Statistics and Probability

IB HL Study Guide

Contents

How to Use This Guide

Statistics and Probability is Topic 4 of the IB Math AA HL syllabus and reliably accounts for a substantial block of marks across Paper 2 and Paper 3. This guide follows the IB AA HL syllabus point by point: probability rules, conditional probability and Bayes' theorem, discrete random variables (binomial and Poisson), continuous random variables (normal distribution), hypothesis testing (chi-squared, t-tests), and correlation and regression. Every section includes fully worked examples drawn from past IB papers, exam alerts for the mistakes that cost marks most often, and a complete formula reference at the end.

IB TIP

How to approach Statistics on exams: The IB rewards structured method. For any probability question, always define your events and state the formula before you substitute values. For hypothesis tests, always write H_0 and H_1 explicitly, state the significance level, calculate or identify the test statistic, compare to the critical value or p-value, and write a conclusion in context. For distributions, always name the distribution and its parameters — for example, "Let $X \sim B(12, 0.3)$" — before calculating any probabilities. Your GDC can compute binomial, Poisson, and normal probabilities directly; know how to use these functions and show the setup clearly in your working.

MEMORISE THIS

What is and is not in the formula booklet: The booklet gives: $P(A \cup B)$, $P(A \mid B)$, $E(X)$, $\text{Var}(X)$, binomial mean and variance, Poisson formula and mean/variance, normal standardisation $Z = \frac{X-\mu}{\sigma}$, the chi-squared statistic, and the Pearson correlation coefficient formula. **NOT given:** Bayes' theorem (you must derive it from first principles or recognise the pattern), the method for constructing tree diagrams, the decision rule for hypothesis tests, and how to interpret the regression coefficient b.

Section 1: Probability (4.5–4.7)

Probability is a measure of how likely an event is to occur, expressed as a number between 0 (impossible) and 1 (certain). An **event** is a subset of the **sample space** S, the set of all possible outcomes.

Notation:

- $P(A)$ — probability that event A occurs
- A' or $\bar{A}$ — the complement of A (event A does NOT occur)
- $A \cup B$ — A or B (or both)
- $A \cap B$ — A and B simultaneously

1.1 Combined Probability

The **addition rule** gives the probability that at least one of two events occurs:

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

The subtraction removes the double-counting of outcomes in both A and B.

Special case — mutually exclusive events: If A and B cannot both occur, $P(A \cap B) = 0$, so:

$$P(A \cup B) = P(A) + P(B) \qquad \text{(mutually exclusive only)}$$

Complement rule:

$$P(A') = 1 - P(A)$$

WORKED EXAMPLE

Combined Probability — Addition Rule

In a class of 30 students, 18 study French, 12 study Spanish, and 6 study both. A student is chosen at random. Find the probability they study French or Spanish.

Define events: Let F = studies French, S = studies Spanish.

$$P(F) = \tfrac{18}{30}, \quad P(S) = \tfrac{12}{30}, \quad P(F \cap S) = \tfrac{6}{30}$$

Apply the addition rule:

$$P(F \cup S) = \tfrac{18}{30} + \tfrac{12}{30} - \tfrac{6}{30} = \tfrac{24}{30} = \tfrac{4}{5}$$

WORKED EXAMPLE

Venn Diagram — Finding Unknown Probabilities

Given $P(A) = 0.5$, $P(B) = 0.4$, and $P(A \cup B) = 0.7$, find $P(A \cap B)$.

Rearrange the addition rule:

$$P(A \cap B) = P(A) + P(B) - P(A \cup B) = 0.5 + 0.4 - 0.7 = 0.2$$

Hence find $P(A \text{ only})$ — the probability of A but not B.

$$P(A \cap B') = P(A) - P(A \cap B) = 0.5 - 0.2 = 0.3$$

1.2 Conditional Probability

The **conditional probability** of A given that B has occurred is:

$$P(A \mid B) = \tfrac{P(A \cap B)}{P(B)}, \qquad P(B) > 0$$

Rearranging gives the **multiplication rule:**

$$P(A \cap B) = P(A \mid B) \cdot P(B) = P(B \mid A) \cdot P(A)$$

⚠ EXAM ALERT

Confusing $P(A \mid B)$ with $P(B \mid A)$ is one of the most costly errors in IB Statistics. These are generally different. $P(\text{has disease} \mid \text{test positive})$ is NOT the same as $P(\text{test positive} \mid \text{has disease})$. Always ask: "which event is the condition (after the bar)?"

WORKED EXAMPLE

Conditional Probability — Two-Way Table

A survey records whether 100 students exercise regularly and whether they sleep more than 7 hours per night.

	Sleep ≥ 7 h	Sleep < 7 h	Total
Exercises regularly	42	18	60
Does not exercise	23	17	40
Total	65	35	100

(a) Find the probability a randomly selected student exercises regularly given they sleep at least 7 hours.

(b) Find the probability a student sleeps fewer than 7 hours given they do not exercise regularly.

Part (a): Let E = exercises, G = sleeps ≥ 7 h.

$$P(E \mid G) = \frac{P(E \cap G)}{P(G)} = \frac{42/100}{65/100} = \frac{42}{65} \approx 0.646$$

Part (b): Let L = sleeps < 7 h, E' = does not exercise.

$$P(L \mid E') = \frac{17/100}{40/100} = \frac{17}{40} = 0.425$$

1.3 Independence

Two events A and B are **independent** if knowing one occurred gives no information about the other:

$$A \text{ and } B \text{ are independent} \iff P(A \cap B) = P(A) \cdot P(B)$$

Equivalently, $P(A \mid B) = P(A)$ and $P(B \mid A) = P(B)$.

⚠ **EXAM ALERT**

Independence $\neq$ mutual exclusivity. Mutually exclusive events (cannot both occur) with non-zero probabilities are actually the most extreme form of *dependence* — if A happens, B definitely cannot, so $P(B \mid A) = 0 \neq P(B)$. Students frequently confuse these two concepts.

WORKED EXAMPLE

Testing Independence

$P(A) = 0.4$, $P(B) = 0.5$, $P(A \cap B) = 0.2$. *Are A and B independent?*

Check: $P(A) \times P(B) = 0.4 \times 0.5 = 0.20$

Since $P(A \cap B) = 0.20 = P(A) \cdot P(B)$, the events are **independent**.

Also verify: $P(A \mid B) = \dfrac{0.2}{0.5} = 0.4 = P(A)$. Confirmed.

1.4 Tree Diagrams

Tree diagrams organise sequential probability problems. Each branch carries a conditional probability, and probabilities along a path are multiplied (multiplication rule). Probabilities on branches from the same node must sum to 1.

WORKED EXAMPLE

Tree Diagram — Two-Stage Problem

A box contains 4 red and 6 blue balls. Two balls are drawn without replacement. Find the probability that both balls are the same colour.

Stage 1 branch probabilities:

- $P(\text{Red}_1) = \frac{4}{10}$, $P(\text{Blue}_1) = \frac{6}{10}$

Stage 2 branch probabilities (conditional on Stage 1):

- After Red: $P(\text{Red}_2 \mid \text{Red}_1) = \frac{3}{9}$, $P(\text{Blue}_2 \mid \text{Red}_1) = \frac{6}{9}$
- After Blue: $P(\text{Red}_2 \mid \text{Blue}_1) = \frac{4}{9}$, $P(\text{Blue}_2 \mid \text{Blue}_1) = \frac{5}{9}$

Path probabilities (same colour):

$P(\text{RR}) = \frac{4}{10} \cdot \frac{3}{9} = \frac{12}{90}$

$P(\text{BB}) = \frac{6}{10} \cdot \frac{5}{9} = \frac{30}{90}$

$P(\text{same colour}) = \frac{12}{90} + \frac{30}{90} = \frac{42}{90} = \frac{7}{15}$

1.5 Bayes' Theorem HL

Bayes' theorem allows you to reverse conditional probabilities — to find the probability of a cause given an observed effect. It arises naturally from the multiplication rule.

For two complementary events A and A':

$$P(A \mid B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

where the **total probability** $P(B)$ is expanded as:

$$P(B) = P(B \mid A) \cdot P(A) + P(B \mid A') \cdot P(A')$$

Combining these:

$$P(A \mid B) = \frac{P(B|A) \cdot P(A)}{P(B|A) \cdot P(A) + P(B|A') \cdot P(A')}$$

IB TIP

Bayes' theorem is not in the formula booklet. You are expected to derive it or apply it directly using a tree diagram. Drawing the tree and labelling all branches is the safest approach — read off the answer by dividing the target path probability by the total probability of the observed outcome.

WORKED EXAMPLE

Bayes' Theorem — Medical Test

A disease affects 1% of a population. A diagnostic test has a 95% sensitivity (true positive rate) and a 2% false positive rate. A randomly selected person tests positive. Find the probability they actually have the disease.

Define events:

- D = has the disease, D' = does not have the disease
- T^+ = tests positive

Given: $P(D) = 0.01$, $P(D') = 0.99$, $P(T^+ \mid D) = 0.95$, $P(T^+ \mid D') = 0.02$

Total probability of testing positive:

$$P(T^+) = P(T^+ \mid D) \cdot P(D) + P(T^+ \mid D') \cdot P(D')$$

$$= (0.95)(0.01) + (0.02)(0.99) = 0.0095 + 0.0198 = 0.0293$$

Apply Bayes:

$$P(D \mid T^+) = \frac{P(T^+|D) \cdot P(D)}{P(T^+)} = \frac{0.0095}{0.0293} \approx 0.324$$

Despite a positive test, there is only a 32.4% chance the person actually has the disease. This counter-intuitive result arises because the disease is rare — most positive tests come from the large pool of healthy people with false positives.

EXAM ALERT

In Bayes problems, the denominator is always $P(\text{observed outcome})$, computed via the law of total probability over all mutually exclusive "causes." The most common error is forgetting to include all branches in this denominator, or using $P(D)$ in place of $P(T^+)$.

WORKED EXAMPLE

Bayes' Theorem — Factory Quality Control

Machine A produces 60% of a factory's output; machine B produces the remaining 40%. Machine A has a 3% defect rate; machine B has a 5% defect rate. A randomly selected item is found to be defective. What is the probability it came from machine A?

Setup: $P(A) = 0.6$, $P(B) = 0.4$, $P(D \mid A) = 0.03$, $P(D \mid B) = 0.05$

$$P(D) = (0.03)(0.6) + (0.05)(0.4) = 0.018 + 0.020 = 0.038$$

$$P(A \mid D) = \frac{(0.03)(0.6)}{0.038} = \frac{0.018}{0.038} \approx 0.474$$

There is about a 47.4% chance the defective item came from machine A.

Section 2: Discrete Random Variables (4.8–4.9)

A **discrete random variable (DRV)** X takes a countable set of values, each with a defined probability. The probability distribution of X is a complete list of all possible values x_i and their probabilities $P(X = x_i)$.

Requirements for a valid probability distribution:

1. $0 \leq P(X = x_i) \leq 1$ for all i
2. $\sum_i P(X = x_i) = 1$

2.1 Expected Value and Variance

The **expected value** (mean) of X is the long-run average outcome:

$$E(X) = \sum x \cdot P(X = x)$$

The **variance** measures the spread around the mean:

$$\text{Var}(X) = E(X^2) - [E(X)]^2$$

where $E(X^2) = \sum x^2 \cdot P(X = x)$.

The **standard deviation** is $\text{SD}(X) = \sqrt{\text{Var}(X)}$.

Linear transformations: For $Y = aX + b$:

$$E(aX + b) = a \cdot E(X) + b$$

$$\text{Var}(aX + b) = a^2 \cdot \text{Var}(X)$$

⚠ **EXAM ALERT**

$\text{Var}(aX + b) = a^2\text{Var}(X)$, not $a^2\text{Var}(X) + b$. The constant b shifts the distribution but does not affect spread. Students frequently add b or b^2 to the variance.

WORKED EXAMPLE

Expected Value and Variance

A biased die has the following distribution:

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

(a) Find $E(X)$. *(b) Find* $\text{Var}(X)$. *(c) Find* $E(3X - 2)$ *and* $\text{Var}(3X - 2)$.

Part (a):

$E(X) = 1(0.1) + 2(0.3) + 3(0.4) + 4(0.2) = 0.1 + 0.6 + 1.2 + 0.8 = 2.7$

Part (b): First compute $E(X^2)$:

$E(X^2) = 1^2(0.1) + 2^2(0.3) + 3^2(0.4) + 4^2(0.2) = 0.1 + 1.2 + 3.6 + 3.2 = 8.1$

$\text{Var}(X) = 8.1 - (2.7)^2 = 8.1 - 7.29 = 0.81$

Part (c):

$E(3X - 2) = 3(2.7) - 2 = 8.1 - 2 = 6.1$

$\text{Var}(3X - 2) = 3^2 \cdot \text{Var}(X) = 9 \times 0.81 = 7.29$

2.2 Binomial Distribution

The **binomial distribution** models the number of successes in n independent trials, each with probability p of success.

Conditions for a binomial model:

1. Fixed number of trials n
2. Each trial has exactly two outcomes (success / failure)
3. Constant probability p of success on each trial
4. Trials are independent

If $X \sim B(n, p)$, then:

$P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}, \qquad x = 0, 1, 2, \ldots, n$

$E(X) = np \qquad \text{Var}(X) = np(1 - p)$

MEMORISE THIS

Binomial Quick Reference

Quantity	Formula
Distribution	$X \sim B(n, p)$
P.M.F.	$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$
Mean	$E(X) = np$
Variance	$\text{Var}(X) = np(1-p)$
GDC (Casio)	`BinomPD(x, n, p)` for exact; `BinomCD(x, n, p)` for $P(X \leq x)$

WORKED EXAMPLE

Binomial — Exact Probability

A fair coin is tossed 8 times. Find the probability of getting exactly 5 heads.

State the distribution: Let X = number of heads. Since trials are independent with $p = 0.5$ and $n = 8$: $X \sim B(8, 0.5)$.

$P(X = 5) = \binom{8}{5}(0.5)^5(0.5)^3 = 56 \times (0.5)^8 = \frac{56}{256} = \frac{7}{32} \approx 0.219$

WORKED EXAMPLE

Binomial — Cumulative and Complement

In a production line, 15% of items are defective. A batch of 20 items is selected. Find: (a) $P(X \leq 3)$, (b) $P(X \geq 4)$, (c) $P(2 \leq X \leq 5)$.

State: $X \sim B(20, 0.15)$

Part (a): Use GDC cumulative binomial: $P(X \leq 3) \approx 0.6477$

Part (b): Complement rule:

$P(X \geq 4) = 1 - P(X \leq 3) \approx 1 - 0.6477 = 0.3523$

Part (c):

$P(2 \leq X \leq 5) = P(X \leq 5) - P(X \leq 1) \approx 0.9327 - 0.1756 = 0.7571$

EXAM ALERT

When NOT to use binomial: The binomial requires (1) fixed n, (2) constant p, (3) independence. Drawing without replacement from a small population violates independence — use the hypergeometric distribution or direct probability instead. "Selecting cards from a deck" without replacement is not binomial.

WORKED EXAMPLE

Binomial — Finding Parameters from Mean and Variance

A random variable $X \sim B(n, p)$ has mean 6 and variance 4.2. Find n and p.

Set up equations:

$np = 6 \qquad np(1-p) = 4.2$

Divide the second by the first:

$1 - p = \frac{4.2}{6} = 0.7 \implies p = 0.3$

Substitute back:

$n(0.3) = 6 \implies n = 20$

2.3 Poisson Distribution HL

The **Poisson distribution** models the number of occurrences of a random event in a fixed interval of time or space, when events occur independently at a constant average rate.

Conditions for a Poisson model:

1. Events occur independently
2. Events occur at a constant average rate m per unit interval
3. Two events cannot occur simultaneously

If $X \sim \text{Po}(m)$, then:

$P(X = x) = \frac{e^{-m} m^x}{x!}, \qquad x = 0, 1, 2, \ldots$

$E(X) = \text{Var}(X) = m$

The equality of mean and variance is a diagnostic property — if a dataset has mean $\approx$ variance, a Poisson model is plausible.

MEMORISE THIS

Poisson Quick Reference

Quantity	Formula
Distribution	$X \sim \text{Po}(m)$
P.M.F.	$P(X = x) = \dfrac{e^{-m} m^x}{x!}$
Mean	$E(X) = m$
Variance	$\text{Var}(X) = m$
GDC (Casio)	`PoissonPD(x, m)` for exact; `PoissonCD(x, m)` for $P(X \leq x)$

WORKED EXAMPLE

Poisson — Direct Calculation

Calls arrive at a helpdesk at an average rate of 3 per hour. Find the probability that: (a) exactly 2 calls arrive in one hour, (b) fewer than 4 calls arrive in one hour.

State: $X \sim \text{Po}(3)$

Part (a):

$P(X = 2) = \frac{e^{-3} \cdot 3^2}{2!} = \frac{e^{-3} \cdot 9}{2} \approx \frac{9}{2} \times 0.04979 \approx 0.224$

Part (b):

$P(X < 4) = P(X \leq 3) = P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3)$

$= e^{-3}\left(1 + 3 + \frac{9}{2} + \frac{27}{6}\right) = e^{-3} \cdot 13 \approx 0.04979 \times 13 \approx 0.647$

WORKED EXAMPLE

Poisson — Changing the Interval

Emails arrive at an average rate of 6 per hour. Find the probability that: (a) no emails arrive in 20 minutes, (b) more than 2 emails arrive in 30 minutes.

Key step: Rescale the rate to match the interval.

- In 20 minutes: $m = 6 \times \frac{20}{60} = 2$, so $X \sim \text{Po}(2)$
- In 30 minutes: $m = 6 \times \frac{30}{60} = 3$, so $Y \sim \text{Po}(3)$

Part (a):

$P(X = 0) = \frac{e^{-2} \cdot 2^0}{0!} = e^{-2} \approx 0.135$

Part (b):

$P(Y > 2) = 1 - P(Y \leq 2) = 1 - e^{-3}\left(1 + 3 + \frac{9}{2}\right) = 1 - e^{-3}(8.5) \approx 1 - 0.423 = 0.577$

EXAM ALERT

When the Poisson interval changes, multiply the rate proportionally. If the rate is "m per hour" and you want a 15-minute interval, use $m_{\text{new}} = m \times \frac{1}{4}$. Forgetting to rescale is one of the most common Poisson errors.

Poisson as an approximation to binomial HL: When n is large and p is small (rule of thumb: $n \geq 50, p \leq 0.1$), $B(n, p) \approx \text{Po}(np)$.

✎ **WORKED EXAMPLE**

Poisson Approximation to Binomial HL

A rare genetic mutation occurs in 1 in 500 births. In a sample of 400 births, find the approximate probability that at most 2 have the mutation.

Check conditions: $n = 400$ (large), $p = 0.002$ (small). Approximate with $\text{Po}(m)$ where $m = np = 400 \times 0.002 = 0.8$.

$P(X \leq 2) = e^{-0.8}\left(1 + 0.8 + \frac{0.64}{2}\right) = e^{-0.8}(2.12) \approx 0.4493 \times 2.12 \approx 0.953$

Section 3: Continuous Random Variables (4.10–4.12)

A **continuous random variable (CRV)** takes values in a continuous range. Unlike DRVs, we cannot assign probability to individual values — instead we work with a **probability density function (pdf)** $f(x)$.

Properties of a pdf:

1. $f(x) \geq 0$ for all x
2. $\displaystyle\int_{-\infty}^{\infty} f(x)\,dx = 1$ (total area under the curve equals 1)
3. $P(a \leq X \leq b) = \displaystyle\int_a^b f(x)\,dx$

For a CRV: $P(X = x) = 0$ for any single value. Therefore $P(X \leq a) = P(X < a)$.

Mean and variance of a CRV:

$E(X) = \int_{-\infty}^{\infty} x \cdot f(x)\,dx \qquad \text{Var}(X) = \int_{-\infty}^{\infty} x^2 f(x)\,dx - [E(X)]^2$

3.1 The Normal Distribution

The **normal distribution** is the most important continuous distribution. It models many natural phenomena — heights, measurement errors, exam scores — where data clusters symmetrically around a mean.

If $X \sim N(\mu, \sigma^2)$, its bell-shaped pdf is:

$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$

Key properties:

- Symmetric about the mean μ
- Mean = Median = Mode $= \mu$
- Inflection points at $\mu \pm \sigma$
- $P(\mu - \sigma < X < \mu + \sigma) \approx 0.683$
- $P(\mu - 2\sigma < X < \mu + 2\sigma) \approx 0.954$
- $P(\mu - 3\sigma < X < \mu + 3\sigma) \approx 0.997$

MEMORISE THIS

The 68-95-99.7 Rule

Interval	Approximate probability
Within 1σ of μ	68.3%
Within 2σ of μ	95.4%
Within 3σ of μ	99.7%

These are approximations. Use your GDC for exact values.

3.2 Standardisation and Z-Scores

The **standard normal distribution** $Z \sim N(0,1)$ has mean 0 and variance 1. Any normal random variable can be converted to Z via:

$Z = \frac{X-\mu}{\sigma}$

A **z-score** tells you how many standard deviations an observation is from the mean. $Z = 1.5$ means the value is 1.5 standard deviations above the mean.

WORKED EXAMPLE

Normal Distribution — Finding Probabilities

Heights of adults are normally distributed with mean 170 cm and standard deviation 8 cm. Find: (a) $P(X > 180)$, (b) $P(160 < X < 185)$.

State: $X \sim N(170, 64)$

Part (a): $Z = \dfrac{180-170}{8} = 1.25$

Using GDC or standard normal table: $P(Z > 1.25) = 1 - P(Z \leq 1.25) \approx 1 - 0.8944 = 0.1056$

Part (b): Lower bound: $Z_1 = \dfrac{160-170}{8} = -1.25$; Upper bound: $Z_2 = \dfrac{185-170}{8} = 1.875$

$P(-1.25 < Z < 1.875) = P(Z < 1.875) - P(Z < -1.25) \approx 0.9696 - 0.1056 = 0.8640$

EXAM ALERT

When using GDC for normal probabilities, use `normalcdf(lower, upper, μ, σ)`. For $P(X > a)$, use a very large upper bound (e.g., 10^{99}), or compute $1 - P(X \leq a)$ using the complement. Show your GDC setup in working — write out which distribution, the parameters, and the bounds.

WORKED EXAMPLE

Normal Distribution — Symmetry Shortcut

$X \sim N(50, 25)$. *Find* $P(45 < X < 55)$ *without a table*.

The interval $[45, 55]$ is symmetric about $\mu = 50$, each end $\frac{50-45}{5} = 1$ standard deviation away.

By the 68-95-99.7 rule: $P(\mu - \sigma < X < \mu + \sigma) \approx 0.683$.

For an exact answer: Z bounds are -1 and 1, so using GDC: $P(-1 < Z < 1) = 0.6827$.

3.3 Normal Approximation to the Binomial

When n is large, the binomial distribution $B(n, p)$ can be approximated by a normal distribution:

$$X \sim B(n,p) \quad \text{approximates to} \quad Y \sim N(np, np(1-p))$$

Conditions for the approximation:

- n is sufficiently large (typically $n \geq 30$)
- $np \geq 5$ and $n(1-p) \geq 5$

Continuity correction: Because we are approximating a discrete distribution with a continuous one, we apply a continuity correction:

- $P(X = k) \approx P(k - 0.5 < Y < k + 0.5)$
- $P(X \leq k) \approx P(Y \leq k + 0.5)$
- $P(X \geq k) \approx P(Y \geq k - 0.5)$
- $P(X < k) \approx P(Y < k - 0.5)$
- $P(X > k) \approx P(Y > k + 0.5)$

⚠ EXAM ALERT

Forgetting the continuity correction is one of the most common errors when using the normal approximation to binomial. Always add or subtract 0.5 to account for the discrete-to-continuous transition. Without it, your answer can be significantly wrong.

WORKED EXAMPLE

Paper 2 Style: Binomial vs Normal Approximation [8 marks]

A manufacturer produces electronic components, of which 5% are defective. A quality inspector tests a random sample of 200 components.

(a) Let X be the number of defective components found. Write down the exact distribution of X. [1]

(b) Verify that X can be approximated by a normal distribution and state the parameters of this normal distribution. [2]

(c) Using the normal approximation with continuity correction, find the probability that the inspector finds:

- (i) exactly 12 defective components [2]
- (ii) fewer than 8 defective components [3]

▶ Show Solution

3.5 Inverse Normal

The **inverse normal** problem asks: given a probability p, find the value x such that $P(X \leq x) = p$.

On the GDC, use `invNorm(p, μ, σ)` to find x directly.

WORKED EXAMPLE

Inverse Normal — Finding a Threshold

Test scores are distributed as $X \sim N(65, 100)$*. The top 10% of students receive a distinction. Find the minimum score needed for a distinction.*

We need x such that $P(X > x) = 0.10$, i.e., $P(X \leq x) = 0.90$.

Using GDC: `invNorm(0.90, 65, 10)` $= 77.8$

A student needs at least **77.8** (round up to 78) to receive a distinction.

WORKED EXAMPLE

Inverse Normal — Symmetric Interval

$X \sim N(\mu, \sigma^2)$ *with* $\mu = 40$ *and* $\sigma = 5$. *Find the value* k *such that* $P(\mu - k < X < \mu + k) = 0.90$.

By symmetry, $P(X < \mu - k) = 0.05$.

Using GDC: `invNorm(0.05, 40, 5)` ≈ 31.78, so $k = 40 - 31.78 = 8.22$.

Check: $P(40 - 8.22 < X < 40 + 8.22) = P(31.78 < X < 48.22) = 0.90$. Confirmed.

WORKED EXAMPLE

Finding μ or σ from a Normal Probability

$X \sim N(\mu, 16)$. *It is given that* $P(X > 20) = 0.2$. *Find* μ.

Standardise: $P\left(Z > \dfrac{20 - \mu}{4}\right) = 0.2$

Find the z-score: $P(Z > z) = 0.2 \Rightarrow P(Z \leq z) = 0.8 \Rightarrow z \approx 0.8416$

Solve:

$\frac{20-\mu}{4} = 0.8416 \implies 20 - \mu = 3.366 \implies \mu \approx 16.6$

Section 4: Hypothesis Testing (4.13–4.15)

Statistical hypothesis testing provides a formal framework for deciding whether observed data provide sufficient evidence against a default assumption. Every test follows the same logical structure.

The universal procedure:

1. **State H_0 and H_1** — the null and alternative hypotheses
2. **State the significance level** α (typically 0.05 or 0.01)
3. **Identify the test statistic** and its distribution under H_0
4. **Calculate the test statistic** (or use GDC)
5. **Find the p-value** (or compare test statistic to critical value)
6. **Make a decision:** reject H_0 if p-value $< \alpha$
7. **Write a conclusion in context**

⚠ **EXAM ALERT**

Never omit H_0 and H_1, and never write them in terms of sample statistics. Hypotheses are always about **population** parameters. Write $H_0 : \mu = 50$, not "H_0: the sample mean is 50." Losing marks for missing hypotheses is one of the most avoidable errors in IB exams.

4.1 The p-value and Type I/II Errors

The **p-value** is the probability of obtaining a result at least as extreme as the observed data, assuming H_0 is true. A small p-value means the observed data would be very unlikely under H_0, providing evidence against it.

Decision rule: Reject H_0 if p-value $< \alpha$.

IB TIP

The p-value is NOT the probability that H_0 is true. It is the probability of the observed data (or more extreme) given H_0 is true. This distinction matters in written conclusions — say "there is sufficient evidence to reject H_0" rather than "we have proved H_0 is false."

Type I and Type II Errors:

	H_0 is actually TRUE	H_0 is actually FALSE
We reject H_0	**Type I error** — false positive (probability $= \alpha$)	Correct decision (power $= 1 - \beta$)
We do not reject H_0	Correct decision (probability $= 1 - \alpha$)	**Type II error** — false negative (probability $= \beta$)

- **Type I error:** Rejecting H_0 when it is true (false alarm). The significance level α is the probability of making this error.
- **Type II error:** Failing to reject H_0 when it is false (missed detection). The probability is denoted β.
- **Power of a test:** The probability of correctly rejecting a false H_0. Power $= 1 - \beta$.

⚠ **EXAM ALERT**

A common misconception: lowering α (making the test more conservative) **increases** β — you become less likely to reject H_0, so you miss more true effects. There is a trade-off between Type I and Type II error rates. To reduce both simultaneously, you need a larger sample size.

WORKED EXAMPLE

Paper 2 Style: Type I and Type II Errors [6 marks]

A pharmaceutical company develops a new drug claimed to lower blood pressure. A clinical trial will test $H_0 : \mu = 140$ (no effect) vs. $H_1 : \mu < 140$ (drug lowers blood pressure), where μ is the mean systolic blood pressure (in mmHg) after treatment.

(a) Describe what a Type I error would mean in the context of this trial. [2]

(b) Describe what a Type II error would mean in the context of this trial. [2]

(c) The significance level is set at $\alpha = 0.05$. The medical regulator is concerned about approving ineffective drugs. Should they increase or decrease α? Explain the trade-off. [2]

▶ Show Solution

4.2 Chi-Squared Goodness-of-Fit Test

The **chi-squared goodness-of-fit test** assesses whether observed frequency data are consistent with a proposed theoretical distribution.

Test statistic:

$$\chi^2 = \sum_{\text{all cells}} \frac{(O-E)^2}{E}$$

where O is the observed frequency and E is the expected frequency under H_0.

Under H_0, $\chi^2 \sim \chi^2_\nu$ where the degrees of freedom $\nu = k - 1 - m$, with k = number of categories and m = number of parameters estimated from the data.

Conditions: All expected frequencies $E \geq 5$. If some are too small, combine adjacent categories.

MEMORISE THIS

Chi-Squared GOF Procedure

Step	Action
H_0	The data follow the proposed distribution
H_1	The data do not follow the proposed distribution
ν	$k - 1$ (if no parameters estimated); $k - 1 - m$ (if m estimated)
Reject H_0 if$\chi^2_{\text{calc}} > \chi^2_{\text{crit}}$ or $p < \alpha$	

WORKED EXAMPLE

Chi-Squared Goodness-of-Fit

A die is rolled 120 times. The observed frequencies are:

Face	1	2	3	4	5	6
Observed	17	22	18	25	19	19

Test at the 5% significance level whether the die is fair.

Step 1 — Hypotheses:

- H_0: The die is fair (each face has probability $\frac{1}{6}$)
- H_1: The die is not fair

Step 2 — Significance level: $\alpha = 0.05$

Step 3 — Expected frequencies: If fair, $E = \frac{120}{6} = 20$ for each face.

Step 4 — Test statistic:

$$\chi^2 = \frac{(17-20)^2}{20} + \frac{(22-20)^2}{20} + \frac{(18-20)^2}{20} + \frac{(25-20)^2}{20} + \frac{(19-20)^2}{20} + \frac{(19-20)^2}{20}$$

$$= \frac{9+4+4+25+1+1}{20} = \frac{44}{20} = 2.2$$

Step 5 — Degrees of freedom: $\nu = 6 - 1 = 5$

Step 6 — Critical value: $\chi^2_{5,0.05} = 11.07$

Step 7 — Decision: $\chi^2_{\text{calc}} = 2.2 < 11.07$, so we **fail to reject** H_0.

Conclusion: At the 5% significance level, there is insufficient evidence to conclude the die is unfair.

EXAM ALERT

The chi-squared test requires all **expected** frequencies to be at least 5, not the observed frequencies. If any $E < 5$, merge that category with an adjacent one before calculating χ^2.

WORKED EXAMPLE

Paper 2 Style: Chi-Squared Goodness-of-Fit Test [7 marks]

A biologist is studying the distribution of flower colors in a population of wildflowers. Mendelian genetics predicts the ratio Red : Pink : White should be 1:2:1. The biologist counts 240 flowers and observes:

Color	Red	Pink	White
Observed	52	136	52

(a) Calculate the expected frequencies under the genetic model. [2]

(b) State appropriate null and alternative hypotheses. [1]

(c) Calculate the χ^2 test statistic and state the degrees of freedom. [3]

(d) The critical value at the 5% significance level is 5.991. State your conclusion. [1]

▶ Show Solution

4.3 Chi-Squared Test for Independence

The **chi-squared test for independence** tests whether two categorical variables are associated in a contingency table.

H_0 : The two variables are independent H_1 :
The two variables are not independent (there is an association)

Expected frequency for cell (i, j):

$E_{ij} = \frac{(\text{row } i \text{ total}) \times (\text{column } j \text{ total})}{\text{grand total}}$

Degrees of freedom: $\nu = (r-1)(c-1)$ where r = number of rows and c = number of columns.

WORKED EXAMPLE

Chi-Squared Test for Independence

Use the table from Section 1.2 (exercise vs. sleep). Test at 5% significance whether exercise and sleep are independent.

	Sleep ≥ 7 h	Sleep < 7 h	Total
Exercises	42	18	60
Does not exercise	23	17	40
Total	65	35	100

Hypotheses: H_0: Exercise and sleep hours are independent. H_1: They are associated.

Expected frequencies: $E_{11} = \frac{60 \times 65}{100} = 39$, $E_{12} = \frac{60 \times 35}{100} = 21$, $E_{21} = \frac{40 \times 65}{100} = 26$, $E_{22} = \frac{40 \times 35}{100} = 14$

Test statistic:

$$\chi^2 = \frac{(42-39)^2}{39} + \frac{(18-21)^2}{21} + \frac{(23-26)^2}{26} + \frac{(17-14)^2}{14}$$

$$= \frac{9}{39} + \frac{9}{21} + \frac{9}{26} + \frac{9}{14} \approx 0.231 + 0.429 + 0.346 + 0.643 = 1.649$$

Degrees of freedom: $\nu = (2-1)(2-1) = 1$

Critical value: $\chi^2_{1,0.05} = 3.841$

Decision: $1.649 < 3.841$, fail to reject H_0.

Conclusion: At the 5% level, there is insufficient evidence of an association between exercise habits and sleep duration.

4.4 t-Test for the Mean

The **one-sample t-test** tests whether the population mean μ equals a specified value μ_0 , when the population standard deviation is unknown and the population is approximately normal.

Hypotheses:

- Two-tailed: $H_0 : \mu = \mu_0$ vs. $H_1 : \mu \neq \mu_0$
- One-tailed: $H_0 : \mu = \mu_0$ vs. $H_1 : \mu > \mu_0$ (or $\mu < \mu_0$)

Test statistic:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

where $\bar{x}$ is the sample mean, s is the sample standard deviation, and n is the sample size. Under H_0, $t \sim t_{n-1}$ (t-distribution with $n-1$ degrees of freedom).

WORKED EXAMPLE

One-Sample t-Test

A manufacturer claims their light bulbs last a mean of 1000 hours. A random sample of 15 bulbs gives a mean of 985 hours with a standard deviation of 30 hours. Test at the 5% level whether the mean lifetime is less than claimed.

Hypotheses: $H_0 : \mu = 1000$; $H_1 : \mu < 1000$ (one-tailed)

Significance level: $\alpha = 0.05$

Test statistic:

$t = \frac{985-1000}{30/\sqrt{15}} = \frac{-15}{7.746} \approx -1.936$

Degrees of freedom: $\nu = 15 - 1 = 14$

Critical value (one-tailed): $t_{14,0.05} = -1.761$

Decision: $t = -1.936 < -1.761$, so we **reject** H_0.

Conclusion: At the 5% significance level, there is sufficient evidence to conclude the mean bulb lifetime is less than 1000 hours.

WORKED EXAMPLE

Paper 2 Style: Hypothesis Testing with t-distribution [8 marks]

A nutritionist claims that a new diet reduces cholesterol levels. A study measures cholesterol (in mmol/L) for 10 participants before and after 8 weeks on the diet.

Participant	1	2	3	4	5	6	7	8	9	10
Before	6.2	5.8	6.5	5.9	6.8	6.1	5.7	6.4	6.0	6.3
After	5.8	5.5	6.0	5.6	6.2	5.9	5.4	6.1	5.7	6.0

(a) State suitable null and alternative hypotheses to test whether the diet reduces cholesterol. [2]

(b) Calculate the differences for each participant (After – Before) and find the mean difference $\bar{d}$ and standard deviation s_d. [3]

(c) Perform a paired t-test at the 5% significance level and state your conclusion. [3]

▶ Show Solution

IB TIP

On a GDC (Casio), the t-test is under `STAT → TEST → 1-Sample tTest`. Enter μ_0, $\bar{x}$, s_x, and n, then select the correct tail. The GDC outputs the t-statistic and p-value directly — always report both in your working.

4.5 Two-Sample t-Test and Paired t-Test HL

Two-sample t-test: Tests whether two independent populations have the same mean.

$$H_0 : \mu_1 = \mu_2 \qquad \text{(equivalently, } \mu_1 - \mu_2 = 0\text{)}$$

Test statistic (assuming equal but unknown variances — pooled):

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \qquad s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$$

with $\nu = n_1 + n_2 - 2$ degrees of freedom.

Paired t-test HL: Used when observations come in matched pairs (before/after, two measurements on the same subject).

Define the differences $d_i = x_{1i} - x_{2i}$. Then:

$$\bar{d} = \frac{1}{n}\sum d_i, \qquad s_d = \text{SD of the } d_i \text{ values}$$

$$t = \frac{\bar{d}}{s_d/\sqrt{n}} \sim t_{n-1}$$

⚠ **EXAM ALERT**

Use a **paired** t-test when the same subject is measured twice (before/after). Use a **two-sample** t-test when comparing two different, independent groups. Applying the wrong test to paired data ignores the correlation between measurements and leads to incorrect inference.

WORKED EXAMPLE

Paper 2 Style: Two-Sample t-Test HL **[7 marks]**

An ecologist compares plant heights (in cm) between two different soil treatments. The data are:

Treatment A (organic): 24, 28, 22, 30, 26, 25, 29

Treatment B (synthetic): 32, 30, 35, 28, 34, 31

(a) State suitable hypotheses to test whether the mean height differs between treatments. [1]

(b) Calculate the sample means and sample standard deviations for both treatments. [2]

(c) The pooled variance is $s_p^2 = 11.27$. Perform a two-sample t-test at the 5% significance level. State the test statistic, degrees of freedom, and your conclusion. [4]

▶ Show Solution

WORKED EXAMPLE

Paired t-Test HL

Eight students take a memory test before and after a training programme. Their scores are:

Student1	2	3	4	5	6	7	8
Before	6271588065735570						
After	6876608467786174						

Test at the 5% level whether the training improves scores.

Step 1 — Compute differences $d_i = \text{After} - \text{Before}$:

| d_i | 6 | 5 | 2 | 4 | 2 | 5 | 6 | 4 |

Step 2 — Statistics of d**:**

$$\bar{d} = \frac{6+5+2+4+2+5+6+4}{8} = \frac{34}{8} = 4.25$$

$$s_d = \sqrt{\frac{\sum(d_i - \bar{d})^2}{n-1}} \approx 1.5811$$

Step 3 — Hypotheses: $H_0 : \mu_d = 0$; $H_1 : \mu_d > 0$ (one-tailed)

Step 4 — Test statistic:

$$t = \frac{4.25}{1.5811/\sqrt{8}} = \frac{4.25}{0.5590} \approx 7.60$$

Step 5 — Critical value: $t_{7,0.05} = 1.895$

Step 6 — Decision: $t = 7.60 \gg 1.895$, reject H_0.

Conclusion: At the 5% level, there is very strong evidence that the training programme significantly improves scores.

Section 5: Correlation and Regression (4.1–4.4)

Bivariate data involves two variables measured on the same individual. We explore whether changes in one variable are associated with changes in the other.

5.1 Scatter Diagrams

A **scatter diagram** plots pairs (x_i, y_i) to visualise the relationship between two variables. Key features to comment on:

- **Direction:** positive (up-right) or negative (down-right) trend
- **Strength:** how closely do points follow a line?
- **Form:** linear or non-linear?
- **Outliers:** points far from the general pattern

5.2 Pearson's Correlation Coefficient

The **Pearson product-moment correlation coefficient (PMCC)** r measures the strength and direction of the **linear** relationship between two variables:

$$r = \frac{S_{xy}}{\sqrt{S_{xx} \cdot S_{yy}}}$$

where:

$$S_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - n\bar{x}\bar{y}$$

$$S_{xx} = \sum (x_i - \bar{x})^2 = \sum x_i^2 - n\bar{x}^2 \qquad S_{yy} = \sum (y_i - \bar{y})^2 = \sum y_i^2 - n\bar{y}^2$$

Properties of r:

- $-1 \leq r \leq 1$
- $r = 1$: perfect positive linear correlation
- $r = -1$: perfect negative linear correlation
- $r = 0$: no linear correlation (but there may be a non-linear relationship)

MEMORISE THIS

Interpreting r — Approximate Guidelines

$\lvert r \rvert$ value	Interpretation
0.00–0.20	Very weak or no linear correlation
0.20–0.40	Weak linear correlation
0.40–0.60	Moderate linear correlation
0.60–0.80	Strong linear correlation
0.80–1.00	Very strong linear correlation

These are guidelines, not rigid rules. Context matters.

EXAM ALERT

$r = 0.9$ **does NOT mean "90% of the variation is explained."** The coefficient of determination $r^2 = 0.81$ tells you that 81% of the variation in y is explained by the linear relationship with x. Students regularly confuse r with r^2.

EXAM ALERT

Correlation does not imply causation. Even a very high value of r does not mean that changes in x cause changes in y. There may be a lurking variable, or the relationship may be coincidental.

WORKED EXAMPLE

Computing PMCC

Five students' hours of study (x) and exam marks (y) are:

Student	x	y
A	2	50
B	4	65
C	6	72
D	8	80
E	10	90

Calculate r.

Step 1 — Means: $\bar{x} = \frac{2+4+6+8+10}{5} = 6, \bar{y} = \frac{50+65+72+80+90}{5} = 71.4$

Step 2 — Sums of squares:

	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x-\bar{x})(y-\bar{y})$	$(x-\bar{x})^2$	$(y-\bar{y})^2$
A	-4	-21.4	85.6	16	457.96
B	-2	-6.4	12.8	4	40.96
C	0	0.6	0	0	0.36
D	2	8.6	17.2	4	73.96
E	4	18.6	74.4	16	345.96
Sum			**190**	**40**	**919.2**

$r = \frac{S_{xy}}{\sqrt{S_{xx} \cdot S_{yy}}} = \frac{190}{\sqrt{40 \times 919.2}} = \frac{190}{\sqrt{36768}} = \frac{190}{191.75} \approx 0.991$

Very strong positive linear correlation.

5.3 Spearman's Rank Correlation Coefficient HL

Spearman's rank correlation coefficient r_s measures the strength and direction of a **monotonic** relationship between two variables. It is appropriate when:

- Data are ordinal (ranked categories), or
- The bivariate data are not normally distributed, or
- The relationship may be monotonic but not necessarily linear

Formula:

$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2-1)}$$

where d_i is the difference in ranks for the i-th pair and n is the number of pairs.

Tied ranks: If two or more values are equal, assign each the average of the ranks they would have occupied. For example, if the 3rd and 4th values are tied, both receive rank 3.5.

IB TIP

On Paper 3, your GDC can calculate Spearman's rank correlation directly — but you must also be able to rank data by hand and apply the formula step by step. Expect to show the ranking process and the d_i^2 table in your written working.

EXAM ALERT

Two common errors: (1) forgetting to rank the raw data before computing differences — d_i is the difference in **ranks**, not in raw values; (2) confusing Spearman's r_s with Pearson's r. Pearson's measures linear association and requires roughly bivariate-normal data; Spearman's measures monotonic association and makes no distributional assumption.

WORKED EXAMPLE

Spearman's Rank — Step-by-Step

Six athletes are assessed by two judges. Their scores are:

Athlete	A	B	C	D	E	F
Judge 1 score	85	70	92	78	65	88
Judge 2 score	80	74	90	75	68	85

Calculate Spearman's rank correlation coefficient.

Step 1 — Rank each judge's scores (rank 1 = highest):

Athlete	Judge 1 score	Rank 1	Judge 2 score	Rank 2
A	85	3	80	3
B	70	5	74	4
C	92	1	90	1
D	78	4	75	5
E	65	6	68	6
F	88	2	85	2

Step 2 — Compute d_i and d_i^2:

Athlete	Rank 1	Rank 2	d_i	d_i^2
A	3	3	0	0
B	5	4	1	1
C	1	1	0	0
D	4	5	-1	1
E	6	6	0	0
F	2	2	0	0
Sum				**2**

Step 3 — Apply the formula ($n = 6, \sum d_i^2 = 2$):

$r_s = 1 - \frac{6 \times 2}{6(36-1)} = 1 - \frac{12}{210} = 1 - 0.0571 \approx 0.943$

Conclusion: $r_s \approx 0.943$ indicates very strong positive agreement between the two judges' rankings.

5.4 Linear Regression

The **regression line of y on x** (also called the least-squares regression line) minimises the sum of squared residuals. Its equation is:

$y = ax + b$

where the slope and intercept are:

$a = \frac{S_{xy}}{S_{xx}} \qquad b = \bar{y} - a\bar{x}$

The regression line **always passes through the point of means** $(\bar{x}, \bar{y})$.

IB TIP

In IB notation the regression line is written $y = ax + b$ (not $\hat{y} = \beta_0 + \beta_1 x$). The GDC (Casio) gives a and b directly via `STAT → REG → ax+b`. Always write out the full equation with the numerical values of a and b, not just "the regression line."

WORKED EXAMPLE

Finding and Using the Regression Line

Using the data from the PMCC example above, find the regression line $y = ax + b$ *and estimate the exam mark for a student who studies for 7 hours.*

Step 1 — Calculate a:

$$a = \frac{S_{xy}}{S_{xx}} = \frac{190}{40} = 4.75$$

Step 2 — Calculate b:

$$b = \bar{y} - a\bar{x} = 71.4 - 4.75 \times 6 = 71.4 - 28.5 = 42.9$$

Regression line: $y = 4.75x + 42.9$

Prediction at $x = 7$:

$$y = 4.75(7) + 42.9 = 33.25 + 42.9 = 76.15 \approx 76$$

EXAM ALERT

Do not extrapolate beyond the data range. The regression line is only reliable for x values within the range of the original data (here, $2 \leq x \leq 10$). Predicting outside this range — for example, estimating the score for 20 hours of study — may give absurd or meaningless results. State this limitation explicitly if asked.

WORKED EXAMPLE

Coefficient of Determination r^2

For the study data, $r \approx 0.991$. *Interpret* r^2.

$$r^2 \approx (0.991)^2 \approx 0.982$$

Interpretation: Approximately 98.2% of the variation in exam marks is explained by the linear relationship with hours of study. This suggests the linear model is an excellent fit for these data.

WORKED EXAMPLE

Regression with GDC — Full Workflow

The following data shows temperature (x, in °C) and ice cream sales per day (y, in units):

x	15	18	22	25	28	30
y	80	110	145	170	200	220

(a) Find r and the regression line. (b) Estimate sales when temperature is 20°C. (c) Comment on reliability.

Part (a) — Using GDC (Casio):

1. Enter x values in List 1, y values in List 2
2. `STAT → CALC → 2VAR` to obtain: $\bar{x} = 23, \bar{y} = 154.17, r \approx 0.999$
3. `STAT → REG → ax+b`: $a \approx 9.77, b \approx -70.5$

Regression line: $y = 9.77x - 70.5$

Part (b): $y = 9.77(20) - 70.5 = 195.4 - 70.5 = 124.9 \approx 125$ units

Part (c): $r \approx 0.999$, which indicates a very strong positive linear correlation. Since $x = 20$ lies within the data range $[15, 30]$, this is interpolation and the prediction is reliable. The model explains $r^2 \approx 99.8\%$ of the variation in sales.

Section 6: Quick Reference

MEMORISE THIS

Probability Rules

Rule	Formula
Addition	$P(A \cup B) = P(A) + P(B) - P(A \cap B)$
Complement	$P(A') = 1 - P(A)$
Conditional	$P(A \mid B) = \dfrac{P(A \cap B)}{P(B)}$
Multiplication	$P(A \cap B) = P(A \mid B) \cdot P(B)$
Independence	$P(A \cap B) = P(A) \cdot P(B)$
Bayes' theorem	$P(A \mid B) = \dfrac{P(B \mid A) \cdot P(A)}{P(B \mid A) \cdot P(A) + P(B \mid A') \cdot P(A')}$

MEMORISE THIS

Discrete Distributions

Distribution	Parameters	P.M.F.	Mean	Variance
General DRV	—	$P(X=x)$ given	$\sum x \cdot P(X=x)$	$E(X^2)-[E(X)]^2$
Binomial	n, p	$\binom{n}{x}p^x(1-p)^{n-x}$	np	$np(1-p)$
Poisson	m	$\frac{e^{-m}m^x}{x!}$	m	m

MEMORISE THIS

Continuous Distributions

Distribution	Parameters	Key formula	Mean	Variance
Normal	μ, σ^2	$Z = \frac{X-\mu}{\sigma}$	μ	σ^2
Standard Normal	$0, 1$	$P(Z \leq z)$ from table/GDC	0	1

68-95-99.7 rule: $P(\mu - k\sigma < X < \mu + k\sigma) \approx 68.3\%, 95.4\%, 99.7\%$ for $k = 1, 2, 3$.

MEMORISE THIS

Hypothesis Testing Summary

Test	H_0	Test statistic	ν	Conditions
χ^2 GOF	Distribution fits	$\chi^2 = \sum \frac{(O-E)^2}{E}$	$k-1-m$	$E \geq 5$ all cells
χ^2 independence	Variables independent	$\chi^2 = \sum \frac{(O-E)^2}{E}$	$(r-1)(c-1)$	$E \geq 5$ all cells
1-sample t	$\mu = \mu_0$	$t = \frac{\bar{x}-\mu_0}{s/\sqrt{n}}$	$n-1$	Normal population or n large
2-sample t	$\mu_1 = \mu_2$	pooled t	n_1+n_2-2	Independent samples
Paired t HL	$\mu_d = 0$	$t = \frac{\bar{d}}{s_d/\sqrt{n}}$	$n-1$	Matched pairs

MEMORISE THIS

Correlation and Regression

Quantity	Formula
PMCC	$r = \frac{S_{xy}}{\sqrt{S_{xx} \cdot S_{yy}}}$
Spearman's rank HL	$r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$
Regression slope	$a = \frac{S_{xy}}{S_{xx}}$
Regression intercept	$b = \bar{y} - a\bar{x}$
Regression line	$y = ax + b$ passes through $(\bar{x}, \bar{y})$
Coefficient of determination	r^2 = proportion of variance in y explained by x
S_{xy}	$\sum x_i y_i - n\bar{x}\bar{y}$
S_{xx}	$\sum x_i^2 - n\bar{x}^2$
S_{yy}	$\sum y_i^2 - n\bar{y}^2$

MEMORISE THIS

Linear Transformation Rules

Property	Formula
$E(aX+b)$	$a \cdot E(X)+b$
$\text{Var}(aX+b)$	$a^2 \cdot \text{Var}(X)$
$\text{SD}(aX+b)$	$
$E(X+Y)$	$E(X)+E(Y)$
$\text{Var}(X+Y)$ (independent)	$\text{Var}(X)+\text{Var}(Y)$

Related Topics

Looking to deepen your understanding or see how statistics connects to other areas? These guides complement the concepts covered here:

Calculus — Integration provides the mathematical foundation for continuous probability distributions. The normal distribution probabilities you calculate in Section 3 are actually definite integrals of the probability density function. If you want to understand where the formulas come from rather than just applying them, review Section 4 (Integration) and Section 5 (Applications of Integration) in the calculus guide.

Functions — Probability density functions, cumulative distribution functions, and regression equations are all special types of functions. Understanding domain, range, composition, and inverse functions (Sections 1–2) helps you interpret statistical models and work confidently with percentiles and quantiles.

Number & Algebra — Sigma notation and series appear throughout expected value and variance calculations. If you need to strengthen your algebra skills for manipulating summations or working with binomial expansions in probability, that guide provides the foundation.

Mixed Practice — Exam Style

IB TIP

How to use this section: Unlike topic-specific practice, these questions are interleaved — they mix all topics from this guide in random order. Before answering, identify *which concept or topic area* the question is testing. This is exactly the skill you need on Paper 2 and Paper 3, where you don't know in advance which topic each question covers.

1. **[Normal Distribution]** A continuous random variable $X \sim N(20, 16)$. Find $P(16 < X < 28)$.

   A. 0.7745

   B. 0.8186

   C. 0.9772

D. 0.6827

2. **[Conditional Probability]** A bag contains 4 red and 6 blue balls. Two balls are drawn without replacement. Given that the first ball drawn is red, what is the probability the second is also red?

   A. $\frac{4}{10}$

   B. $\frac{16}{100}$

   C. $\frac{3}{9}$

   D. $\frac{4}{9}$

3. **[Poisson Distribution]** A radioactive source emits on average 3 particles per second. Find the probability that exactly 5 particles are emitted in a given second. Leave your answer in exact form.

   A. $\frac{3^5 e^{-3}}{5!}$

   B. $\frac{5^3 e^{-5}}{3!}$

   C. $\binom{5}{3}(0.3)^3(0.7)^2$

   D. e^{-3}

4. **[Hypothesis Testing — Chi-Squared]** A chi-squared test of independence at the 5% significance level gives $\chi^2_{\text{calc}} = 6.12$ with 2 degrees of freedom. The critical value is 5.991. What is the correct conclusion?

   A. Accept H_0 — there is sufficient evidence of association

   B. Reject H_0 — there is sufficient evidence of association at the 5% level

   C. Reject H_0 — the variables are definitely not independent

   D. Accept H_0 — the test is inconclusive because the p-value is unknown

5. **[Bayes' Theorem]** A medical test for a disease has sensitivity 95% (probability of a positive result given disease) and specificity 90% (probability of a negative result given no disease). The disease prevalence is 2%. A patient tests positive. What is the approximate probability they have the disease?

   A. 95%

   B. 16%

   C. 50%

D. 2%

6. **[Binomial Distribution]** A fair coin is tossed 8 times. Find the probability of obtaining fewer than 3 heads.

A. $\binom{8}{2}\left(\frac{1}{2}\right)^8$

B. $\sum_{k=0}^{2}\binom{8}{k}\left(\frac{1}{2}\right)^8$

C. $3\left(\frac{1}{2}\right)^8$

D. $1-\sum_{k=3}^{8}\binom{8}{k}\left(\frac{1}{2}\right)^8$

7. **[Correlation and Regression]** The regression line of y on x is $y = 2.4x - 1.8$ and $\bar{x} = 5$. Find $\bar{y}$.

A. 10.2

B. 13.8

C. 1.8

D. 12

8. **[Normal Distribution — Inverse]** $X \sim N(\mu, \sigma^2)$ with $P(X > 72) = 0.10$ and $P(X < 48) = 0.05$. Which system of equations is correct?

A. $\frac{72-\mu}{\sigma} = 1.282$ and $\frac{48-\mu}{\sigma} = -1.645$

B. $\frac{72-\mu}{\sigma} = 0.10$ and $\frac{48-\mu}{\sigma} = -0.05$

C. $\sigma = 72 - \mu$ and $\sigma = \mu - 48$

D. $72\mu = 1.282$ and $48\mu = 1.645$

9. **[Conditional Probability — Independence]** Events A and B satisfy $P(A) = 0.4$, $P(B) = 0.5$, and $P(A \cap B) = 0.2$. Which statement is correct?

A. A and B are mutually exclusive

B. A and B are independent

C. A and B are neither mutually exclusive nor independent

D. A and B are both mutually exclusive and independent

10. **[Hypothesis Testing — Interpretation]** A student performs a one-tailed t-test and obtains $p = 0.032$. At the 5% significance level, the correct interpretation

is:

A. There is a 3.2% probability that H_0 is true

B. There is a 3.2% probability that the result occurred by chance if H_0 is true; reject H_0

C. There is a 96.8% probability that H_1 is true

D. The result is not statistically significant at the 5% level

▶ Show Answers

IB Math IA Ideas — Statistics and Probability

IB TIP

Exploration topics from this chapter:

- **Does home advantage exist in football?** — Collect win/draw/loss records for home and away matches across a season and use a chi-squared test of independence to determine whether venue is statistically associated with result. Extend by comparing leagues across different countries to investigate whether the effect size varies.
- **Modelling goal-scoring with the Poisson distribution** — Goals per match in football (or other sports) often follow a $\mathrm{Po}(\lambda)$ distribution. Estimate λ from real data, perform a goodness-of-fit chi-squared test, and investigate whether the Poisson assumption holds equally well for high- and low-scoring teams.
- **The birthday problem and simulation** — Derive analytically the probability that at least two people in a group of n share a birthday, then verify with a Monte Carlo simulation. Extend to non-uniform birthday distributions using real birth-rate data (e.g., from national statistics offices) to see how much the real probability deviates from the uniform-distribution model.
- **Income inequality and the Gini coefficient** — Obtain income-distribution data from the World Bank or OECD. Fit a log-normal distribution to model incomes, compute the theoretical Gini coefficient from the distribution parameters, and compare to the empirical value. Investigate how the Gini coefficient has changed over time for a country of your choice.
- **Regression analysis in sport or health** — Choose two quantitative variables with a plausible causal link (e.g., hours of sleep and reaction time, training load and performance, or diet and cholesterol). Collect or source real data, compute the regression line and r^2, test the significance of the correlation, and critically evaluate confounding factors.

- **Bayesian updating and medical testing** — Use Bayes' theorem to model how the probability that a patient has a disease changes as successive independent tests come back positive. Investigate how sensitivity, specificity, and prevalence interact, and calculate the number of positive tests needed to exceed a 95% posterior probability of disease.
- **Does music tempo affect heart rate? A hypothesis test** — Design a small experiment: measure resting heart rate, play fast and slow music, measure again. Use a paired t-test to test whether tempo has a significant effect. Discuss Type I and Type II errors and how sample size affects the power of the test.

Tip: A strong IA has a clear personal engagement angle. Pick a topic that connects to something you genuinely find interesting — sport, health, economics, or psychology — and let the mathematics serve your question, not the other way around.

November 2026 Prediction Questions

⚠ EXAM ALERT

These are NOT official IB questions. These are trend-based practice questions written to reflect the topic areas and question styles most likely to appear on the November 2026 IB Math AA HL Paper 2. Based on recent exam patterns (2022-2025), expect heavy weighting on: hypothesis testing (chi-squared and t-tests), normal distribution calculations, Bayes' theorem, and linear regression.

WORKED EXAMPLE

Question 1 [Hypothesis Testing] [~8 marks]

A factory claims that the mean mass of its cereal boxes is 500 g. A quality inspector takes a random sample of 12 boxes and records the following masses (in grams):

$498, 502, 495, 501, 497, 503, 496, 499, 504, 498, 500, 497$

(a) State appropriate null and alternative hypotheses for a two-tailed test.

(b) The sample mean is $\bar{x} = 499.17$ g and the sample standard deviation is $s = 2.89$ g. Calculate the t-statistic for this test.

(c) The critical values for a two-tailed t-test at the 5% significance level with 11 degrees of freedom are ± 2.201. State the conclusion of the test, justifying your answer.

(d) State one assumption required for this test to be valid.

▶ Show Solution

WORKED EXAMPLE

Question 2 [Bayes' Theorem] [~7 marks]

A medical screening test for a disease has the following properties:

- The probability of a positive result given the patient has the disease (sensitivity) is 0.95.
- The probability of a negative result given the patient does not have the disease (specificity) is 0.90.
- The prevalence of the disease in the population is 0.02.

(a) Construct a tree diagram or define the events and their probabilities.

(b) Find the probability that a randomly selected person tests positive.

(c) Find the probability that a person who tests positive actually has the disease.

(d) Comment on the practical implications of your answer to part (c).

▶ Show Solution

WORKED EXAMPLE

Question 3 [Regression and Correlation] [~6 marks]

A researcher collects data on the number of hours spent studying (x) and exam score (y) for 8 students. The regression line of y on x is found to be:

$\hat{y} = 3.2x + 42.5$

The Pearson correlation coefficient is $r = 0.87$ and $\bar{x} = 6.5$.

(a) Interpret the value of the gradient (3.2) in context.

(b) Calculate the predicted exam score for a student who studies for 6.5 hours.

(c) Explain why it would be unreliable to use this model to predict the exam score for a student who studies for 20 hours.

(d) The coefficient of determination is r^2. Calculate r^2 and interpret it in context.

▶ Show Solution