Statistics 1

Dr Oliver
Mathematics

Principles of Statistical Modelling

Stage 1: The recognition of a real-world problem.
Stage 2: A statistical model is devised.
Stage 3: The statistical model is used to make predictions.
Stage 4: Experimental data from real-world testing is collected.
Stage 5: Comparisons are made between the statistical model's predictions and real-world outcomes.
Stage 6: Statistical concepts are used to test how well the statistical model describes the real-world problem.
Stage 7: The statistical model should be refined as necessary.

Some Reasons for Using Models

- To simplify a real-world problem.
- To improve our understanding of a real-world problem.
- To describe a real-world problem.
- To analyse a real-world problem.
- Statistical models are quicker and cheaper than using the real thing.
- To make predictions about future outcomes.
- To refine the model, possibly by changing the parameters.

Box Plots

Box plots are a five-figure summary of a set of data which:

- indicate the minimum value, the lower quartile, the median, the upper quartile, and the maximum value,
- indicate spread, range, and IQR,
- show if and how the data is skewed,
- show outliers (if they exist),
- allow comparisons to be made between different data sets.

Quantiles

We always use $n+1$ rather than n although Edexcel will give full marks for a correct solution using either approach. For lists of raw data and frequency tables, there are three possibilities that you can get for the position of the quantile:

- the positional value could be a whole number—in which case use that,
- the positional value could involve the fraction $\frac{1}{2}$ — for example, if the positional value is $5\frac{1}{2}$ then take the average of the fifth and sixth values,
- the positional value could involve a fraction other than $\frac{1}{2}$ — in which case round to the nearest whole number. For example, if the positional value is $11\frac{3}{4}$ then take the twelfth value; if the positional value is $29\frac{37}{100}$ then take the twenty-ninth value.
- For grouped frequency tables, use the unrounded positional value in your linear interpolation calculation.

Histograms

- Appropriate for continuous data.
- The area of a bar is proportional to the frequency.
- The vertical axis represents frequency density.

Probability

- $\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \cap B)$
- Two events A and B are *mutually exclusive* if
$$\mathrm{P}(A \cap B) = 0.$$
- Two events A and B are *independent* if
$$\mathrm{P}(A \cap B) = \mathrm{P}(A) \times \mathrm{P}(B).$$
- The *conditional probability* of A given B, denoted by $\mathrm{P}(A|B)$, is defined to be
$$\mathrm{P}(A|B) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)}.$$

Correlation

Let (x_1, y_1), (x_1, y_1), $\ldots$, (x_n, y_n) be n pairs of values. We define the following three quantities:

$$\begin{aligned} S_{xx} &= \sum_{i=1}^{n} (x_i - \overline{x})^2 \\ &= \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}, \\ S_{yy} &= \sum_{i=1}^{n} (y_i - \overline{y})^2 \\ &= \sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}, \\ S_{xy} &= \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}) \\ &= \sum_{i=1}^{n} x_i^2 y_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}. \end{aligned}$$

We then define the *product moment correlation coefficient*, usually abbreviated to PMCC and denoted by r, to be

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}.$$

Regression

Let (x_1, y_1), (x_1, y_1), $\ldots$, (x_n, y_n) be n pairs of values. The *line of regression of y on x* is given by

$$y = a + bx \text{ where } b = \frac{S_{xy}}{S_{xx}} \text{ and } a = \overline{y} - b\overline{x}.$$

Expectation

- The *expectation* of X, denoted by $\mathrm{E}(X)$, is defined to be
$$\mathrm{E}(X) = \sum_{i=1}^{n} p_i x_i = p_1 x_1 + p_2 x_2 + \cdots + p_n x_n$$
and the expectation of X^2, denoted by $\mathrm{E}(X^2)$, is defined to be
$$\mathrm{E}(X^2) = \sum_{i=1}^{n} p_i x_i^2 = p_1 x_1^2 + p_2 x_2^2 + \cdots + p_n x_n^2.$$
- For any real constants a and b,
$$\mathrm{E}(aX + b) = a\,\mathrm{E}(X) + b.$$

Variance

- The *variance* of X, denoted by $\mathrm{Var}(X)$, is defined to be
$$\mathrm{Var}(X) = \mathrm{E}(X - \mu)^2 = \mathrm{E}(X^2) - \mathrm{E}(X)^2.$$
- For any real constants a and b,
$$\mathrm{Var}(aX + b) = a^2\,\mathrm{Var}(X).$$

Normal Distribution

We use

$$Z \sim \mathrm{N}(0, 1)$$

to represent the standard normal distribution. The standard normal distribution has a mean of 0 and a standard deviation of 1.

- The mean, median, and mode of the standard normal distribution are all 0.
- The standard normal distributions is symmetrical about its mean, median, and mode.
- Approximately 68.3% of the distribution lies within one standard deviation of the mean.
- Approximately 95.4% of the distribution lies within two standard deviations of the mean.
- Approximately 99.7% of the distribution lies within three standard deviations of the mean.
- The total area under the graph is exactly 1 and this represents a probability. The function that we use is
$$\Phi(z) = \mathrm{P}(Z < z).$$
- For negative values,
$$\Phi(-z) = 1 - \Phi(z).$$

We use

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

to represent the standard normal distribution that has a mean of μ and a standard deviation of σ. Suppose that $X \sim \mathrm{N}(\mu, \sigma^2)$, i.e., X has mean μ and variance σ^2. The random variable

$$\frac{X - \mu}{\sigma}$$

has mean 0 and variance 1, i.e., this is the standard normal distribution. This process is called *standardisation* and will allow us to convert any normal distribution to $Z \sim \mathrm{N}(0, 1)$ so that we can use our tables.