Statistics and Probability

IB SL Study Guide

Contents

IB Math AI SL — Statistics and Probability

Complete Study Guide

Topics Covered

1. Descriptive Statistics — measures of central tendency and spread
2. Data Presentation — histograms, box plots, cumulative frequency
3. Probability — combined events, conditional probability, tree diagrams
4. Probability Distributions — binomial and normal
5. Statistical Tests — chi-squared test for independence
6. Correlation and Regression — linear regression, r and R^2
7. Practice Questions and Exam Alerts

Topic 4 of the IB Math AI SL syllabus — this is the largest topic at 36 SL hours.

IB TIP

The heart of Math AI: Statistics and probability makes up the largest portion of the syllabus and is heavily represented on both papers. Expect at least one full extended-response question on Paper 2 to be purely statistical. Master your GDC's statistics functions — they are essential.

MEMORISE THIS

Key statistics formulas

Measure	Formula
Mean	$\bar{x} = \frac{\sum x_i}{n}$ or $\bar{x} = \frac{\sum f_i x_i}{\sum f_i}$
Standard deviation	$\sigma = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n}}$ (population)
Interquartile range	$\text{IQR} = Q_3 - Q_1$
Probability	$P(A) = \frac{n(A)}{n(U)}$
Combined events	$P(A \cup B) = P(A) + P(B) - P(A \cap B)$
Conditional probability	$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$

Section 1: Descriptive Statistics

1.1 Measures of Central Tendency

Measure	What it tells you	When to use
Mean ($\bar{x}$)	Average value	Symmetric data, no extreme outliers
Median	Middle value	Skewed data or data with outliers
Mode	Most frequent value	Categorical data

1.2 Measures of Spread

Measure	Formula / Description
Range	$\max - \min$
Interquartile range (IQR)	$Q_3 - Q_1$ (middle 50% of data)
Standard deviation (σ or s)	Measures spread around the mean
Variance	σ^2 (the square of standard deviation)

IB TIP

GDC for statistics: Enter data into a list and use 1-Var Stats (TI-84) or STAT CALC (Casio). This gives you the mean, median, quartiles, standard deviation, and more instantly. Never calculate these by hand on Paper 2.

1.3 Outliers

An **outlier** is typically defined as any value:

- Below $Q_1 - 1.5 \times \text{IQR}$, or
- Above $Q_3 + 1.5 \times \text{IQR}$

WORKED EXAMPLE

Descriptive statistics in context

The daily rainfall (mm) in a city over 10 days: 0, 0, 2, 3, 5, 7, 8, 12, 15, 48.

Using GDC: $\bar{x} = 10.0$, median $= 6.0$, $Q_1 = 2$, $Q_3 = 12$, $\sigma = 13.3$.

$\text{IQR} = 12 - 2 = 10$.

Outlier boundaries: $2 - 15 = -13$ and $12 + 15 = 27$. The value 48 is above 27, so it is an **outlier**.

The median (6.0) is a better measure of centre than the mean (10.0) because the outlier pulls the mean up.

Section 2: Data Presentation

2.1 Histograms

A **histogram** shows the distribution of continuous data. The area of each bar is proportional to the frequency.

For **unequal class widths**, the y-axis shows **frequency density** = $\dfrac{\text{frequency}}{\text{class width}}$.

2.2 Cumulative Frequency Curves

Plot cumulative frequency against the **upper boundary** of each class. Use to read off:

- Median: at $\frac{n}{2}$
- Q_1: at $\frac{n}{4}$
- Q_3: at $\frac{3n}{4}$
- Percentiles: at the appropriate fraction of n

2.3 Box Plots (Box-and-Whisker Diagrams)

A box plot shows: minimum, Q_1, median, Q_3, maximum. Outliers are shown as individual points.

Comparing distributions: When two box plots are shown side by side, compare:

- Centre (median) — which group has higher/lower values
- Spread (IQR and range) — which group is more variable
- Skewness — if the median is closer to Q_1, data is positively skewed

> ⚠ **EXAM ALERT**
>
> **Box plot comparison questions** are almost guaranteed on the exam. Always make **two comparisons** (one about centre, one about spread) and refer to the **context** (not just "Dataset A has a higher median" but "Students in Class A scored higher on average").

Section 3: Probability

3.1 Basic Probability

$P(A) = \frac{\text{number of favourable outcomes}}{\text{total number of outcomes}}$

- $0 \leq P(A) \leq 1$
- $P(\text{not } A) = 1 - P(A)$

3.2 Combined Events

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

If A and B are **mutually exclusive** ($A \cap B = \emptyset$): $P(A \cup B) = P(A) + P(B)$

Multiplication rule: $P(A \cap B) = P(A) \times P(B \mid A)$

If A and B are **independent**: $P(A \cap B) = P(A) \times P(B)$

3.3 Conditional Probability

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

WORKED EXAMPLE

Conditional probability — medical testing

A disease affects 2% of a population. A test correctly identifies the disease 95% of the time (sensitivity) and correctly identifies healthy people 90% of the time (specificity). If a person tests positive, what is the probability they have the disease?

Let D = has disease, T = tests positive.

$P(D) = 0.02$, $P(T \mid D) = 0.95$, $P(T \mid D') = 0.10$.

$P(T) = P(T \mid D) \times P(D) + P(T \mid D') \times P(D') = 0.95 \times 0.02 + 0.10 \times 0.98 = 0.019 + 0.098 = 0.117$

$$P(D \mid T) = \frac{P(T \mid D) \times P(D)}{P(T)} = \frac{0.019}{0.117} = 0.162$$

Only a 16.2% chance of actually having the disease, despite the positive test. This is a classic result that highlights the importance of base rates.

3.4 Tree Diagrams and Two-Way Tables

Tree diagrams are useful for sequential events. Multiply along branches, add between branches.

Two-way tables organize data for two categorical variables.

IB TIP

Which tool to use? If the events are sequential (first this, then that), use a **tree diagram**. If you have data classified by two categories, use a **two-way table**. On the exam, drawing the correct diagram usually earns a method mark even if the final answer is wrong.

Section 4: Probability Distributions

4.1 Discrete Random Variables

A **discrete random variable** X takes specific values with known probabilities. The probabilities must sum to 1: $\sum P(X = x) = 1$.

Expected value (mean): $E(X) = \mu = \sum x \cdot P(X = x)$

4.2 Binomial Distribution

Use when:

- Fixed number of **independent** trials (n)
- Two outcomes only (success/failure)
- Constant probability of success (p)

$X \sim B(n, p)$

$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$

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

WORKED EXAMPLE

Binomial distribution — quality control

In a factory, 8% of items are defective. A sample of 20 items is tested. Find (a) the probability that exactly 2 are defective, (b) the probability that at most 1 is defective, (c) the expected number of defective items.

$X \sim B(20, 0.08)$

(a) Using GDC: $P(X = 2) = \binom{20}{2}(0.08)^2(0.92)^{18} = 0.2711$

(b) $P(X \leq 1) = P(X = 0) + P(X = 1) = (0.92)^{20} + 20(0.08)(0.92)^{19}$

Using GDC binomcdf: $P(X \leq 1) = 0.1887 + 0.3282 = 0.5169$

(c) $E(X) = np = 20 \times 0.08 = 1.6$ defective items.

4.3 Normal Distribution

The **normal distribution** is a continuous, bell-shaped, symmetric distribution. It is fully described by its mean μ and standard deviation σ.

$X \sim N(\mu, \sigma^2)$

Key properties:

- Symmetric about the mean
- 68% of data within 1σ of the mean
- 95% within 2σ
- 99.7% within 3σ

MEMORISE THIS

The 68-95-99.7 rule

Range	Percentage
$\mu \pm \sigma$	68%
$\mu \pm 2\sigma$	95%
$\mu \pm 3\sigma$	99.7%

4.4 Calculating Normal Probabilities with GDC

Forward problem (given x, find probability):

- TI-84: `normalcdf(lower, upper, mean, sd)`
- Casio: `P(lower < X < upper)` in DIST menu

Inverse problem (given probability, find x):

- TI-84: `invNorm(area, mean, sd)`
- Casio: `InvN` in DIST menu

WORKED EXAMPLE

Normal distribution — exam scores

Exam scores are normally distributed with mean 65 and standard deviation 12. Find (a) the probability a student scores above 80, (b) the score that 90% of students exceed.

$X \sim N(65, 12^2)$

(a) $P(X > 80) = \text{normalcdf}(80, 10^{99}, 65, 12) = 0.1056$

About 10.6% of students score above 80.

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

$x = \text{invNorm}(0.10, 65, 12) = 49.6$

90% of students score above 49.6.

EXAM ALERT

Always sketch the normal curve. Draw the bell curve, mark the mean, shade the area you are finding. This helps you check whether your answer is reasonable (e.g., a probability above 0.5 for a value below the mean).

Section 5: Chi-Squared Test for Independence

The χ^2 test determines whether two categorical variables are **independent** (unrelated) or **associated**.

5.1 Setting Up the Test

1. **State hypotheses:**
   - H_0: The variables are independent
   - H_1: The variables are not independent
2. **Create observed frequency table** from data
3. **Calculate expected frequencies:** $E = \dfrac{\text{row total} \times \text{column total}}{\text{grand total}}$
4. **Calculate the test statistic:** $\chi^2_{\text{calc}} = \sum \dfrac{(O-E)^2}{E}$
5. **Find degrees of freedom:** $\text{df} = (\text{rows} - 1)(\text{columns} - 1)$
6. **Compare** χ^2_{calc} with the critical value at the given significance level, **or** compare the p-value with α.
7. **Conclude** in context.

> **IB TIP**
>
> **GDC does it all:** Enter the observed frequencies as a matrix. Run the χ^2 test. The GDC gives you the test statistic, p-value, degrees of freedom, and expected frequencies. You just need to set up hypotheses, run the test, and write the conclusion.

WORKED EXAMPLE

Chi-squared test — favourite subject by gender

A survey asked 200 students their favourite subject. The results:

	Science	Arts	Sport	Total
Male	45	25	30	100
Female	35	40	25	100
Total	80	65	55	200

Test at the 5% significance level whether favourite subject is independent of gender.

H_0: Favourite subject is independent of gender.

H_1: Favourite subject is not independent of gender.

Expected frequencies: For Male-Science: $E = \dfrac{100 \times 80}{200} = 40$.

Full expected table:

	Science	Arts	Sport
Male	40	32.5	27.5
Female	40	32.5	27.5

$$\chi^2_{\text{calc}} = \frac{(45-40)^2}{40} + \frac{(25-32.5)^2}{32.5} + \frac{(30-27.5)^2}{27.5} + \frac{(35-40)^2}{40} + \frac{(40-32.5)^2}{32.5} + \frac{(25-27.5)^2}{27.5}$$

$$= 0.625 + 1.731 + 0.227 + 0.625 + 1.731 + 0.227 = 5.17$$

$\text{df} = (2-1)(3-1) = 2$. Critical value at 5%: $\chi^2_{0.05,2} = 5.991$.

Since $5.17 < 5.991$, we **do not reject** H_0.

There is insufficient evidence at the 5% level to conclude that favourite subject depends on gender.

EXAM ALERT

Writing the conclusion: Always state the conclusion in the **context** of the question, not in statistical jargon. Say "There is insufficient evidence that favourite subject depends on gender" rather than "We fail to reject the null hypothesis." Also state whether you are comparing with a critical value or using the p-value.

5.2 Conditions for the Chi-Squared Test

- Data must be **frequencies** (counts), not percentages or proportions
- **Expected frequencies** should all be at least 5 (the IB may ask you to check this)
- Observations must be **independent**

Section 6: Correlation and Regression

6.1 Scatter Plots

A scatter plot shows the relationship between two quantitative variables. Describe the relationship using:

- **Direction:** positive (both increase), negative (one increases as other decreases)
- **Strength:** strong, moderate, weak
- **Form:** linear, non-linear, no correlation

6.2 Pearson's Correlation Coefficient (r)

r measures the strength and direction of a **linear** relationship.

Value of r	Interpretation
$r = 1$	Perfect positive linear
$0.7 \leq r < 1$	Strong positive
$0.4 \leq r < 0.7$	Moderate positive
$0 < r < 0.4$	Weak positive
$r = 0$	No linear correlation
Negative values	Same interpretation, negative direction

> ⚠ **EXAM ALERT**
>
> **r only measures linear correlation.** Two variables can have a strong non-linear relationship (e.g., quadratic) but $r \approx 0$. Always check the scatter plot.

6.3 Linear Regression

The **least squares regression line** $\hat{y} = a + bx$ minimizes the sum of squared residuals.

- $b = \dfrac{S_{xy}}{S_{xx}}$ (gradient — in the formula booklet)
- $a = \bar{y} - b\bar{x}$ (the line passes through $(\bar{x}, \bar{y})$)

Use the regression line for **interpolation** (predicting within the data range). Be cautious with **extrapolation**.

WORKED EXAMPLE

Regression — advertising and sales

A company records weekly advertising spend (x thousands) and sales (y thousands):

x	2	4	6	8	10
y	15	22	28	35	40

Using GDC linear regression: $\hat{y} = 3.1x + 9.1$, $r = 0.998$.

Interpretation: For each additional 1000 spent on advertising, sales increase by approximately 3100 (thousands of dollars). The very high r value indicates a strong positive linear relationship.

Predict sales for $x = 7$: $\hat{y} = 3.1(7) + 9.1 = 30.8$ thousand dollars. This is **interpolation** (within the data range) and is reliable.

Predict sales for $x = 20$: $\hat{y} = 3.1(20) + 9.1 = 71.1$ thousand. This is **extrapolation** (far beyond the data) and is unreliable — the linear trend may not continue.

6.4 Coefficient of Determination (R^2)

$R^2 = r^2$ represents the proportion of variation in y explained by the linear relationship with x.

For example, if $r = 0.9$, then $R^2 = 0.81$, meaning 81% of the variation in y is explained by x.

Section 7: Practice Questions

Paper 1 Style (Short Answer)

▶ **Q1.** A dataset has $Q_1 = 23$, $Q_3 = 41$. (a) Find the IQR. (b) Determine the outlier boundaries.

▶ **Q2.** Two fair dice are rolled. Find the probability that the sum is at least 10.

▶ **Q3.** Heights of students are normally distributed with mean 168 cm and standard deviation 7 cm. Find the probability that a randomly selected student is between 160 cm and 175 cm tall.

Paper 2 Style (Extended Response)

▶ **Q4.** A researcher records the hours of study (x) and exam score (y) for 8 students. The GDC gives: $\bar{x} = 5.5$, $\bar{y} = 68$, $r = 0.92$, regression line $\hat{y} = 4.8x + 41.6$. (a) Describe the correlation. (b) Interpret the gradient. (c) Predict the score for a student who studies 7 hours. (d) A student claims this proves studying causes higher scores. Comment.

▶ **Q5.** A company claims that 85% of its deliveries arrive on time. In a random sample of 25 deliveries, 18 arrived on time. (a) Using a binomial model, find the probability of 18 or fewer on-time deliveries if the claim is true. (b) Does this provide evidence against the company's claim at the 5% significance level?

▶ **Q6.** A survey of 150 employees classified by department and lunch preference gives the following data. Test at the 10% significance level whether lunch preference is independent of department.

⚠ EXAM ALERT

Hypothesis testing checklist: (1) State H_0 and H_1 in context. (2) State the significance level. (3) Calculate the test statistic or p-value using GDC. (4) Compare with critical value or α. (5) State conclusion in context. Missing any step loses marks.

May 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 May 2026 IB Math AI SL Paper 2. Based on recent exam patterns (2022–2025), expect heavy weighting on: normal distribution (finding probabilities and inverse values with GDC), chi-squared test of independence with a full contingency table, and linear regression including Pearson's r interpretation and prediction reliability.

✎ WORKED EXAMPLE

Question 1 — Normal Distribution [~8 marks]

The masses of apples from an orchard are normally distributed with mean 182 g and standard deviation 24 g. An apple is selected at random.

(a) Find the probability that the apple has a mass greater than 200 g.

(b) Find the probability that the apple has a mass between 150 g and 210 g.

(c) The lightest 10% of apples are classified as "grade C" and sold at a discount. Find the maximum mass of a grade C apple.

(d) A crate holds 30 apples selected at random. Find the expected number of apples in the crate with mass greater than 200 g.

▶ Show Solution

WORKED EXAMPLE

Question 2 — Chi-Squared Test of Independence [~8 marks]

A sports club surveyed 180 of its members on their preferred activity, categorized by age group. The results are shown below.

	Swimming	Tennis	Gym	Total
Under 30	22	18	40	80
30–50	28	24	18	70
Over 50	20	8	2	30
Total	70	50	60	180

(a) State the null and alternative hypotheses for a chi-squared test.

(b) Calculate the expected frequency for the "Under 30 / Swimming" cell. Show your working.

(c) Use your GDC to find the chi-squared test statistic and the p-value.

(d) State the number of degrees of freedom.

(e) At the 5% significance level, determine whether preferred activity is independent of age group. State your conclusion in context.

▶ Show Solution

WORKED EXAMPLE

Question 3 — Linear Regression and Prediction [~7 marks]

A researcher collects data on the average daily temperature (x degrees C) and the number of visitors (y) to an outdoor museum on 8 selected days.

x	12	15	17	19	22	24	27	29
y	210	265	290	330	410	450	500	545

(a) Use your GDC to find the equation of the regression line $\hat{y} = ax + b$.

(b) Find Pearson's correlation coefficient r and describe the correlation.

(c) Use your model to predict the number of visitors on a day when the temperature is 20 degrees C.

(d) The researcher uses the model to predict visitor numbers on a day forecast to reach 38 degrees C. State whether this is interpolation or extrapolation, and comment on the reliability of this prediction.

▶ Show Solution