

S1 Formulae

Mean, Spread and Skewness of Data

You should know these measures:

$$\text{Mean} = \mu = \frac{\sum x}{n}$$

$$\text{Variance} = \sigma^2 = \frac{\sum x^2}{n} - \mu^2$$

$$\text{StandardDeviation} = \sqrt{\text{Variance}} = \sigma = \sqrt{\frac{\sum x^2}{n} - \mu^2}$$

We will often want to look at the skewness of the the data too. To do this we need to look at the 3 Quartiles Q_1 , Q_2 and Q_3 , these are the lower quartile, median and upper quartile respectively. To find these we must first order the data into ascending value and then find the value a quarter, half and three quarters of the way through the list. This may involve taking the average of two separate values. We can then use these values to measure skew by comparing $Q_2 - Q_1$ and $Q_3 - Q_2$:

- $Q_2 - Q_1 < Q_3 - Q_2$ implies negative skew as the median is closer to the upper quartile than the lower quartile.
- $Q_2 - Q_1 = Q_3 - Q_2$ implies no skew since the median lays in the middle of the two other quartiles.
- $Q_2 - Q_1 > Q_3 - Q_2$ implies positive skew as the median is closer to the lower quartile than the upper quartile.

Correlation and Regression

First we must define three different values:

$$S_{xx} = \sum x^2 - \frac{(\sum x)^2}{n}$$

$$S_{yy} = \sum y^2 - \frac{(\sum y)^2}{n}$$

$$S_{xy} = \sum xy - \frac{\sum x \sum y}{n}$$

Using these three values we can then calculate the following:

- Product Moment Correlation Coefficient

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

There are three possible results for r:

1. $r < 0$ - The data shows negative correlation
2. $r = 0$ - The data shows no correlation
3. $r > 0$ - The data shows positive correlation

The closer the values are to the extreme values of ± 1 , then the more correlated the data is.

- Least Squares Regression

We can estimate values by assigning a straight line $y = a + bx$ through the points, where:

$$b = \frac{S_{xy}}{S_{xx}}$$

and then $a = \bar{y} - b\bar{x}$, where \bar{x} and \bar{y} are the means of x and y respectively.

Probability

First we must know all the correct terminology and notation:

- $P(A)$ = Probability of event A happening.
- $P(A')$ = Probability of A complement which means A *not* happening.
- $P(A \cap B)$ = Probability of A "intersect" B which mean A *and* B.
- $P(A \cup B)$ = Probability of A "union" B which means A *or* B *or* both.
- $P(A|B)$ = Conditonal Probability of A *given* B.

Using these we can define a few rules:

- $P(A') = 1 - P(A)$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- $P(A|B) = \frac{P(A \cap B)}{P(B)}$

These values can be viewed using a venn diagram which is something you should be comfortable constructing and reading.

Random Variables

Random variables are very important when modelling physical actions using probability. We usually define the random variable as capital letter such as X , which can take various values x , defined by some probability function:

$P(X = x)$ The Probability that random variable X takes value x

By the laws of Probability we know that:

$$\sum_x P(X = x) = 1$$

Now to estimate the random variable we calculate the expected value which is defined:

$$E(X) = \sum_x xP(X = x)$$

We can also look at the variance of our random variable so that we can say how likely it is for the actual result to be somewhere near our expected value. To find the variance we must first find $E(X^2)$, then use this to find $Var(X)$ like so:

$$E(X^2) = \sum_x x^2P(X = x)$$

$$Var(X) = E(X^2) - [E(X)]^2$$

Sometimes our random variables may be transformed slightly before we have to find the above values. To deal with this we use these rules (where a and b are constants):

$$E(aX + b) = aE(X) + b$$

$$Var(aX + b) = a^2Var(X)$$

Normal Distribution

We say that a continuous random variable is normally distributed if its values follow a normal "bell shaped" curve. Physical examples of such a variable are height, weight etc. If something is normally distributed then $Mean = Mode = Median$ since the distribution is symmetrical. If a variable X is normally distributed with mean μ and variance σ^2 then we write:

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

When we try to analyse normal random variables, we need to turn to standard normal tables to look up the probabilities. In the normal tables, they always use a *standardised* normal distribution which means it has mean = and variance 1. The standard normal variable is always called Z and so:

$$Z \sim N(0, 1)$$

This means that whenever we use the tables to look at a general normal distribution, we must first standardise it so that our distribution has mean 0 and variance 1. To do this we must make some alterations to our variable:

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

The tables give us the $P(Z < z) = \Phi(z)$, therefore if we are looking at $X \sim N(\mu, \sigma^2)$, then:

$$P(X < x) = P\left(\frac{X - \mu}{\sigma} < \frac{x - \mu}{\sigma}\right) = P\left(Z < \frac{x - \mu}{\sigma}\right) = \Phi\left(\frac{x - \mu}{\sigma}\right)$$

which is a value that can be read directly from the tables.

It is also important to realise that since the standard normal is symmetrical about the origin:

$$P(Z < -z) = P(Z > z) = 1 - P(Z < z)$$

Linear Interpolation

Interpolation is a way of estimating a value when we haven't been given data for that value, but have data for either side of it. For example, when looking up a value in the Normal tables and we want the probability at 0.855 but only have values for 0.85 and 0.86. The easiest way to deal with this is to treat the function as linear between these two points.

So for some point x which lies between x_1 and x_2 , we can estimate y if we know y_1 and y_2 by using this equation:

$$y = y_1 + (x - x_1) \frac{y_2 - y_1}{x_2 - x_1}$$