Statistics and Probability for Data Science And Data Analyst

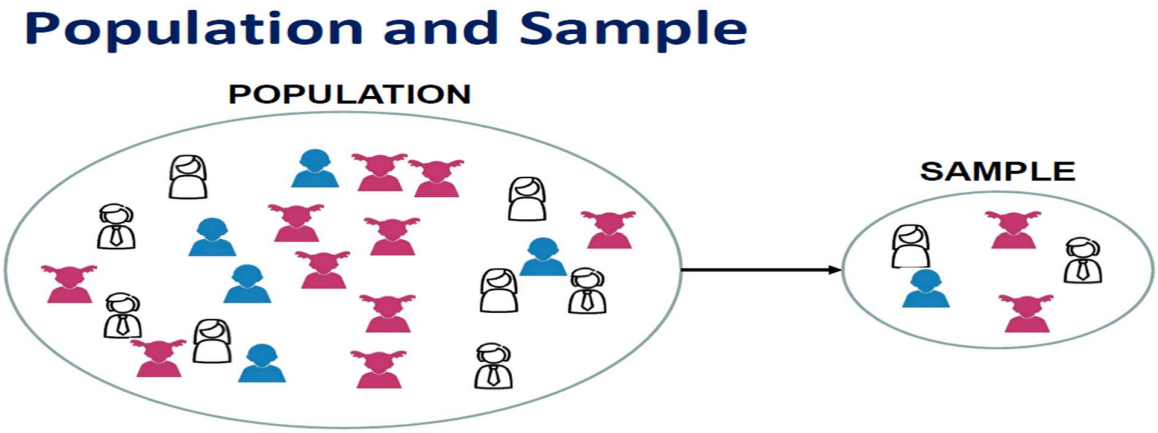
Chapter 1 - Introduction to Statistics:

❖ **What is data?**
Data is a collection of information gathered by observations, measurements, research or analysis.

❖ **What is statistics?**
Statistics is the science of planning studies and experiments, obtaining data, and organising, summarizing, presenting, analysing, interpreting, and drawing conclusions based on the data.

❖ **What is Population?**
A population is the complete collection of all individuals (scores, people, measurements, and so on). The collection is complete in the sense that it includes all of the individuals to be studied.

❖ **What is Sample data?**
A sample is a subcollection of members selected from a population.



❖ **What kind of number are parameter and statistics?**
Statistics and parameters are numbers that summarize any measurable characteristic of a sample or a population.

*For categorical variables (e.g., political affiliation), the most common statistic or parameter is a proportion.
*For numerical variables (e.g., height), the mean or standard deviation are commonly reported statistics or parameters.

❖ **There are various sampling techniques: -**

A. **Probability sampling: -** means that every member of the population has a chance of being selected. It is mainly used in quantitative research.

1. **Simple random sampling: -** every member of the population has an equal chance of being selected.

Eg :- 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.

Statistics and Probability for Data Science And Data Analyst

2. **Systematic sampling: -** Every member of the population is listed with a number, but instead of randomly generating numbers, individuals are chosen at regular intervals.

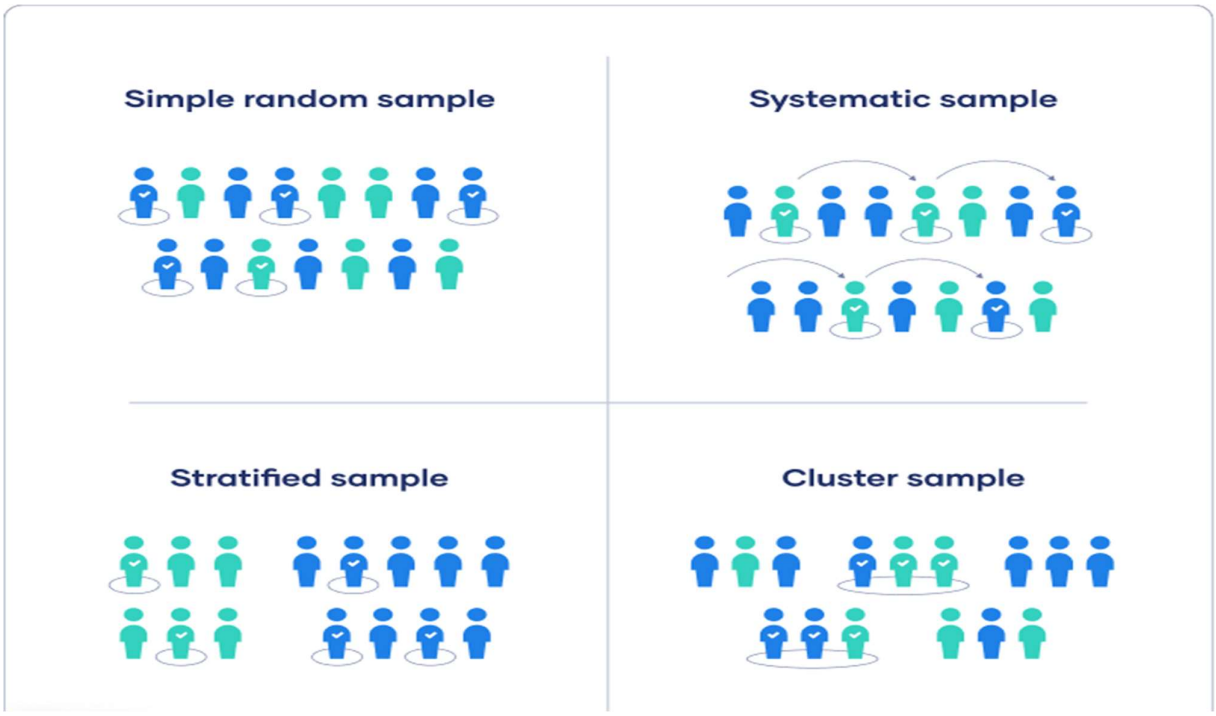
   Eg:- All employees of the company are listed in alphabetical order. From the first 10 numbers, you randomly select a starting point: number 6. From number 6 onwards, every 10th person on the list is selected (6, 16, 26, 36, and so on), and you end up with a sample of 100 people.

3. **Stratified sampling: -** You divide the population into subgroups (called strata) based on the relevant characteristic (e.g., gender identity, age range, income bracket, job role).

   Eg: - 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: -** Dividing the population into subgroups, but each subgroup should have similar characteristics to the whole sample. you randomly select entire subgroups.

   Eg: - 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.



Statistics and Probability for Data Science And Data Analyst

B. Non-Probability Sampling: - Individuals are selected based on non-random criteria, and not every individual has a chance of being included. Non-probability sampling techniques are often used in exploratory and qualitative research.

1) **Convenience sampling: -** With convenience sampling we simply use results that are very easy to get.

Eg: - 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.

❖ Types Of Data

- **Qualitative or categorical data: -** Data consists of names or labels that are not numbers representing counts or measurements**.**

  Eg: - Gender, Religion, colours.

- **Quantitative data: -** It represents the numerical values where we can count or measurements.

  Eg: - Height, weight, temperature.

Quantitative data can be further described by Discrete And Continuous types.

1] Discrete data: - A data which is represented with a finite number or a count number.

Eg:- Number of employee in organization.
Number of boats in port.

2] Continuous data: - A data which is represented with finite number of values that can subdivide into fraction and decimals and it can be measure.

Eg:- Weight, Height

- **Interval data: -** Interval Data are measured and ordered with the nearest items but have no meaningful zero. Interval data can be negative.

  Eg:- Temperature (°C or F, but not Kelvin).

- **Ratio data: -** Ratio Data are measured and ordered with equidistant items and a meaningful zero and never be negative.

  Eg:- Age (from 0 years to 100+).

Statistics and Probability for Data Science And Data Analyst

Qualitative data can be further described by Nominal and Ordinal types.

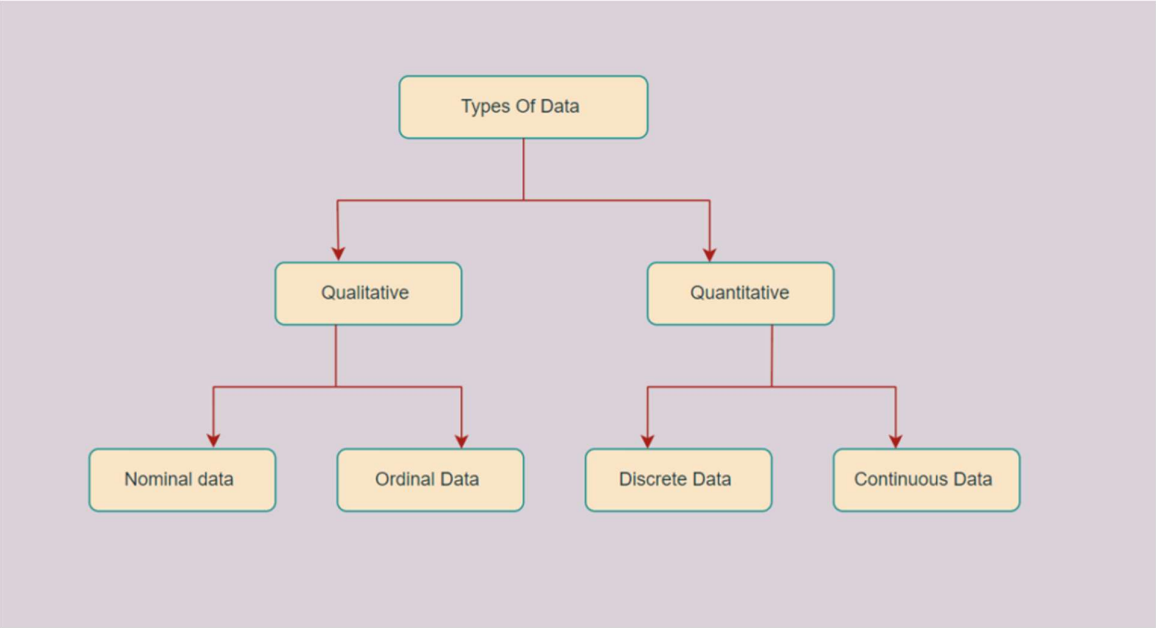
1] Nominal data: - Data that consist of names, labels, or categories only. The data cannot be arranged in an order.

Eg:- Nationality, country

2] Ordinal data: - Ordinal type of data involves variables that follow a natural order.

Eg:- Class ranking in a school: 8th, 9th, 10th, and so on.

Ranking of athletes in a race: first, second, third.



https://linkedin.com/in/jignesh-shah-b64bb1184

https://github.com/Jigs1696

By Jignesh shah