Session-13 Statistics

02 August 2024 01:55 PM

Agenda

- **Introduction to Statistics**
- **Type of Statistics**
- **Descriptive Stats Vs Inferential Stats**
- **Population and Sample data**
- **Sampling and their techniques.**
- **Simple Random Sampling**
- **Systematic Sampling**
- **Stratified Sampling**
- **Cluster Sampling**

1. What is Statistics? Statistics is the science of collecting, organising & analysing data. → Piece of information.

Height = {190, 170, 165, 175, 160} cm

Weight → {65, 60, 70, 50, 50} kg

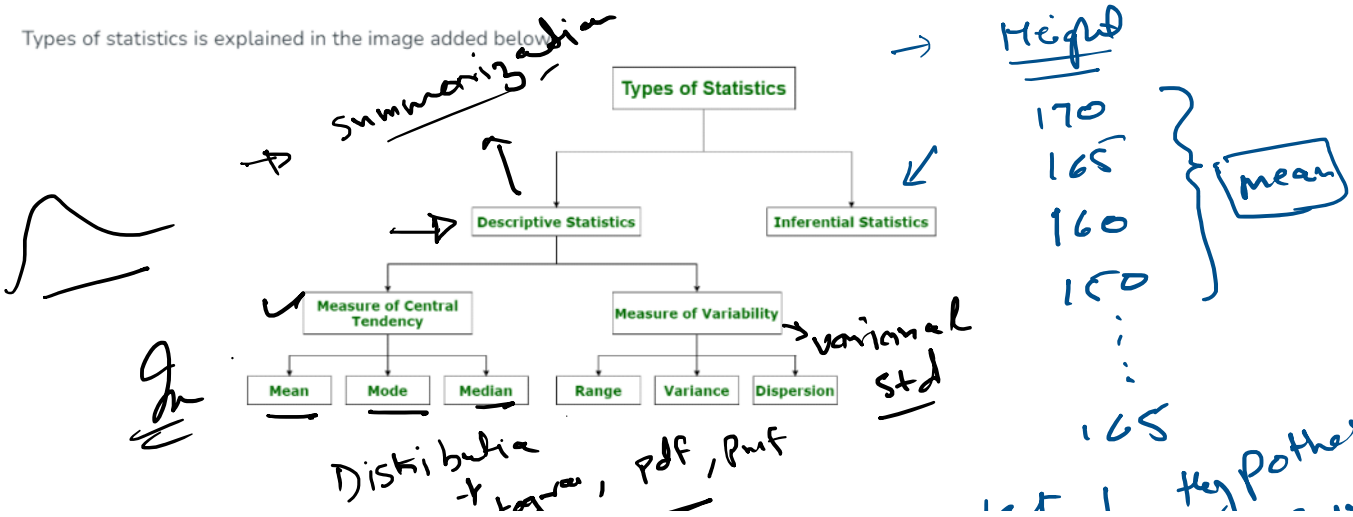
IQ = {120, 100, 90, 95}

Types of Statistics

There are 2 types of statistics:

- *Descriptive Statistics*
- *Inferential Statistics*

Types of statistics is explained in the image added below



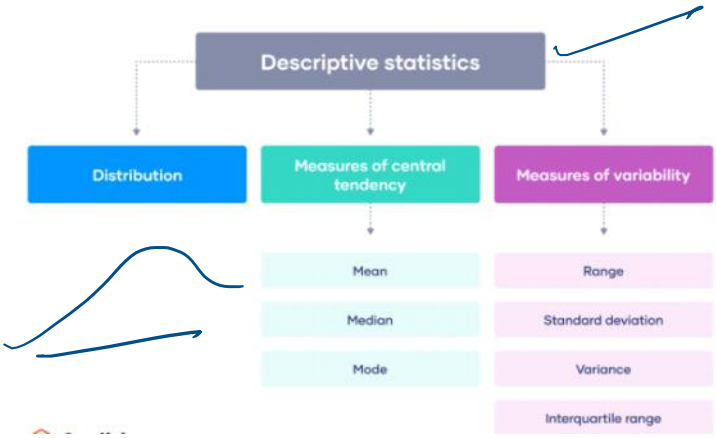
Descriptive Statistics

Descriptive statistics uses data that provides a description of the population either through numerical calculated graphs or tables. It provides a graphical summary of data.

Types of descriptive statistics

There are 3 main types of descriptive statistics:

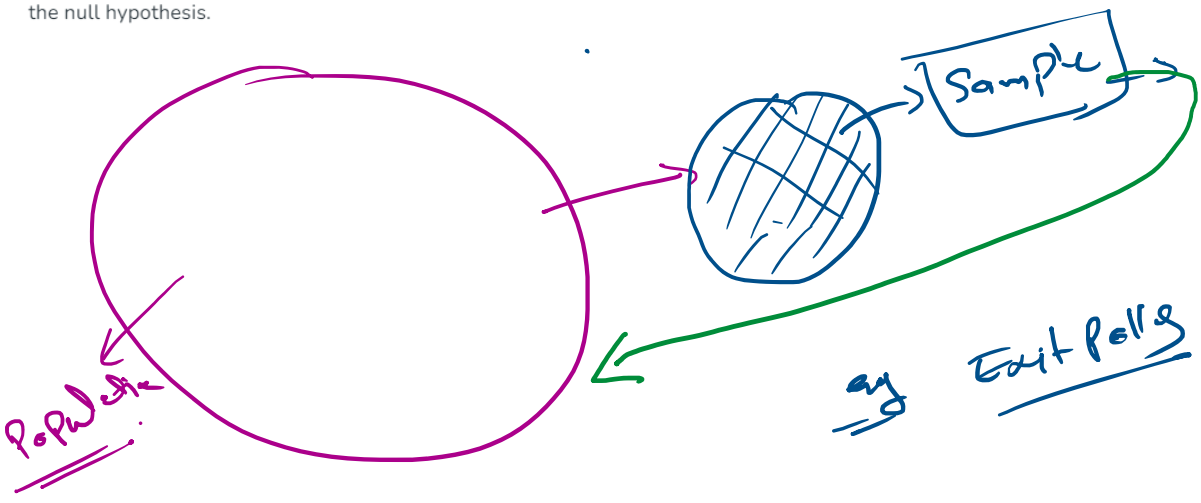
- The **distribution** concerns the frequency of each value.
- The **central tendency** concerns the averages of the values.
- The **variability** or dispersion concerns how spread out the values are.



Inferential Statistics

Inferential Statistics **makes inferences and predictions about the population based on a sample of data taken from the population.** It generalizes a large dataset and applies probabilities to draw a conclusion.

It is simply used for explaining the meaning of descriptive stats. It is simply used to analyze, interpret results, and draw conclusions. Inferential Statistics is mainly related to and associated with hypothesis testing whose main target is to reject the null hypothesis.



Population vs. sample

First, you need to understand the difference between a population and a sample, and identify the target population of your research.

- The **population** is the entire group that you want to draw conclusions about.
- The **sample** is the specific group of individuals that you will collect data from.

The population can be defined in terms of geographical location, age, income, or many other characteristics.

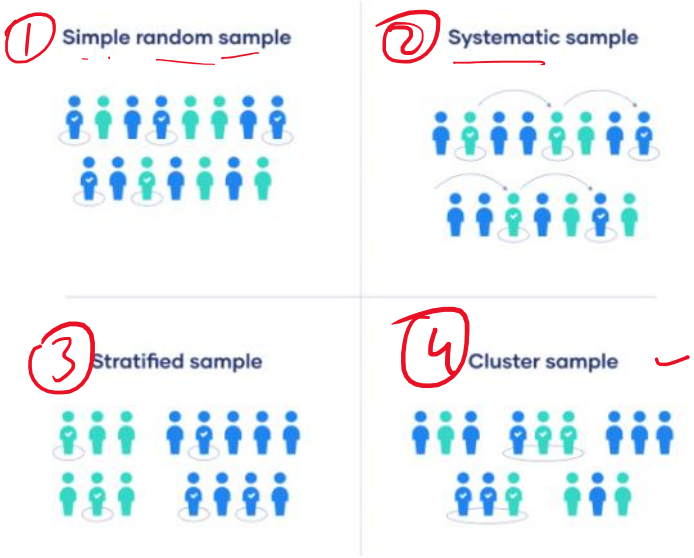
It can be very broad or quite narrow: maybe you want to make inferences about the whole adult population of your country; maybe your research focuses on customers of a certain company, patients with a specific health condition, or students in a single school.

It is important to carefully define your target population according to the purpose and practicalities of your project.

Types of Samling

Probability sampling methods

There are four main types of probability sample.

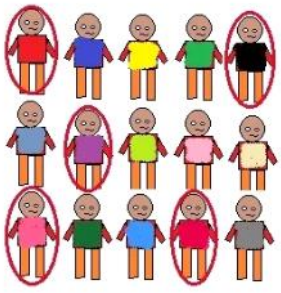


1. Simple random sampling

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.

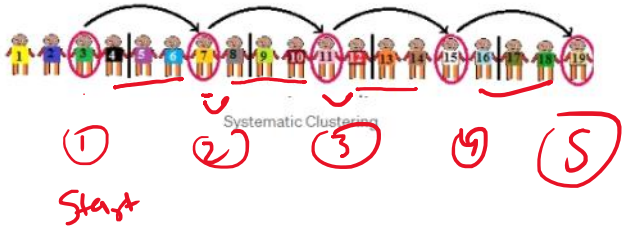


Single Random Sampling

For example: Random selection of 20 students from class of 50 student. Each student has equal chance of getting selected. Here probability of selection is 1/50

2. Systematic 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.



Systematic Clustering

Example: Systematic sampling

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.

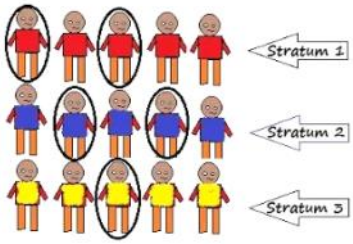
3. Stratified sampling

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

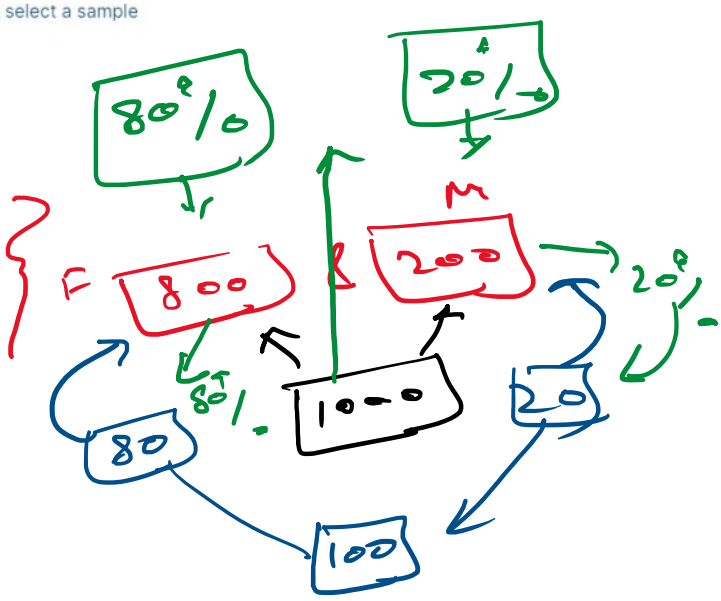
To use this sampling method, you divide the population into subgroups (called strata) based on the relevant characteristic (e.g., gender identity, age range, income bracket, job role).

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.

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.



Stratified Sampling



Example: Stratified sampling

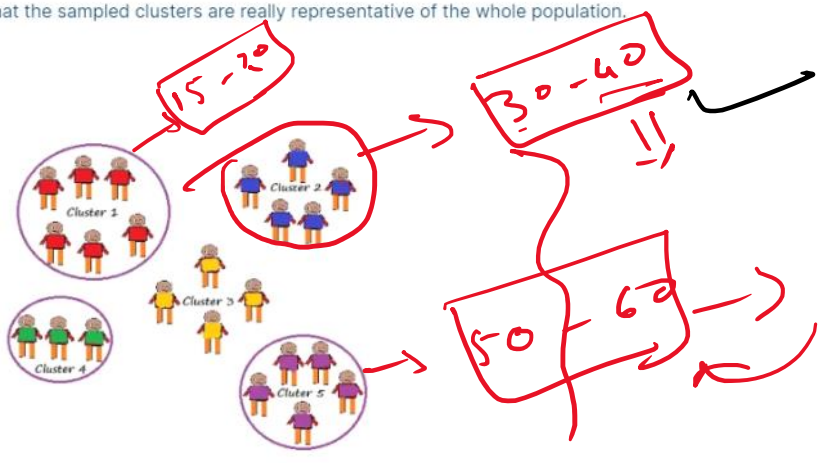
The company has 800 female employees and 200 male employees. You want to ensure that the sample reflects the gender balance of the company, so you sort the population into two strata based on gender. Then you use random sampling on each group, selecting 80 women and 20 men, which gives you a representative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



Single Stage Cluster Sampling

Example: Cluster sampling

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.

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.

Measure of Central tendency

① Mean or Average

$$Avg = \frac{\sum_{i=1}^{N} x_i}{N}$$

Population → N

$$Avg = \frac{\sum_{i=1}^{n} x_i}{n}$$

Sample → n

Q $X = \{1, 1, 2, 2, 3, 4, 5, 5, 6\}$

$$Avg = \frac{1+1+2+2+3+4+5+5+6}{9} = \frac{32}{9}$$

$$avg = 3.2$$

② Median

② Median

ex $\{5, 4, 2, 3, 2, 1\}$

Step 1 -> Sort the random variable

$x = \{1, 2, 2, 3, 4, 5\}$

Step 2 No. of elements count = 6

if count = even

= $\{1, 2, \boxed{2, 3}, 4, 5\}$

median = $\frac{2+3}{2} = 2.5$

if count = odd

=> $\{1, 2, \boxed{3}, 4, 5\}$

median = 3

Q Why median

Q = Why median

$x = \{1, 2, 3, 4, 5\}$

$avg(\bar{x}) = \frac{1+2+3+4+5}{5} = \frac{15}{5}$

$avg(\bar{x}) = 3$

outlier

$x = \{1, 2, 3, 4, 5, 100\}$

$\bar{x} = \frac{1+2+3+4+5+100}{6}$

$\bar{x} = \frac{115}{6} \approx 19$

$x_{median} = 1, 2, 3, 4, 5, 100$

$x_{median} = \frac{3+4}{2} = \frac{7}{2} = 3.5$

Note: Median is used to find Central ... outliers

Note Median is used to find tendency of data when outliers are present

(3) Mode { Frequency man }

eg { 2, 1, 1, 1, 1, 4, 5, 7, 8, 9, 9, 10 }

Mode = 1

eg

Age	weight	Salary	Gender	Degree
24	70	10K	M	—
25	mean —	20K	F	—
				—
27	mean —	median	M	
30	60	median	M	PHD
mean — m	0	—	= {	BA
— m		1.2L		
— m	53		M	B. Tech
40	70	3L		

mode