Research Methodology

Sampling

UNIVERSITY
OF WOLLONGONG
AUSTRALIA

Analysing data

- For qualitative data, the researcher might analyse as the research progresses, continually refining and reorganising in light of the emerging results.
- For quantitative data, the analysis can be left until the end of the data collection process,
  - if it is a large survey, statistical software is the easiest and most efficient method to use.
  - However, once this has been done the analysis is quick and efficient, with most software packages producing well presented graphs, pie charts and tables which can be used for the final report.

Quantitative Data Analysis

- For a large survey,
  - You might collect data via questionnaire
    - properly constructed and worded,
    - proper sample size.
- Data from experiments
- Data from simulations
- …

Quantitative Data Analysis

- Statistical analysis
  - Using statistical formulas
  - By statistical software
    - SPSS
    - R
    - JMP
    - …
  - Software for graphs, tables and pie charts which can be used in your final report

Population

- A population is a collection of people, items, or events about which you want to make inferences.
- Universe/Population:
  - From a statistical point of view, the term ‘Universe’ refers to the total of the items or units in any field of inquiry,
  - whereas the term ‘population’ refers to the total of items about which information is desired.
  - Quite often, we do not find any difference between population and universe, and as such the two terms are taken as interchangeable.

Population

- The population can be finite or infinite.
  - In finite population the number of items is certain
    - The population of a city, the number of workers in a factory and the like are examples of finite population,
  - In case of an infinite population the number of items is infinite,
    - i.e., we cannot have any idea about the total number of items
    - the number of stars in the sky, listeners of a specific radio programme, throwing of a dice etc. are examples of infinite population.

Why sample

- Example: A company needs to do research about people's potential attitude about one product.
- Population: Every person in a nation
- Q: Is that possible to contact every person in a nation?
- A: Impossible!

What is sampling

- Selection of a number of study units from a defined study elements (population)
  - A sample could be people, statistical variables, etc.
- Questions to be asked in sampling
  - What is a set of elements from which we want to draw a sample?
  - How many elements we need?
  - How will these elements be selected?

Sample

- The main difference between a population and a sample has to do with how observations are assigned to the data set.
  - A population includes all of the elements from a set of data.
  - A sample consists of one or more observations from the population.
    - A sample is a subset of people, items, or events from a larger population that you collect and analyse to make inferences.
    - To represent the population well, a sample should be randomly collected and adequately large.

Sampling Frame (source list)

- The elementary units or the group or cluster of such units may form the basis of sampling process in which case they are called as sampling units.
  - A list containing all such sampling units is known as sampling frame.
  - Sampling frame consists of a list of items from which the sample is to be drawn.
  - If the population is finite and the time frame is in the present or past, then it is possible for the frame to be identical with the population.

https://www.surveymonkey.com/blog/tag/sampling-frame/

Sampling Frame

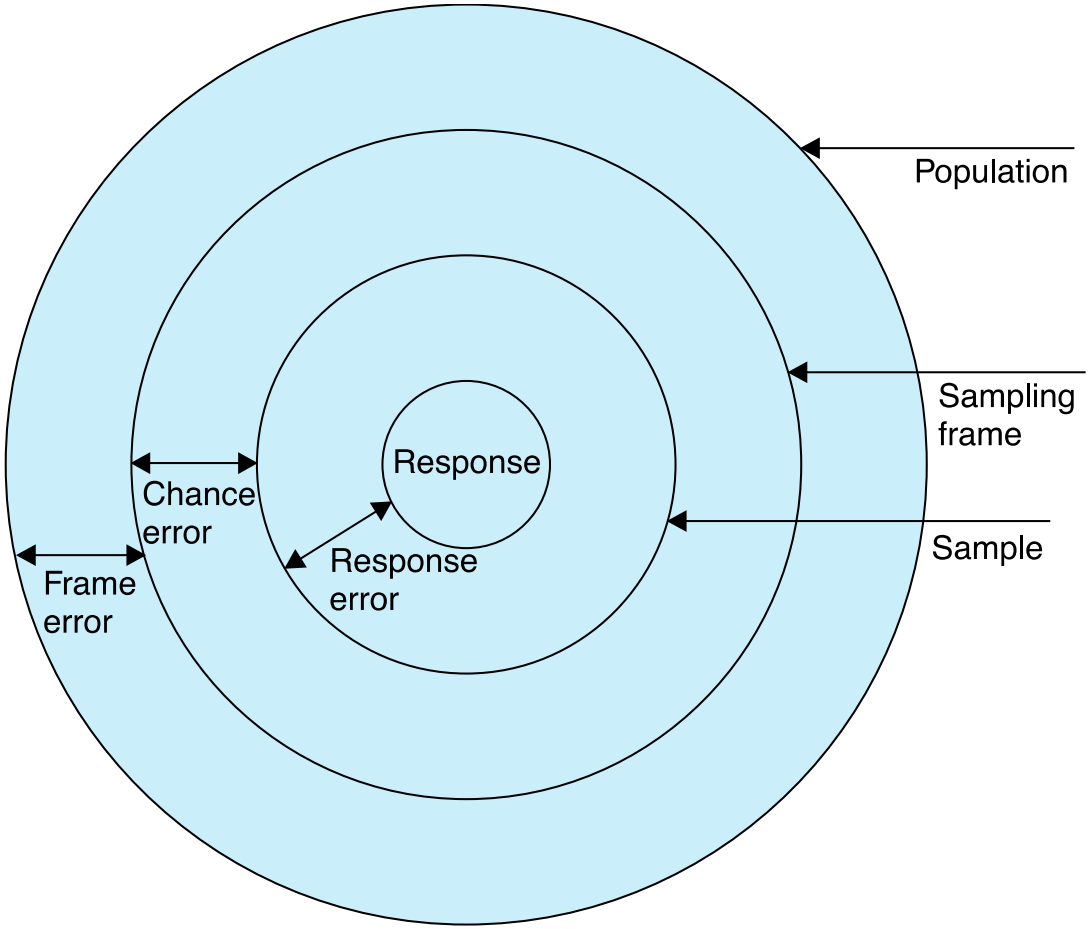
- Sample frame is either constructed by a researcher for the purpose of his study or may consist of some existing list of the population.
  - For instance, one can use telephone directory as a frame for conducting opinion survey in a city.
  - Whatever the frame may be, it should be a good representative of the population.

Statistic and Parameter

- A statistic is a characteristic of a sample, whereas a parameter is a characteristic of a population.
  - When we work out certain measures such as mean, median, or mode from samples, then they are called statistic(s)
  - When such measures describe the characteristics of a population, they are known as parameter(s)

Sampling Error

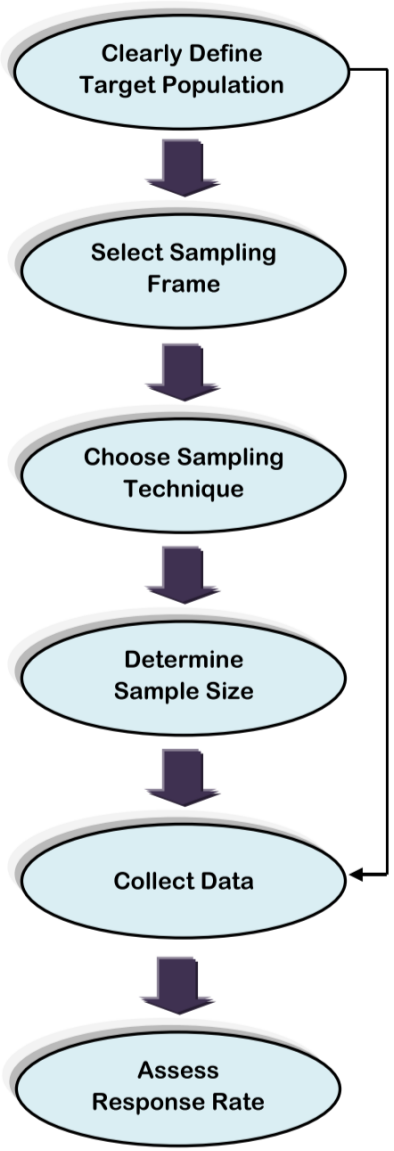
- Sample surveys do imply the study of a small portion of the population
- There would naturally be a certain amount of inaccuracy in the information collected.
  - This inaccuracy may be termed as sampling error or error variance.
- Sampling errors generally happen to be random variations (in case of random sampling) in the sample estimates around the true population values.



Sampling error = Frame error
+ chance error + response error.
(If we add measurement error or the non-sampling error to sampling error, we get total error)

Sampling error = Frame error + Chance error + Response error

Sampling process steps



Sample Design

- A definite plan for obtaining a sample from a given population
- Has to be determined before data is collected.
- Sampling unit:
  - A decision has to be taken concerning a sampling unit before selecting sample.
  - Sampling unit may be a geographical one such as state, district, village, etc., or a construction unit such as house, flat, etc., or it may be a social unit such as family, club, school, etc., or it may be an individual.
- Sampling Frame

Sample Design

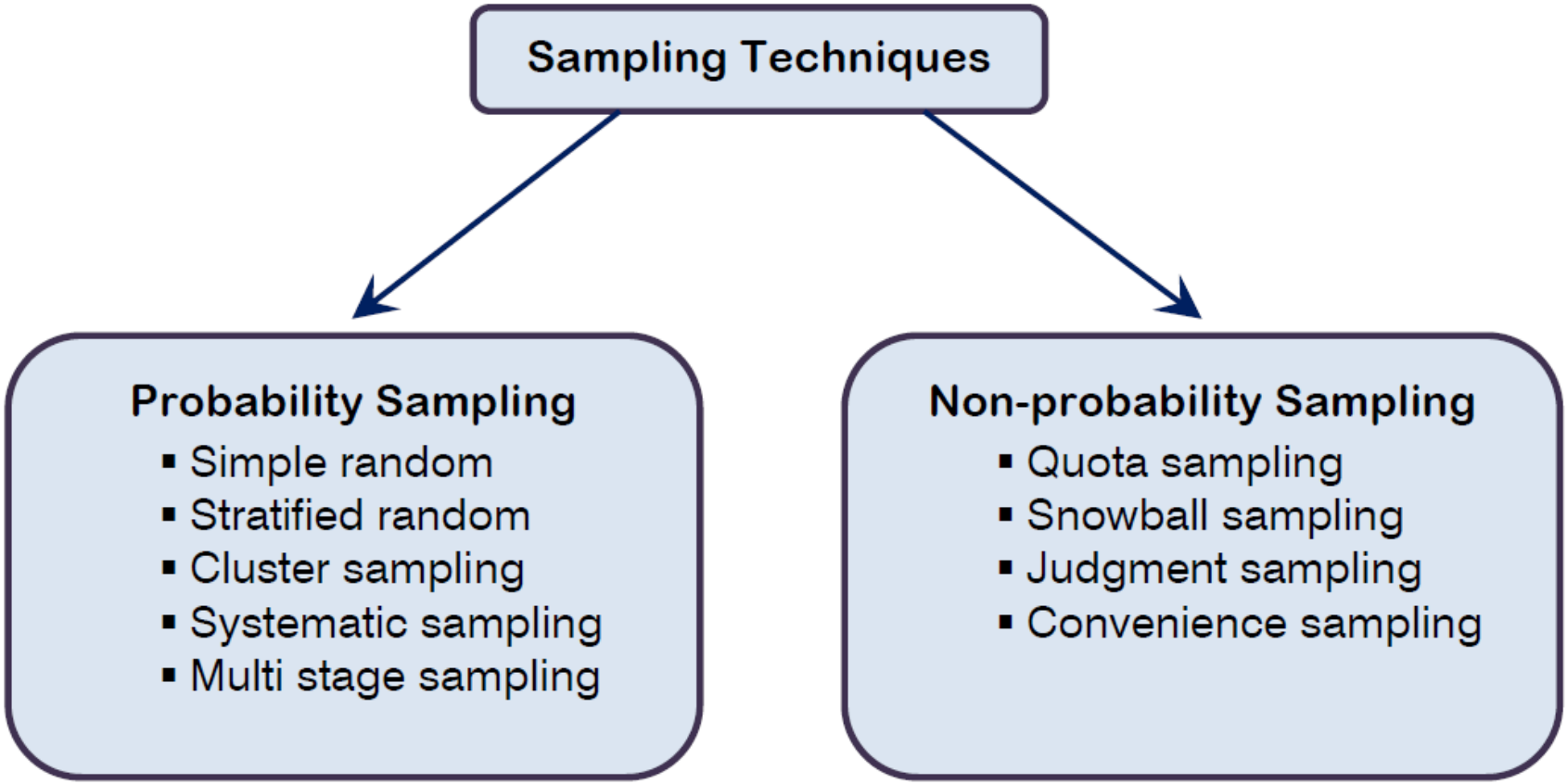
- Size of sample:
  - This refers to the number of items to be selected from the population to constitute a sample.
  - The size of sample should neither be excessively large, nor too small. It should be optimum.
  - An optimum sample is one which fulfils the requirements of efficiency, representativeness, reliability and flexibility.

Sample Design

- Size of sample:
  - ...
  - While deciding the size of sample, researcher must determine the desired precision as also an acceptable **confidence level** for the estimate.
  - The size of population variance needs to be considered as in case of larger variance usually a bigger sample is needed.

A **confidence level** refers to the percentage of all possible samples that can be expected to include the true population parameter. For example, suppose all possible samples were selected from the same population, and a confidence interval were computed for each sample. A 95% confidence level implies that 95% of the confidence intervals would include the true population parameter.

Sampling techniques



Probability Sampling

- Probability sampling is also known as ‘random sampling’ or ‘chance sampling’.
- Every item of the population has an equal chance of inclusion in the sample.
- Construct a sampling frame first and then used a random number generation computer program to pick a sample.

Probability Sampling

- The results obtained from probability or random sampling can be assured in terms of probability
  - i.e., we can measure the errors of estimation or the significance of results obtained from a random sample,
- Random sampling ensures the law of Statistical Regularity which states that if on an average the sample chosen is a random one, the sample will have the same composition and characteristics as the population.
- Considered to be the best technique of selecting a representative sample.
- Has the greatest freedom from bias but may represent the most costly sample in terms of time and energy for a given level of sampling error (Brown, 1947).

Probability Sampling

- Simple random sampling
  - The simple random sample means that every case of the population has an equal probability of inclusion in sample.
  - Disadvantages include (Ghauriand Gronhaug, 2005):
  - A complete frame (a list of all units in the whole population) is needed
  - In some studies, such as surveys by personal interviews, the costs of obtaining the sample can be high if the units are geographically widely scattered
  - The standard errors of estimators can be high.

Probability Sampling

- Systematic sampling
  - Selecting items from an ordered population using a skip or sampling interval.
  - Example: random select 1,000 people from 50,000. Place all the participants in a list and a starting point would be selected. Every 50th person on the list would be chosen, since 50,000/1,000=50.
  - Systematic sampling is better than random sampling when data does not exhibit patterns and there is a low risk of data manipulation by a researcher.

Probability Sampling

- Stratified random sampling
  - The population is divided into strata (or subgroups) and a random sample is taken from each subgroup.
  - Subgroups might be based on company size, gender or occupation (to name but a few).
  - Often used where there is a great deal of variation within a population.
  - The purpose is to ensure that every stratum is adequately represented.

Probability Sampling

- Cluster sampling
  - Whole population is divided into clusters or groups.
  - A random sample is taken from these clusters, all of which are used in the final sample.
  - Advantageous for researchers whose subjects are fragmented over large geographical areas as it saves time and money.

Non-probability sampling

- The sampling procedure which does not afford any basis for estimating the probability that each item in the population has of being included in the sample.
- Non-probability sampling is also known by different names such as deliberate sampling, purposive sampling and judgement sampling.
- Items are selected deliberately by the researcher; his choice concerning the items remain supreme.
- Purposively choose the particular units of the population for constituting a sample on the basis that the small mass will be typical or representative of the whole.
  - For instance, if economic conditions of people living in a state are to be studied, a few towns and villages may be purposively selected for intensive study on the principle that they can be representative of the entire state. Thus the judgement of the organizers of the study plays an important part in this sampling design.

Non-probability sampling

- Personal element has a great chance to be selected (if It happens, the entire inquiry may get vitiated)
- In case the investigators are impartial, the result obtained maybe tolerable reliable. However, there is no assurance that every element has some specifiable chance of being included.
- This sampling is rarely adopted in large inquiries of important.
- In small inquiries, this research may be adopted because of the relative advantage of time and money inherent in this method of sampling (e.g. Quota Sampling : interviewers are given quotas to be filled with some restriction, actual selection of items is left to the interviewer's discretion)

Non-probability sampling

- Quota sampling
  - A non random sampling technique in which participants are chosen on the basis of predetermined characteristics so that the total sample will have the same distribution of characteristics as the wider population (Davis, 2005).

Non-probability sampling

- Convenience sampling
  - Selecting participants because they are often readily and easily available.
  - Inexpensive and an easy option. For example, using friends or family as part of sample is easier than targeting unknown individuals.

Non-probability sampling

- Snowball sampling
  - Anon random sampling method that uses a few cases to help encourage other cases to take part in the study, thereby increasing sample size.
  - This approach is most applicable in small populations that are difficult to access due to their closed nature, e.g. secret societies and inaccessible professions.

Example

Consider the population consisting of four elements (say A,B,C,D). Suppose we want to take a sample of size 2 from it.

What are the samples? What is the probability of being randomly chosen of an item?

Determine sample size

- Sample size refers to the number of items to be selected from the population to constitute a sample.
  - The size of sample should neither be excessively large, nor too small. It should be optimum.
  - An optimum sample is one which fulfils the requirements of efficiency, representativeness, reliability and flexibility.

Determine sample size

- What is important is not the proportion of the research population that gets sampled, but the absolute size of the sample selected relative to the complexity of the population, the aims of the researcher.
  - Larger sample sizes reduce sampling error but at a decreasing rate.

Determine sample size

- While deciding the size of sample, researcher must determine the **desired precision** as also an acceptable **confidence level** for the estimate.
  - The size of population variance needs to be considered as in case of larger variance usually a bigger sample is needed.

A **confidence level** refers to the percentage of all possible samples that can be expected to include the true population parameter.

Population to sample

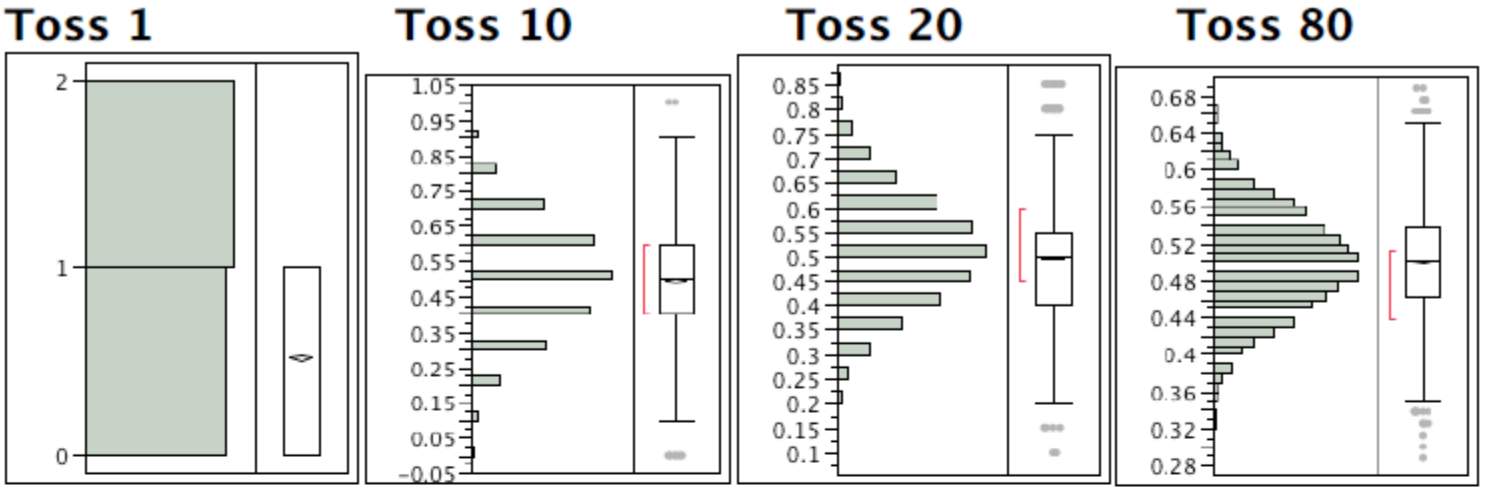
- What can we say about samples from that population?
- The average of the sample means tends towards the population mean as the sample size increases as long as we have a probability sample.
- The sample variance is a good estimator of the population variance for sample sizes greater than about 30.

Population to sample

- The standard error (the standard deviation of the distribution of the sample mean for repeated samples) is the population standard deviation divided by the square root of the sample size.
- The distribution of the sample mean is approximately Normal (a known distribution) for sample sizes greater than about 30.
- None of these statements require us to know the full distribution of values in the population
  - however, the population distribution is important if we deal with small sample sizes (less than 30), when we cannot use some of our approximations without careful checking.

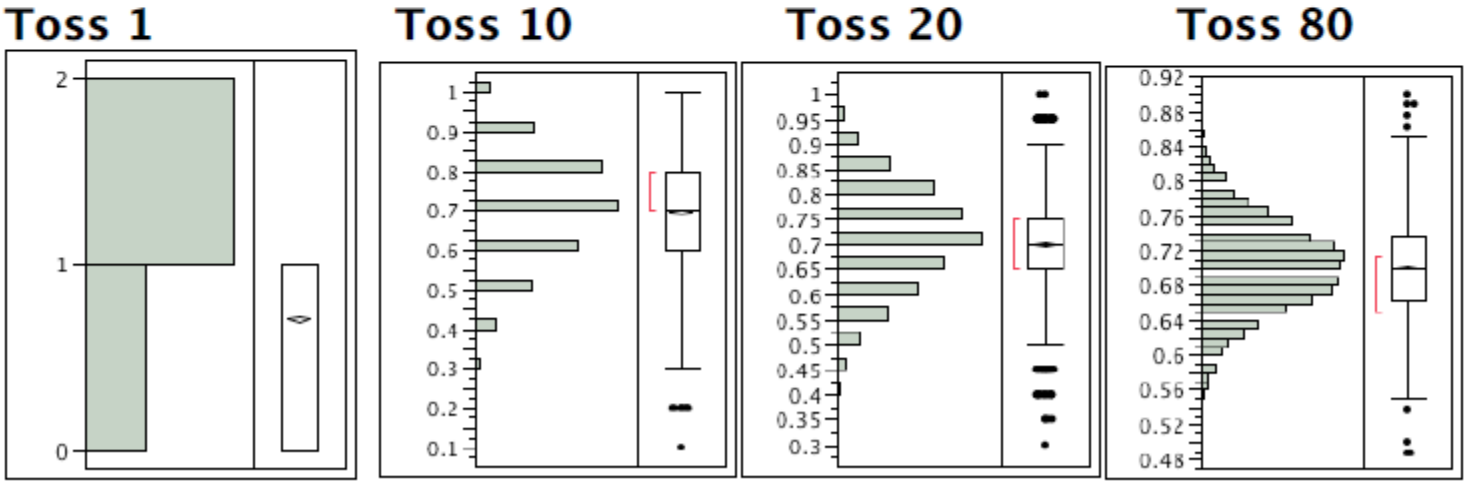
Example

We now do a simulation of tossing a coin. We will do 5000 simulations of different experiments. First, we consider a fair coin (i.e. Pr(H) = 0.5). Let us compare the histogram and box plot we get for the proportion of Heads in our 5000 simulations if we do 1 toss, 10 tosses, 20 tosses or 80 tosses in each experiment



Even with 20 tosses, histogram is bell-shaped; as sample size increases, spread (width) decreases, centre of the distribution close to 0.5.

Now unfair coin (Pr(H)=0.7)



As sample size increases, get symmetrical histogram, decreasing spread, centre close to 0.7.

Sample to population

- As the sample mean gets very close to the population mean for large samples, we say that the sample mean is a good estimator of the population mean.
- The sample mean can be used as:
    1. a point estimator for the population mean
    2. the centre of an interval estimator (confidence interval) for the population mean
    3. the basis for a hypothesis test of whether the population mean has a particular value.
- In all cases, we are using our knowledge about how the population relates to the sample to make reverse statements about the population from a selected sample.