Lecture 1

Statistics

For ML, DL, NLP, Vision, Data Analyst

Definition

It is the branch of science which involves collecting, analysing, data in large quantities so that we can come up with various use cases, exploring and visualising and coming out with meaning information and conclusions

Topics which we will see:

Descriptive statistics
Inferential statistics
Population
Sample
Sampling techniques
Measure of central tendency
Measure of dispersion or Variance
Probability
Permutation and combination

Statistic

Descriptive

1 Analyse, Explore
Visualizing → Techniques
⇓

To understand the
data

Eg; Histograms, bar,
pie, scatterplot

Inferential

1 Population of
data
⇓
Sample of data
⇓
Experiment
⇓

Conclusion
Find Out

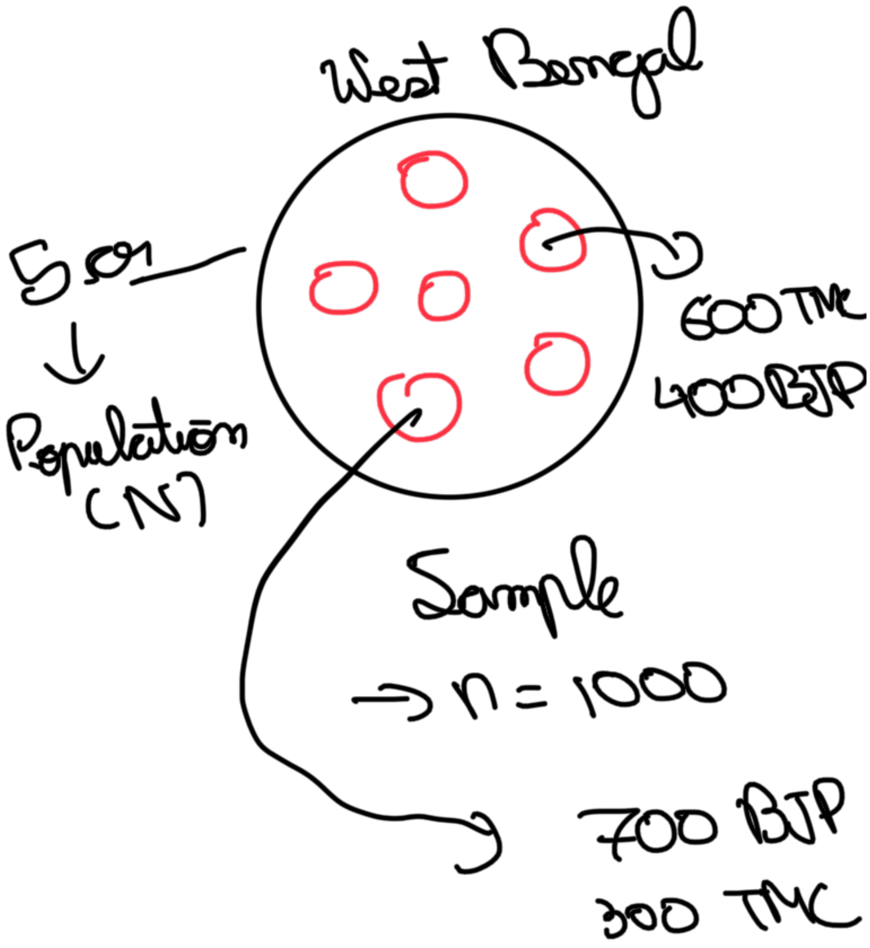
Eg; Hypothesis
Testing

P - Test,
T - Test,
F - Test
Z - Test

Null Hypothesis
Alternative

Decision Making

Population (N)



From these samples we can say

{ Max seats will go to TMC, and second highest will be BJP }

Different sampling methods

1. ***Random sampling-*** randomly getting selected

Disadvantage :-

a) overlapping- some cases are being repeated

b) For specific use case -

2. ***Stratified sampling***-population is divided into Stratus(male and female)

Male for exit pole

Household for females

So we are dividing them into subgroups.

3. ***Systematic sampling-*** doing sampling systematically, that is if we take the second person then 4th then 6th then 8th. We are following a system

It might lead to bias which means more than 1 type in a sample

4. ***Clustering sample-*** Clusters of groups

Every group will give different answers
We will take groups instead of individuals

Targeting different clusters of customers(based on expenses we will target different sets of rich and poor customers)

Measure of Central Tendency

① Mean ② Median ③ Mode

↓

$$\bar{x} = \frac{\Sigma n}{n}$$

When we have outliers then we should never take mean

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

mean = 3

If $\{1, 2, 3, 4, 5, 100\}$

mean = 19.6

100 has become outlier

Q If we should not use mean what should we use?

⇒ We can use median.

$\{1, 2, \boxed{3, 4}, 5, 100\}$

Median = 3.5

$\{1, 3, 100\}$

Median = 3

Median does not get impacted with outliers ⊛

In median we need to always sort the elements. ⊛

---

③ Mode - Most frequently occuring element

334 → Mode 3

1333444 → 2 modes

$\boxed{\text{3 and 4}}$

In latest python the first by default which is present more times that it is the mode.

***

Random variable - whose value depends on the outcome of a random phenomena

Tossing a coin :

⇒ we can get $\{H, T\}$ which is a random phenomena

a) Categorical Variables
Qualitative Variables

b) Continuous Variable
Quantitative Variables

a) ⇒ Gender	Weekdays
↳ M	Sun
↳ F	Mon
	⋮
	Sat

---

a) → Nominal → Rank not important
eg F and M, etc

a) → Ordinal } → Ranking needs to be considered

For IT Professional

Sunday , Sat , Fri , ...
7 1 2

Ranking

Customers Ratings

5, 4, 3, 2, 1

---

b) Continuous Variable

eg; Height = {170.3, 171.4, 180.
25}

→ Discrete Quantative Variables

→ Continuous Quantitative Variable

Temperature	– Continuous Quantitative Variable
Pincode	– Nominal Catogorical
Age	– Descriptive Quantative

<u>Variables</u>

1) Quantative Variables
2) Qualatative Variables
3) Random Variables

When there is NaN value we cannot replace it with mean or median coz we need numerical value in that case so we use mode

For mean → Remove Outliers
→ Or else use median

① Independent Samples
② Dependent Samples

Σ Temp	Rainfall	CO_2	NO_2	Humidity

Independent features

With this we calculate

Output AQI

Dependent features