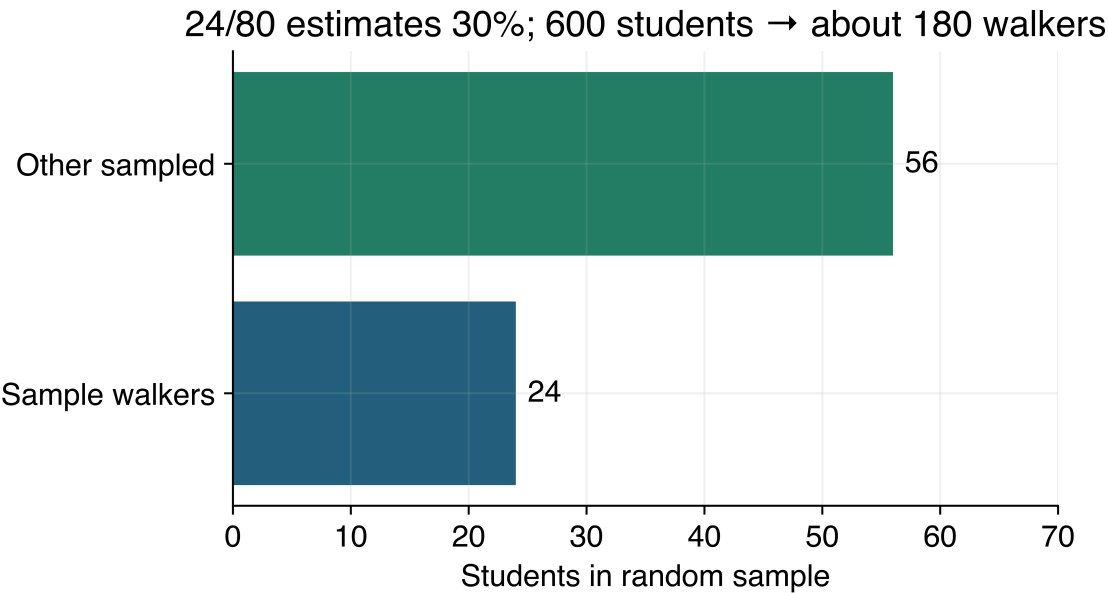
$$\hat{N} = N \frac{f}{n}$$

In a random sample of 80 students, 24 walk to school, so the observed proportion is $24/80=0.3$. Applied to a school of 600 it suggests about 180 walkers, with sampling uncertainty: it is an estimate rather than an exact count. A sports-club convenience sample may overrepresent active students. A voluntary online poll can miss people who do not respond; adding responses does not necessarily remove that bias. A study should record who could be selected, missing responses and the question wording. Number of siblings is discrete; travel time is continuous even if recorded to whole minutes. A school's published attendance records are secondary data for a new project; measuring new travel times produces primary data.

Sample size alone cannot fix biased selection. A precise calculated estimate need not be accurate for the target population. Recording continuous measurements as integers does not change the underlying variable type.

AQA S1/S4/S5 uses samples to describe populations and recognises limitations/data types. State what the sample supports and what might make extrapolating to the

whole population unreliable.



Original native-lesson illustration; labels belong to its worked example.

■ **Choosing displays, pie charts and time series · Foundation**

Use separate bars for categories, vertical line charts for discrete numerical values and time-ordered lines for a time series. Pie-chart sector angle is $\text{frequency}/\text{total} \times 360^\circ$. A pictogram needs a stated key and honest fractional symbols. Label axes/units and choose a scale that does not conceal the relevant variation. Continuous grouped distributions need histograms rather than categorical bars.

$$\theta = \frac{f}{n} 360^\circ$$

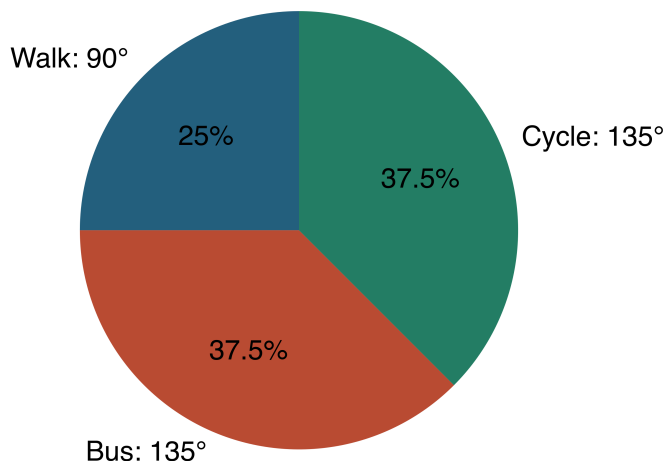
Of 40 students, 10 walk, 15 take the bus and 15 cycle. Pie angles are $90^\circ, 135^\circ, 135^\circ$, summing to 360° . With one pictogram symbol representing five students, the groups use 2,3,3 symbols. If one symbol instead means four students, ten walkers need 2.5 symbols. A daily attendance time series 32,35,33,36,34 can be joined in weekday order; the largest count is 36 and its range is 4. A frequency table for number of siblings 0,1,2,3 uses those numbers as positions on a vertical line chart, rather than treating widths as probabilities.

A pie chart must represent a complete total with non-overlapping categories. Pictogram keys control the count, not decorative icon size. A truncated axis may exaggerate a small difference; read the scale before comparing.

AQA S2 covers tables, categorical bars/pies/pictograms, discrete vertical lines and time-series tables/graphs. Explain why the representation fits the variable and the

question.

40 students: frequencies determine sector angles



Original native-lesson illustration; labels belong to its worked example.

■ Frequency summaries and grouped estimates · Foundation

For exact value frequencies, mean is $\text{sum}(\text{value} \times \text{frequency}) / \text{total frequency}$. Locate the median using cumulative counts. For interval data, use class midpoints to estimate a mean; the exact values are unknown. The modal class has greatest frequency, which need not be the tallest unequal-width histogram bar. Compare a suitable average and spread, in context, and state the limitations of grouping or outliers.

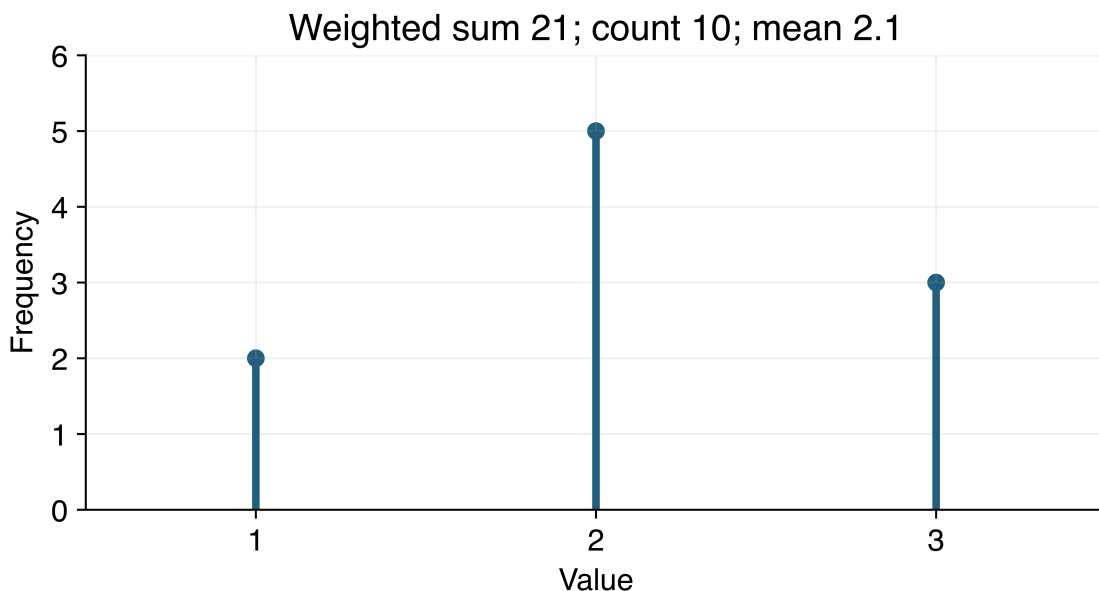
$$\bar{x} \approx \frac{\sum fm}{\sum f}$$

Values 1,2,3 with frequencies 2,5,3 give total 10 and weighted sum $2+10+9=21$, so mean is 2.1. The fifth and sixth observations are both 2, giving median 2 and mode 2. For continuous classes $0 \leq x < 10, 10 \leq x < 20, 20 \leq x < 30$ with frequencies 2,5,3, midpoints 5,15,25 give estimated sum $10+75+75=160$ and estimated mean 16. The modal class is $10 \leq x < 20$. The range of individual observations cannot be recovered exactly from these intervals. For values 2,4,4,6,24, mean is 8 and median 4; the unusually large value raises the mean. Comparing two groups should describe both a typical value and variation, rather than selecting whichever summary favours a claim.

The grouped mean is an estimate because all members are represented by a midpoint. A median is not found by averaging the class labels. Outliers can alter the mean/range substantially.

AQA S4/S5 Foundation includes appropriate mean/median/mode/modal class and

range, with grouped data and outlier awareness. Higher quartiles/box plots are taught separately.



Original native-lesson illustration; labels belong to its worked example.

■ Scatter graphs, correlation and cautious predictions · Foundation

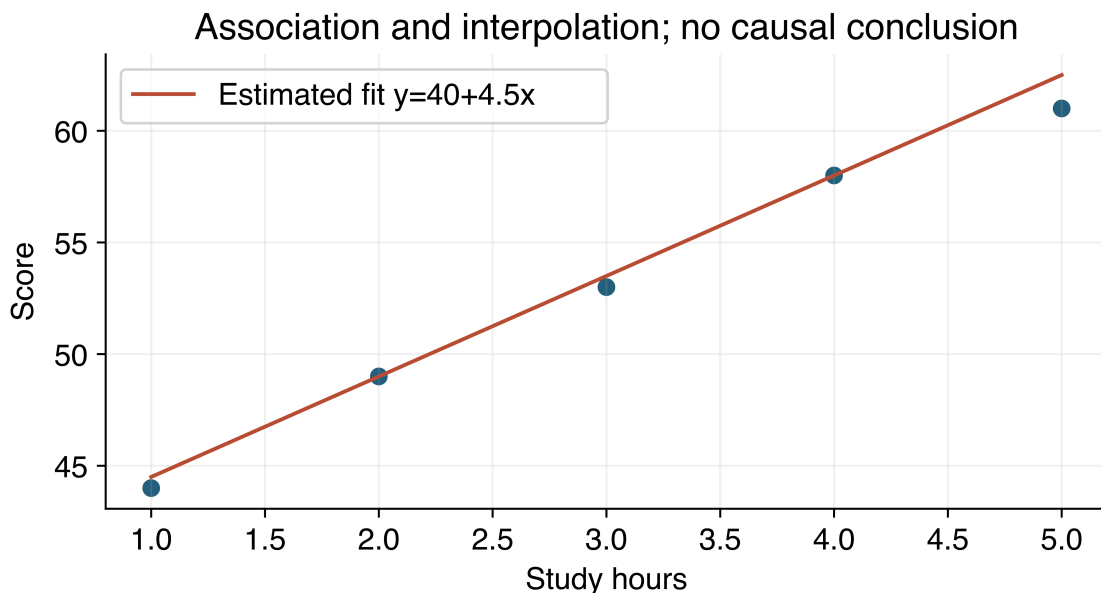
Plot paired observations as points, with one variable on each axis. Positive correlation rises, negative falls and no correlation has no clear trend. Strong/weak describes how tightly points follow a trend, not its steepness. Draw an estimated line of best fit through the middle of the pattern, balancing points around it; it need not pass through the origin. Interpolation stays within observed inputs; extrapolation goes outside them.

$$\hat{y} = 40 + 4.5x$$

Observed study hours 1,2,3,4,5 with scores 44,49,53,58,61 show a positive trend. An estimated line $y=40+4.5x$ predicts 53.5 at $x=3$ and 56.2 at $x=3.6$. These inputs are inside 1–5, so the predictions interpolate. At $x=10$ the line gives 85, but this extrapolation may fail if gains flatten or the group differs. Sleep hours and tiredness may show a negative correlation; an almost horizontal cloud with no ordered pattern may show little correlation. A shared cause such as prior preparation can affect both study time and score; the plot alone does not establish causation.

Do not join scatter points in observation order as though they formed a time series. A steep line need not imply strong correlation. An extrapolated formula value is a prediction whose assumptions need scrutiny.

AQA S6 Foundation includes correlation, estimated best-fit lines, predictions and dangers of extrapolation. State the observed input range, association direction/strength and the limits of a causal claim.



Original native-lesson illustration; labels belong to its worked example.

■ Histogram density and cumulative-frequency construction · Higher

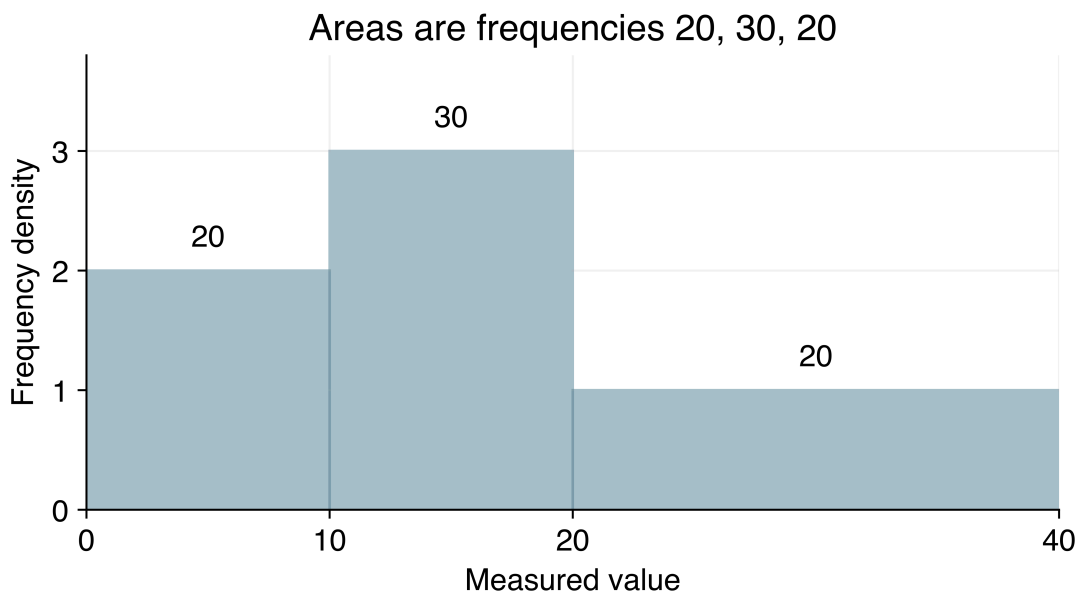
Use continuous adjoining class boundaries on the horizontal axis. Density=frequency/class width, so bar area recovers frequency. Unequal widths require this adjustment. A cumulative-frequency graph plots each upper class boundary against the running total, beginning at the first lower boundary with zero. Read percentiles at fractions of the total; values within classes are estimates.

$$d = \frac{f}{w}, \quad f = dw$$

For classes $0 \leq x < 10$, $10 \leq x < 20$, $20 \leq x < 40$ with frequencies 20, 30, 20, widths are 10, 10, 20 and densities 2, 3, 1. The last two bars have different heights but areas 30 and 20. Cumulative points are (0, 0), (10, 20), (20, 50), (40, 70). The median is at cumulative count 35; a straight-line estimate in the second class gives $10 + (35 - 20) / 30 \times 10 = 15$. Q1 is at 17.5 and Q3 at 52.5, giving corresponding interpolated estimates 8.75 and 22.5. Class midpoints 5, 15, 30 give estimated mean $(100 + 450 + 600) / 70 \approx 16.43$. A sketch that joins upper boundaries with a smooth curve can give slightly different readings, so use the stated precision and graph.

A histogram uses density on its vertical axis when widths differ. Plot cumulative totals at boundaries, not midpoints. Grouped percentiles depend on interpolation assumptions and are not exact individual measurements.

AQA S3 Higher includes equal/unequal-class histograms and cumulative-frequency graphs, with suitable interpretation. Explain why bar area represents frequency and label the graph axes/units.



Original native-lesson illustration; labels belong to its worked example.

■ Quartiles, box plots and distribution comparisons · Higher

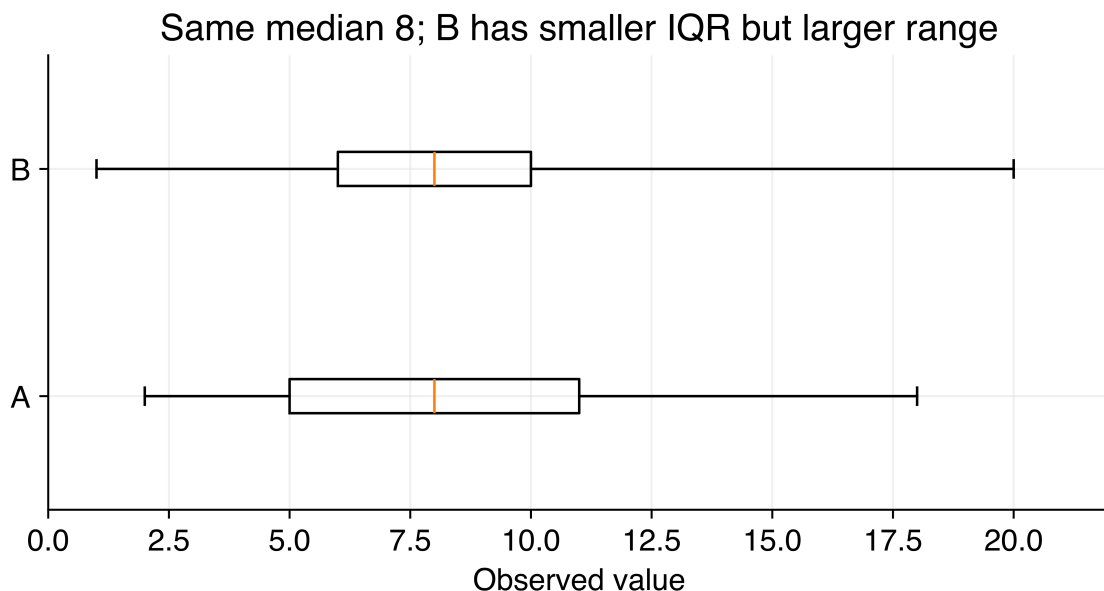
A box plot marks minimum, lower quartile Q_1 , median, upper quartile Q_3 and maximum, or identifies separately shown outliers according to the stated convention. $IQR = Q_3 - Q_1$ describes the middle half. Compare median and IQR in context; a smaller IQR suggests less middle-half variation, not necessarily a smaller full range. Quartile conventions vary for a finite raw list, so follow the task or supplied summaries.

$$IQR = Q_3 - Q_1$$

Group A has summary 2,5,8,11,18 and group B has 1,6,8,10,20. Both medians are 8. A has $IQR = 6$ and $range = 16$; B has $IQR = 4$ and $range = 19$. B's middle half is more consistent although its full range is larger. Under a stated $1.5 \times IQR$ outlier rule, A's fences are $5 - 9 = -4$ and $11 + 9 = 20$; a new value 23 lies above the upper fence. This rule is a specified convention, not a compulsory rule for every box plot. When using cumulative frequency, read Q_1 at $N/4$ and Q_3 at $3N/4$, then subtract; small reading errors affect the estimated IQR.

The median line need not be halfway along the box. Equal medians do not mean identical distributions. Whisker meanings depend on whether outliers are drawn separately; read the stated convention.

AQA S4 Higher adds box plots, quartiles and IQR to distribution comparisons. Name both centre and spread, with units and context, and avoid claims about all observations from just the box.



Original native-lesson illustration; labels belong to its worked example.

Original independent transfer

■ Foundation

A sample contains ten values with mean 6 and fifteen values with mean 10. Find the combined mean. A voluntary online survey suggests 70% of respondents cycle; explain why a much larger voluntary response need not establish that 70% of the whole school cycles.

■ Foundation worked solution

The totals are 60 and 150, so combined mean is $\bar{x} = (10 \cdot 6 + 15 \cdot 10) / (10 + 15) = 210 / 25 = 8.4$. Weight each group mean by its size. A voluntary survey can overrepresent people interested in cycling and miss nonrespondents. More responses reduce some random variation but do not automatically remove selection bias. State the target population and how students could be selected before generalising.

■ Higher

Grouped continuous data have classes $0 \leq t < 10$, $10 \leq t < 30$, $30 \leq t < 35$, with frequencies 8, 12, 10. Give histogram heights, cumulative frequencies and an estimated mean. Explain why the estimates cannot recover every original observation.

■ Higher worked solution

Class widths are 10,20,5. Frequency density is frequency divided by width, giving 0.8,0.6,2; bar areas return 8,12,10. Cumulative frequencies at upper class boundaries 10,30,35 are 8,20,30, starting from (0,0). Midpoints 5,20,32.5 give estimated total $8(5) + 12(20) + 10(32.5) = 605$, hence estimated mean $605/30 = 121/6$, about 20.17. Within-class positions are unknown; the midpoint model estimates their total. Height alone is not frequency when widths differ, and a cumulative curve depends on an interpolation assumption between boundaries.

Terms

range .
frequency density .
interquartile range .
population .
time series .
modal class .
correlation .