Summary

Data are **measurements or observations that are collected as a source of information**.

Quantitative variables are any variables where the data represent amounts (e.g. height, weight, or age).

Categorical variables are any variables where the data represent groups. This includes rankings (e.g. finishing places in a race), classifications (e.g. brands of cereal), and binary outcomes

Continuous and Discrete data

Continuous data refers to data that can be measured. This data has values that are not fixed and have an infinite number of possible values. These measurements can also be broken down into smaller individual parts.

Discrete data also referred to as discrete values, is data that only takes certain values. Commonly in the form of whole numbers or integers, this is data that can be counted and has a finite number of values. These values must be able to fall within certain classifications and are unable to be broken down into smaller parts.

Measures of central tendency

A measure of central tendency (also referred to as measures of centre or central location) is a summary measure that attempts to describe a whole set of data with a single value that represents the middle or centre of its distribution.

There are three main measures of central tendency:

- mode
- median
- mean

Summation Notation

- Many statistical formulas involve adding a series of numbers. The notation for adding a series of numbers is the capital Greek letter sigma. The sigma stands for "add up everything that follows." Therefore, if the sigma is followed by the letter X, it means that you should add up all of the X scores.

$$\sum X$$

Measures of spread

Measures of spread describe how similar or varied the set of observed values are for a particular variable (data item). Measures of spread include the range, quartiles and the interquartile range, variance and standard deviation.

Range

The range is the difference between the smallest value and the largest value in a dataset.

Quartiles

Quartiles divide an ordered dataset into four equal parts, and refer to the values of the point between the quarters. A dataset may also be divided into quintiles (five equal parts) or deciles (ten equal parts).

Interquartile range

The interquartile range (IQR) is the difference between the upper (Q3) and lower (Q1) quartiles, and describes the middle 50% of values when ordered from lowest to highest. The IQR is often seen as a better measure of spread than the range as it is not affected by outliers.

Variance and standard deviation

The variance and the standard deviation are measures of the spread of the data around the mean. They summarise how close each observed data value is to the mean value.

Inferential Statistics

Inferential statistics is a branch of statistics that makes the use of various analytical tools to draw inferences about the population data from sample data.

Simpson's paradox

In statistics, an effect that occurs when the marginal association between two categorical variables is qualitatively different from the partial association between the same two variables after controlling for one or more other variables.

Probability is the likelihood of one or more events occurring. It represents the possibility of getting a certain outcome. Probability can also be described as the probability of an event occurring divided by the number of expected outcomes of the event.

Ex: flip coin n times

There are 2^n sequences of Heads and Tails of length n

exactly k heads and n−k tails

so the formula is going to be n!/(n-k)!*k!

p= (n!/(n-k)!*k!) * (p^k)*(1-p)^(n-k)