IB DP MATHS AA · TOPIC 4 STATISTICS AND PROBABILITY · SL 4.1–4.3, 4.5, 4.6, 4.11 · AHL 4.13 · ~9 MIN READ

# Statistics and probability

Sampling and data, mean, median, mode, quartiles, standard deviation, outliers, box plots and cumulative frequency, the effect of transforming data, probability rules, Venn and tree diagrams, conditional probability and independence, and Bayes' theorem for HL.

Sampling & bias · Mean, median, mode · Quartiles & IQR · Standard deviation · Outliers · Box plots · Cumulative frequency · Venn diagrams

RS by Riyaz Sir
mentorriyaz.in

No tag = **SL and HL** · HL ONLY HL students only · FORMULA BOOKLET given in the exam
M1 A1 how marks are awarded · highlight = comes up again and again · Test yourself at the end

## IN 30 SECONDS

- **Centre:** mean, median, mode. **Spread:** range, IQR = $Q_3 - Q_1$, standard deviation σ. An **outlier** is more than 1.5 × IQR beyond a quartile.
- **Adding a constant** to every value shifts the mean and median but leaves the spread alone. **Multiplying** by k multiplies both the centre and the spread by k.
- **Probability:** $P(A \cup B) = P(A) + P(B) - P(A \cap B)$, $P(A \mid B) = P(A \cap B)/P(B)$. A and B are **independent** ⇔ $P(A \cap B) = P(A)P(B)$.
- **HL Bayes:** reverse a conditional probability with a tree diagram: $P(A \mid B) = P(A \cap B)/P(B)$, with P(B) summed over all branches.

## BY THE END YOU CAN

- Identify types of data, sampling methods and sources of bias.
- Find and interpret measures of centre and spread, including from grouped data.
- Draw and interpret box plots and cumulative frequency graphs, and identify outliers.
- Describe the effect of adding or multiplying every value by a constant.
- Use Venn diagrams, tree diagrams and sample space diagrams to find probabilities.
- Use conditional probability and test for independence.
- Use Bayes' theorem with two or three events. HL

**SYLLABUS NOTE**

From first exams in 2029, bivariate statistics (scatter diagrams, correlation and regression lines) leaves the AA syllabus. If your exams are before 2029, you still need it: the short section near the end covers it.

## 1 Data and sampling

- **Discrete** data is counted (number of siblings); **continuous** data is measured (height, time).
- **Sampling methods:** simple random, systematic (every kth), stratified (proportional to group sizes), quota, convenience. Know one advantage and one problem for each.
- **Bias** comes from the method, not the size: a large convenience sample is still biased.

**COMMON MISTAKE**

Calling a sample "random" just because you didn't choose it carefully. Random means every member had an equal chance of being picked.

## 2 Centre and spread

$$\bar{x} = \frac{\Sigma f x}{n}$$ FORMULA BOOKLET

mean; for grouped data use the class midpoints

- **σ** (population standard deviation) measures the typical distance from the mean. In IB exams, use the GDC for σ; you are not expected to use the formula by hand.

**WORKED EXAMPLE**

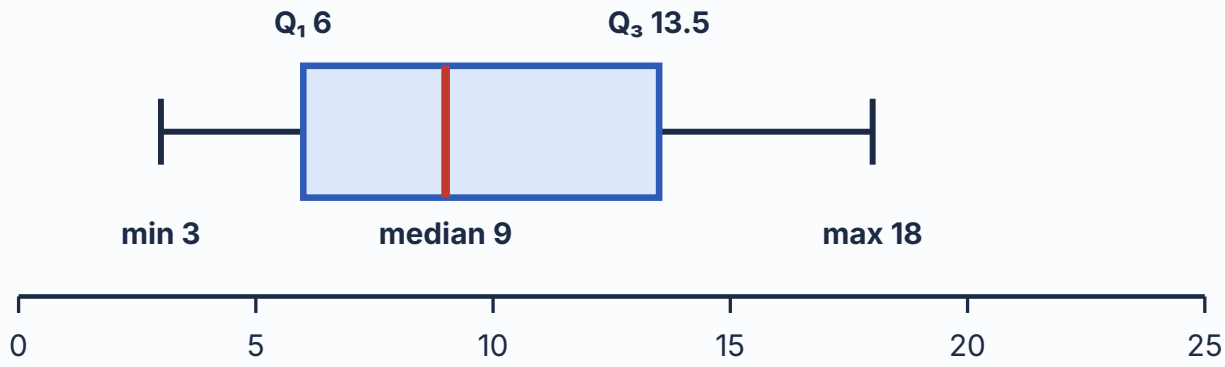
**Find the mean, median, quartiles and IQR of 3, 5, 7, 8, 9, 11, 12, 15, 18.**

1. mean = 88/9 = **9.78** A1
2. median = 5th value = **9** A1
3. **$Q_1 = 6$, $Q_3 = 13.5$, so IQR = 7.5** A1 A1

**WORKED EXAMPLE**

**Are there any outliers in that data set?**

1. $Q_1 - 1.5(7.5) = -5.25$ and $Q_3 + 1.5(7.5) = 24.75$ M1
2. **every value is between these, so no outliers** A1




IQR = 13.5 − 6 = 7.5. Outliers lie beyond $Q_1 - 1.5 \times IQR$ or $Q_3 + 1.5 \times IQR$.

The box holds the middle 50% of the data; the whiskers reach the smallest and largest non-outliers.

WORKED EXAMPLE

**The ages of 30 people are grouped: 0–10 (4), 10–20 (10), 20–30 (12), 30–40 (4). Estimate the mean.**

1. midpoints 5, 15, 25, 35 M1
2. (20 + 150 + 300 + 140)/30 M1
3. **≈ 20.3 years** A1

- **Why "estimate"?** You only know which class each value is in, so the midpoint stands in for every value in the class.

| Change to every value | Mean, median | SD, IQR |
|---|---|---|
| add b | + b | no change |
| multiply by a | × a | × lal |
| y = ax + b | a x̄ + b | lalσ |

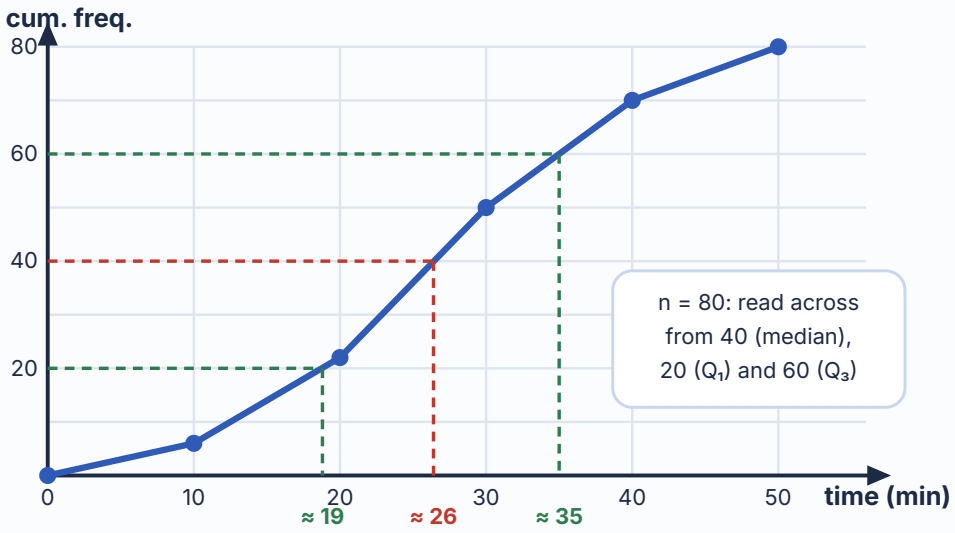
WORKED EXAMPLE

**The mean of a data set is 12 and the SD is 3. Each value is doubled and then 5 is added. Find the new mean and SD.**

1. new mean = 2(12) + 5 = **29** A1
2. **new SD = 2(3) = 6 (adding 5 doesn't change spread)** A1

Riyaz Sir

## 3 Cumulative frequency




Plot cumulative frequency at the upper class boundary, join with a smooth curve or straight lines, then read across and down.

- **Reading off:** median at n/2, $Q_1$ at n/4, $Q_3$ at 3n/4. For "how many took longer than 30 minutes", read up from 30 and subtract from n.
- **Percentiles:** the 90th percentile is read from 0.9n.

## 4 Probability rules

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

complement

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

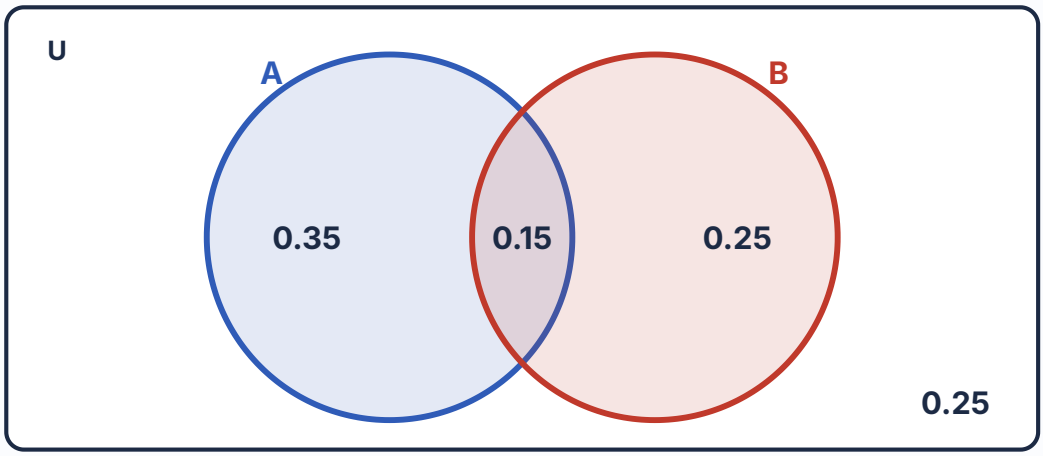
addition rule; for mutually exclusive events $P(A \cap B) = 0$

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

conditional probability

$$P(A \cap B) = P(A)P(B)$$ FORMULA BOOKLET

independent events only




**P(A) = 0.5, P(B) = 0.4, P(A ∩ B) = 0.15, so P(A ∪ B) = 0.75**

Fill a Venn diagram from the middle outwards, then make sure everything adds to 1.

WORKED EXAMPLE

**P(A) = 0.5, P(B) = 0.4 and P(A ∩ B) = 0.15. Find P(A ∪ B) and P(A | B), and decide whether A and B are independent.**

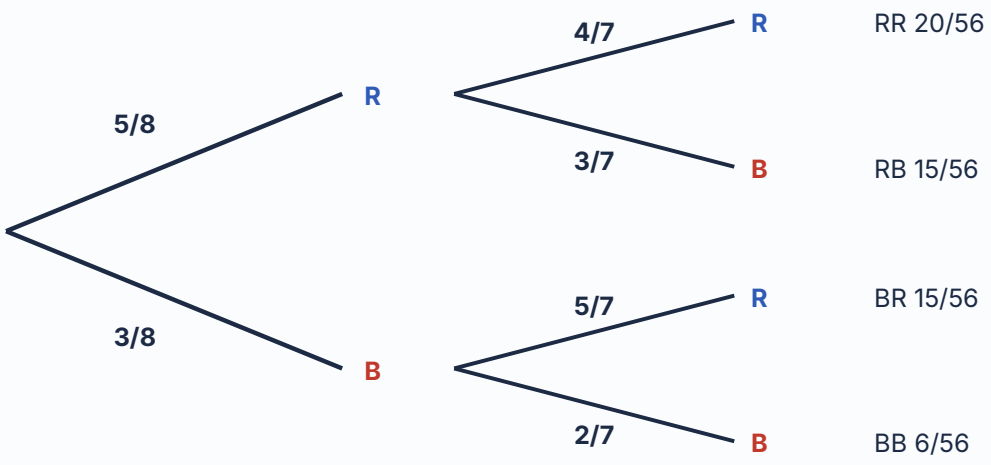
1. P(A ∪ B) = 0.5 + 0.4 − 0.15 = **0.75** A1
2. P(A | B) = 0.15/0.4 = **0.375** A1
3. **P(A)P(B) = 0.2 ≠ 0.15, so not independent** R1

COMMON MISTAKE

Confusing "mutually exclusive" with "independent". Mutually exclusive events can't happen together (P(A ∩ B) = 0); independent events don't affect each other. Two events with non-zero probability can't be both.

## 5 Tree diagrams

**5 red, 3 blue: two picks without replacement**




Multiply along the branches; add the end results you want.

WORKED EXAMPLE

**A bag has 5 red and 3 blue counters. Two are taken without replacement. Find P(both the same colour).**

1. $P(RR) = \frac{5}{8} \times \frac{4}{7} = \frac{20}{56}$ M1
2. $P(BB) = \frac{3}{8} \times \frac{2}{7} = \frac{6}{56}$ M1
3. **total** $= \frac{26}{56} = \frac{13}{28}$ A1

WORKED EXAMPLE

**Given that both counters are the same colour, find the probability that they are both red.**

1. $P(RR \mid \text{same}) = \frac{20/56}{26/56}$ M1
2. $= \frac{10}{13}$ A1

EXAM TIP

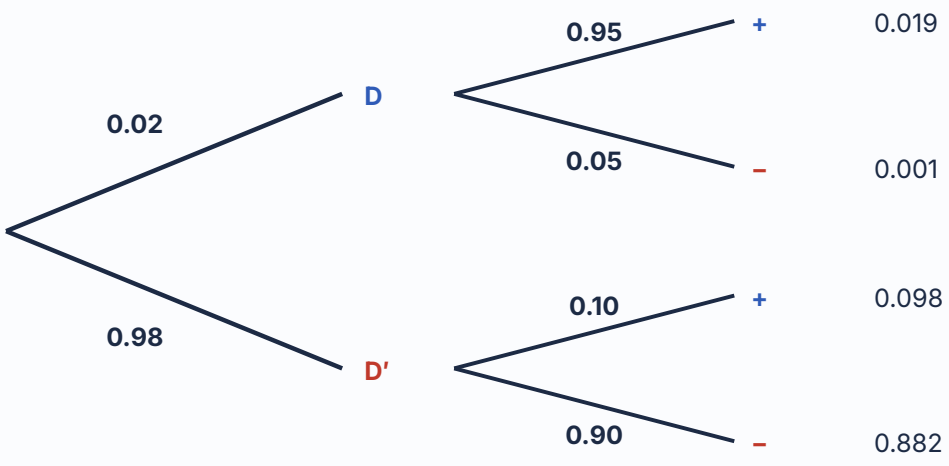
"Given that" means conditional probability: the denominator is the probability of the condition, here P(same colour).

## 6 Bayes' theorem HL

$$P(B \mid A) = \frac{P(B)P(A \mid B)}{P(B)P(A \mid B) + P(B')P(A \mid B')}$$ FORMULA BOOKLET

the tree-diagram version of $P(B \mid A) = P(A \cap B)/P(A)$

**Disease test: 2% have it, 95% detected, 10% false positives**




Most positive tests come from the large healthy group, which is why the answer is so low.

WORKED EXAMPLE

**2% of people have a disease. A test detects 95% of cases but gives a positive result for 10% of healthy people. Find P(disease | positive).**

1. $P(+) = 0.02(0.95) + 0.98(0.10) = 0.019 + 0.098 = 0.117$ M1 A1
2. $P(D \mid +) = \frac{0.019}{0.117}$ M1
3. **= 0.162** A1

- **Three or more events:** the same idea works with three first-stage branches (say, three factories making a part): the denominator adds three products.

## 7 Bivariate statistics (before 2029 only)

- **Pearson's r** ($-1 \le r \le 1$) measures linear correlation; use the GDC. r close to ±1 is strong, close to 0 is weak. Correlation is not causation.
- **Regression line y on x** predicts y from x; **x on y** predicts x from y. Both pass through $(\bar{x}, \bar{y})$. Only interpolate: predictions outside the data range are unreliable.

GDC (CALCULATOR) METHOD

Paper 2: enter data as a list (with a frequency list for grouped data) and use 1-variable statistics for $\bar{x}$, $\sigma$, $Q_1$, median and $Q_3$. Write down the values you use; "from GDC" with the numbers is accepted.

LINKING QUESTION

Probability distributions (Random variables, binomial and normal) are built on these rules, and the mean and SD of a distribution are the theoretical versions of $\bar{x}$ and $\sigma$ here.

IDEA FOR YOUR IA / EXPLORATION

Collect your own data (reaction times, commute times, sleep) and compare two groups with box plots and summary statistics, or test a claim using conditional probability. Explain the sampling method and possible bias.

## Test yourself — cover the answers first

**QUICK CHECK**

1. Is "time to run 100 m" discrete or continuous?
2. Find the median of 4, 9, 2, 7, 5, 1.
3. Every value is increased by 10. What happens to the IQR?
4. P(A) = 0.3. Find P(A′).
5. A and B are independent, P(A) = 0.4 and P(B) = 0.5. Find P(A ∩ B).
6. HL Give the denominator in Bayes' theorem in words.

**EXAM-STYLE**

7. In a class of 30, 18 study French, 12 study Spanish and 5 study both. A student is chosen at random. Find the probability that they study French given that they study Spanish, and the probability that they study neither.
8. HL Machines A, B and C make 50%, 30% and 20% of the parts in a factory. 2%, 3% and 5% of their parts are faulty. A part is faulty. Find the probability it came from C.

**Show answers**

1. continuous
2. 4.5
3. no change
4. 0.7
5. 0.2
6. the total probability of the given event
7. $P(F \mid S) = \frac{5}{12}$
   neither = 30 − (18 + 12 − 5) = 5
   P(neither) = $\frac{1}{6}$
8. P(F) = 0.5(0.02) + 0.3(0.03) + 0.2(0.05) = 0.01 + 0.009 + 0.01 = 0.029
   P(C | F) = 0.01/0.029
   **= 0.345**


**Stuck on a question?**

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

Share with a friend | Ask Riyaz Sir

Want regular help? Send this to your parent →