Representation of data

Stem-and-leaf diagrams, box-and-whisker plots, histograms, cumulative frequency graphs, measures of centre and spread, and the mean and standard deviation from totals, including coded totals.

Representation of data

Choosing a diagram

Diagram	Best for	Weakness
Stem-and-leaf	Small sets of data; keeps every value	Hard to use with a large set
Box-and-whisker plot	Comparing data sets; shows the median and spread	Loses the individual values
Histogram	Grouped continuous data; shows shape	Loses the individual values
Cumulative frequency graph	Estimating medians, quartiles and percentiles in large sets	Loses the shape at a glance

Stem-and-leaf diagrams

Values are split into a stem and a leaf, with the leaves in order and a key.

1	2 5 8
2	0 3 3 7
3	1 6

Key: 2 | 3 means 23

A **back-to-back** diagram shows two sets on a shared stem, with each one's leaves increasing outward. Always give a key for each side.

Median and quartiles

For n ordered values, the median is at position $\frac{n+1}{2}$. The lower and upper quartiles are the medians of the lower and upper halves.

> **TRY IT**

Note Academy

Find the median and quartiles of 3, 5, 7, 8, 9, 10, 12, 13, 15, 18, 20.

There are 11 values. The median is the 6th, so 10. The lower half is 3, 5, 7, 8, 9, so $Q_1 = 7$. The upper half is 12, 13, 15, 18, 20, so $Q_3 = 15$. The interquartile range is $15 - 7 = 8$.

Box-and-whisker plots

A box plot shows the five-number summary: least value, Q_1, median, Q_3 and greatest value.

When comparing data sets, make one statement on the centre and one on the spread, in context: "The median mark of group A is higher, so it did better on average. Its interquartile range is smaller, so its marks are more consistent."

Histograms

In a histogram, the **area** of each bar represents the frequency. The bars can have different widths, and the height is:

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

Continuous data has no gaps, so use the true class boundaries. Ages given as "10–19" years run from 10 to 20, and a length to the nearest cm given as "10–19" runs from 9.5 to 19.5.

TRY IT

Find the frequency density of a class of 30 with boundaries 10 and 20, and of a class of 12 with boundaries 20 and 24.

$\frac{30}{10} = 3$ and $\frac{12}{4} = 3$. The bars are the same height, although the frequencies differ.

Cumulative frequency graphs

Plot the cumulative frequency against the **upper class boundary**, and join the points with a smooth curve.

From the curve you can estimate the median, the quartiles and any percentile, as well as how many values lie above or below any point.

Note Academy

Measures of centre and spread

Measure	Use it when
Mean	The data has no extreme values; it uses every value
Median	The data has extreme values or is skewed; these do not affect it much
Mode	The data is categories, or the most common value matters
Range	A quick look at spread; affected by a single extreme value
Interquartile range	Spread of the middle half; not affected by extremes
Standard deviation	Spread using every value; pairs with the mean

Mean and standard deviation

Mean $\bar{x} = \frac{\sum x}{n}$. Standard deviation $\sigma = \sqrt{\frac{\sum x^2}{n} - \bar{x}^2}$.

The variance is the square of the standard deviation.

TRY IT

$n = 10$, $\sum x = 120$ and $\sum x^2 = 1640$. Find the mean and standard deviation.

$\bar{x} = 12$, and $\sigma = \sqrt{164 - 144} = \sqrt{20} = 4.47$.

Grouped data

Use the midpoint of each class for x, with f for frequency:

$$\bar{x} = \frac{\sum fx}{\sum f} \qquad \sigma = \sqrt{\frac{\sum fx^2}{\sum f} - \bar{x}^2}$$

These are estimates, because the midpoints stand in for the actual values.

Coded totals

Subtracting a constant a moves the mean by a but does not change the spread.

Note Academy

$\bar{x} = a + \frac{\sum(x-a)}{n}$ **and** $\sigma = \sqrt{\frac{\sum(x-a)^2}{n} - \left(\frac{\sum(x-a)}{n}\right)^2}$

TRY IT

$n = 20$, $\sum(x-50) = -40$ and $\sum(x-50)^2 = 480$. Find the mean and standard deviation of x.

$\bar{x} = 50 + \frac{-40}{20} = 48$, and $\sigma = \sqrt{\frac{480}{20} - (-2)^2} = \sqrt{20} = 4.47$.

Combining two data sets

Totals add, but means and standard deviations do not. Go back to $\sum x$ and $\sum x^2$ for each set, then add them.

TRY IT

Group A has $n = 10$, mean 12 and standard deviation 2. Group B has $n = 15$, mean 20 and standard deviation 3. Find the mean of the combined group, and its standard deviation.

For A: $\sum x = 120$ and $\sum x^2 = 10(2^2 + 12^2) = 1480$.

For B: $\sum x = 300$ and $\sum x^2 = 15(3^2 + 20^2) = 6135$.

Combined: $n = 25$, $\sum x = 420$ and $\sum x^2 = 7615$.

$\bar{x} = \frac{420}{25} = 16.8$ $\quad \sigma = \sqrt{\frac{7615}{25} - 16.8^2} = \sqrt{22.36} = 4.73$

Note Academy