IB DP MATHEMATICS

AA Topic 4 · Statistics and Probability

MARK SCHEME

4.1 Descriptive Statistics · 4.2 Correlation & Regression
4.3 Probability · 4.4 Probability Distributions

Standard and Higher Level · Syllabus 2025

Fahad H. Ahmad

+92 323 509 4443 | Megalecture.com

Original exam-style questions written for the IB Diploma Programme Mathematics course (first assessment 2025)

Section A — Multiple Choice: Answer Key (15 marks)

1. Answer: **B** — The interquartile range is IQR = Q3 − Q1, the range of the middle 50% of the data.

2. Answer: **B** — The line inside the box of a box-and-whisker plot represents the median (the middle value of the ordered dataset).

3. Answer: **B** — Number of standard deviations = $(28-20) \div 4 = 8 \div 4 = 2$.

4. Answer: **B** — A common rule for identifying outliers is any value more than $1.5 \times IQR$ below Q1 or above Q3.

5. Answer: **B** — Pearson's correlation coefficient r always satisfies $-1 \le r \le 1$, where -1 and 1 represent perfect negative and positive linear correlation respectively.

6. Answer: **B** — A value of r close to -1 (here -0.95) indicates a strong negative linear correlation between the two variables.

7. Answer: **B** — The line of best fit is used to make predictions (interpolation, and sometimes cautious extrapolation) for values based on the observed linear relationship, though correlation does not imply causation.

8. Answer: **C** — A value of r close to 0 (here 0.05) indicates essentially no linear correlation between the two variables.

9. Answer: **B** — The addition rule is $P(A \cup B) = P(A) + P(B) - P(A \cap B)$, which avoids double-counting the overlap between A and B.

10. Answer: **C** — Two events are independent if and only if $P(A \cap B) = P(A) \times P(B)$, i.e. the occurrence of one does not affect the probability of the other.

11. Answer: **A** — Conditional probability is defined as $P(A|B) = P(A \cap B) \div P(B)$, the probability of A occurring given that B has already occurred.

12. Answer: **B** — The overlapping region of a two-circle Venn diagram represents the intersection, $A \cap B$ (outcomes belonging to both A and B).

13. Answer: **A** — A binomial distribution applies to a fixed number n of independent trials, each with exactly two outcomes (success/failure) and a constant probability of success p.

14. Answer: **B** — For a binomial distribution $X \sim B(n,p)$, the mean is $E(X) = np$.

15. Answer: **B** — For a normal distribution, approximately 68% of values lie within ± 1 standard deviation of the mean (the empirical / 68-95-99.7 rule).

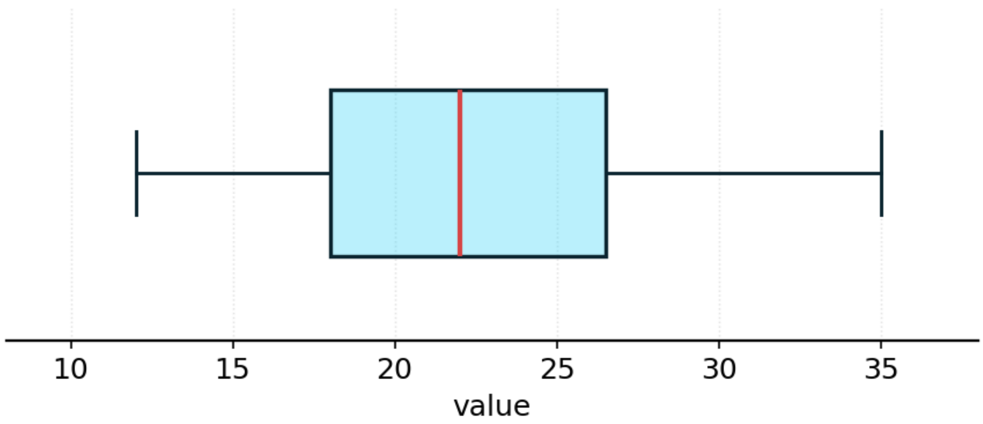
Section B — Structured Questions: Worked Solutions (65 marks)

Mark-scheme notation: M = method mark, A = accuracy mark (dependent on the preceding M mark), R = reasoning mark.

4.1 Descriptive Statistics

Question 1 [7 marks]

The box-and-whisker plot shows a dataset of 11 values (computed).



Box-and-whisker plot of an 11-value dataset (computed).

(a) Using the box plot, write down the median, the lower quartile (Q1), and the upper quartile (Q3) of the dataset. **[2]**

Median = **22**; Q1 = **18**; Q3 = **28**. **A1 A1**

(b) Calculate the interquartile range (IQR) of the dataset. **[2]**

IQR = Q3−Q1 = 28−18 = **10**. **M1 A1**

(c) Using the box plot, write down the minimum and maximum values, and hence calculate the range of the dataset. **[3]**

Minimum = **12**, maximum = **35**. Range = maximum−minimum = 35−12 = **23**. **A1 M1 A1**

Question 2 [6 marks]

Measures of central tendency and spread summarise the key features of a dataset.

(a) State one advantage of using the median (rather than the mean) to describe the centre of a dataset that contains an outlier. **[2]**

The median is **not affected (resistant) to extreme values (outliers)**, whereas the mean can be significantly skewed by a single very large or small value. **R1 R1**

(b) A dataset has mean 45 and standard deviation 6. Calculate how many standard deviations the value 57 is above the mean. **[2]**

Number of standard deviations = (57−45)÷6 = 12÷6 = **2**. **M1 A1**

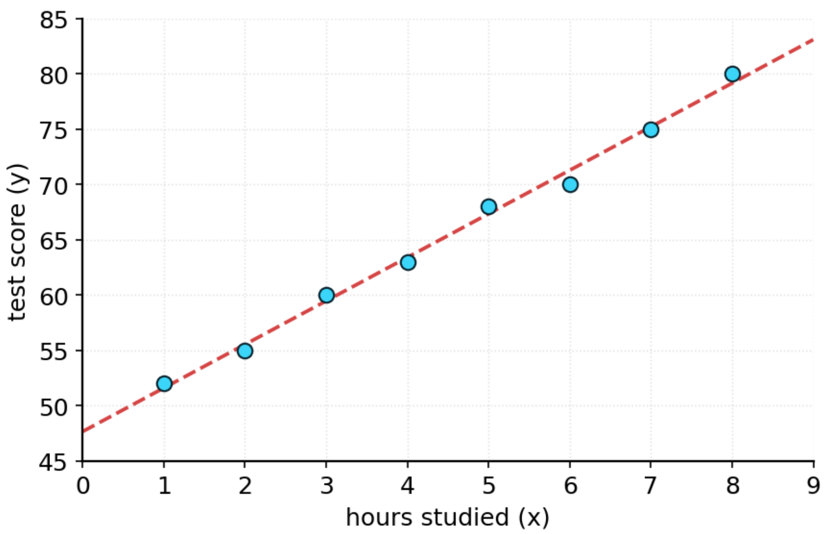
(c) State the rule commonly used to identify an outlier, in terms of Q1, Q3, and the IQR. **[2]**

A value is typically considered an outlier if it is **more than 1.5×IQR below Q1, or more than 1.5×IQR above Q3**. **A1 A1**

4.2 Correlation & Regression

Question 3 [7 marks]

The scatter diagram shows the number of hours studied (x) and test score (y) for 8 students, with a line of best fit (computed).



Hours studied vs test score, with line of best fit (computed).

(a) Using the scatter diagram, describe the correlation between hours studied and test score. **[2]**

The scatter diagram shows a **strong, positive** correlation: as the number of hours studied increases, the test score also tends to increase, and the points lie close to a straight line. **A1 A1**

(b) The value of Pearson's correlation coefficient for this data is r = 0.997. Interpret this value. **[2]**

Since r = 0.997 is **very close to +1**, this confirms a **very strong, positive linear correlation** between hours studied and test score. **A1 R1**

(c) The equation of the line of best fit is y = 3.94x + 47.64. Use this equation to predict the test score for a student who studies for 9 hours. **[3]**

y = 3.94(9) + 47.64 = **83.1** (predicted test score, noting this is a short extrapolation just beyond the given data range). **M1 M1 A1**

Question 4 [6 marks]

Pearson's correlation coefficient, r, quantifies the strength and direction of a linear relationship between two variables.

(a) State the range of possible values of Pearson's correlation coefficient, r, and state what the values r = −1, r = 0, and r = 1 each represent. **[2]**

$-1 \le r \le 1$. $r = -\mathbf{1}$: perfect negative linear correlation; $r = \mathbf{0}$: no linear correlation; $r = \mathbf{1}$: perfect positive linear correlation. **A1 A1**

(b) A dataset has a correlation coefficient of $r = 0.85$. Describe the strength and direction of this correlation. **[2]**

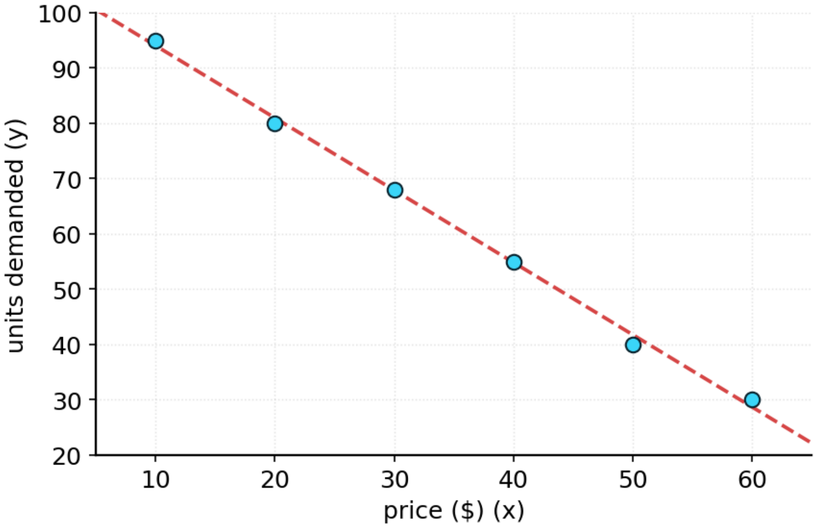
$r = 0.85$ indicates a **strong, positive** linear correlation between the two variables. **A1 A1**

(c) Explain why a strong correlation between two variables does not necessarily imply that one causes the other. **[2]**

A strong correlation only shows that two variables tend to **change together**; it does not show that a change in one variable **directly causes** a change in the other. The relationship could instead be due to a **third (confounding) variable**, coincidence, or reverse causation. **R1 R1**

Question 5 [7 marks]

The scatter diagram shows the price ($) of a product (x) and the number of units demanded (y), with a line of best fit (computed).



Price vs units demanded, with line of best fit (computed).

(a) Using the scatter diagram, describe the correlation between price and units demanded. **[2]**

The scatter diagram shows a **strong, negative** correlation: as price increases, the number of units demanded tends to decrease, and the points lie close to a straight line with negative gradient. **A1 A1**

(b) The value of Pearson's correlation coefficient for this data is $r = -0.999$. Interpret this value. **[2]**

Since $r = -0.999$ is **very close to -1**, this confirms a **very strong, negative linear correlation** between price and units demanded. **A1 R1**

(c) State the sign of the gradient of the line of best fit shown, and explain how this relates to the value of r found in part (b). **[3]**

The gradient of the line of best fit is **negative** (the line slopes downward from left to right). This is consistent with r being **negative** in part (b): both the sign of the gradient and the sign of r indicate the **direction** of the linear relationship between the two variables. **A1 R1 R1**

4.3 Probability

Question 6 [6 marks]

Two events, A and B, have P(A) = 0.4, P(B) = 0.5, and P(A∩B) = 0.2.

(a) State the addition rule of probability for P(A∪B). **[2]**

P(A∪B) = P(A) + P(B) − P(A∩B). **A1 A1**

(b) Calculate P(A∪B). **[2]**

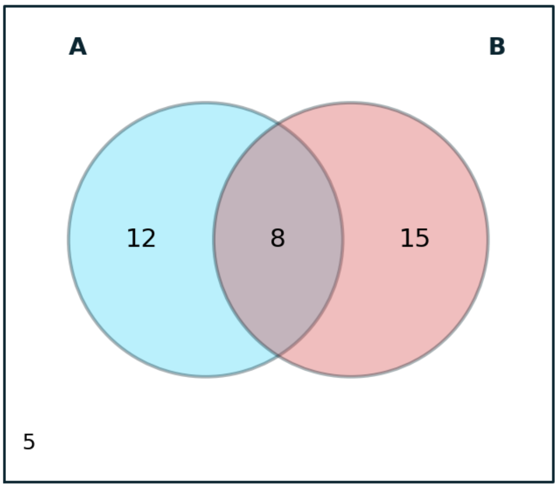
P(A∪B) = 0.4 + 0.5−0.2 = **0.7**. **M1 A1**

(c) Determine, showing your working, whether events A and B are independent. **[2]**

For independence, P(A∩B) should equal P(A)×P(B) = 0.4×0.5 = 0.2. Since this **equals the given P(A∩B) = 0.2**, events A and B **are independent**. **M1 A1**

Question 7 [7 marks]

The Venn diagram shows the number of students (out of 40) who study French (A) and/or Spanish (B) (given data).



Venn diagram showing the number of students studying French (A) and/or Spanish (B) (given data).

(a) Using the Venn diagram, calculate P(A), the probability that a randomly selected student studies French. **[2]**

P(A) = (12+8)÷40 = 20÷40 = **0.50**. **M1 A1**

(b) Calculate P(A∩B), the probability that a randomly selected student studies both French and Spanish. **[2]**

P(A∩B) = 8÷40 = **0.20**. **M1 A1**

(c) Calculate P(A∪B), the probability that a randomly selected student studies French or Spanish (or both). **[3]**

P(A∪B) = (12+15+8)÷40 = 35÷40 = **0.875** (equivalently, using the addition rule, P(A)+P(B)−P(A∩B)). **M1 M1 A1**

Question 8 [6 marks]

Using the same events A and B as in Question 6, where P(A) = 0.4, P(B) = 0.5, and P(A∩B) = 0.2.

(a) State the formula for the conditional probability P(A|B). **[2]**

P(A|B) = P(A∩B)÷P(B). **A1 A1**

(b) Calculate P(A|B). **[2]**

P(A|B) = 0.2÷0.5 = **0.4**. **M1 A1**

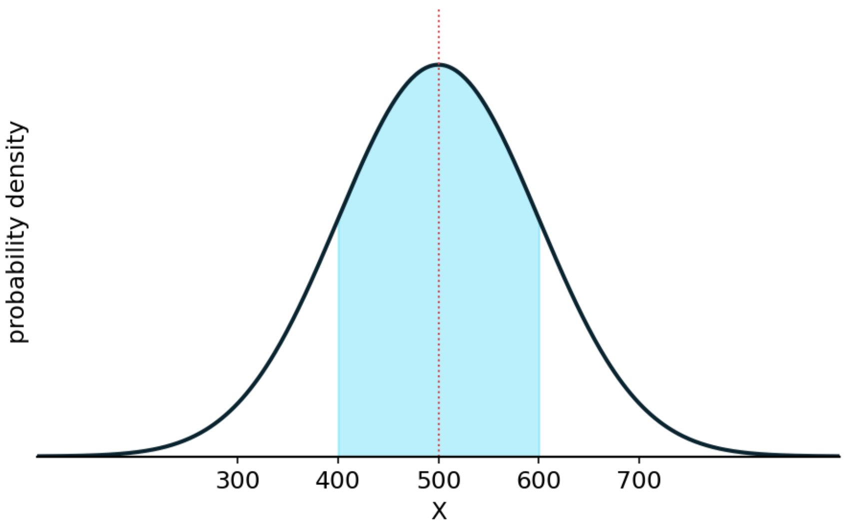
(c) Compare your answer to part (b) with P(A), and explain what this confirms about events A and B. **[2]**

P(A|B) = 0.4, which is **equal to P(A) = 0.4**. This confirms that knowing B has occurred **does not change the probability of A**, consistent with A and B being **independent** (as found in Question 6(c)). **R1 R1**

4.4 Probability Distributions

Question 9 [7 marks]

The graph shows a normal distribution with mean μ = 500 and standard deviation σ = 100, with the region between 400 and 600 shaded (computed).



Normal distribution, μ=500, σ=100, with P(400<X<600) shaded (computed).

(a) Using the graph, state the mean and standard deviation of this distribution, and state how the shaded region relates to μ and σ. **[2]**

μ = **500**, σ = **100**. The shaded region represents the interval from μ−σ to μ+σ (**within 1 standard deviation of the mean**). **A1 A1**

(b) Using the empirical rule (or z-scores), estimate P(400 < X < 600), giving your answer to 3 significant figures. **[2]**

$z_1 = (400-500)\div 100 = -1$, $z_2 = (600-500)\div 100 = 1$. $P(-1<Z<1) =$ **0.683** (≈68.3%, consistent with the empirical rule). **M1 A1**

(c) Calculate P(X > 600), giving your answer to 3 significant figures. **[3]**

$z = (600-500)\div 100 = 1$. $P(X>600) = P(Z>1) = 1-P(Z\leq 1) =$ **0.159** (using the standard normal distribution). **M1 M1 A1**

Question 10 [6 marks]

A random variable X follows a binomial distribution, $X \sim B(10, 0.3)$.

(a) State three conditions required for a random variable to follow a binomial distribution. **[2]**

Any three of: there is a **fixed number of trials, n**; each trial has only **two possible outcomes** (success/failure); the trials are **independent**; the probability of success, p, is **constant** for every trial. **A1 A1**

(b) Calculate the mean and the variance of X. **[2]**

Mean = $np = 10\times 0.3 =$ **3.0**. Variance = $np(1-p) = 10\times 0.3\times 0.7 =$ **2.1**. **M1 A1**

(c) Calculate P(X = 3), giving your answer to 3 significant figures. **[2]**

$P(X=3) = {}^{10}C_3(0.3)^3(0.7)^7 =$ **0.267**. **M1 A1**