Statistics- The science of learning from data. It involves collecting, presenting, analysing and interpreting data. It also involves determining population c
haracteristics, using data from samples

Objectives of statistical data

A primary objective of the statistician is to obtain information about a target population, using a sample.

Target population- Comprises all relevant subjects of interest

Sample- A manageable subset, selected to make a feasible study

Therefore, we use a sample to answer questions about a target population.

Scope of statistical data

Study design → Data analysis → Interpretation of results → Next study?
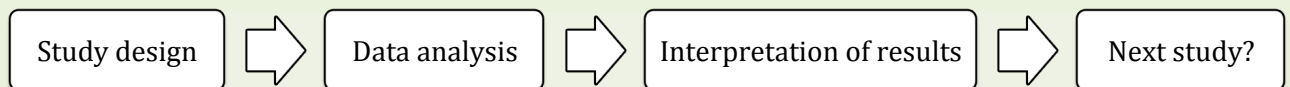
Study design

- ❖ Formulate the question of interest (who, what, why etc)

- ❖ Specify the target population (who/what, where, when

    The target population should be well defined

    The **sample** should be representative of the target population (not biased), and large enough to give accurate information about the population

    Ideally, the observations should be independent to each other/ stratified

- ❖ Determine the measurements to be collected (the variables)

- ❖ Define the method of data collection (how when where)

Selecting a sample

Only a representative sample should be used to make inferences about the target population.
One way to ensure that a sample is representative of the target population is to obtain a random sample.

A simple random sample, of given size n, is one in which each set of that size has the same chance of being selected from the target population.

A random sample- Where **each member** of the (whole) pop. has the **same chance** of being selected
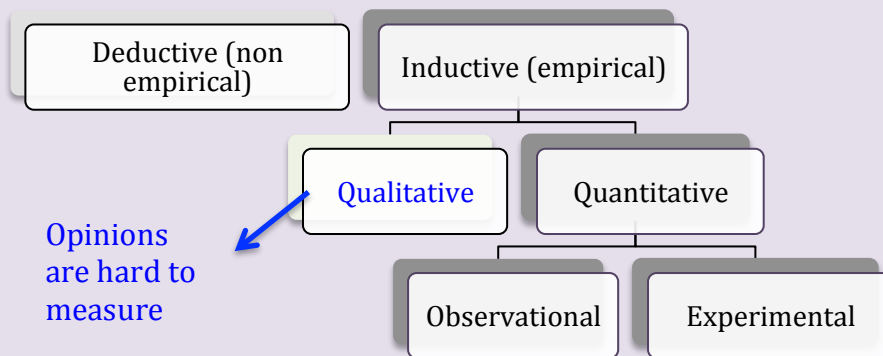
A representative sample- It is often difficult or even impossible to obtain a simple random sample. However, researchers should ensure that the sample is at least a representative sample. That is, its characteristics should represent those of the target population without bias.

Eg- If opinion on same sex marriage was sought from people living in Aust., our sample should have people living in Aust. From different backgrounds, age groups, occupations etc.

Types of Studies

Studies are needed to resolve questions of interest.
Empirical- Based on observation or experience rather than theory or pure logic.

```
┌─────────────────┐   ┌─────────────────────┐
│ Deductive (non  │   │ Inductive (empirical)│
│   empirical)    │   └─────────────────────┘
└─────────────────┘        │
              ┌────────────┴──────────┐
         ┌──────────┐         ┌──────────────┐
         │Qualitative│        │ Quantitative │
         └──────────┘         └──────────────┘
                                  │
                        ┌─────────┴────────┐
                   ┌──────────────┐ ┌──────────────┐
                   │ Observational│ │ Experimental │
                   └──────────────┘ └──────────────┘
```
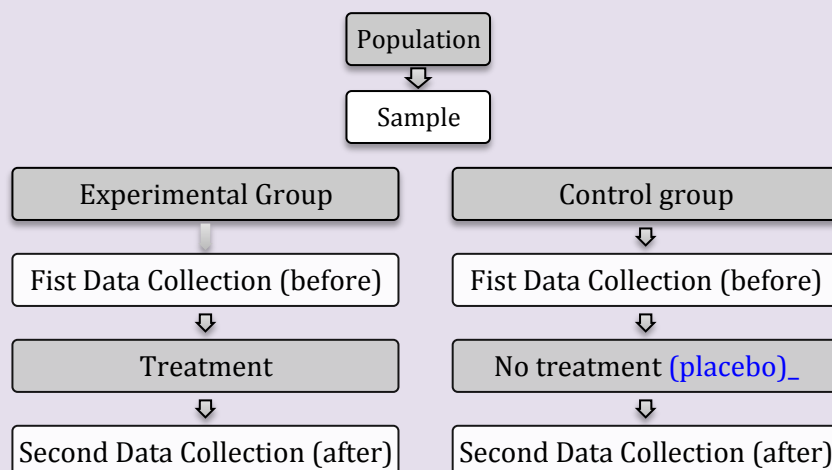
Opinions are hard to measure

❖ **Observational Study**- One in which there is **no intervention by the investigator nor is there any treatment imposed.** (a link but not proven)

Eg- The association between the use of mobile phones and a rare form of eye cancer called uveal melanoma. It was shown that subjects with this eye cancer had significantly higher mobile phone usage.

❖ **Experimental Study**- One in which **the investigator has some control over the determinan**t.

Eg- A researcher randomly assigns law students into 2 groups. Members of one group are all given an alcoholic drink- the other does not. Each student is asked to argue on a topic and the quality of their arguments were rated. Ratings are compared for the 2 groups

Design of an experimental study:

```
                    ┌────────────┐
                    │ Population │
                    └────────────┘
                          ⇩
                    ┌────────────┐
                    │   Sample   │
                    └────────────┘
```

Experimental Group	Control group
⇩	⇩
Fist Data Collection (before)	Fist Data Collection (before)
⇩	⇩
Treatment	No treatment (placebo)_
⇩	⇩
Second Data Collection (after)	Second Data Collection (after)

When collecting data, we are collecting "values" of variables.

Measurements are taken on subjects in a study according to the variables of interest. The measurements will vary from one subject to another.

In any study, variables take on specific roles and these roles may be classified as:

- ❖ Outcomes (responses)
- ❖ Determinants (what may influence the response)

Determinants	→	Influence	→	Outcomes
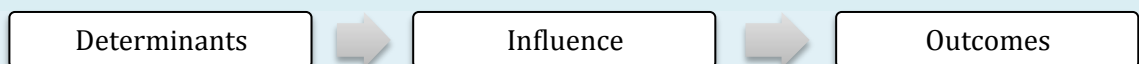
Eg- outcomes may be: no. of hours sleep the previous night/ difference between no. of hours sleep with and without alcohol the previous night

Possible determinant- the number of alcoholic drinks

Eg- Outcome may be- whether client uses a pet boarding service

Possible determinant- How many pets/ How often client goes on holidays.

Note- Any other measurements recorded on the subject Eg.- age, sex etc may be possible determinants

---

Bias- samples should be unbiased- They should represent the target population.

Bias- any systematic error (not occurring randomly)

This results in an incorrect estimate of a parameter or an incorrect association between variables in a study. Studies can be affected by various types of bias, including:

❖ Selection Bias-

Any systematic differences occurring in the way that subjects are selected for study

Eg- Suppose we wish to estimate the proportion of 18 – 25 year olds in Aust. Who have private health insurance.

Selecting a sample from a student database could produce a biased result, since the proportion of students with private health insurance may differ from the proportion of other young adults with private health insurance.

❖ Measurement Bias

Systematic differences in the measurement of variables.

Eg- In a comparison of influenza rates among people with and without chronic illnesses, the responses for people with chronic illnesses may be more accurate as their past illnesses may be better documented and/or recalled

People collecting info from subjects may do so more carefully in the morning than in the afternoon

❖ Response Bias	❖ Confounding.
When the response rate to a service is too low.	A confounder is a variable that distorts (increases or decreases) the apparent effect of one variable (determinant) on another (outcome)
It is well known that those who respond to a survey often have different characteristics than those who don't respond.	
	Eg- It's suggested that watching > 4 hours of TV/day is associated with an > risk of heart disease. However, it is more likely that those who watch a lot of TV don't exercise much & it's the lack of exercise which leads to the > risk of heart disease.
Ideally, the response rate should be at least 75% to ensure that a study is not significantly affected by response bias	
	A confounder- ice-cream sales/ drownings.

Sample size and sample size issues

A sample size needs to be sufficiently large to give an accurate representation of the target population.

The accuracy of a sample for determining a population characteristic depends on 2 factors-
   ⇒ The sample size (n) used for study (which depends on the kind of data in interest) and the
   ⇒ Variability (spread) of the measurements

Sample size requirements-

Sample size for determining a proportion:

Most opinion polls are based on surveys of at least 500 persons (large samples). We need this number to ensure a reasonable degree of accuracy.

Sample size for determining a mean:

Smaller samples are often sufficient for estimating characteristics of populations of numerical (measured) data

The sample size needs to be large to have accuracy but a small sample may be sufficient when the population is homogenous (i.e the population does not vary much).