MEGA LECTURE

IB DP MATHEMATICS

Applications and Interpretation (AI)

Topic 4 Statistics and Probability

Revision Notes · Standard and Higher Level

Fahad H. Ahmad

+92 323 509 4443 | Megalecture.com

Original notes prepared for the IB Diploma Programme Mathematics: Applications and Interpretation course (first assessment 2021)

What Topic 4 requires

Statistics and probability is the largest topic in Mathematics: Applications and Interpretation, and it is examined heavily on both papers with a graphic display calculator (GDC) assumed throughout. Use this checklist as a final sweep before the exam; items marked (HL) are additional Higher Level content.

Area	You should be able to...
Sampling and data	Describe sampling techniques, recognise bias, and classify data as discrete or continuous.
Summary statistics	Find central tendency and dispersion for raw and grouped data; identify outliers with the 1.5 × IQR rule.
Presenting data	Read and draw histograms, cumulative frequency curves and box-and-whisker plots.
Bivariate data	Interpret Pearson's r and Spearman's r_s; use the least-squares regression line to predict.
Probability	Apply the addition and conditional rules using Venn, tree and table diagrams.
Distributions	Work with E(X), the binomial B(n, p) and the normal distribution on the GDC.
Hypothesis testing	Run the χ^2 tests and the t-test; state hypotheses and interpret the p-value.
(HL) Further work	Variance of a random variable, the Poisson distribution, combining random variables, confidence intervals and Type I / II errors.

> **Calculator note:** Almost every numerical answer in this topic comes from a GDC menu (1-Var Stats, regression, distribution and test menus). Marks are awarded for correct set-up, correct notation and a correct interpretation — so always write the values you feed in and a sentence explaining the output.

1. Sampling and types of data

A **population** is the entire group we want to study; a **sample** is the subset we actually measure. We use a sample because studying a whole population is usually too slow or too expensive. A good sample is **representative**, so that conclusions drawn from it generalise reliably to the population.

Sampling techniques

Technique	How it works	Watch out for
Simple random	Every member has an equal, independent chance of selection (e.g. numbered names drawn at random).	Needs a full list of the population.
Systematic	Order the population and take every k-th member from a random start.	Hidden periodic patterns can bias it.
Stratified	Split the population into strata (groups) and sample each in proportion to its size.	Requires strata sizes to be known.
Quota	Fill fixed quotas for each group, but choose members non-randomly.	Selection within a quota is subjective.

Technique	How it works	Watch out for
Convenience	Sample whoever is easiest to reach.	Usually strongly biased; least reliable.

Bias and reliability

Bias is any systematic tendency for a sample to over- or under-represent part of the population. Common causes are non-random selection, non-response, leading questions and self-selection. A **reliable** result is one that would recur if the study were repeated; larger, well-designed random samples are more reliable.

Discrete and continuous data

- **Discrete** data can take only separate values, usually from counting (number of siblings, goals scored).
- **Continuous** data can take any value in a range, usually from measuring (height, time, mass) and is recorded to a chosen accuracy.

> **Exam note:** To describe a stratified sample fully you must state the sampling fraction. For a sample of size n from a population of size N, take a fraction n/N from every stratum — see practice question 1.

2. Central tendency and dispersion

A data set is summarised by a measure of the centre and a measure of the spread. Choosing the right pair, and reading them from the GDC, is a routine exam skill.

Measure	Meaning	When to prefer it
Mean, $\bar{x}$	Sum of values divided by how many: $\bar{x} = (\Sigma f\, x) / (\Sigma f)$	Symmetric data with no extreme outliers.
Median	Middle value once the data is ordered	Skewed data or when outliers are present.
Mode	Most frequent value; the modal class for grouped data	Categorical data or the most 'typical' value.

Grouped data

When data is grouped into class intervals we no longer know the exact values, so we estimate the mean using the **mid-interval value** of each class as x. The **modal class** is the interval with the highest frequency (or highest frequency density if class widths differ). The median class is the one containing the (n/2)-th value.

Measures of dispersion

- **Range** = largest − smallest value — simple but sensitive to outliers.
- **Quartiles** Q_1, Q_2 (median), Q_3 split ordered data into quarters; the **interquartile range** IQR = $Q_3 - Q_1$ measures the spread of the middle 50%.
- **Standard deviation** σ measures the typical distance of values from the mean. Read it directly from the GDC 1-Var Stats output (the σ_x value); you are not expected to compute it by hand.

Outliers

A value is treated as an **outlier** if it lies below $Q_1 - 1.5 \times \text{IQR}$ or above $Q_3 + 1.5 \times \text{IQR}$. Always test both fences and quote the boundary you used.

Worked example 1 — grouped mean, modal class and standard deviation

The table shows the time t (minutes) that 40 students spent on homework one evening.

Time t (min)	$0 \le t < 10$	$10 \le t < 20$	$20 \le t < 30$	$30 \le t < 40$	$40 \le t < 50$
Frequency	4	9	13	10	4

Mid-interval values are 5, 15, 25, 35, 45. Entering these against the frequencies in the GDC 1-Var Stats gives:

Mean $\bar{x} = (\Sigma f x)/(\Sigma f) = (20 + 135 + 325 + 350 + 180)/40 = 1010/40 =$ **25.25 min**.

Modal class = **$20 \le t < 30$** (highest frequency, 13). The 20th value also falls here, so the median class is $20 \le t < 30$.

Standard deviation from the GDC: $\sigma_x =$ **11.3 min** (3 s.f.). [Check: $\Sigma f x^2 = 30600$, so variance = $30600/40 - 25.25^2 = 127.44$, and $\sqrt{127.44} = 11.29$.]

3. Presenting data

Each diagram answers a different question, and part of the skill is reading values back off a chart someone else has drawn.

Diagram	Best for	Read off it
Frequency table	Organising raw counts	Totals, modal value, cumulative counts.
Histogram	Shape of a continuous distribution	Modal class, skew; area proportional to frequency.
Cumulative frequency curve	Finding position-based measures	Median, quartiles, percentiles, IQR.
Box-and-whisker plot	Comparing spread of two or more sets	Minimum, Q_1, median, Q_3, maximum and outliers.

Using a cumulative frequency curve

Plot the cumulative frequency against the **upper** boundary of each class and join with a smooth curve. To find the median, go up from n/2 on the vertical axis, across to the curve and down to the value; use n/4 and 3n/4 for Q_1 and Q_3. To find how many data points lie below a value, reverse the process.

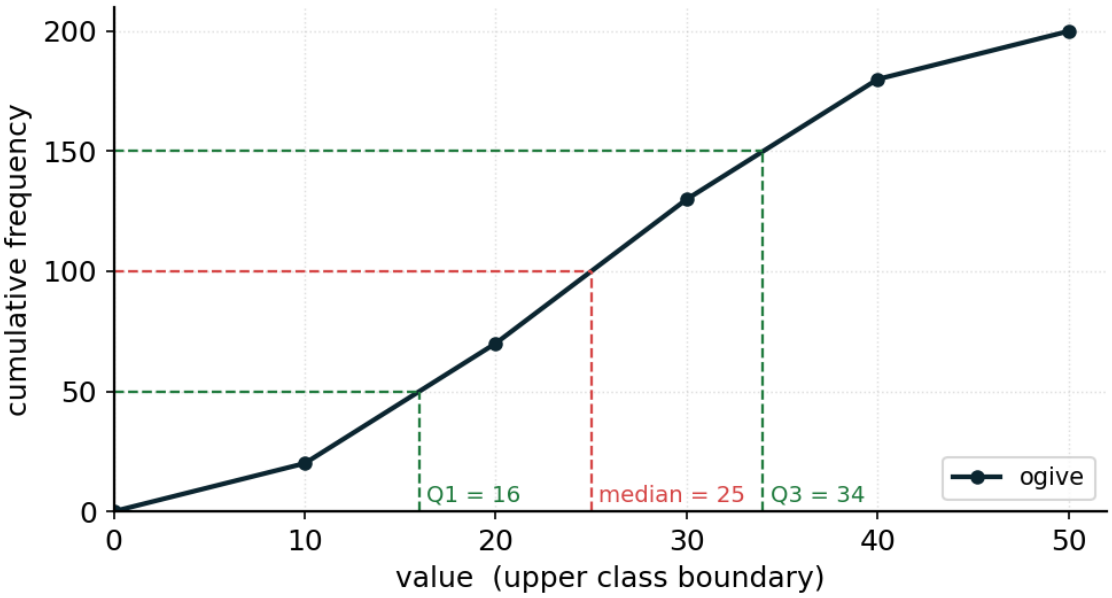

Figure 1. Cumulative frequency ogive for 200 measurements. Reading across from n/2 = 100, n/4 = 50 and 3n/4 = 150 gives median ≈ 25, Q_1 ≈ 16 and Q_3 ≈ 34, so IQR = 18.

Box-and-whisker plots

A box plot displays the five-number summary. The box spans the IQR, the line inside marks the median, and the whiskers reach the smallest and largest non-outlier values; outliers are plotted as separate points. A longer right whisker or a median sitting left of centre indicates positive (right) skew.

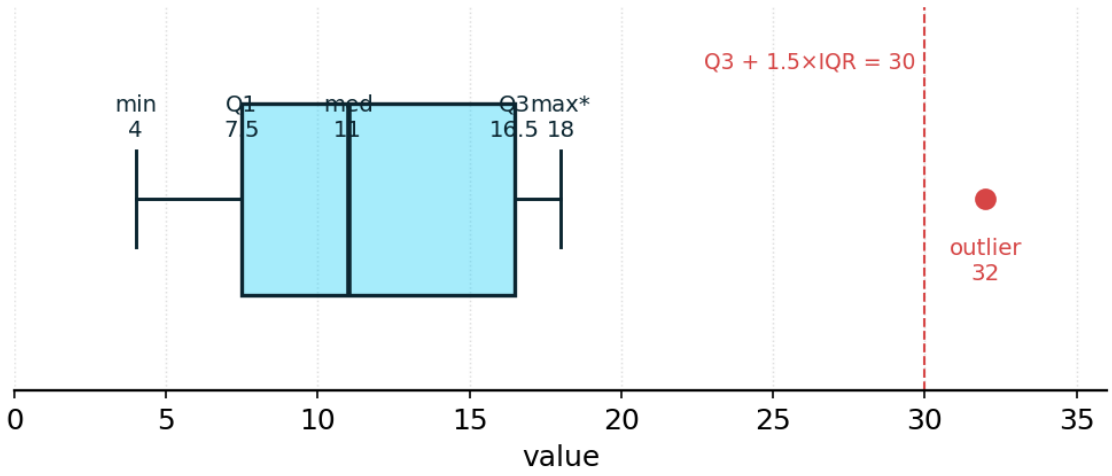

Figure 2. Box-and-whisker plot of 4, 7, 8, 9, 11, 12, 15, 18, 32. Five-number summary: min 4, Q_1 7.5, median 11, Q_3 16.5, largest non-outlier 18; IQR = 9. Since 32 > Q_3 + 1.5 × IQR = 30 it is plotted as an outlier (red).

Worked example 2 — reading a cumulative frequency curve

The cumulative frequencies of 200 measurements are given at each upper class boundary.

Value ≤	10	20	30	40	50
Cumulative freq.	20	70	130	180	200

With n = 200, read the median at the 100th value, Q_1 at the 50th and Q_3 at the 150th, interpolating within the relevant class:

Median ≈ 20 + (100−70)/(130−70) × 10 = **25**; Q_1 ≈ 10 + (50−20)/(70−20) × 10 = **16**; Q_3 ≈ 30 + (150−130)/(180−130) × 10 = **34**.

Hence IQR = 34 − 16 = **18**.

4. Bivariate data, correlation and regression

Bivariate data records two variables for each item. A **scatter diagram** is always the first step: it reveals the direction, form and strength of any relationship, and flags outliers before you calculate anything.

Pearson's product-moment correlation coefficient, r

The value r measures the strength and direction of a **linear** relationship. It satisfies $-1 \le r \le 1$: values near ±1 mean a strong linear relationship, values near 0 mean little linear relationship. Read r from the GDC regression output.

| r | value	Interpretation of linear correlation
0.00 – 0.25	Very weak or none
0.25 – 0.50	Weak
0.50 – 0.75	Moderate
0.75 – 1.00	Strong (near ±1 = very strong)

Spearman's rank correlation coefficient, r_s

Spearman's r_s measures the strength of a **monotonic** relationship (consistently increasing or decreasing, not necessarily linear). Rank each variable separately, find the difference d in ranks for each item, and use:

$$r_s = 1 - [\,6\,\Sigma d^2\,] / [\,n(n^2 - 1)\,]$$

Because it uses ranks, r_s is resistant to outliers and works with ordinal data. It is also less affected by a non-linear (but still monotonic) curve than r is. If two values tie, give each the mean of the ranks they share.

Worked example 3 — Spearman's rank correlation

Seven students' weekly study hours (x) and test scores (y) are recorded. Find r_s.

Hours x	5	8	3	9	6	2	7
Score y	60	72	50	88	65	40	58
Rank x	5	2	6	1	4	7	3
Rank y	4	2	6	1	3	7	5
$d = R_x - R_y$	1	0	0	0	1	0	–2
d^2	1	0	0	0	1	0	4

Ranks use 1 for the largest value. Then $\Sigma d^2 = 1+0+0+0+1+0+4 = 6$ and $n = 7$.

$r_s = 1 - (6 \times 6)/(7 \times (49 - 1)) = 1 - 36/336 = 1 - 0.1071 =$ **0.893** (3 s.f.).

A strong positive monotonic association: students who study more tend to score higher.

Least-squares regression line (y on x)

When a scatter diagram looks linear and | r | is reasonably large, fit the least-squares regression line of y on x. It minimises the sum of the squared vertical distances from the points to the line and always passes through the mean point $(\bar{x}, \bar{y})$. The GDC gives it as $y = a + bx$ (or $y = mx + c$).

- Use the line to **predict** a y-value from an x-value.
- **Interpolation** (predicting inside the range of the data) is reliable.
- **Extrapolation** (predicting outside the range) is unreliable, and so is any prediction when | r | is small.
- The gradient b is the change in y per unit increase in x; the intercept a is the predicted y when x = 0 (often meaningless in context).

Worked example 4 — regression and prediction

Daily maximum temperature x (°C) and ice-creams sold y are recorded on seven days.

x (°C)	14	16	18	20	22	24	26
y (sold)	20	27	35	38	47	52	60

The GDC gives r = 0.997 (very strong positive) and the y-on-x line $y = 3.25x - 25.1$.

Predict sales at 21 °C: $y = 3.25 \times 21 - 25.1 = 68.25 - 25.1 \approx$ **43 ice-creams**. This is interpolation (21 lies inside 14–26), so it is reliable.

Predicting at 35 °C would give ≈ 89, but this is **extrapolation** well outside the data — treat it with caution. Interpretation of b: each extra 1 °C is associated with about 3.25 more ice-creams sold.

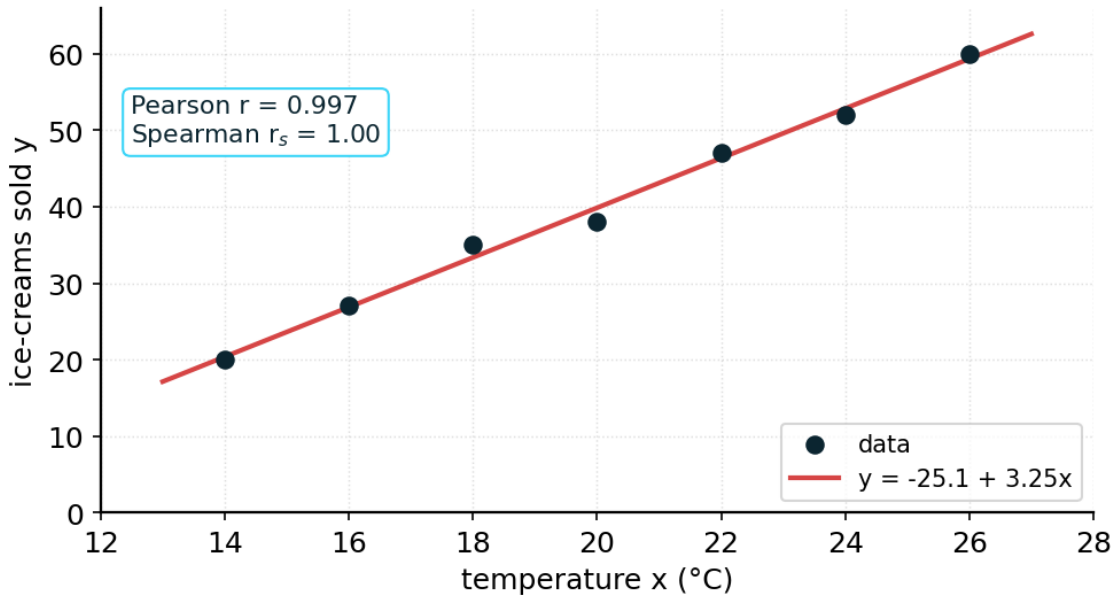

Figure 3. Scatter plot of the ice-cream data with the least-squares line y = a + bx computed with numpy: a = −25.1, b = 3.25, Pearson r = 0.997. Spearman's r_s = 1.00 here (the ranks of x and y match exactly), confirming a perfect monotonic — and near-perfect linear — relationship.

> **Correlation is not causation:** A large | r | shows the variables move together, not that one causes the other. A third (confounding) variable may drive both. Never claim cause from a correlation coefficient alone.

5. Probability

The **sample space** U is the set of all possible outcomes. For an event A,

P(A) = (number of outcomes in A) / (total number of equally likely outcomes), with 0 ≤ P(A) ≤ 1.

The **complement** A′ (A does not happen) satisfies P(A′) = 1 − P(A). This is often the quickest route: for 'at least one', work out P(none) and subtract from 1.

Combining events

The addition rule (for any two events):

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Relationship	Definition	Consequence
Mutually exclusive	A and B cannot both occur: P(A ∩ B) = 0	P(A ∪ B) = P(A) + P(B)
Independent	One occurring does not change the other's probability	P(A ∩ B) = P(A) × P(B)
Conditional	Probability of A given B has occurred	P(A | B) = P(A ∩ B) / P(B)

Diagrams

- **Venn diagrams** show overlaps and complements; fill in the intersection first, then work outwards.

- **Tree diagrams** suit sequential events; multiply along branches, add between branches, and each set of branches sums to 1.
- **Two-way tables** suit two categorical variables and feed directly into conditional probabilities and the χ^2 test.

Worked example 5 — conditional probability with a tree

A factory makes bolts on two machines. Machine A produces 60% of output with a 3% defect rate; machine B produces 40% with a 5% defect rate. A bolt is chosen at random.

(a) P(defective) = P(A)P(def | A) + P(B)P(def | B) = 0.60 × 0.03 + 0.40 × 0.05

= 0.018 + 0.020 = **0.038**.

(b) Given the bolt is defective, the probability it came from machine A is

P(A | def) = P(A ∩ def)/P(def) = 0.018 / 0.038 = **0.474** (3 s.f.).

Notice P(A | def) ≠ P(A): learning the bolt is defective changes the probability, so 'defective' and 'from A' are not independent.

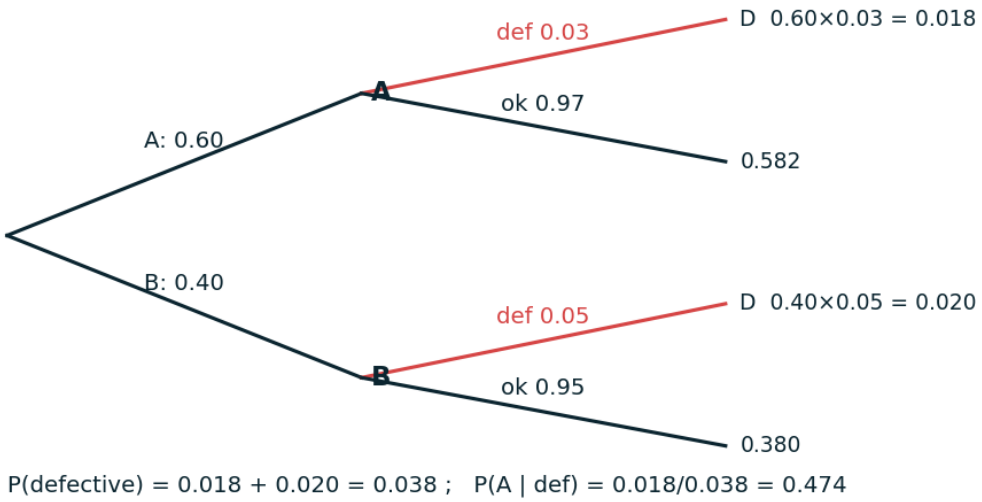

Figure 4. Tree diagram for the two-machine problem. The branch probabilities at each node sum to 1 (0.60 + 0.40; 0.03 + 0.97; 0.05 + 0.95). Multiplying along and adding the defective paths gives P(defective) = 0.018 + 0.020 = 0.038, and P(A | defective) = 0.018/0.038 = 0.474.

Mutually exclusive ≠ independent: These are different ideas that are easy to confuse. Mutually exclusive events have no overlap, so if one happens the other cannot — which actually makes them dependent. Two events with non-zero probability cannot be both mutually exclusive and independent.

6. Discrete random variables and distributions

A **discrete random variable** X takes separate values with probabilities that sum to 1: ΣP(X = x) = 1. The **expected value** is the long-run average outcome,

$$E(X) = \Sigma\, x \cdot P(X = x)$$

A game is **fair** when its expected gain is zero. E(X) need not be a value the variable can actually take.

Worked example 6 — expected value and variance from a table

A spinner scores X points with the probability distribution below.

x	1	2	3	4
P(X = x)	0.4	0.3	0.2	0.1

The probabilities sum to 1, as required. E(X) = Σ x P(X = x) = 1(0.4) + 2(0.3) + 3(0.2) + 4(0.1) = **2.0**.

(HL) $E(X^2) = 1(0.4) + 4(0.3) + 9(0.2) + 16(0.1) = 5.0$, so $\text{Var}(X) = E(X^2) - [E(X)]^2 = 5.0 - 2.0^2 =$ **1.0**, and the standard deviation is $\sqrt{1.0} = 1.0$.

The binomial distribution B(n, p)

Use B(n, p) when there are n independent trials, each with only two outcomes (success / failure) and a constant probability p of success. Then X, the number of successes, has **mean E(X) = np**. Compute probabilities on the GDC with binompdf (for P(X = r)) and binomcdf (for P(X ≤ r)).

Worked example 7 — binomial distribution

A multiple-choice quiz has 10 questions, each with 4 options. A student guesses every answer, so X ~ B(10, 0.25).

Mean number correct: E(X) = np = 10 × 0.25 = **2.5**.

P(exactly 3 correct) = binompdf(10, 0.25, 3) = **0.250** (3 s.f.).

P(at least 3 correct) = 1 − P(X ≤ 2) = 1 − binomcdf(10, 0.25, 2) = 1 − 0.526 = **0.474**.

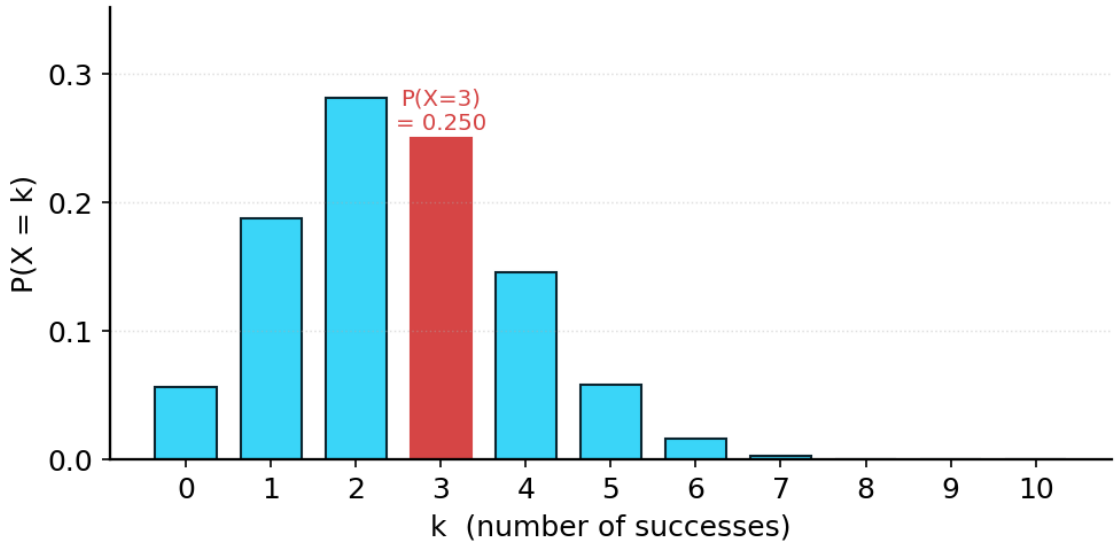

Figure 5. The binomial distribution B(10, 0.25). Each bar height is P(X = k) from binompdf; the probabilities sum to 1 and the mode is k = 2. The highlighted bar shows P(X = 3) = 0.250, matching the worked example.

The normal distribution

A continuous variable X ~ N(μ, σ^2) has a symmetric bell-shaped curve centred on the mean μ, with total area 1. About 68% of data lies within 1σ of the mean, 95% within 2σ and 99.7% within 3σ. Standardise with the **z-score**,

$$\mathbf{z = (x - \mu) / \sigma}$$

which counts how many standard deviations x is from the mean. Find probabilities with the GDC normalcdf (lower, upper, μ, σ); to go from a probability back to a value, use the **inverse normal** invNorm.

Worked example 8 — normal distribution and z-scores

Adult heights are modelled by X ~ N(170, 8^2) cm.

(a) P(X < 180): z = (180 − 170)/8 = 1.25, giving normalcdf ≈ **0.894**.

(b) P(X > 160): z = (160 − 170)/8 = −1.25; by symmetry the probability is again **0.894**.

(c) The tallest 10% exceed what height? Solve P(X > k) = 0.10, i.e. invNorm(0.90) = z = 1.282, so k = 170 + 1.282 × 8 = **180 cm** (3 s.f.).

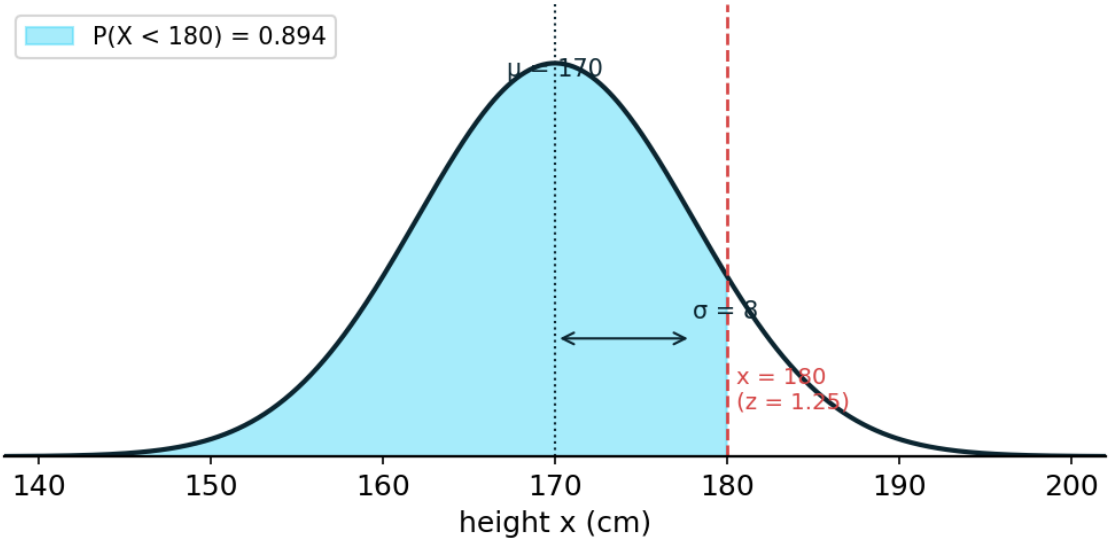

Figure 6. The normal model $X \sim N(\mu, \sigma^2)$ with $\mu = 170$ and $\sigma = 8$. The shaded region $X < 180$ corresponds to $z = 1.25$; its area, computed with math.erf, is $P(X < 180) = 0.894$, agreeing with part (a).

> **Sketch it first:** Before any normal calculation, draw the bell curve and shade the region you want. It stops the classic error of finding the complement of the required area, and tells you at once whether the z-score should be positive or negative.

7. Hypothesis testing (SL)

A significance test decides whether sample evidence is strong enough to reject a default claim. Every test follows the same skeleton:

- State the **null hypothesis** H_0 (the 'no effect / independent' claim) and the **alternative** H_1.
- Fix the **significance level** (usually 5% or 1%).
- Use the GDC to get the **p-value** (and the test statistic).
- **Decision rule:** if p < significance level, **reject H_0**; otherwise there is insufficient evidence to reject it.

The χ^2 test for independence

This tests whether two categorical variables in a contingency table are independent. H_0: the variables are independent; H_1: they are not (they are associated). For each cell the **expected frequency** is

$$\mathbf{f_e = (row\ total \times column\ total) / grand\ total,\ and\ \chi^2_{calc} = \Sigma\,(f_o - f_e)^2 / f_e}$$

The **degrees of freedom** for an r × c table are $(r - 1)(c - 1)$. Reject H_0 when p < significance level, or equivalently when χ^2_{calc} > the critical value.

Worked example 9 — χ^2 test for independence

A survey of 100 people records drink preference and gender. Test at the 5% level whether preference and gender are independent.

Observed f_o	Coffee	Tea	Total
Male	30	20	50
Female	15	35	50
Total	45	55	100

H_0: preference is independent of gender. H_1: they are associated.

Expected values f_e = row × column / 100: Male&Coffee = 50 × 45/100 = 22.5; Male&Tea = 27.5; Female&Coffee = 22.5; Female&Tea = 27.5.

$\chi^2_{calc} = (30-22.5)^2/22.5 + (20-27.5)^2/27.5 + (15-22.5)^2/22.5 + (35-27.5)^2/27.5$

$= 2.50 + 2.05 + 2.50 + 2.05 =$ **9.09**. Degrees of freedom = $(2-1)(2-1) = 1$.

The GDC gives p = 0.0026. Since p = 0.0026 < 0.05 (equivalently 9.09 > 3.841, the critical value), we **reject H_0**: there is significant evidence that drink preference and gender are associated.

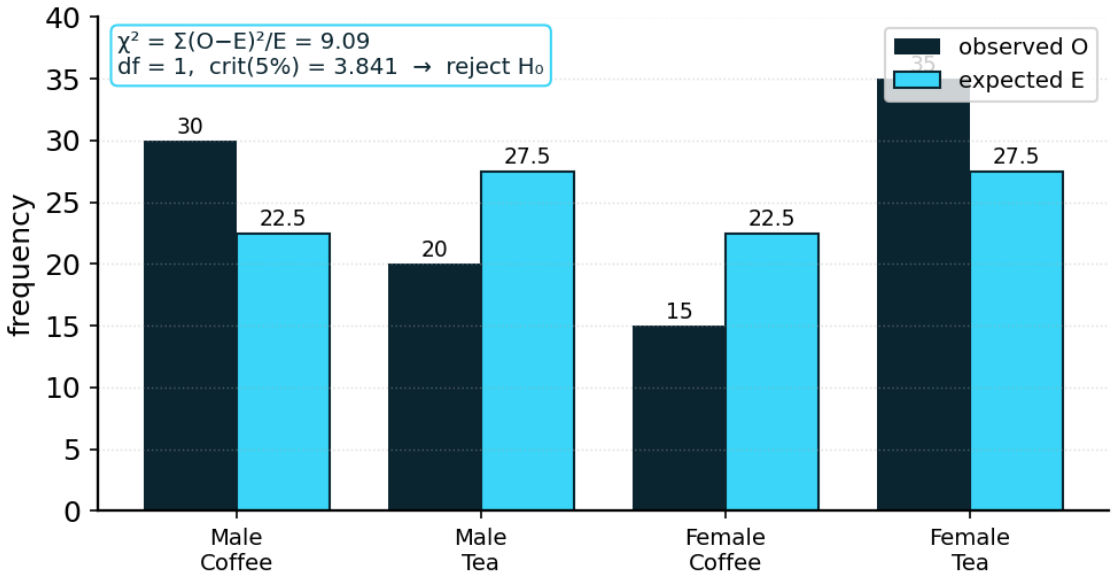

Figure 7. Observed (dark) versus expected (blue) frequencies for the drink-preference contingency table. The test statistic $\chi^2_{calc} = \Sigma(f_o - f_e)^2/f_e = 9.09$ with df = 1 exceeds the 5% critical value 3.841, so H_0 is rejected.

The χ^2 goodness-of-fit test

This checks whether observed frequencies fit a claimed distribution (for example, a die being fair, or data following a given model). H_0: the data fits the distribution. Expected frequencies come from the claimed proportions, χ^2_{calc} uses the same formula, and for k categories the degrees of freedom are k − 1 (subtract one more for each parameter estimated from the data). The decision rule is unchanged: p < significance level → reject H_0.

The t-test (comparing two means)

Use a two-sample t-test to decide whether two populations have different means (for example, exam marks from two teaching methods). H_0: $\mu_1 = \mu_2$; H_1 is either $\mu_1 \neq \mu_2$ (two-tailed) or a one-sided version. The GDC t-Test returns the p-value; again reject H_0 when p < significance level. The samples should be drawn from (approximately) normal populations.

> **Expected values must be large enough:** The χ^2 test is unreliable if any expected frequency is below 5. If that happens, combine adjacent classes before testing. Also note the test uses *expected* frequencies, never the observed ones, in the denominator.

8. (HL) Expectation and variance of a discrete variable

At HL you also quantify the spread of a discrete random variable. The **variance** is the expected squared deviation from the mean and is found most easily from

$$\mathbf{Var(X) = E(X^2) - [E(X)]^2, \text{ where } E(X^2) = \Sigma\, x^2 \cdot P(X = x)}$$

The standard deviation is $\sqrt{\text{Var}(X)}$. For useful reference, the binomial B(n, p) has mean np and **variance np(1 − p)**. These results are also the basis for combining variables in section 10.

Two properties worth memorising, for constants a and b: E(aX + b) = aE(X) + b, and $\text{Var}(aX + b) = a^2\text{Var}(X)$. Adding a constant shifts the mean but leaves the spread unchanged; the '+ b' therefore disappears from the variance.

9. (HL) The Poisson distribution

The Poisson distribution models the number of events occurring in a fixed interval of time or space when events happen **independently** at a constant average rate. Write X ~ Po(m), where m is the mean number of events in the interval, and

$$\mathbf{P(X = x) = e^{-m} \cdot m^x / x!, \text{ for } x = 0, 1, 2, \ldots}$$

A defining feature is that the **mean equals the variance**: E(X) = Var(X) = m. Conditions to check before using it: events are independent, occur singly, and at a constant average rate. If $X_1 \sim \text{Po}(m_1)$ and $X_2 \sim \text{Po}(m_2)$ are independent, then their **sum** $X_1 + X_2 \sim \text{Po}(m_1 + m_2)$. Scaling the interval scales the mean in proportion. Use poissonpdf and poissoncdf on the GDC.

Worked example 10 — Poisson distribution (HL)

A help-desk receives calls at an average rate of 3 per hour, modelled by X ~ Po(3).

(a) P(exactly 5 calls in an hour) = $e^{-3} \times 3^5/5! = 0.0498 \times 243/120 =$ **0.101**.

(b) P(at least one call) = 1 − P(X = 0) = 1 − e^{-3} = 1 − 0.0498 = **0.950**.

(c) Over a 2-hour interval the mean scales to m = 6, so X ~ Po(6): P(exactly 4 calls) = $e^{-6} \times 6^4/4!$ = **0.134** (3 s.f.).

Poisson or binomial? Use the binomial when you count successes in a fixed number of trials n. Use the Poisson when you count events over an interval with no fixed n, given only an average rate. The tell-tale sign of a Poisson is 'mean = variance'.

10. (HL) Linear combinations of random variables

For independent random variables the means always add, and for **independent** variables the variances add too (never the standard deviations). For constants a, b:

$$\mathbf{E(aX \pm bY) = aE(X) \pm bE(Y);\ Var(aX \pm bY) = a^2Var(X) + b^2Var(Y)}$$

Notice the variance uses a **plus** sign for both aX + bY and aX − bY, and squares the coefficients. In particular Var(X − Y) = Var(X) + Var(Y) for independent X and Y — subtracting variables still increases the spread.

Combining normal variables

Any linear combination of independent normal variables is itself normal. If $X \sim N(\mu_X, \sigma_X^2)$ and $Y \sim N(\mu_Y, \sigma_Y^2)$ are independent, then, for example,

$$\mathbf{X + Y \sim N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2).}$$

A common application is the mean of a sample of n independent copies of $X \sim N(\mu, \sigma^2)$: the sample mean is distributed as $N(\mu, \sigma^2/n)$, so larger samples give a mean that clusters more tightly around μ.

11. (HL) Confidence intervals and unbiased estimators

A sample gives a single **point estimate** of a population parameter; a **confidence interval** gives a range of plausible values with a stated level of confidence. The sample mean $\bar{x}$ is an **unbiased estimator** of the population mean μ (its long-run average equals μ). The unbiased estimate of the population variance uses divisor (n − 1):

$$\mathbf{s_{n-1}^2 = [\, n / (n - 1)\,]\, \sigma_n^2}$$

A c% confidence interval for the mean is centred on $\bar{x}$ and produced directly by the GDC (t-Interval when the population variance is unknown). Interpreting it: a 95% confidence interval means that if we repeated the sampling many times, about 95% of the intervals constructed this way would contain the true mean μ. Two things widen the interval — a higher confidence level and a smaller sample; a larger sample narrows it.

Worked example 11 — confidence interval for a mean (HL)

A random sample of 25 components has mean lifetime $\bar{x} = 52.4$ hours and unbiased standard deviation $s_{n-1} = 6.0$ hours. Construct a 95% confidence interval for the population mean μ.

The population variance is unknown, so use the t-interval with n − 1 = 24 degrees of freedom ($t^* \approx 2.064$).

Margin = $t^* \times s_{n-1} / \sqrt{n} = 2.064 \times 6.0/\sqrt{25} = 2.064 \times 1.2 = 2.48$.

95% CI: 52.4 ± 2.48 = **(49.9, 54.9) hours** (3 s.f.). Since 50 lies inside this interval, a claim that μ = 50 would not be rejected at the 5% level (two-tailed).

> **Reading a confidence interval:** If a claimed mean lies *outside* a 95% confidence interval, that value is implausible at the 5% level — the interval and a two-tailed test at the matching level agree.

12. (HL) More hypothesis testing

Choosing the right test

Situation	Test
Two categorical variables, contingency table	χ^2 test for independence
One categorical variable vs a claimed distribution	χ^2 goodness-of-fit test
Comparing the means of two samples	Two-sample t-test
Testing a single population mean (variance known)	z-test on the mean (HL)
Testing a single population mean (variance unknown)	one-sample t-test (HL)

One-tailed and two-tailed tests

Use a **two-tailed** test when H_1 is 'different from' ($\mu \neq \mu_0$); use a **one-tailed** test when H_1 has a direction ($\mu > \mu_0$ or $\mu < \mu_0$). Decide the direction from the context **before** seeing the data, and keep the tail consistent with H_1.

Testing a population mean

To test H_0: $\mu = \mu_0$, enter the sample data (or its summary statistics) into the GDC z-Test (population variance known) or t-Test (variance unknown). Read off the p-value and apply the usual rule: p < significance level → reject H_0.

Type I and Type II errors

	H_0 is actually true	H_0 is actually false
We reject H_0	Type I error (probability = significance level)	Correct decision
We do not reject H_0	Correct decision	Type II error

A **Type I error** rejects a true H_0 (a false alarm); its probability equals the significance level you chose. A **Type II error** fails to reject a false H_0 (a miss). Lowering the significance level reduces Type I errors but increases Type II errors — a trade-off you cannot escape without a larger sample.

Common pitfalls

- **Correlation ≠ causation:** a strong r never proves one variable causes the other.
- **Pearson vs Spearman:** Pearson's r measures *linear* association; Spearman's r_s measures *monotonic* association and suits ranked or ordinal data and outliers.
- **Expected values in χ^2:** divide by the expected frequency f_e, not the observed one, and combine classes if any $f_e < 5$.

- **Degrees of freedom:** $(r-1)(c-1)$ for independence, $k-1$ for goodness-of-fit — do not use the number of cells.
- **Decision rule direction:** $p <$ significance level → reject H_0; a large p means insufficient evidence, not proof of H_0.
- **Mutually exclusive vs independent:** exclusive events with non-zero probability are dependent — do not multiply their probabilities.
- **Extrapolation:** a regression line is only trustworthy inside the data range and only when $|r|$ is large.
- **Which standard deviation:** use σ_n to describe the data; use s_{n-1} (HL) to estimate a population.

Quick reference

Result	Formula / statement
Mean (grouped)	$\bar{x} = (\Sigma f x) / (\Sigma f)$, using mid-interval x
Interquartile range	$IQR = Q_3 - Q_1$
Outlier fences	below $Q_1 - 1.5 \times IQR$ or above $Q_3 + 1.5 \times IQR$
Pearson / Spearman	$r \in [-1, 1]$ (linear); $r_s = 1 - 6\Sigma d^2/[n(n^2-1)]$ (monotonic)
Addition rule	$P(A \cup B) = P(A) + P(B) - P(A \cap B)$
Independence / conditional	$P(A \cap B) = P(A)P(B)$; $P(A \mid B) = P(A \cap B)/P(B)$
Expected value	$E(X) = \Sigma\, x\, P(X = x)$
Binomial B(n, p)	mean np; variance $np(1-p)$ (HL)
Normal z-score	$z = (x - \mu) / \sigma$
χ^2 test	$\chi^2 = \Sigma(f_o - f_e)^2/f_e$; $df = (r-1)(c-1)$
(HL) Variance of X	$Var(X) = E(X^2) - [E(X)]^2$
(HL) Poisson Po(m)	$P(X=x) = e^{-m}m^x/x!$; mean = variance = m
(HL) Combining (indep.)	$Var(aX \pm bY) = a^2Var(X) + b^2Var(Y)$

Test yourself

Attempt all ten without notes or a solution in view, then check against the full worked answers that follow. Questions marked (HL) are Higher Level.

1. A school has 1200 students: 500 in Year 1, 400 in Year 2 and 300 in Year 3. A stratified sample of 60 is required. How many students are taken from each year?

2. For the data 4, 7, 8, 9, 11, 12, 15, 18, 32 find the median, the quartiles and the IQR, and determine whether 32 is an outlier.

3. The mean of the five values 5, 8, x, 12, 15 is 10. Find x.

4. A study finds a Pearson correlation of $r = -0.87$ between hours of screen time and hours of sleep. Interpret this value and state what it does not tell you.

5. A regression line of y on x is $y = 2.4x + 5$, based on data with x from 2 to 12. Predict y when $x = 10$, and comment on the reliability of the prediction.

6. For events A and B, $P(A) = 0.5$, $P(B) = 0.4$ and $P(A \cap B) = 0.2$. Find $P(A \cup B)$. Are A and B independent? Are they mutually exclusive?

7. A biased coin lands heads with probability 0.3. It is tossed 8 times. Find the expected number of heads and P(exactly 2 heads).

8. Test scores are modelled by $X \sim N(50, 5^2)$. Find $P(X < 58)$ and the score below which 90% of students lie.

9. A χ^2 test of independence gives $\chi^2_{calc} = 4.2$ with 2 degrees of freedom at the 5% level (critical value 5.991). State the conclusion.

10. (HL) Emails arrive at a mean rate of 4 per hour, $X \sim Po(4)$. Find P(exactly 2 in an hour) and P(at least 1 in an hour).

Worked answers

1. Sampling fraction = 60/1200 = 0.05. Year 1: $0.05 \times 500 =$ **25**; Year 2: $0.05 \times 400 =$ **20**; Year 3: $0.05 \times 300 =$ **15** (total 60).

2. Ordered, n = 9, so the median is the 5th value = **11**. Lower half 4, 7, 8, 9 → $Q_1 = (7+8)/2 =$ **7.5**; upper half 12, 15, 18, 32 → $Q_3 = (15+18)/2 =$ **16.5**. IQR = $16.5 - 7.5 =$ **9**. Upper fence = $16.5 + 1.5 \times 9 = 30$; since $32 > 30$, **32 is an outlier**.

3. Sum = $5 \times 10 = 50$, so $5 + 8 + x + 12 + 15 = 50$, i.e. $40 + x = 50$ and **x = 10**.

4. $r = -0.87$ is a **strong negative linear correlation**: more screen time is associated with less sleep. It does **not** establish that screen time causes reduced sleep — correlation is not causation, and a third variable may be involved.

5. $y = 2.4 \times 10 + 5 =$ **29**. Since x = 10 lies inside the data range 2–12, this is interpolation and is **reliable** (assuming $|r|$ is large).

6. $P(A \cup B) = 0.5 + 0.4 - 0.2 =$ **0.7**. Independent? $P(A)P(B) = 0.5 \times 0.4 = 0.2 = P(A \cap B)$, so **yes, independent**. Mutually exclusive? $P(A \cap B) = 0.2 \neq 0$, so **no**.

7. $X \sim B(8, 0.3)$. $E(X) = np = 8 \times 0.3 =$ **2.4**. $P(X = 2) =$ binompdf(8, 0.3, 2) = **0.296** (3 s.f.).

8. $P(X < 58)$: $z = (58 - 50)/5 = 1.6$, normalcdf ≈ **0.945**. For the 90th percentile, invNorm(0.90) = 1.282, so score = $50 + 1.282 \times 5 =$ **56.4**.

9. Since $\chi^2_{calc} = 4.2 < 5.991$ (equivalently $p \approx 0.12 > 0.05$), we **do not reject H_0**: there is insufficient evidence of an association — the variables are consistent with being independent.

10. (HL) $P(X = 2) = e^{-4} \times 4^2/2! = 0.0183 \times 8 =$ **0.147**. $P(X \geq 1) = 1 - e^{-4} = 1 - 0.0183 =$ **0.982** (3 s.f.).