Sourav Kapil
Data Scientist
1st Jan, 2023

Statistics Notes for Data Science

Basic Stats

Statistics

Statistics is the science of collecting, organising and analysing the data. We are specifically doing this for better decision making. Based on the representation of data such as using pie charts, bar graphs, or tables, we analyse and interpret it.

- **Data -** Facts or peace of information that can be measured. For example: Age of students of a class - **Data : {25, 21, 22, 20, 23}**

Types of Statistics

The types of statistics are categorised based on these features:

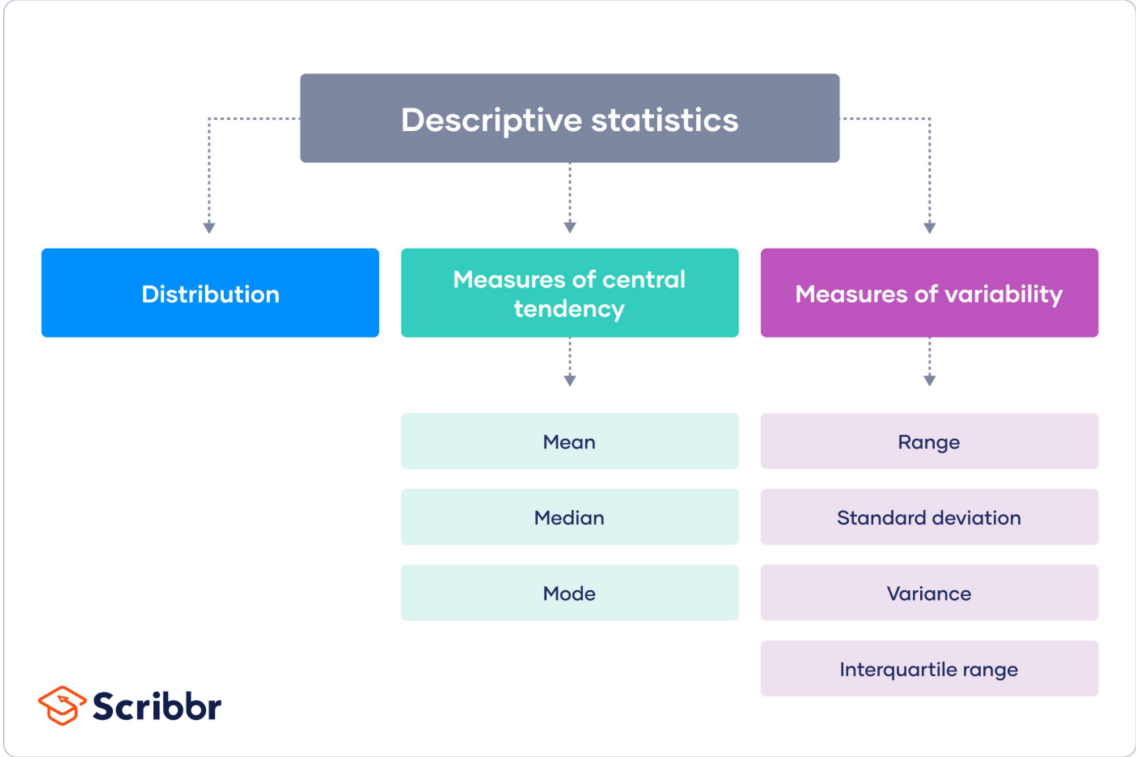
- **Descriptive Statistics**
- **Inferential Statistics**

1. Descriptive Statistics -

Descriptive statistics is a means of describing features of a dataset by generating summaries about the data samples. There are 4 major types of descriptive stats -

- ➢ Measure of **Frequency** (Count, Percent, Frequency)
- ➢ Measure of **Central Tendency** (Mean, Median, Mode)
- ➢ Measure of **Dispersion and Variation** (Range, Variance, Standard Deviation)
- ➢ Measure of **Position** (Percentile/Quartile Ranks)

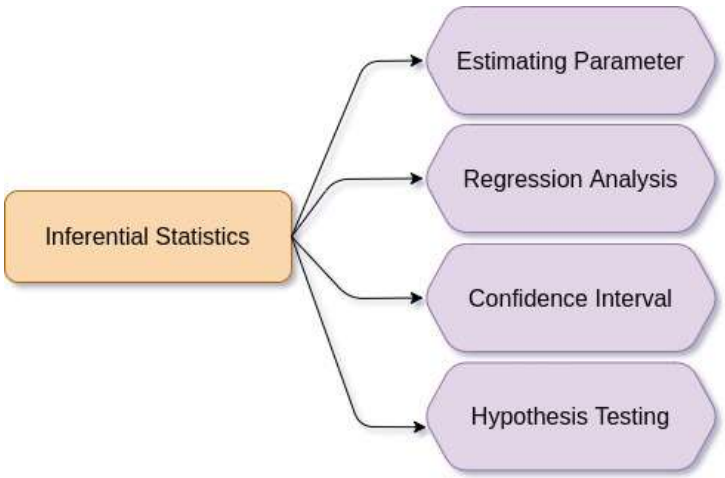
- ❖ The distribution concerns the frequency of each value.
- ❖ The central tendency concerns the averages of the values.
- ❖ The variability or dispersion concerns how spread out the values are.



2. Inferential Statistics -

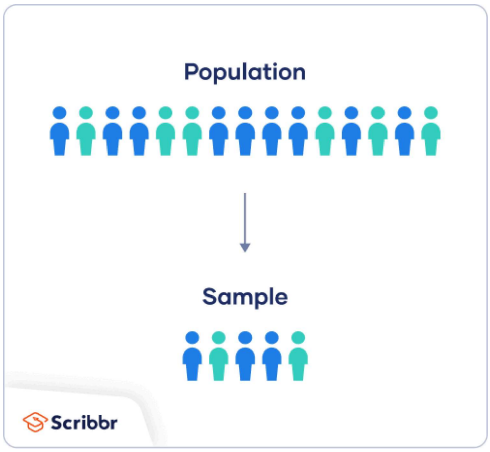
Inferential statistics is a way of making inferences about population based on samples. It's a technique where we use the data that we have measured to form conclusions. There are 3 major types of inferential stats -

- ➢ Confidence Interval
- ➢ Hypothesis Testing
- ➢ Regression Analysis



Population Vs Sample

- A population is the entire group that you want to draw conclusions about.
- A sample is the specific group that you will collect data from. The size of the sample is always less than the total size of the population.
- Samples are used to make inferences about populations.
- Samples are easier to collect data from because they are practical, cost-effective, convenient, and manageable.
- In research, a population doesn't always refer to people.



Population Mean And Sample Mean

★ The sample mean is the mean calculated from a group of random variables, drawn from the population. Compared to the population, the sample size is small. The sample size is represented by 'n.' The sample mean is calculated as under :-

$$Sample\ Mean\ \bar{x} = \frac{1}{n}\sum_{i=1}^{n} a_i$$

where, n = Size of sample
∑ = Add up
ai = All the observations

★ Population mean represents the actual mean of the whole population.The population size is large, and the sample size is denoted by 'N.' The population mean is calculated as under :-

$$Population\ Mean\ \mu = \frac{1}{N}\sum_{i=1}^{N} a_i$$

where N = Size of the population
∑ = Add up
ai = All the observations

Sampling

Sampling means selecting the group of data as a sample from the entire data for your research. It allows you to test hypotheses about the characteristics of a population.

Reasons for sampling -

- ➢ **Necessity:** Sometimes it's simply not possible to study the whole population due to its size or inaccessibility.
- ➢ **Practicality:** It's easier and more efficient to collect data from a sample.
- ➢ **Cost-effectiveness:** There are fewer participant, laboratory, equipment, and researcher costs involved.
- ➢ **Manageability:** Storing and running statistical analyses on smaller datasets is easier and reliable.

Sampling Techniques

When you conduct research about a group of people, it's rarely possible to collect data from every person in that group. Instead, you select a sample. The sample is the group of individuals who will actually participate in the research.

To draw valid conclusions from your results, you have to carefully decide how you will select a sample that is representative of the group as a whole. This is called a sampling method. There are two primary types of sampling methods that you can use in your research:

- ★ **Probability Sampling** involves random selection, allowing you to make strong statistical inferences about the whole group.
- ★ **Non-probability Sampling** involves non-random selection based on convenience or other criteria, allowing you to easily collect data.

Notes :-

Sampling Frame

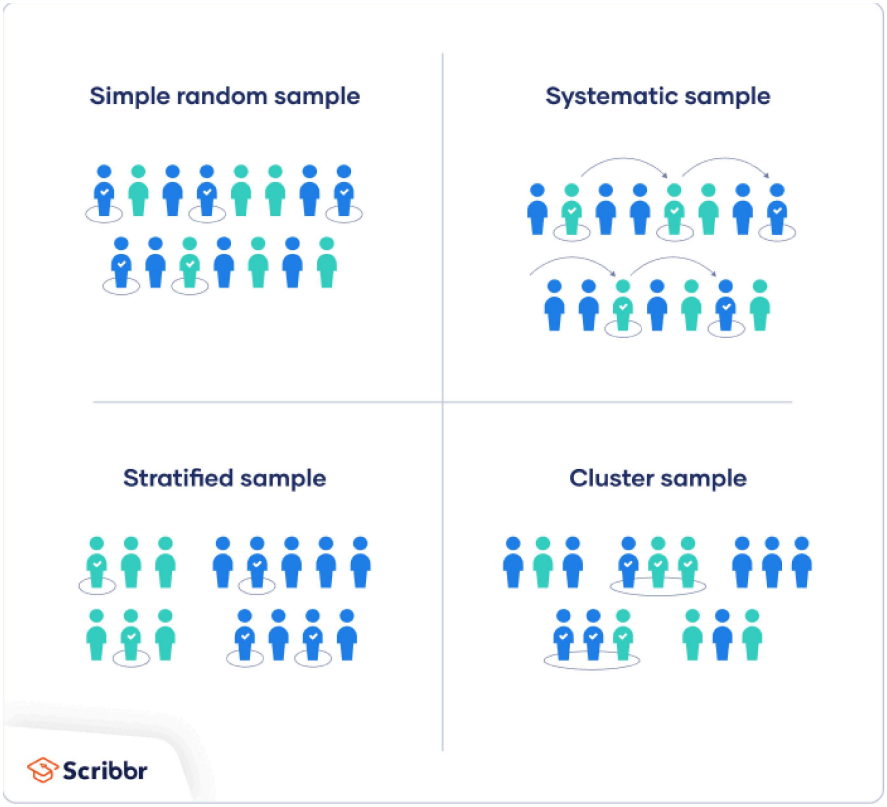
The sampling frame is the actual list of individuals that the sample will be drawn from. Ideally, it should include the entire target population (and nobody who is not part of that population).

Example: Sampling Frame
You are doing research on working conditions at a social media marketing company. Your population is all 1000 employees of the company. Your sampling frame is the company's HR database, which lists the names and contact details of every employee.

Probability Sampling

There are four main types of probability samples -

- ➢ **Simple Random Sampling**
- ➢ **Systematic Sampling**
- ➢ **Stratified Sampling**
- ➢ **Cluster sampling**



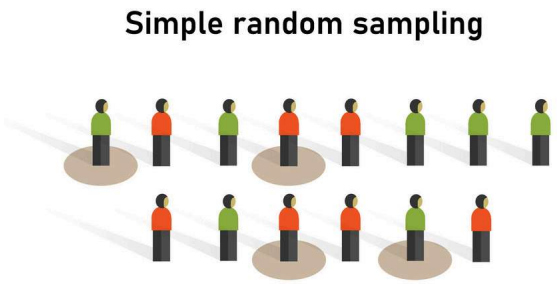
1. Simple Random Sampling

★ In a simple random sample, every member of the population has an equal chance of being selected. Your sampling frame should include the whole population.

★ To conduct this type of sampling, you can use tools like random number generators or other techniques that are based entirely on chance.

Example:

You want to select a simple random sample of 1000 employees of a social media marketing company. You assign a number to every employee in the company database from 1 to 1000, and use a random number generator to select 100 numbers.

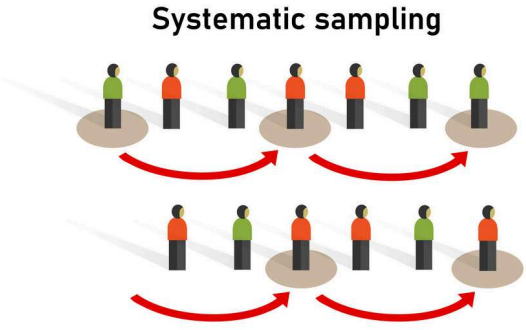


2. Systematic Random Sampling

Systematic sampling is similar to simple random sampling, but it is usually slightly easier to conduct. Every member of the population is listed with a number, but instead of randomly generating numbers, individuals are chosen at regular intervals.

Example:

All employees of the company are listed in alphabetical order. From the first 10 numbers, you randomly select a starting point: number 6. From number 6 onwards, every 10th person on the list is selected (6, 16, 26, 36, and so on), and you end up with a sample of 100 people.



3. Stratified Sampling

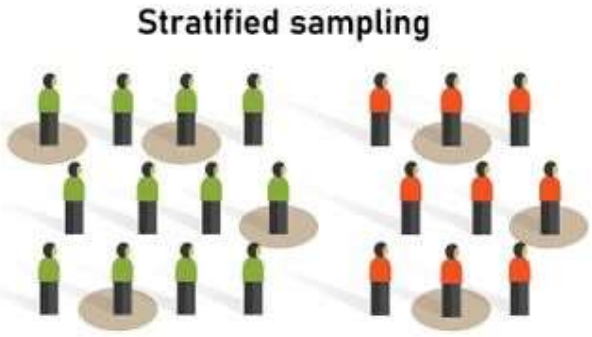
★ Stratified sampling involves dividing the population into subpopulations that may differ in important ways. It allows you to draw more precise conclusions by ensuring that every subgroup is properly represented in the sample.

★ To use this sampling method, you divide the population into subgroups (called strata) based on the relevant characteristics (e.g., gender identity, age range, income bracket, job role).

★ Based on the overall proportions of the population, you calculate how many people should be sampled from each subgroup. Then you use random or systematic sampling to select a sample from each subgroup.

Example:

The company has 800 female employees and 200 male employees. You want to ensure that the sample reflects the gender balance of the company, so you sort the population into two strata based on gender. Then you use random sampling on each group, selecting 80 women and 20 men, which gives you a representative sample of 100 people.



4. Cluster Sampling

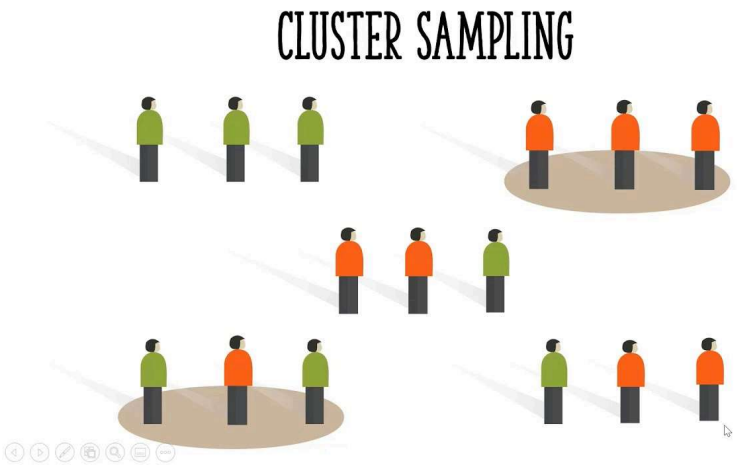
★ Cluster sampling also involves dividing the population into subgroups, but each subgroup should have similar characteristics to the whole sample. Instead of sampling individuals from each subgroup, you randomly select entire subgroups.

★ If it is practically possible, you might include every individual from each sampled cluster. If the clusters themselves are large, you can also sample individuals from within each cluster using one of the techniques above. This is called multistage sampling.

★ This method is good for dealing with large and dispersed populations, but there is more risk of error in the sample, as there could be substantial differences between clusters. It's difficult to guarantee that the sampled clusters are really representative of the whole population.

Example:

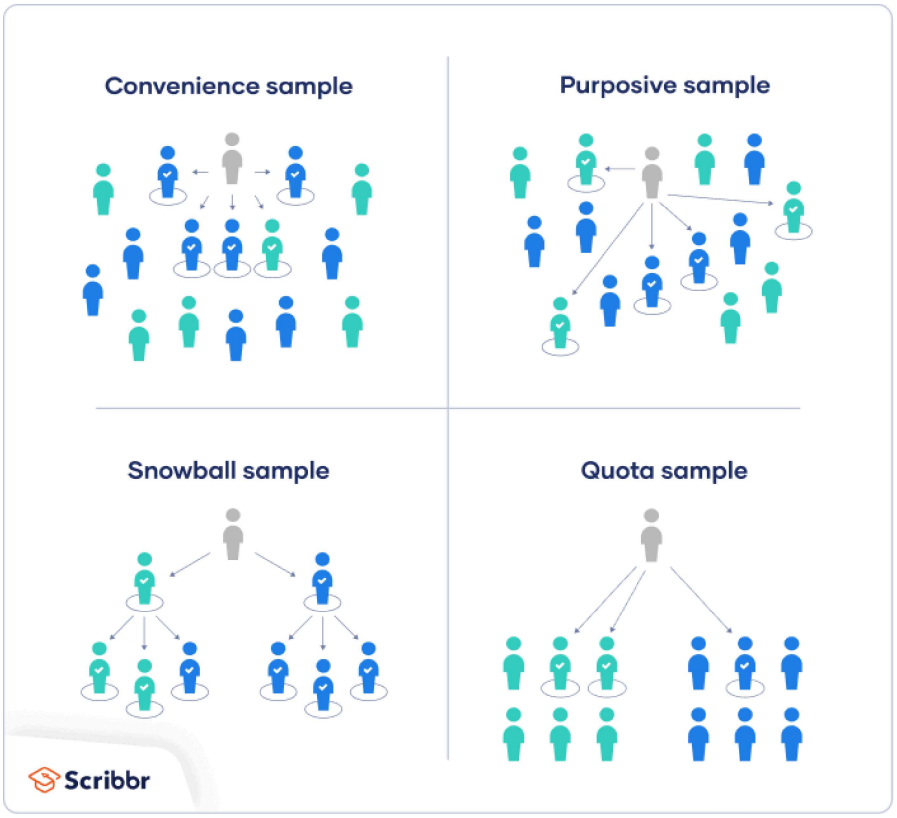
The company has offices in 10 cities across the country (all with roughly the same number of employees in similar roles). You don't have the capacity to travel to every office to collect your data, so you use random sampling to select 3 offices – these are your clusters.



Non-Probability Sampling

There are also 4 types of non-probability sampling -

- ➢ **Convenience Sampling**
- ➢ **Purposive sampling**
- ➢ **Snowball Sampling**
- ➢ **Quota Sampling**



1. Convenience Sampling

★ A convenience sample simply includes the individuals who happen to be most accessible to the researcher.

★ This is an easy and inexpensive way to gather initial data, but there is no way to tell if the sample is representative of the population, so it can't produce generalizable results. Convenience samples are at risk for both sampling bias and selection bias.

Example:

You are researching opinions about student support services in your university, so after each of your classes, you ask your fellow students to complete a survey on the topic. This is a convenient way to gather data, but as you only surveyed students taking the same classes as you at the same level, the sample is not representative of all the students at your university.