Population and sample

- The aim of research in psychology is to answer questions about human behaviour.
- **Population** – the entire set/collection of individuals or events of interest in a particular study. I.e. all individuals that you are interested in studying.
- **Sample** – a subset of individuals selected from the entire population.
- **Representative sample** – a sample that shares key characteristics of the population from which it has been taken.
- Example of population vs sample – a psychologist wants to determine whether motivational programs are effective in improving work performance in Australian office workers. The population here is every Australian office worker and a sample would be 100 office workers from every state.
- **Parameter** – a value that describes a key characteristic of the population. E.g. average income of all Australian office workers.
- **Statistic** – a value that describes a key characteristic of the sample. Statistics are used to estimate values that exist in the population. E.g. measure the age of the 100 depressed patients. There are two main types of statistics:
  - ✓ **Descriptive statistics** – simply used to describe the data. It summarises the data and makes it more manageable. It includes averages and range of scores as well as examining the data for extreme scores.
  - ✓ **Inferential statistics** – generally used after we have described the data and are interested in answering specific research questions. It allows us to make generalisations from the sample to the population. E.g. infer the behaviour of all office workers based on the data collected from the sample of 100 office workers, given statistically significant results.
- Statistics are used as an estimate of the actual population parameters, they are not perfect values as there is some error in the estimation. This results in sampling error.
  - ✓ **Sampling error** – the difference between the sample statistic and the corresponding population parameter. There will always be a discrepancy between your sample characteristics and population characteristics due to sampling error (particularly in small samples).
- **Variable** – something about a construct/object/event that can take on different values (i.e. it can vary). E.g. gender, age, socioeconomic status, motivation and even personality. Different types of variables have different labels. The distinction between different types of variables is very important for deciding how we statistically analyse the data we obtain.
  - ✓ **Discrete variables** have only a limited number of values (e.g. gender). When we are dealing with discrete variables we have **categorical data**, i.e. people are in categories (high/low anxiety, male/female etc). it is measured in percentages/frequencies/proportions.
  - ✓ **Continuous variables** can take on many different values (e.g. age, anxiety score, IQ, reaction time, height). Scores on continuous variables deals with **measurement (or quantitative) data** to reflect the fact that the scores are obtained due to some measurement that was taken (e.g. questionnaire), i.e. scores have been measured along a continuum. It is measured in averages. The null hypothesis for continuous variables is that there is no difference between the sample mean and the population mean.

- Whether a variable is discrete or continuous often depends on how the measurement is set up. E.g. anxiety levels of HPS201 students would generally be continuous (score on anxiety questionnaire) but it can also be categorised into groups such as high anxiety and low anxiety, making it discrete.

- **The type of data we collect influences the type of statistical approach we use.**

  - ✓ **Measurement data** (depicts continuous variables) are usually summarised using means/variances/standard deviations.

  - ✓ **Categorical data** (depicts discrete variables) are usually summarised using percentages/frequencies.

- **Independent variables** – are those that are manipulated by the researcher. In our office worker example, the researcher assigned participants to either the experimental treatment group (motivational program) or the control group. Thus, 'group membership' would be the independent variable here.

- **Dependent variables** – the data we obtain from the study, which might be levels of positive feelings and work performance for office worker example. I.e. the factor we expect will be influenced by the IV during the study.

- **Measurement scales** – variables can be measured on a variety of different scales. These include:
    - ✓ **Nominal scale** – which are labels for categories of data. There is no meaningful underlying order to these labels; we cannot order the variables or talk about distance between them, i.e. the order that they are in is not important and doesn't mean anything, they are just different labels for the different categories. Religion is an example of a variable that is measured on a nominal scale (e.g. Muslim, Christian, Jewish, other). Another example is hair colour (e.g. brown, black, blonde, red).
    - ✓ **Ordinals scale** – which are categories with different names AND are organised into an ordered sequence, measuring magnitude. E.g. degree of illness – none, mild, moderate, severe. This allows us to determine the direction of the difference, that is, we can say severe is greater than moderate and moderate is greater than mild. However, distances between the categories are unknown.
    - ✓ **Interval scale** – with an interval scale such as temperature we have meaningful differences between points on the scale. There is an equal distances between points on an interval scale (e.g. 10-20 degrees is the same distance as 50-60 degrees). There are also many more points on the interval scale than there is on an ordinal scale. However, there is no true zero point (0 degrees is not an absence of temperature)—so we cannot say that 40 degrees is half as hot as 80 degrees. Therefore, with interval scales we cannot talk in terms of ratios.
    - ✓ **Ratio scale** – this is the 'highest' scale of measurement, in that it is the most informative. Ratio scales have all the characteristics of an interval scale plus a true zero point (e.g. time, age, weight and length).

Summary of scales:

Type of scale	Scale qualities	Example
Nominal	Categories for labels (order is meaningless)	Gender, religion, hair colour
Ordinal	Ordered categories, measures magnitude	Small, medium, large, moderate, none
Interval	Measures magnitude, distances are equal, no true zero point.	Temperature
Ratio	Measures magnitude, distances are equal, true zero point.	Age, time, weight, height.

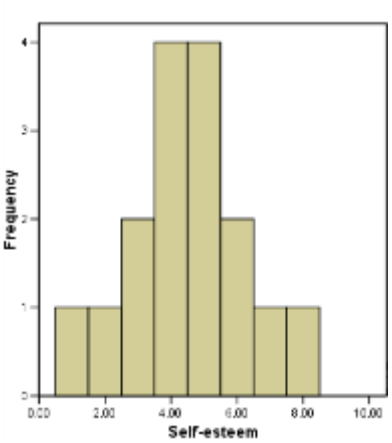
- There are a number of ways that we can describe and explore data from psychological studies. These methods help to simplify the large amounts of raw data that are collected.
    - ✓ **Frequency distributions** – reports on how often each score was recorded in the data set. It also helps to summarise complex data as it allows us to quickly see how many people reported each score. When presented in table form, a frequency distribution lists all values for a variable along with the frequency of each. Adding up all the values under 'frequency' gives total particpants. For example, imagine a researcher obtains the following self-esteem scores from 16 people (measured as a continuous variable ranging from 0–4):

4 1 3 4 3 2 2 4 1 4 3 4 4 1 3 3

Frequency distribution table:

Self-esteem score	Frequency
0	0
1	3
2	2
3	5
4	6

- ✓ **Histogram -** a histogram essentially provides a graphical illustration of the frequency distribution. When a data set has a large range of many different values, the data can be plotted as a grouped distribution. In the case of a smaller data set we can simply plot the data as collected.



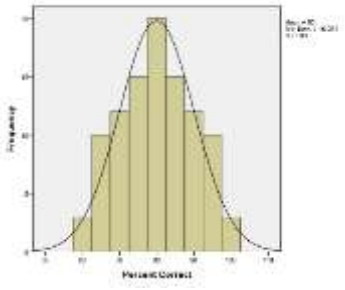
- ✓ **Stem and leaf plot** – with more complex data it may be better to illustrate the information using stem-and-leaf displays. Scores are presented in table form using ***leading digits*** **(stem)** and ***trailing digits*** **(leaves)**.

Raw data	Stem	Leaf
	0	000235679
0 0 0 2 3 5 6 7 9	1	011279
10 11 11 12 17 19	2	022336677777
20 22 22 23 23 26	3	0012225
26 27 27 27 27 27	4	29
30 30 31 32 32 32	5	07
35 42 49 50 57 62	6	28
68 71 72 72 79 80	7	1229
	8	0

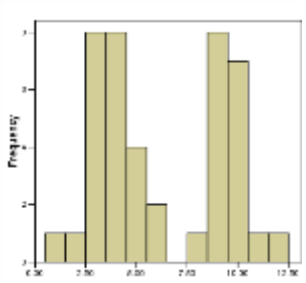
This table illustrates a stem-and-leaf display for some hypothetical data. The stem values are the ten's digits. The leaf values opposite the stem correspond to all the responses that occur within that category. Thus, next to the stem of 7 we have 1, 2, 2, 9, which indicates that one person scored 71, two people scored 72 and one scored 79.

- **Types of distribution** –
  - ✓ **Normal distribution** – where most of the scores are in the middle with fewer scores in the extremes. The assumption of normal distribution is that the DVs are normally distributed in the population.
  - ✓ **Bimodal distribution** – has two peaks (modes)**,** where the majority of scores are grouped at two points (different locations on the scale).

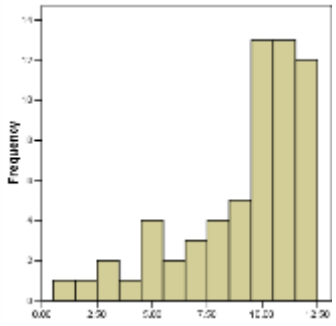
e.g. Normal distribution:



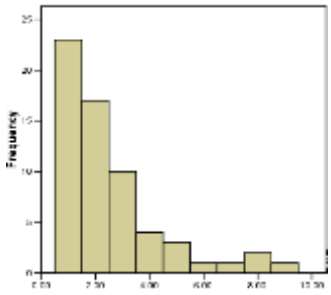
e.g. Bimodal distribution:



- ✓ **Skewed distribution –** skewness is the degree of symmetry in the variable distribution.
  - ➢ **Negatively skewed distribution (skewed to the left)** – when more people score at the high end of a scale, i.e. data consists of more highs scores, the distribution is said to be negatively skewed**.**



  - ➢ **Positively skewed distribution (skewed to the right)** – when more people score at the low end of a scale i.e. data consists of more low scores, the distribution is said to be positively skewed.



- **Kurtosis** – refers to how 'flat' or 'peaked' the distribution appears. I.e. degree of peakedness/flatness in the variable distribution. There are two types of kurtosis distributions:
  - ✓ **Platykurtic** – the distribution is flatter; less scores in the centre. I.e. low degree of peakedness.
  - ✓ **Leptokurtic** – the distribution is characterised by a high peak in the centre of the scale; more scores in the centre. I.e. high degree of peakedness.