PHOTON ACADEMY · IB MATH AI HL CHEATSHEET

Bivariate Statistics

SL 4.4 · 4.10

ibtuition.sg · free

FORMULAS YOU MUST KNOW

Two variables recorded on each individual, displayed as points on the scatter diagram.
Bivariate data:

Positive (points rise to the right), negative (fall to the right), or none.
Direction:

Linear (points hug a straight line) or non-linear (they follow a curve).
Form:

How tightly the points cluster about that trend — weak, moderate or strong.
Strength:

$-1 \leq r \leq 1.$
Range:

$r > 0$ positive association, $r < 0$ negative.
Sign = direction:

$|r|$ near 1 is a strong linear fit; near 0 is weak or none.
Size = strength:

Enter the two lists, run 2-variable statistics / LinReg, and read r directly.
On the GDC:

$R^2 = r^2.$
Definition:

The proportion (or percentage) of the variation in y that the model explains.
Meaning:

$0 \leq R^2 \leq 1.$
Range:

A hidden third (confounding) variable can drive both x and y, manufacturing a correlation with no direct link between them.
Lurking variable:

KEY METHODS

- Describe the scatter before you calculate: it tells you whether a linear model (Pearson's r and a straight regression line) is even appropriate. A clear curve means a straight line is the wrong tool.
- These bands are descriptive, not official IB cut-offs. Always pair the strength word with the direction, e.g. "a strong positive linear correlation."
- Read R^2 as a percentage: $R^2 = 0.92$ means about 92% of the variation in y is explained by the regression on x, and the remaining 8% is due to other factors.
- To bank the mark, either name a plausible lurking variable or state that a causal link "cannot be concluded from correlation alone."
- Don't grind the sums by hand — enter the two lists, run LinReg, and the GDC returns a, b, r and R^2 in

EXAM TRAPS TO AVOID

- "Correlation" describes only the linear pattern. A strong curved relationship (say a neat parabola) can still give a correlation coefficient near 0 — the points genuinely are related, just not by a straight line.
- r measures linear association only. So $r \approx 0$ does not mean "no relationship" — it means "no straight-line relationship." Check the scatter for a curve before concluding the variables are unrelated.
- R^2 is always positive, so it loses the direction that r carries: a strong negative fit and a strong positive fit can share the same R^2. To go back from R^2 to r, take $\pm\sqrt{R^2}$ and choose the sign from the slope of the scatter. And an R^2 below 0 or above 1 always means a slip.
- "There is a strong correlation, therefore x causes y" is a guaranteed lost mark. Only a properly controlled

one screen.

- Interpret with units and context: e.g. "each extra hour of revision raises the predicted score by a marks." A bare number scores nothing.
- Predict from the full-accuracy line stored in the GDC, not from a and b chopped to one or two significant figures — small coefficient rounding compounds into a visibly wrong prediction.

experiment — not observational data — can support a causal claim.

- The line of y on x (for predicting y) is not the line of x on y. Use y on x whenever you predict y from a given x; the two lines differ and only agree when $|r| = 1$.
- The intercept b is only meaningful when $x = 0$ sits in or near the data. If $x = 0$ is far outside the range, reading y there is extrapolation and can be nonsense.
- Extrapolation can produce impossible answers — a test score above 100%, a negative mass. Always check your x lies within the data before trusting the output.