Getting Started with Statistics (DAY-6)

L, Statistics is the Science of Collecting, Organizing and Analyzing data
↓
Decision Making

Data → "facts or pieces of info"
(measured, Calculated, Analyzed)

Eg:- "~~facts or~~,

① weights of students in a class

{60, 50, 45, 30 ---}

② IQ of the students in a clas

{100, 90, 95, 80 --- }

Data set :- House price Dataset

City	Area	No. of rooms	Price
Bangalore	1000	2	45Lakhs
New york	1250	25	50 lakhs
mumbai	–	–	–

} Analyze Date
↓
Data Scient → Mo
↓
Price

Data Analyst → Report → Visualization → meaning Decision
L, project →

Application

① Data exploration & Summarization

② Model building & validation

③ statistical Analysis → Sample data → population data.

④ Hypothensis testing

⑤ optimization & Efficiency.

⑥ Reporting.

Types of Statistics

① Descriptive Statistics

↳ It 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 ~~Ten~~ 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

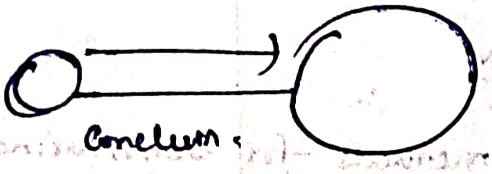
④ Summary statistics

① Five Number Summary

Q_1, Q_2, Q_3, Map Value.

② Inferential Statistics.

↳ It involves methods for making predictions or inferences about a population based on a sample of data. It allows for hypothesis testing, estimation & drawing Conclusions.

Sample data 

① Hypothesis testing

② P Value

③ Confidence Interval

④ statistical Analysis test

① Z-test

② t-test

③ F-test (ANOVA)

④ chi Square test

Eg:- Let Say there are 20 statistics class in your college,
and you have Collected the height of students in the Class.

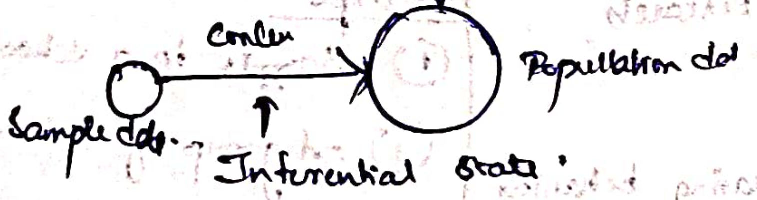
Heights are recorded {175cm, 180cm, 140cm, 135cm, 165cm, 120cm

Descriptive statistics

" what is the Avg height of the entire class"
↑ measure of Central tendency.

Inferential Statistics!

" Are the height of sample student in class similar to what you expect in the entire college"?



* population and sample data →

* It is the entire set of objects of interest in a particular study.

* A Sample is a subset of the population that is used to represent the entire group.

Population

haracteristics!

1) Complete set

2) Parameter
↓
A Numerical value summarize the entire the populati

[Eg:- population mean, population variance]
(μ) (σ^2)

nt- 1) population in a school study

- All students enrolled in a school)

~~tot~~ Avg height of studen
↑
Population mean

2) population in market Research

All Consumer in a City

To understand the purchasing behavior of all consumer!

3) population in a medical study

* All the patient with a specific disease

* To study the effectiveness of a drug.

Sample data

Characteristic

(1) Subset of population

(2) Statistic

A numerical value summarize the Sample data.

(i) Sample mean

(2) Sample variance

(3) Random Sampling!

Samples should be randomly selected, to avoid bias.

example

(1) Sample in a school study

(i) A group of 50 student from scho

(ii) estimate the Avg height of student in a school.

(2) Sample in market research

(a) 500 Consumer from the city

usea - Behaviour → population

(3) sample medical study

200 patients

use ca Test the effectiveness of the drug

Types of sample Techniques :-

① ~~po'~~ probability sampling

ⓐ simple Random sampling

↓

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

Eg:- selecting people randomly.

Draw names random from a class of student.

ⓑ ~~Sysmatic~~ Systamatic sampling :-

↳ select every nth member of the population after a random starting point.

Eg:- Airport → Credit card → 5th person, 10th person, 15th person,

Feedback survey.

ⓒ Stratified sampling :-

Divide the population into strata (groups) based on the specific characteristic & then randomly sampling from each Group.

Eg:- Divide employees by department in a Company & select a proportional number from dept to from a survey sample, Eg:- Age —— <12 12-18 >18 {Poultry}

↑ ↑ ↑

④ **Cluster sampling**

↳ Divide the population into cluster, randomly selecting clusters, then sampling all the member from the selected clusters.

Eg:- randomly selecting several schools from a district, and surveying all trade within those schools.

⑤ **Multi stage sampling** :-

Combining several sampling method, usually involve selecting cluster, then randomly sampling within the cluster

Eg:- randomly selecting cities, each selected city randomly selecting household to survey.

(1) Non - probability Sampling :

↳ select individual who are easiest to reach

Eg:- Surveying people in the mall

(a) Convenience sampling :-

↳ selecting individual who are easiest to reach

(b) Judgemental Sampling

↳ select individual ~~who~~ are based on the researcher's Judgement

↑ useful for representation.

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

(c) Snowball Sampling :-

existing study subjects future subjects from among their acquirements.

(5) Types of Data

(i) Quantitative		(2) Qualitative	
Discrete	Continuous	Nominal	Ordinal
↓ (1) Whole Numbers	→ Any value	Eg: Gender (M, F)	Eg: Consumer feedb
Eg: No. of bank accounts	Eg: weight, height, temperature, speed.	bloodgroup.	Good, bad, best
No. of children in a family		Pincode	↳ Contains rank
		↳ Cathegorical value.	
		↳ No rank	

Scales of Measurement of Data

↳ Describes the nature of info within the value assigned to ~~measurement~~ Variable.

(1) Nominal Scale

↳ This scale classifies the Data into Distinct Categori that do not have an intrinsic order.

Qualitative / Categorical data.

Characteristics

(i) Cathegorized based on labels, names or qualities.

(ii) Categorized are Mutually Exclusive.

(iii) No logical order among category [No rank]

Ex:- Gender · Colour

— M

— F

red	— 5	50%
Blue	— 4	40%
Pink	— 1	10%
	10	

* Ex:- Types of cuisines
{ Italian
chinese
mexica }

(2) Ordinal Scale :-

↳ Classifies the data that can be ranked or ordered.

Characteristics

(1) Have logical order among category [rank]

(2) Interval b/w rank are not necessarily equal.

Ex: Educational level = F F — 1 rank

High school	1
Bachelor	2
masters	3
Doctorate	4

Ex: Customer feedback
Satisfied — 1, Very satisfied 2, unsatisfied 0

③ Interval Scale :

↳ The interval scale not only categories and order but also specify the exact distance, b/w intervals.

→ It. lacks a true zero point.

Characteristics:

① ordered with consistent interval b/w vale.

② Allows for meaningful Comparision of difference.

③ No true zero point.

Eg:- Temp in Farenheit

10°F , 20°F , 30°F

⇒ 0°F → zero temp

IQ score :-

90, 100, 110

↳ Diff 100 − 90 = 10

IQ ≠ 10

Calendar year

2024, 2020, 2016 ⇒ 0 yrs.

(4) <u>Ratio Scale</u> :

* Order matters

* Differences are measurable (ratio can be measured,

* Contains a 0 starting point

Eg: marks of a the students in a class.