Statistics 1

Edexcel IAL S1 (WST01/01); the statistics unit pairing M1 in IAS Mathematics

Statistics 1: data, probability, regression and distributions

- represent and summarise data, and recognise skewness and outliers;
- calculate probabilities with Venn diagrams and tree diagrams, including **conditional probability**;
- measure correlation and find a **regression line**;

Statistics 1: data, probability, regression and distributions (2)

- work with discrete random variables;
- use the **Normal distribution**.

Mathematical models in probability and statistics

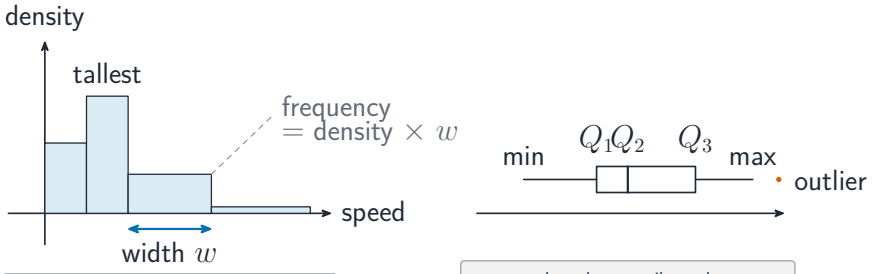
- it is quicker and cheaper than collecting real data every time;
- it can be used to make predictions;
- it can be tested against new data and improved.

Representing data: histograms, stem and leaf, box plots

- In a **histogram 直方图**, the area of each bar is proportional to the frequency.
- The height is the frequency density:
- So when classes have different widths, compare the heights of the bars, not their frequencies.
- A **stem and leaf 茎叶图** diagram keeps the original values; a back-to-back one compares two groups.

$$\text{frequency density} = \frac{\text{frequency}}{\text{class width}}$$

Figure: Representing data: histograms, stem and leaf, box plots



- Left: a **histogram** with unequal classes, where frequency = density × width. Right: a **box plot** with its quartiles and an **outlier**.
- Read the axes and units first.

Measures of location and dispersion

- **Median**: the 112.5th value, in 25–30. $Q_2 = 25 + 5 \times \frac{112.5-54}{90} = 28.25$.
- Lower quartile: the 56.25th value, also in 25–30.
  $Q_1 = 25 + 5 \times \frac{56.25-54}{90} = 25.125$.
- Upper quartile: the 168.75th value, in 30–40. $Q_3 = 30 + 10 \times \frac{168.75-144}{60} = 34.125$.

$$Q_2 = a + w \times \frac{\frac{n}{2} - F}{f}$$

Measures of location and dispersion (2)

- IQR $= 34.125 - 25.125 = 9$ mph.
- $\sum x = 54 \times 45 = 2430$, so class B's total is $2430 - 24 \times 47 = 1302$ and $b = \frac{1302}{30} = 43.4$.
- Combined **standard deviation**: $\sigma = \sqrt{\frac{66\,876 + 73\,826}{54} - 45^2} = \sqrt{580.6} = 24.1$.
- If every mark in class B goes up by 2, class B's standard deviation does not change. Its mean moves from 43.4 to 45.4, closer to class A's mean of 47, so the combined mean goes up and the combined standard deviation goes down.

Skewness

- **positively skewed 右偏**: $Q_3 - Q_2 > Q_2 - Q_1$, or mean $>$ **median**, with a long tail to the right;
- negatively skewed: $Q_3 - Q_2 < Q_2 - Q_1$, or mean $<$ median;
- symmetrical: they are about equal.

Probability

- mutually exclusive if they cannot both happen: $P(A \cap B) = 0$;
- independent if one does not affect the other: $P(A \cap B) = P(A)P(B)$.
  Independence 独立 also means $P(B \mid A) = P(B)$.
- $P(C \cap D) = P(C \mid D)P(D) = 0.2 \times 0.45 = 0.09$.

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

Probability (2)

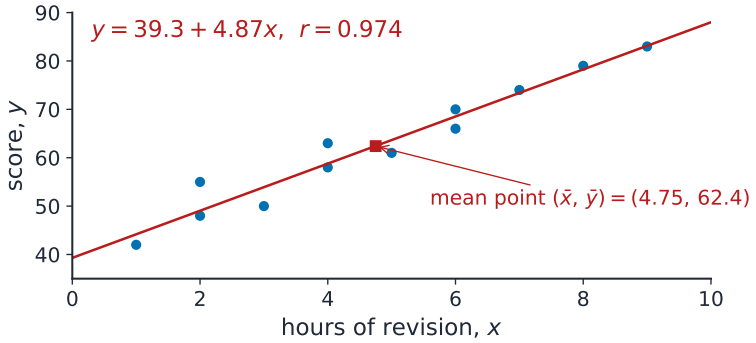
- $P(C) = P(C \cup D) - P(D) + P(C \cap D) = 0.59 - 0.45 + 0.09 = 0.23.$
- Exactly one yellow: $\frac{6}{9} \times \frac{3}{8} + \frac{3}{9} \times \frac{6}{8} = \frac{1}{2}$.
- At least one white: $1 - P(\text{both yellow}) = 1 - \frac{6}{9} \times \frac{5}{8} = \frac{7}{12}$.
- Both white, given at least one white: $\dfrac{\frac{3}{9} \times \frac{2}{8}}{\frac{7}{12}} = \dfrac{1}{12} \times \dfrac{12}{7} = \dfrac{1}{7}$.

Correlation and regression

- $S_{ww} = 10\,036.45 - \frac{339.25^2}{12} = 445.57.$
- $r = \dfrac{-91.55}{\sqrt{26.43 \times 445.57}} = -0.844$: strong negative correlation, so on warmer days less gas is used.
- Regression of w on t: $b = \frac{-91.55}{26.43} = -3.464$ and $a = \frac{339.25}{12} + 3.464 \times \frac{9.1}{12} = 30.90$, so $w = 30.90 - 3.464t$.
- Divide by 11.5: $d = 2.69 - 0.301t$. Each extra degree means about 0.3 m^3 less gas. When $t = 2$, $d = 2.08$ m^3.

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}, \qquad -1 \le r \le 1$$

Figure: Correlation and regression



- **Regression line** of y on x through the mean point
- Read the axes and units first.

Discrete random variables

- **Expectation 期望**: $E(X) = \sum xP(X = x)$, and $E(X^2) = \sum x^2 P(X = x)$.
- **Variance 方差**: $\text{Var}(X) = E(X^2) - [E(X)]^2$.
- The cumulative distribution function is $F(x) = P(X \leq x)$.

Discrete random variables (2)

- If X is equally likely to take each of the values $1, 2, \ldots, n$ (the discrete uniform distribution), $E(X) = \frac{n+1}{2}$ and $\text{Var}(X) = \frac{n^2-1}{12}$.
- $F(3.5) = P(X \leq 3) = 0.3 + 0.2 + 0.35 = 0.85$.
- $P(3X - 3 < X + 2)$: solve first, $2X < 5$, so $X < 2.5$, and $P(X \leq 2) = 0.5$.

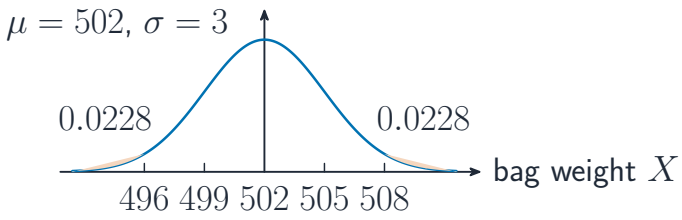
Discrete random variables (3)

- $E(X) = 0.3 + 0.4 + 1.05 + 0.6 = 2.35$ and $E(X^2) = 0.3 + 0.8 + 3.15 + 2.4 = 6.65$.
- $\text{Var}(X) = 6.65 - 2.35^2 = 1.1275$, so $\text{Var}(5 - 2X) = 4 \times 1.1275 = 4.51$.

The Normal distribution

- $P(X > 508) = P(Z > 2) = 1 - 0.9772 = 0.0228.$
- In a sample of 4 bags, the probability that exactly 2 are over 508 g is $\binom{4}{2}(0.0228)^2(0.9772)^2 = 0.00298.$
- Large bags have mass $Y \sim N(1024, \sigma^2)$, and 85% weigh less than 1038.51 g. From the table, $z = 1.0364$, so $\sigma = \frac{1038.51-1024}{1.0364} = 14.$
- If $P(X > k) = P(Y < 2k)$, then by symmetry $\frac{k-502}{3} = \frac{1024-2k}{14}$. So $14k - 7028 = 3072 - 6k$, giving $k = 505$, and the probability is $P(Z > 1) = 0.159.$

Figure: The Normal distribution



standardise $Z = \dfrac{X - \mu}{\sigma}$: tails at $z = \pm 2$ hold 0.0228 each

- The Normal curve for worked example 9, mean 502 g and **standard deviation** 3 g: each tail beyond 496 g and 508 g holds about 2.28%.
- Read the axes and units first.

Sources for this lesson

- Pearson Edexcel, *International Advanced Level Mathematics, Further Mathematics and Pure Mathematics specification* (2018), unit S1.
- F. Galton, "Regression towards mediocrity in hereditary stature", *Journal of the Anthropological Institute*, 15 (1886).
- Pearson Edexcel, WST01/01 June 2025, questions 1 to 7 (adapted).

Common errors

- Drawing **histogram** bars with heights equal to the frequencies when the class widths differ.
- Using the class width instead of the class boundaries in interpolation.
- Forgetting to take the square root of the **variance** to get the **standard deviation**.

Common errors (2)

- Confusing "mutually exclusive" with "independent".
- Saying that correlation proves cause and effect.
- Using a **regression line** to predict x from y, or outside the range of the data.

Common errors (3)

- Writing $\mathrm{Var}(5 - 2X) = -2\,\mathrm{Var}(X)$ or $5 - 4\,\mathrm{Var}(X)$; constants added do not change the **variance**, and the multiplier is squared.
- Forgetting that $P(Z < -a) = 1 - P(Z < a)$.

Key terms

- **Histogram**: a chart in which the area of each bar is proportional to the frequency.
- **Interquartile range**: $Q_3 - Q_1$, the range of the middle half of the data.
- **Conditional probability**: the probability of an event, given that another has happened.

Key terms (2)

- **Regression line**: the line of best fit found by least squares, used to predict y from x.
- **Expectation**: the mean value of a random variable, $E(X)$.
- **Normal distribution**: a continuous, symmetrical, bell-shaped distribution.

Check yourself

- 1.
- A class of width 20 has frequency 50.
- Find its frequency density. 2.
- Five values have $\sum x = 40$ and $\sum x^2 = 360$.