Statistics

collect → present → summarise → locate → read grouped data → relate two variables

I Populations, samples & sampling

AA 4.1

Population = the whole group studied
Sample = a subset, used to conclude about the whole.

	population	sample
mean	μ	$\bar{x}$
s.d.	σ	s_n

Discrete = values you can list (shoe sizes) · **continuous** = values that fill an interval (times, masses).

The five techniques — only these five



- **Simple random**: everyone has the same chance · **Systematic**: random start, then every *k*th · **Convenience**: easiest to reach.
- **Quota**: fixed numbers from each subgroup · **Stratified**: from each subgroup in proportion to its size.

Watch out systematic + list with a periodic pattern → biased · convenience is almost always biased · a **large** sample does not repair a **biased** one.
Missing data and recording errors: a 0 may be a value or a measurement never taken — never average them together.

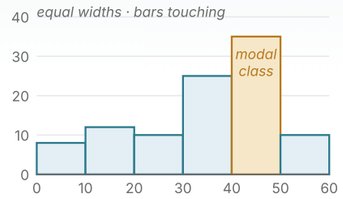
II Presenting data

AA 4.2

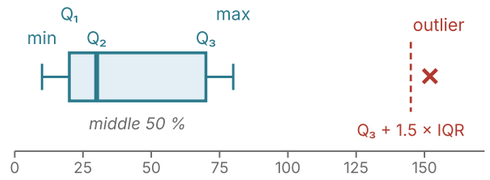
Frequency table

Each value or class with its frequency. Continuous classes as inequalities, **no gaps**: $10 \le x < 20$. **Equal widths only** — frequency density is not required.

Histogram



Box-and-whisker diagram



- Five numbers: min, Q_1, median Q_2, Q_3, max. Each of the four sections ≈ 25 % of the data.
- Outliers marked ×; the whisker stops at the last value that is **not** an outlier. Shows spread — **never** how many data.
- **Compare two box plots** : medians → IQRs → ranges, each in context. A box symmetric about its median suggests normal data.

III Central tendency & spread

AA 4.3

Where the data sit

- **mean** — sum ÷ number of values
- **median** — middle value, at position $\frac{n+1}{2}$ (not $\frac{n}{2}$)
- **mode** — most frequent · grouped: **modal class**

How spread out

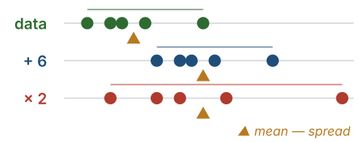
- **range** = max − min
- **IQR** = $Q_3 - Q_1$ booklet 4.2
- **standard deviation** σ GDC
- **variance** = σ^2

$\bar{x} = \frac{\Sigma f_i x_i}{n}$ with $n = \Sigma f_i$ booklet 4.3
check n first: total frequency, not number of rows

Grouped data ① mid-interval values (10 ≤ *x* < 20 → 15) ② use them as x_i ③ the mean is an **estimate**. gives $\bar{x} = 34.7$.

Effect of a change

	mean	σ	variance
$+ a$	$+ a$	unchanged	unchanged
$\times k$	$\times k$	$\times k$	$\times k^2$

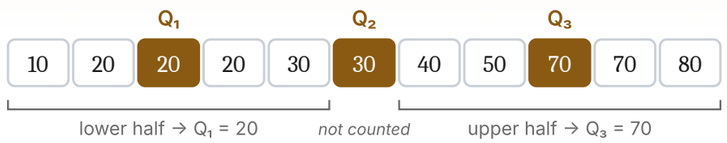


Variance from the GDC : square the **stored** σ, never the rounded one — 22.962…² = 527.

IV Quartiles, percentiles & outliers

AA 4.1 · 4.3

Quartiles by hand



① order, find the median Q_2 ② Q_1 = median of the values before it, Q_3 = median of those after ③ Q_2 is in **neither** half. Here IQR = 70 − 20 = 50. A GDC may give a slightly different quartile — both accepted.

Outlier — quote the definition

below $Q_1 - 1.5 \times \text{IQR}$ or above $Q_3 + 1.5 \times \text{IQR}$
*the two **fences***

fences 20 − 75 = −55 and 70 + 75 = 145 → every value lies inside: no outlier.

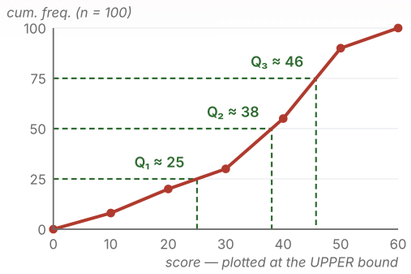
Watch out an outlier is **not automatically a mistake**: genuine value or recording error? Say which, in context.

Percentiles

Q_1 = 25th · **Q_2** = 50th · **Q_3** = 75th · 90th percentile: 90 % of the data lie below it

V Cumulative frequency

AA 4.2



Build & read

1. Running total of the frequencies.
2. Plot it against the **upper bound** of each class — never the mid-interval.
3. Join with a smooth curve.
4. Across, then down: $\frac{n}{4}$ → Q_1, $\frac{n}{2}$ → median, $\frac{3n}{4}$ → Q_3.

with $n = 100$: $Q_1 \approx 25$, median ≈ 38, Q_3 ≈ 46. Scoring more than 46: 100 − 75 = 25 students.

Every session

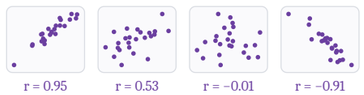
- "Within 15 marks of the median" → read the cumulative frequency at both ends, then **subtract**.
- "Only 10 % scored more than *k*" → 90 % below: across from $0.9n$, then down.

VI Correlation & regression

AA 4.4

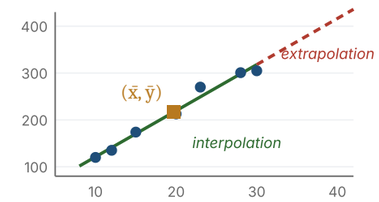
Scatter diagram

Describe **direction** (positive / negative / none) and **strength** (strong / moderate / weak). Pearson's r: $-1 \le r \le 1$ GDC



r near ±1: close to a line · near 0: no **linear** relationship.

Regression line of *y* on *x*



$y = ax + b$ through $(\bar{x}, \bar{y})$ · a, b, r GDC — not in the booklet
a: change in y per unit of x · b: value at x = 0 (may mean nothing)

$r = 0.990$: very strong positive · $y = 9.83x + 23.1$ · $x = 18$ (inside the data) → $y \approx 200$.

Watch out *r* only measures **linear** links · correlation ≠ causation (a third factor) · no **extrapolation** (*x* = 40 → 416: nothing supports it) · never predict *x* from *y* with this line · interpret *r* in context.

Command terms

State / Write down one word or number, no working · **Describe** name the technique and how · **Find** GDC, the value earns the mark · **Comment on / Interpret** a judgement with its reason, in context — "strong" alone earns nothing · **Hence** use the result just found.

Paper 1 · Paper 2

P1, no GDC: read diagrams, quote definitions, arithmetic on Σx (new mean when one value is removed).
P2, GDC: 1-Var Stats ($\bar{x}$, σ), grouped data, regression line. **Check *n* first** — a frequency list left out is the commonest error. 3 s.f.

Booklet or memory?

in the booklet $\bar{x} = \Sigma fx / n$ (4.3) · IQR (4.2)
GDC only σ · r · a · b
by heart median position · outlier fences · upper bounds · the five techniques

 Not affiliated with or endorsed by the International Baccalaureate Organization.