IB DP MATHS AI · TOPIC 4 STATISTICS AND PROBABILITY · SL 4.1–4.4, 4.10 · AHL 4.12, 4.13 · ~8 MIN READ

Statistics, correlation and regression

Sampling methods and bias, presenting data with histograms, cumulative frequency and box plots, mean, median, quartiles, standard deviation and outliers, scatter diagrams, Pearson's r, the regression line, Spearman's rank correlation, and for HL: designing data collection, reliability and validity, and non-linear regression with R^2.

Population & sample · Sampling methods & bias · Histograms · Cumulative frequency · Box plots · Mean, median, mode · Quartiles & IQR · Standard deviation

RS by Riyaz Sir
mentorriyaz.in

No tag = **SL and HL** · HL ONLY HL students only · FORMULA BOOKLET given in the exam

M1 A1 how marks are awarded · highlight = comes up again and again · Test yourself at the end

IN 30 SECONDS

- **Sampling:** a random sample gives every member an equal chance. Bias comes from the method (convenience, self-selection, leading questions), not from the sample being small.
- **Summaries:** mean, median, mode for centre; range, IQR = $Q_3 - Q_1$ and standard deviation σ for spread. An **outlier** is below $Q_1 - 1.5 \times \text{IQR}$ or above $Q_3 + 1.5 \times \text{IQR}$.
- **Pearson's r** ($-1 \le r \le 1$) measures **linear** correlation. The regression line of y on x, $y = ax + b$, predicts y from x, and passes through $(\bar{x}, \bar{y})$.
- **Spearman's r_s** uses ranks, so it measures any **monotonic** relationship (always increasing or always decreasing), linear or not.
- **HL:** design reliable, valid data collection, and choose between models with the coefficient of determination R^2.

BY THE END YOU CAN

- Distinguish populations and samples, and choose sampling methods; identify bias.
- Present data with frequency tables, histograms, cumulative frequency graphs and box plots.
- Find and interpret mean, median, mode, quartiles, IQR and standard deviation, and identify outliers.
- Use scatter diagrams, Pearson's r and the regression line of y on x, and interpret them in context.
- Calculate and interpret Spearman's rank correlation coefficient.
- Design data collection, and discuss reliability and validity. HL
- Fit non-linear regression models and compare them using R^2. HL

Riyaz Sir

1 Sampling and data

Method	How	Watch out for
simple random	every member equally likely (random numbers)	needs a full list of the population
systematic	every kth member from a random start	patterns in the list
stratified	sample from each group in proportion to its size	needs group sizes
quota	fill a fixed number from each group, not at random	interviewer bias
convenience	whoever is easiest to reach	rarely representative

WORKED EXAMPLE

A school has 300 students in Year 11 and 200 in Year 12. A stratified sample of 40 is chosen. How many from each year?

1. Year 11: $\frac{300}{500} \times 40 = \mathbf{24}$ M1 A1
2. **Year 12: $\frac{200}{500} \times 40 = 16$** A1

- **Discrete** data is counted; **continuous** data is measured and can be grouped into classes. Outliers should be checked: a typing error is removed, a genuine extreme value is kept.

2 Summary statistics and box plots

WORKED EXAMPLE

Commute times (minutes) are 12, 15, 15, 18, 20, 22, 25, 31, 45. Find the mean, median, quartiles and decide whether 45 is an outlier.

1. mean = $\frac{203}{9}$ = **22.6 min**; median = **20** A1 A1
2. $Q_1 = 15$, $Q_3 = 28$ (GDC), IQR = 13 A1
3. **upper fence = 28 + 1.5 × 13 = 47.5, so 45 is not an outlier** M1 R1

- **Effect of changes:** adding k to every value adds k to the mean and median and leaves σ and IQR unchanged; multiplying by k multiplies all of them by k.
- **Grouped data:** estimate the mean with class midpoints; read the median and quartiles from a cumulative frequency curve at n/2, n/4 and 3n/4.

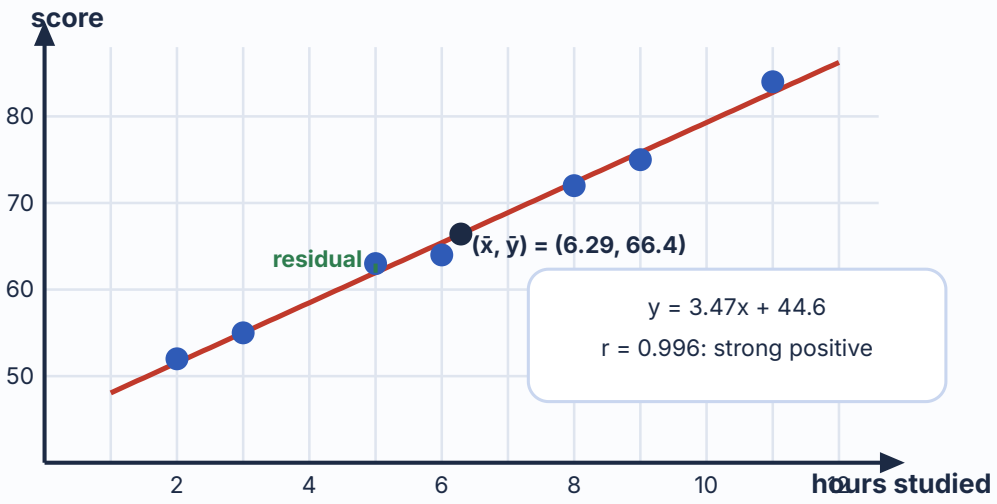
GDC (CALCULATOR) METHOD

Use 1-variable statistics (with a frequency list for tables). Write down $\bar{x}$, σ, Q_1, median and Q_3 with their labels: "from GDC" plus the values earns the marks.

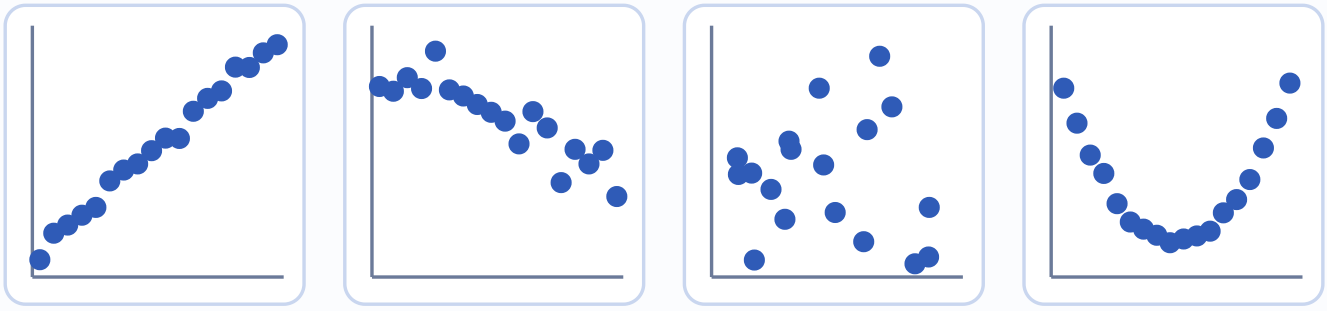
COMMON MISTAKE

Comparing two data sets with one number. Always compare a measure of centre **and** a measure of spread, in context: "Group A's times are longer on average (median 20 vs 16) and more spread out (IQR 13 vs 6)".

3 Correlation and regression



The regression line passes through the mean point. A residual is the vertical gap between a point and the line.



strong + **weak –** **none (r ≈ 0)** **non-linear**

Pearson's r describes linear patterns only. Look at the scatter diagram before trusting r.

r	Description
$0.75 \le \lvert r \rvert \le 1$	strong
$0.5 \le \lvert r \rvert < 0.75$	moderate
$0.25 \le \lvert r \rvert < 0.5$	weak
$\lvert r \rvert < 0.25$	very weak or none

WORKED EXAMPLE

Hours studied x = 2, 3, 5, 6, 8, 9, 11 and scores y = 52, 55, 63, 64, 72, 75, 84. Find r, the regression line, and predict the score for 7 hours.

1. linear regression on the GDC: **r = 0.996**, strong positive linear correlation A1 R1
2. **y = 3.47x + 44.6** A1
3. **y(7) = 3.47(7) + 44.6 = 68.9** A1

- **Interpreting a and b in context:** each extra hour of study adds about 3.47 marks; a student who studies 0 hours is predicted 44.6 (outside the data, so unreliable).

COMMON MISTAKE

Extrapolating. 7 hours is inside the data range (interpolation), so the prediction is reliable. Predicting for 20 hours would give 114, which is impossible on a test out of 100.

- **Correlation is not causation:** ice-cream sales and drownings correlate because both rise in hot weather.
- **Which line?** Use the y-on-x line only to predict y from x. The IB course doesn't ask you to predict x from y with it.

4 Spearman's rank correlation

WORKED EXAMPLE

A judge's scores p = 10, 20, 30, 40, 50, 60 and a second judge's q = 3, 7, 5, 12, 11, 15. Find Spearman's r_s.

1. ranks of p: 1, 2, 3, 4, 5, 6; ranks of q: 1, 3, 2, 5, 4, 6 M1
2. Pearson's r of the ranks (GDC) M1
3. **r_s = 0.886, strong agreement** A1 R1

- **Ties:** give tied values the mean of the ranks they share. Two values tied for 2nd and 3rd both get 2.5.

EXAM TIP

Choose Spearman when the relationship is monotonic but curved, or the data are ranks; choose Pearson for linear relationships with measured data.

5 Designing data collection HL

- **Reliability:** would you get similar results if you repeated it (test–retest, parallel forms)? **Validity:** does it measure what it claims to (content validity, criterion validity)?
- **Choosing a sample size and method:** larger samples reduce random error but not bias. Pilot your questionnaire, avoid leading questions, and decide how to handle missing data.

6 Non-linear regression and R^2 HL

- **Least squares:** the regression model makes the sum of squared residuals as small as possible.
- **R^2** ($0 \le R^2 \le 1$) is the proportion of the variation in y explained by the model. For a linear model $R^2 = r^2$; here $0.996^2 = 0.993$.
- **Comparing models:** fit quadratic, cubic, exponential, power and sinusoidal regressions on the GDC and compare R^2, but also ask whether the model makes sense in context. A high R^2 with impossible predictions is a bad model.

SYLLABUS NOTE

Pearson's r, the regression line and Spearman's rank stay in AI for 2029. Only the **hypothesis test for the correlation coefficient** (HL) is removed; see Estimation, Poisson, hypothesis tests and Markov chains.

LINKING QUESTION

The regression line is a linear model (Lines, functions and graphs), and non-linear regression fits the models of Modelling with functions. Linearising with logs (Composites, transformations and linearising data) turns exponential and power data into straight lines. Chi-squared and t-tests decide whether a pattern is significant.

IDEA FOR YOUR IA / EXPLORATION

Collect bivariate data yourself (arm span against height, screen time against sleep, price against floor area of flats in Mumbai) and analyse the correlation with Pearson's and Spearman's coefficients, the regression line and its limitations.

Test yourself — cover the answers first

QUICK CHECK

1. Which sampling method picks every 10th name from a random start?
2. Add 5 to every value. What happens to σ?
3. r = −0.92. Describe the correlation.
4. Should you use the y-on-x line to predict x?
5. Two values tie for ranks 4 and 5. What rank does each get?
6. HL A linear model has r = 0.8. Find R^2.

EXAM-STYLE

7. The data set 4, 7, 8, 9, 10, 11, 12, 13, 30 has $Q_1 = 7.5$ and $Q_3 = 12.5$. Is 30 an outlier? Find the median and IQR.
8. For a regression line $y = -2.1x + 95$ with $r = -0.85$, interpret the gradient in context if x is the number of absences and y is the exam score, and comment on predicting the score for 60 absences.

Show answers

1. systematic
2. no change
3. strong negative linear correlation
4. no
5. 4.5
6. 0.64
7. IQR = 5; upper fence 12.5 + 7.5 = 20
   30 > 20, so **30 is an outlier**
   median **10**, IQR **5**
8. each extra absence is associated with a score about 2.1 marks lower
   60 absences gives −31, which is impossible: it is far outside the data, so the prediction is not reliable
   (and correlation doesn't prove absences cause lower scores)

Stuck on a question?

Send it to me on WhatsApp. I read every message.

Share with a friend

Ask Riyaz Sir

Want regular help? Send this to your parent →