Statistics

Defn: Statistics is the science of collecting, organizing and analyzing data.

⇓

Decision Making

Data: "facts or pieces of information" ⇒ Measured, Collected, Analyzed

Eg: Weights of students in the class
{60kg, 50kg, 45kg, 30kg, ...}

IQ of the students of in class
{100, 90, 95, 99, ...}.

House Price Dataset

City	Area	No.of Rooms	Price
Bangalore	1000	2	45 Lakhs
New York	1250	2.5	50 lakhs
Mumbai	-	-	-

} Analyzing this Data

Data Scientist → Model ↓↑ Price

Data Analyst → Report → Visualization → Meaning Decisions.
↳ Project →

Application

1. Data Exploration And Summarize
2. Model Building And Validation
3. Statistical Analysis → Sample Data → Population Data.
4. Hypothesis Testing
5. Optimization And Efficiency.
6. Reporting

Types Of Statistics

① Descriptive Statistics

Descriptive statistics involves methods for summarizing and organizing data to make it understandable. This type of statistics helps to describe the basic features of the data in a study.

1) Measure of Central Tendency
   [Mean, Median, mode]

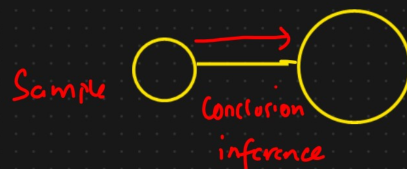
2) Measure of Dispersion [Variance, Standard deviation]

3) Data Distribution
   i) Histograms
   ii) Box plot
   iii) Pie chart
   iv) PDF, PMF

4) Summary Statistics
   i) Five Number Summary
   Q1, Q2, Q3, Maximum

② Inferential Statistics

Inferential statistics involves methods for making predictions or inferences about a population based on a sample of data. It allows for hypothesis testing, estimation, and drawing conclusions.



① Hypothesis Testing
② P value
③ Confidence Interval
④ Statistical Analysis Test
   ① Z test
   ② t test
   ③ ANOVA → F test
   ④ Chi Square.

Summary

Type of Statistics	Key Concepts	Examples
Descriptive Statistics	Measures of Central Tendency (Mean, Median, Mode), Measures of Dispersion (Range, Variance, Standard Deviation), Data Distribution (Histograms, Box Plots), Summary Statistics (Five-number Summary)	Mean score of students, Range of temperatures, Histogram of ages
Inferential Statistics	Hypothesis Testing (Null and Alternative Hypotheses, P-value), Confidence Intervals, Regression Analysis (Simple and Multiple Linear Regression), ANOVA, Chi-Square Test	P-value in test scores comparison, 95% confidence interval for average height, Predicting house prices, Comparing test scores of different schools, Association between gender and product preference

Eg: Let say there are 20 statistics class in your college, and you have collected the height of students in the class.

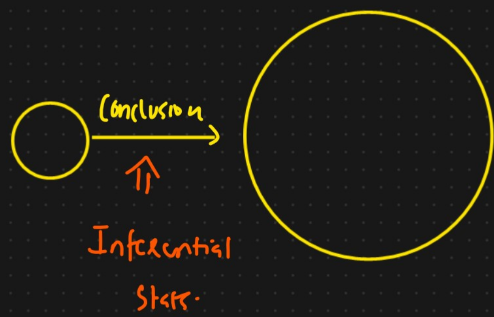
Heights are recorded [175cm, 180cm, 140cm, 135, 160cm, 120cm]

Descriptive Question

"What is the average height of the entire classroom" ← Measure of Central Tendency

Inferential Question

"Are the height of the sample students in classroom similar to what you expect in the entire College"?



③ Population And Sample Data

Population

Definition: A population is the entire set of individuals or objects of interest in a particular study. It includes all members of a defined group that we are studying or collecting information on.

Characteristics

1) Complete Set : Contains all the observation of interest

2) Parameter : A numerical value summari-zing the entire population

   1) Population mean (μ)
   2) Population variance (σ^2)

Examples

1) Population in a School Study

Sample Data

Definition: A sample is a subset of the population that is used to represent the entire group. Sampling involves selecting a group of individuals or observations from the population to draw conclusions about the whole population.

Characteristics

1) Subset : Represents a portion of the population.

2) Statistic : A numerical value summarizing the sample data [Sample mean, Sample variance].

3) Random Sampling : Samples should be randomly selected to avoid bias

Examples

1) Sample In a School Study

* All Students enrolled in a School
* Determine the avg height of Student, population mean

i) A group of 50 Students from School
ii) Usecase: Estimate the average height of Students in a School

2) Population in Market Research

* All consumers in a city.
* To understand the purchasing behaviour of all Consumers,

2) Sample in Market Research

* 500 consumers from the city
⇓
* Behaviour ⟶ Population.

3) Population in a Medical Study

* All the patients with a Specific disease
* To study the effectiveness of a drug.

⟹ 3) Sample in a Medical Study

* 200 patients
* Test the effectiveness of the drug

④ Types of Sampling Techniques

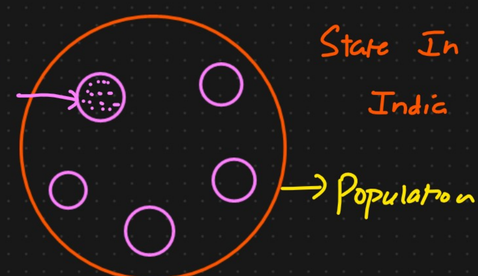
1. Probability Sampling
2. Non Probability Sampling

Eg: Exit Poll

A, B, C, D
↑ ↑ ↑ ↑
7 million

Sample = 10,000

Age ≥ 18



1. Probability Sampling:

a) Simple Random Sampling:

Every member of the population has an equal chance of being selected.

Eg: Selecting people randomly
Draw names random from a class of students.

b) Systematic Sampling

Select every n^{th} member of the population after a random starting point.

Eg: Airport ⟶ Credit Card ⟶ 5^{th} person, 10^{th} person, 15^{th} person

Feedback Survey → Selected every 11^{th} member → Feedback Survey'

c) Stratified Sampling :

Divide the population into Strata (groups) based on specific characteristics and then randomly sampling from each Strata.

Eg: Divide employees by department and then randomly select a proportional number from each department to form a survey sample.

Eg: Age → <12 12 − 18 >18 {Politics}

d) Cluster Sampling

Divide the population into clusters, randomly selecting clusters, then sampling all the members from the selected clusters

Eg: Randomly selecting several schools from a district, and surveying all teachers within those schools.

e) Multi Stage Sampling

Combining several sampling methods. Usually involves selecting clusters, then randomly sampling within those cluster

Eg: Randomly selecting cities, each selected city randomly selecting households to survey.

② Non Probability Sampling

Select individuals who are easiest to reach.

Eg: Surveying people at mall.

ⓐ Convenience Sampling

Selecting individual who are easiest to reach

ⓑ Judgmental (Purposive) Sampling

Select individual based on the researcher's judgement → Useful or Representation

Eg: Choose experts in a field to participate {Data Science}

(c) Snowball Sampling

Existing study subjects recruit future subjects from among their acquaintances.

Eg: Survey members of a rare disease.

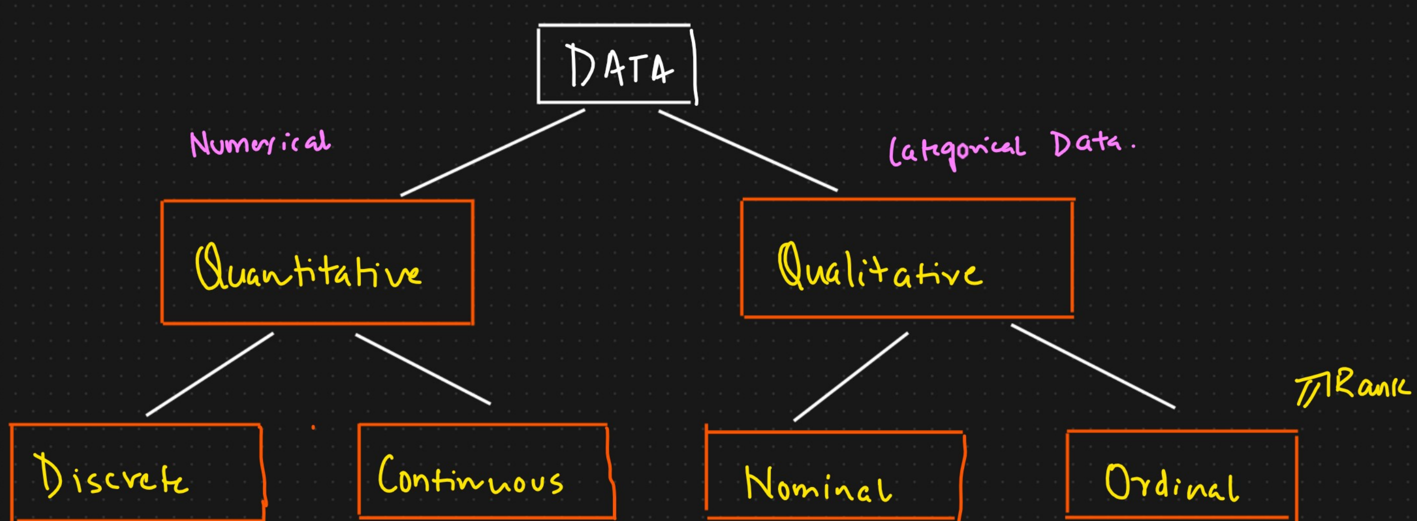
(d) Quota Sampling

Ages, group, genders, caste.

Selecting the sampling technique depends on Usecase

(5) Types Of Data

DMC	DC	ISI	BUI	FWI	Classes	Region
3.4	7.6	1.3	3.4	0.5	not fire	0
4.1	7.6	1	3.9	0.4	not fire	0
2.5	7.1	0.3	2.7	0.1	not fire	0
1.3	6.9	0	1.7	0	not fire	0
3	14.2	1.2	3.9	0.5	not fire	0
5.8	22.2	3.1	7	2.5	fire	0
9.9	30.5	6.4	10.9	7.2	fire	0
12.1	38.3	5.6	13.5	7.1	fire	0



⇓	⇓	⇓ No Rank	
Whole numbers	Any value	Eg: Gender	Eg: Customer Feedback
Eg: No. of bank accounts	Eg: Weights	M, F	Good, Bad, Better
No. of children in a family	Height	Blood Group	1 3 2
	Temperature	Pincode	
	Speed		

⑥ Scales of Measurement of Data

The scales of measurement describe the nature of information within the values assigned to variables.

4 primary scales of Measurement

① Nominal Scale
② Ordinal Scale
③ Interval
④ Ratio

① Nominal Scale

Defn: This scale classifies data into distinct categories that do not have an intrinsic order

Qualitative / Categorical data.

Characteristics

i) Data is categorized based on lables, names or qualities
ii) Categories are multually exclusive
iii) No logical order among categories [No Rank]

Examples

Gender
→ M
→ F

Colors
Red → 5 50%
Blue → 4 40%
Pink → 1 10%
$\overline{10}$

Types of Cuisines
{ Italian
Chinese
Mexican }.

② Ordinal Scale

Defn: This scale classifies the data into categories that can be ranked or ordered

Characteristics

i) Data is categorized and Ranked In a Specific Order
ii) The interval between ranks are not necessarily equal

Example

Education level	⇒ F.E ⇒	Ranks
High School		1
Bachelor		2
Masters		3
Doctorate		4

Customer Feedback	⇒
Satisfied	1
Very satisfied	2
Not Satisfied	0

Socio Economic Status

Low	2
Middle	1
High	0

③ Interval Scale

Defn: The interval scale not only categorizes and orders but also specify the exact difference between intervals. It lacks a true zero point.

Characteristics

1) Data is ordered with consistent interval between values.
2) Allows for meaningful comparison of differences [Ratio cannot be measured].
3) No true zero point.

Example

Temprature in Fahrenheit:

→ $10^{0}F$, $20^{0}F$, $30^{0}F$

$20-10=10$

$30-20=10$

⇒ $0^{0}F$ "No Temprature"

$\frac{30}{10} = \boxed{3:1}$ ←

Calendar years

2024, 2020, 2016 ⇒ 0 year

IQ Scores

90, 100, 110

↳ Difference $100-90=10$

$IQ \neq 0$

4) Ratio Scale

- The order matters
- Differences Are measurable [Ratio can be meaning]
- Contains a 0 Starting point.

Eg: Student marks in a class

0, 90, 60, 30, 75, 45

Asc = 0, 30, 45, 60, 75, 90 ←

$45-30=15$

$60-30=30$

Ratio $= \frac{90}{30} = \frac{3}{1}$

$\boxed{3:1}$

Eg: Weights

→ 10, 20, 30, 40

Income: 10,000, ←

20000,

30000,

40000

Assignment

1. Length of Different Rivers In the World?
2. Favorite food based on Gender?
3. Marital Status?
4. IQ Measurement?