Ep 2 - Fundamentals of Statistics

Nov 27, 2023

Population and Sampling

Population (N):

The population refers to the entire group of individuals, items, or data points that you are interested in studying or analysing. It represents the complete set of subjects that meet specific criteria. For example, if you're studying the heights of all adults in a particular country, the entire adult population of that country would be your population.

Parameter: A parameter is a numerical value that describes a characteristic of a population.

Sample (n):

A sample is a representative subset of the population. It's selected to make inferences or draw conclusions about the entire population. The process of selecting a sample is known as **sampling**.
When selecting a sample, it is crucial to ensure that it is **sizeable**, **random** and **representative** of the population.

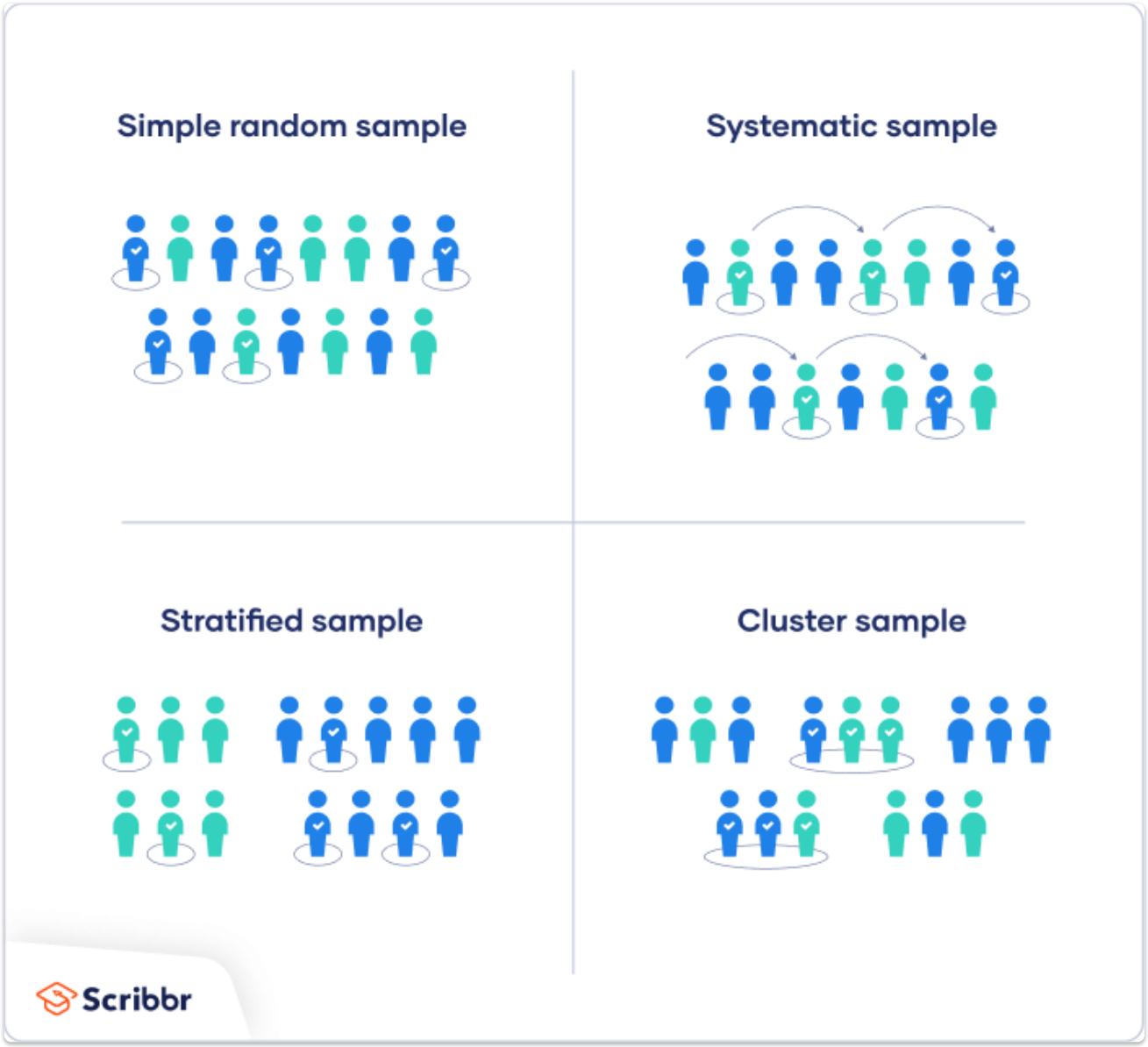
- Random sampling minimizes bias and ensures that every member of the population has an equal chance of being included in the sample.
- Representativeness means that the sample should accurately reflect the characteristics of the population.

Statistic: A statistic is a numerical value that describes a characteristic of a sample, which is a subset of the population.

Sampling Techniques

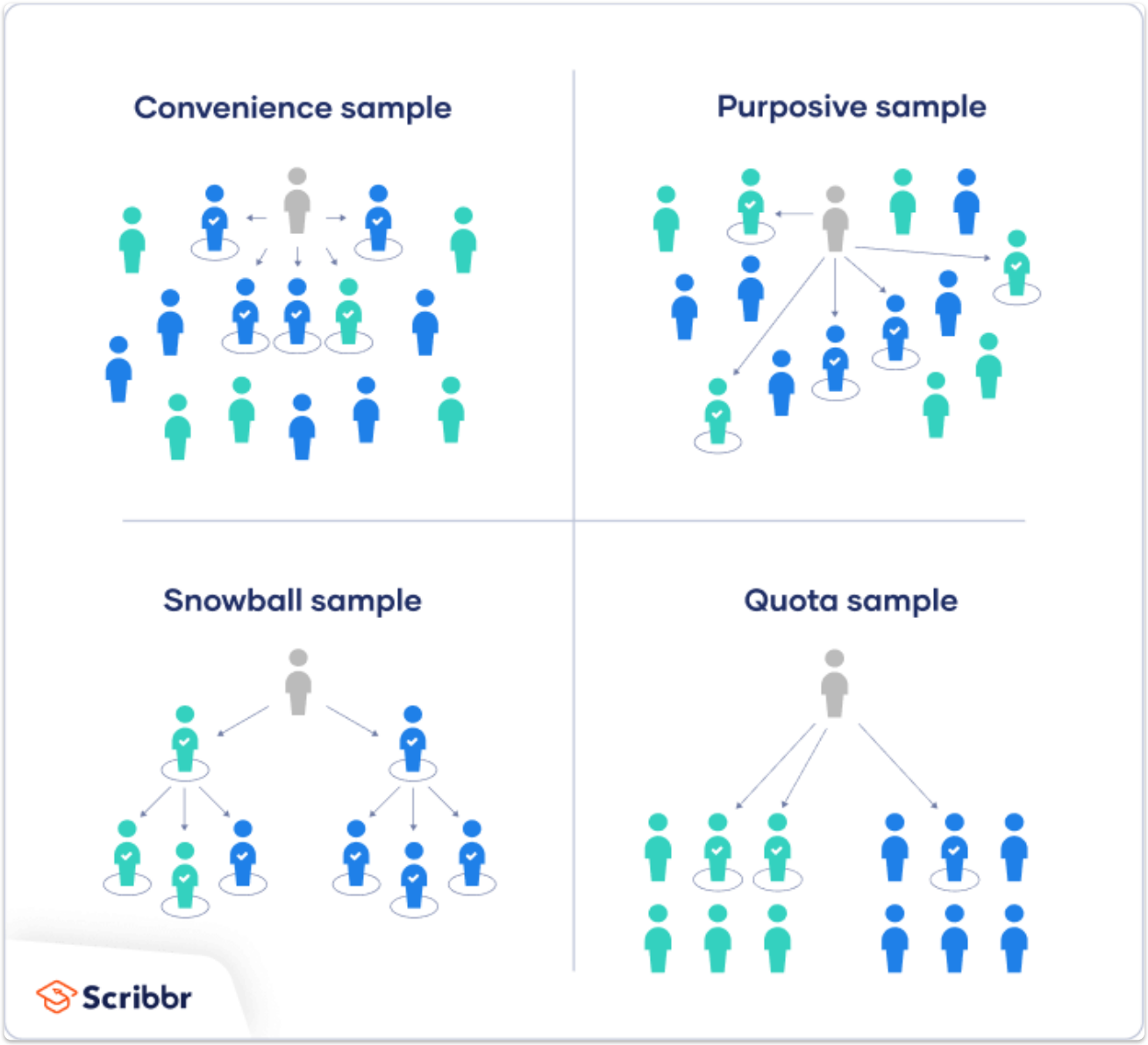
1. **Simple Random Sampling:** In this method, every individual or item in the population has an *equal chance* of being selected. It's often achieved using random number generators or drawing lots.
2. **Systematic Sampling:** Here, every n^{th} individual from a list of the population is selected.

3. **Stratified Random Sampling:** In this method, the population is split into *non-overlapping groups(strata)*. A random sample is then taken from each stratum. This technique ensures that each subgroup is represented in the sample, making it useful for maintaining diversity.
4. **Cluster Sampling:** The population is divided into clusters, and a random sample of clusters is chosen. Then, all individuals within the selected clusters are included in the sample. This technique is useful when it's more practical to sample clusters rather than individuals (e.g., households in a city).



5. **Convenience Sampling:** This method involves selecting individuals who are readily available or convenient to sample. It's *not a highly recommended* technique for obtaining representative samples as it can introduce bias.
6. **Purposive Sampling:** In purposive sampling, the researcher intentionally selects specific individuals or items based on their expertise or knowledge of the subject. This method is used when the researcher believes that *certain individuals have unique insights*.

7. **Snowball Sampling:** Often used in social research, it starts with one or a few participants who, in turn, refer other participants. This method is useful when studying hard-to-reach populations, like marginalized communities.
8. **Quota Sampling:** The researcher defines quotas for different subgroups of the population based on certain characteristics (e.g., age, gender). The sample is then filled by selecting individuals to meet these quotas. It's commonly *used in market research*.



Sampling from a Statistical Distribution
Effective Sample Size

Variables

A variable is a placeholder that can take on any value. There are 2 kinds of variables:

- **Quantitative/Numerical Variable**
- **Qualitative/Categorical Variable**

Numerical Variable

Can be measured and manipulated numerically. Examples are age, sales, height, weight, etc. It can be further classified into 2 more types:

- **Discrete:** These take on a finite set of distinct values or can be counted in whole numbers. Examples include the number of people in a household or the count of items sold.
- **Continuous:** These can take on any value within a range and often involve measurement. Examples include height, weight, temperature, or income.

Categorical Variable

Can be divided into groups on the basis of features. Examples are colours, country, cuisine, etc. It can also be further classified into 2 more types:

- **Ordinal:** These have categories with a meaningful order or rank. For instance, education levels (e.g., high school, bachelor's degree, master's degree) are ordinal because they imply an order of attainment.
- **Nominal:** These have categories with no inherent order or ranking. Examples include colours, species of animals, or types of fruits.

Types of Statistics on the Basis of Variables

Uni-variate, bi-variate, and multivariate are terms used to describe the number of variables or data dimensions involved in statistical analysis. These terms help categorize and define the scope of the analysis.

Uni-variate Analysis

- **Definition:** Uni-variate analysis focuses on a single variable, examining its distribution, characteristics, and statistical properties in isolation. It involves summarizing, visualizing, and drawing inferences about a single data variable.
- **Examples:** Describing the distribution of ages in a population, calculating the mean and standard deviation of test scores, creating histograms of income levels, or examining the frequency of different categories in a survey question.

Bi-variate Analysis

- **Definition:** Bi-variate analysis involves the examination of the relationship between two variables. It explores how one variable (independent variable) affects or is related to another (dependent variable) and often includes correlation and regression analysis.

- **Examples:** Studying the relationship between hours spent studying (independent variable) and exam scores (dependent variable), analysing the correlation between income and education level, or examining how advertising spending affects product sales.

Multivariate Analysis

- **Definition:** Multivariate analysis deals with the simultaneous analysis of three or more variables. It explores complex relationships between multiple variables to uncover patterns, associations, and dependencies among them. It includes a wide range of statistical techniques designed to handle multiple variables simultaneously.
- **Examples:** Using principal component analysis (PCA) to reduce the dimensionality of a dataset with many variables, conducting multivariate regression to model the impact of several predictors on an outcome variable, or employing factor analysis to identify underlying factors explaining the correlations between multiple observed variables.

Summary

Uni-variate analysis focuses on one variable at a time, examining its distribution and characteristics.

Bi-variate analysis involves the examination of relationships between two variables and often includes correlation and regression techniques.

Multivariate analysis deals with three or more variables, aiming to explore complex relationships and patterns among them using various statistical methods.