Vimalathas Vithusan

University of vavuniya

**STATISTIC for data science**

**What are Statistics?**

* Statistic is the science of collecting, organizing and analysing data.

**What is data?**

* Facts or pieces of information that can be measured.

**Types of Statistic**

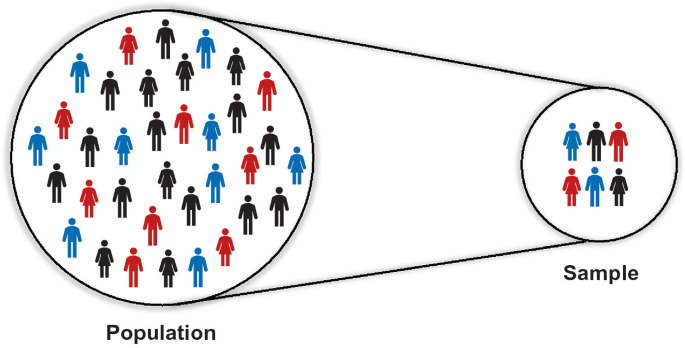
**Inferential Statistic**

* Technique where in we used the data that we have measured to form conclusion.
* These all will come under the inferential stats
  + Hypothesis testing
  + Z Test
  + T Test
  + ANOVA Test
  + Z Table
  + T Table

**Descriptive Statistic**

* It consists of organizing and Summarizing data.
* These all will come under the descriptive stats
  + Measure of central tendency. (*Mean, Median, Mode*)
  + Measure of dispersion. (*Range, Variance, Standard deviation*)
  + Histograms, Probability, Permutation.
  + Gaussian, Lognormal, Binomial distributions.

**Population(N) and Sample(n)**



Eg -:

Population 🡪 Total number of students in a university.

Sample 🡪 Number of students in a specific faculty.

**Sampling Techniques**

1. Random Sampling
   * Definition

* In random sampling, each member of the population has an equal chance of being selected for the sample.
  + Method
    - Select individuals/items randomly without any specific pattern or bias.
  + Advantages
    - Unbiased representation of the population, ensures equal opportunity for all members to be selected.
  + Example
    - Drawing names out of a hat or using random number generators to select participants.

1. Stratified Sampling
   * Definition
     + In stratified sampling, the population is divided into distinct subgroups (strata), and samples are randomly selected from each stratum.
   * Method
     + Allocate proportions of the sample size to each stratum based on their population size.
   * Advantages
     + Ensures representation of all subgroups, useful for studying specific characteristics within the population.
   * Example
     + Sampling students from different grade levels in a school.
2. Cluster Sampling
   * Definition
     + In cluster sampling, the population is divided into clusters or groups, and a random sample of clusters is selected.
   * Method
     + Randomly select clusters and include all individuals within the chosen clusters in the sample.
   * Advantages
     + Practical for large and geographically dispersed populations, reduces data collection efforts.
   * Example
     + Sampling households in randomly selected neighbourhoods.
3. Systematic Sampling
   * Definition
     + In systematic sampling, individuals/items are selected at regular intervals from a list or sequence.
   * Method
     + Choose a starting point randomly and select every nth individual/item until the desired sample size is achieved.
   * Advantages
     + Simple and easy to implement, ensures even coverage of the population.
   * Example
     + Selecting every 10th customer from a list of customers entering a store.
4. Convenience Sampling
   * Definition
     + Convenience sampling involves selecting individuals/items who are readily available or easily accessible to the researcher.
   * Method
     + Choose participants based on convenience or accessibility without any specific sampling procedure.
   * Advantages
     + Quick and inexpensive, suitable for preliminary investigations or when resources are limited.
   * Example
     + Surveying people passing by on the street.

**Variables**

* A variable is a property that can take any values.