

Statistics and probability

IB Maths AI SL · Knowledge organiser

KEY DEFINITIONS

Pearson's r

A number from -1 to 1 measuring the strength and direction of a linear relationship.

Regression line

The line of best fit of y on x , used to predict y from a value of x inside the data range.

Null hypothesis

H_0 , the statement of no effect or no association that a test assumes until the evidence says otherwise.

p-value

The probability, assuming H_0 is true, of a result at least as extreme as the one observed.

YOU SHOULD BE ABLE TO...

- Distinguish populations from samples and discrete from continuous data, compare sampling techniques and identify sources of bias and outliers in data collection.
- Find probabilities from sample spaces and Venn diagrams, use the complement rule and calculate the expected number of times an event occurs.
- Set up a probability distribution for a discrete random variable, use the total probability of 1, find $E(X)$ and decide whether a game is fair.
- Organise data in contingency tables and calculate expected frequencies under the assumption of independence, ready for a χ^2 test.
- Read the test statistic and p-value from technology and write the hypotheses, the decision and the conclusion in context.
- Solve multi-part questions that combine normal probabilities with a binomial model.

COMMON MISTAKES

- ✗ χ^2 and t-tests: hypotheses, the comparison and the conclusion in context
- ✗ Probability: adding instead of multiplying, conditional sample spaces, without replacement
- ✗ Using the class boundary instead of the midpoint when estimating the mean from grouped data
- ✗ Calling a convenience sample random, or labelling a value an outlier without checking the $Q_1 - 1.5 \times IQR$ and $Q_3 + 1.5 \times IQR$ fences

KEY FORMULAS

B = given in the exam (formula booklet or list); learn the rest.

Interquartile range $IQR = Q_3 - Q_1$

B
Mean ($n = \sum f_i$) **B** $\bar{x} = \frac{\sum f_i x_i}{n}$

Outliers $\backslash (x \text{ or } x > Q_3 + 1.5 IQR$

Standard deviation $\sigma = \sqrt{\frac{\sum f(x - \bar{x})^2}{n}}$

Data $\times a$ then $+b$ mean $\rightarrow a\bar{x} + b$, s.d. $\rightarrow |a|\sigma$

Regression: $y = ax + b$ (GDC); use to predict y within the data range only

$|r|$ near 1: strong linear correlation; Spearman's r_s = PMCC of the ranks

Probability **B** $P(A) = \frac{n(A)}{n(U)}$, $P(A) + P(A') = 1$

Expected number of occurrences $n \times P(A)$

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

Mutually exclusive **B** $P(A \cup B) = P(A) + P(B)$

Independent **B** $P(A \cap B) = P(A)P(B)$

Conditional **B** $P(A | B) = \frac{P(A \cap B)}{P(B)}$

Expected value **B** $E(X) = \sum x P(X = x)$

Valid distribution; fair game $\sum P(X = x) = 1$; $E(\text{gain}) = 0$

Binomial $X \sim B(n, p)$ $P(X = r) = \binom{n}{r} p^r (1 - p)^{n-r}$

More on the full formula sheet.

WORKED EXAMPLE

$X \sim N(50, 4^2)$. Find $P(X < 56)$.

1. $z = (56 - 50)/4 = 1.5$

2. $P(X < 56) = P(Z < 1.5)$

Answer: $P(X < 56) = 0.933$ (3 s.f.)