

**FORMULAS YOU MUST KNOW**

Discrete data are counted (whole numbers); continuous data are measured (any value in a range).

Data types:

$k = \frac{\text{population}}{\text{sample size}}$, then take every k -th item from a random start r : $r, r + k, r + 2k, \dots$

Systematic sample:

The middle of the ordered list; for n values it sits in position $\frac{n+1}{2}$.

Median:

$\bar{x} = \frac{\sum fx}{\sum f}$ — multiply each value by its frequency, add, then divide by $\sum f$.

Mean:

Read $\bar{x}$ and σ_x directly; here $\sum f = n$ is the number of data points.

On the GDC:

Q_1 is the median of the lower half; Q_3 is the median of the upper half.

Quartiles:

A sample is a subset used to estimate a feature of the whole population.

Population vs sample:

$\bar{x} = \frac{\sum x}{n}$ — the balance point; every value counts.

Mean:

The most frequent value — a data set can have none, one or several.

Mode:

$\sigma = \sqrt{\frac{\sum f(x - \bar{x})^2}{\sum f}}$ — the typical distance of a value from the mean.

Standard deviation:

$x_{\max} - x_{\min}$ — simple, but wrecked by a single outlier.

Range:

$\text{IQR} = Q_3 - Q_1$ — the spread of the middle 50% of the data.

IQR:

KEY METHODS

- Systematic sampling is quick but becomes biased if the list hides a periodic pattern that lines up with k . Other techniques you should recognise: simple random, stratified (sample each group in proportion), quota and convenience.
- Add first, divide last — and on the GDC, 1-Var Stats returns $\bar{x}$ and the median together. Sanity-check that $\bar{x}$ lies between the smallest and largest value.
- A larger σ means less consistent data. The variance is simply σ^2 .
- The IQR ignores the extreme quarter at each end, so it is a resistant measure of spread — untroubled by

EXAM TRAPS TO AVOID

- The sampling interval is exactly population $\div$ sample size — no rounding up and no $+1$.
- The mean is dragged towards extreme values: a single outlier can pull it well away from the median. For skewed data or data with outliers, the median is the more representative centre.
- IB uses the population standard deviation: take σ_x from the calculator, never s_x (the sample value, which divides by $n - 1$).
- Sort the data first, and build the IQR from the quartiles ($Q_3 - Q_1$) — never from the maximum and minimum.

outliers, unlike the range.

- If the classes have unequal widths, the modal class is the one with the greatest frequency density (frequency $\div$ width), not simply the greatest frequency.
- Go up from the CF axis at the required position, across to the curve, then down to read the value. For grouped data the GDC will not give an exact median — the graph or interpolation is the method.
- Linear interpolation assumes the values are spread evenly across each class, so any percentile read this way is an estimate.

- Divide by $\sum f$ (the total frequency), never by the number of classes. And the answer is only an estimate — the original values are gone.
- It is $1.5 \times$ the IQR, not $1.5 \times Q_3$. And a value sitting exactly on a fence is not an outlier — the inequality is strict.
- Never average two means directly unless the groups are the same size — always weight by the counts n_1 and n_2 .

Cross-reference the **IB Math AI formula booklet** — anything not in it must be memorised.

Photon Academy · ibtuition.sg