manipalglobal
Academy of Data Science

Statistical Techniques for Data Science

Sampling

Introduction to Sampling

Dr. Subhabaha Pal
Manipal Global Academy of Data Science

Objective

manipalglobal
Academy of Data Science

After attending this session, you will be able to –

- **Define Population, Sample, Sampling Frame and Sampling Unit**
- **Define Parameter and Statistic**
- **Explain what is Sampling and distinguish between different sampling methods**
- **Explain what is Sampling Variation**

Population and Sample

manipalglobal
Academy of Data Science

- Population is the set of all entities which are under study
- Sample is the sub-set of the population and it is used to draw inferences about the population under study
- Example – In order to infer about the blood-group of an individual, we generally collect a very small sample of blood from the individual to make inference about the blood-group
- In most of the cases, it is not possible to examine all individual units under study individually as it calls for cost and time and most of the cases may not be feasible also
- Drawing few entities from the whole population as sample facilitates the task of inferring about the population properties in less time and with less cost
- Example – A company wants to know the average monthly consumption of grocery products in the house-holds of a certain area - In order to achieve this, the company will select few house-holds only as sample and will check the average monthly consumptions in those house-holds
- Population Parameters – Population Mean (μ) and Population standard deviation (σ)
- Sample Parameters – Sample Mean ($\bar{x}$) and Sample standard deviation (s)

Parameter and Statistic

manipalglobal
Academy of Data Science

- Parameter refers to a measure that describes population
- Example - The mean height of class 10 students of a region is a parameter
- Statistic refers to a measure that describes a sample which is taken from a population
- Example – The mean height of 100 selected class 10 students is a statistic which is used as an estimate of the parameter – the mean height of all class 10 students of a region

Census, Sampling Frame and Sampling Unit

manipalglobal
Academy of Data Science

- **Census or Enumeration – Collection of information about every member of the population**
- **Sampling – Selecting few entities from the whole population as part of sample**
- **Sampling Frame – Sampling frame is the list of all entities of the population which can be selected in the sample**
- **Sampling Unit – Sampling unit is a member of the sample**

Why Sampling is done?

manipalglobal
Academy of Data Science

- Whole population may be too large to study
- Time-efficient, Cost-efficient and Feasible
- Can provide a close approximation of the population
- Information actually be more accurate when based on carefully drawn samples
- Offer greater scope and flexibility than a census

Sampling Methods

manipalglobal
Academy of Data Science

- **Probability Sampling – Each member of the population has a known non-zero probability of being selected**
  - **Random Sampling**
  - **Systematic Sampling**
  - **Stratified Sampling**
- **Non-Probability Sampling – Members are selected from the population in some non-random manners**
  - **Convenience Sampling**
  - **Judgement Sampling**
  - **Quota Sampling**
  - **Snowball Sampling**

Simple Random Sampling

manipalglobal
Academy of Data Science

- Target Population must be Homogeneous and finite
- Population is relatively small
- Sampling frame is complete and up-to-date.
- Samples are selected unit by unit
- Each sampling unit will have and equal chance of being selected
- The random selection from the sampling frame can be done using a table of random numbers table or Lottery method
- Simple Random Sampling With Replacement (the sample is returned back to sampling frame after noting down features of the entity – can be re-selected)
- Simple Random Sampling Without Replacement (the sample is not returned – can not be re-selected)

Systematic Sampling

manipalglobal
Academy of Data Science

- **Systematic Sampling is often used instead of random sampling**
- **It is also called N-th Name Selection Sampling**
- **After the required sample size has been calculated, every N-th record is selected from a list of population members**
- **As long as the list does not contain any hidden order, this sampling method is as good as the random sampling method**
- **Its only advantage over the random sampling technique is simplicity (and possibly cost effectiveness)**

Stratified Sampling

manipalglobal
Academy of Data Science

- **Stratified sampling** is commonly used probability method that is superior to random sampling because it reduces sampling error

- A stratum is a subset of the population that share at least one common characteristic; such as males and females.

  - Identify relevant stratums and their actual representation in the population so that there is homogeneity within the stratum

  - Random sampling is then used to select a *sufficient* number of subjects from each stratum using Probability Proportional to Population Size (PPPS)

  - Stratified sampling is often used when one or more of the stratums in the population have a low incidence relative to the other stratums

Cluster Sampling

manipalglobal
Academy of Data Science

- Cluster Sample: a probability sample in which each sampling unit is a collection of elements.

- Cluster Sampling is effective under the following conditions -
  - A good sampling frame is not available or costly, while a frame listing clusters is easily obtained
  - The cost of obtaining observations increases as the distance separating the elements increases

- Examples of clusters:
  - City blocks – political or geographical
  - Housing units – college students
  - Hospitals – illnesses

Convenience Sampling

manipalglobal
Academy of Data Science

- **Convenience sampling** is used in exploratory research where the researcher is interested in getting an inexpensive approximation.
- The sample is selected because they are convenient.
- It is a nonprobability method.
  - Often used during preliminary research efforts to get an estimate without incurring the cost or time required to select a random sample

Judgement Sampling

manipalglobal
Academy of Data Science

- **Judgment sampling** is a common nonprobability method.
- The sample is selected based upon judgment.
  - Judgment Sampling is an extension of convenience sampling
- When using this method, the researcher must be confident that the chosen sample is truly representative of the entire population

Quota Sampling

manipalglobal
Academy of Data Science

- **Quota sampling** is the nonprobability equivalent of stratified sampling
  - First identify the stratums and their proportions as they are represented in the population
  - Then convenience or judgment sampling is used to select the required number of subjects from each stratum

Snowball Sampling

manipalglobal
Academy of Data Science

- **Snowball sampling** is a special nonprobability method used when the desired sample characteristic is rare
- It may be extremely difficult or cost prohibitive to locate respondents in these situations
- This technique relies on referrals from initial subjects to generate additional subjects
- It lowers search costs; however, it introduces bias because the technique itself reduces the likelihood that the sample will represent a good cross section from the population

manipalglobal
Academy of Data Science