Probability and Sampling

Interpretation of Probability

We can think of probability in two ways. It could represent the tendency of an event to occur, or it could be a belief in the likelihood of an event occurring. We may categorise these two interpretations in two ways;

Frequency Probability An event occurs in a physical system (such as a roulette wheel) persistently after a very long number of trials.

Subjective probability is Also known as Bayesian Probability. These are degrees of belief based on all of the available evidence.

Probabilistic Models

A probabilistic model always contains a random variable(s). Each time your model is run, it can give different results, even with the same starting conditions. The spin of a roulette wheel is a good example where random variables are at play. Consider the force placed on the thrown ball and the force applied to the wheel. These are always random, even if undertaken by a machine. The outcome of a single bet on the wheel cannot be known.

Empirical Variability

This is based upon observation or experience rather than scientific theory. The results of such observations or experiences are very difficult to interpret.

Events and Sets

We can think of an *event* as the outcome of an experiment with an assigned probability. Consider the probability of selecting a Jack, $P(Jack)$, from a set of 52 playing cards. The probability of drawing a Jack is 4/52, which is 0.077. A probability of 1.0 is the highest achievable (certainty), and 0 is the lowest (impossible). The Jack drawn belongs to a *Set* of Jacks. We can observe other sets, such as Hearts, Blacks, Reds, Odds, Evens, etc. To determine the probability of drawing the Jack of Spades, we have a *singular probability* of 1/52=0.019.

Mutually Exclusive Events

These cannot happen together. For example, you cannot get both heads and tails with one single toss of a coin.

Independent Events

An example of two events that are independent might be to draw a Jack from a card deck, replace the card in the deck and then draw a Queen. The action of replacing the Jack in the deck has not influenced the chances of drawing the Queen.

An example of two events that are not independent might be to draw a Jack from a card deck, remove the card from the deck and then draw a King. The action of removing the Jack from the deck has influenced the chances of drawing the King.
Therefore, we can say that two events, A and B, are independent when event A does not affect the probability of event B occurring.

Conditional Probability

This is the probability of an event occurring after another event has taken place. For example, if a relay has been exposed to extreme heat for an extended period, it is far more likely to fail than a newly manufactured relay.
Given events A and B, we may express the probability of A happening given that B has already happened as…

$$P(A \mid B)$$

Sample Space and Probability

This is the set of all possible outcomes of an experiment. For example, the probability of a coin toss being a tail is 0.5, and the probability of a coin toss being a head is also 0.5. Since there are no other possible outcomes, the sample space is 0.5 + 0.5 = 1.

Addition Law

Given events A and B are mutually exclusive, the probability that A or B will occur is the sum of the probability of each event. For example, the probability of throwing a 2 or a 6 on a die is 1/6+1/6=1⁄3.

Product Law

This is the probability of two *independent* events occurring. For example, tossing two dies simultaneously will give the probability of obtaining a pair of 6's as 1/6×1/6=1/36

Bayes' Theorem

This relates current probability to prior probability and relies on conditional probability. Bayes' Theorem may be expressed mathematically as…

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

One application of Bayes' Theorem is in email spam filtering. Let event A be the probability that an email is a spam. Let event B be a test for certain spam-like words. We may then write…

$$P(spam|words) = \frac{P(words|spam)\,P(spam)}{P(words)}$$

Example

A certain circuit board is installed on petrol station forecourts worldwide and is subject to a wide variation of temperatures.
Let P(A) = 0.002 be the probability of an overheating component on the board and
P(B) = 0.001 be the probability of a failed component on the board.
If the probability of a failed component given an overheated component is $\boldsymbol{P(B|A)=0.15}$ use Bayes' Theorem to calculate the probability that an overheated component has caused a circuit board failure.

Let's gather the data…

$$P(A)=0.002;\ P(B)=0.001;\ P(B|A)=0.15$$

Therefore, we can say…

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)} = \frac{(0.15)(0.002)}{0.001} = 0.3$$

It isn't just extremes of temperature which cause components to fail on circuit boards. Other causal factors are humidity, vibration, ageing, poor manufacture, solar activity, human misuse, and various other reasons

Standard Deviation and Variance

To ascertain the amount of variation in a full given population or a population sample, we use the terms *standard deviation* and *variance*. Let's define these terms…

$$\textit{standard deviation for a full population}, \sigma = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n}}$$

$$\textit{standard deviation sample of population}, \sigma = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}}$$

$$\textit{variance} = \sigma^2$$

The Greek lower case letter sigma (σ) is used for standard deviation, and n is the number of samples. You will notice that the variance is simply the square of the standard deviation. Let's see these in use.

Example

Find the standard deviation and variance for the following population data set…

Sample	1	2	3	4	5	6	7	8	9	10
Age (Years)	19	18	20	19	18	19	19	23	19	19

First, we shall find the standard deviation and then square that answer to get the variance. To find the standard deviation, we normally construct a table as follows…

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
19	19.3-19=0.3	0.09
18	19.3-18=1.3	1.69
20	19.3-20=-0.7	0.49
19	19.3-19=0.3	0.09
18	19.3-18=1.3	1.69
19	19.3-19=0.3	0.09
19	19.3-19=0.3	0.09
23	19.3-23=-3.7	13.69
19	19.3-19=0.3	0.09
19	19.3-19=0.3	0.09
$\bar{x} = \frac{1}{n}\sum_{i=1}^{10} x_i = 19.3$		$\sum (x_i - \bar{x})^2 = 18.1$

$$\therefore \ standard\ deviation, \sigma = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n}} = \sqrt{\frac{18.1}{10}} = \sqrt{1.81} = 1.345$$

$$\therefore \ variance = \sigma^2 = 1.81$$

Question

(a) Calculate the data set's mean, median, and mode below.
(b) Find the standard deviation and variance for the population data set.

Sample	1	2	3	4	5	6	7	8	9	10
Salary (£1000's)	22	24	21	26	32	26	26	54	12	33

ANSWER

(a) Mean = £27,600, Median = £26,000, Mode = £26,000
(b) Standard Deviation = 10.41, Variance = 108.44

Check your answers with this handy online calculator.

Chi-Squared Analysis

Chi-Square analytical tests can be split into two uses.

Chi-square goodness of fit test	Chi-square for independence
This determines if sample data matched a population.	Compares two variables in a contingency table.
Fits one categorical variable into a distribution.	Compares two sets of data to see if they are related.

The goodness for fit test tests if the sample data fits a distribution from a certain population. For example, you ask 100 artists what zodiac sign they belong to. It tells you if your sample data you would expect to find in the general population. You would expect to see them evenly distributed among the 12 zodiac signs. This test will allow you to test this theory. Some disadvantages are:

- Data needs to be put into classes.
- There is a requirement for a sufficient sample size to get valid results.

Formula:

$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

C=degrees of freedom

0=observe value

E=expected value

The summation symbol ∑ means 'add everything up.'

The calculations can be very, very lengthy.

The Chi-Square test for independence, the main difference is you will be summing over positions in a contingency table instead of over categories:

$$c^2 = \sum_{i=1}^{k} \left[\frac{(O_i - E_i)^2}{E_i} \right]$$

C=degrees of freedom

0=observe value

E=expected value

The summation symbol ∑ means 'add everything up.'

I is the 'ith' position in the contingency table.

- A low-value value indicates a high correlation between the two sets of data. In theory, if your observed and your expected values are equal (no difference), chi square=0
- A 'large enough' value isn't as easy as it seems.
- You could compare the statistic to a critical value from the chi-square table.
- You could also use a p-value. Small p-values (under 5%) usually indicate significant differences.

The null hypothesis assumes no significant difference between the observed and expected values. The alternate hypothesis assumes that there is a difference.

Video

Statistical Analysis and Decision Making

A range of statistical tools and tests can be used to determine an outcome. This outcome can assist with your options and your decisions from a management perspective. The video below provides information on which statistical test to use when testing and analysing data.

Video

Statistical Analysis as a Basis for Planning and Implementing Changes to Activities

Changes in an organisation, radical or incremental, can be challenging to implement. You need to consider the implementation process and the people, resources, and the impact on current practices and organisational culture. You can use the evidence from the statistical analysis to gain faith among the team in the best course of action for the decided option for change. Creating buy-in from the team could be smoothed out by communicating the change plan to all the people it will affect. Change management is essential in organisations today, and being fully aware of the impact of the change will allow you to foresee any obstacles that may get in the way. You can then prepare for and plan ways to overcome them to minimize delays in implementation.