Statistics

for ML, DL, NLP, Vision, Data Analyst

Definition

Statistic

Descriptive

① Analyze, Explore Visualizing → Techniques

↓

To understand the data

Eg; Histograms, bar, pie, scatterplot

Inferential

① Population of data

↓

Sample of data

↓

Experiment

↓

Conclusion and Out

Eg; Hypothesis Testing

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

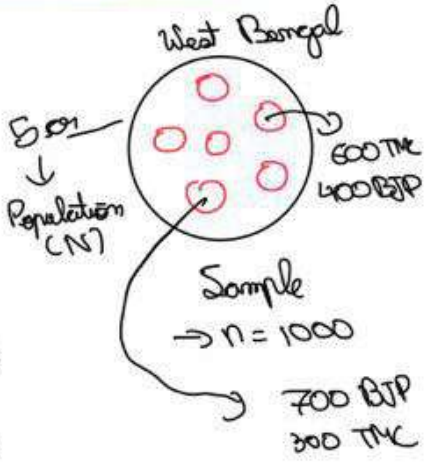
Null Hypothesis
Alternative Hypothesis

Decision Making

Population (N)

From these samples we can say

∑ Most seats will go to TMC, and second highest will be BJP ∑



Measure of Central Tendency

① Mean ② Median ③ Mode

↓

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

{1, 2, 3, 4, 5}
mean = 3

If {1, 2, 3, 4, 5, 100}
mean = 19.6

100 has become outlier

When we have outliers then we should never take mean

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

⇒ We can use median.

{1, 2, [3, 4], 5, 100}
Median = 3.5

{1, 3, 100}
Median = 3

Median does not get impacted with outliers (*)

On median we need to always sort the elements. (*)

③ Mode – Most frequently occurring element

334 → Mode 3

1333444 → 2 modes
[3 and 4]

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

Tossing a coin :

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

a) Categorical Variables
Qualitative Variables

b) Continuous Variables
Quantitative Variables

a) ⇒ Gender | Weekdays
- M
- F
- Sun
- Mon
- ⋮
- Sat

---

a) → Nominal → Rank not important eg F and M, etc
→ Ordinal ⇒ Ranking needs to be considered

Ex: IT Professional

Sunday, Sat, Tue, ...
2, 1, 3
Ranking

Customer Ratings

5, 4, 3, 2, 1

---

b) Continuous Variable

eg; Height = {170.3, 171.4, 180.25 ;

→ Discrete Quantitative Variables
→ Continuous Quantitative Variables

Temperature – Continuous Quantitative Variable

Periods – Nominal Categorical

Age – Descriptive Quantitative

Variables

1) Quantative Variables
2) Qualitative Variables
3) Random Variables

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

---

Ex: mean → Remove Outliers
→ Or else use median

---

1. Independent Samples
2. Dependent Samples

Temp	Rainfall	CO_2	NO_2	Humidity

Independent features

With this we calculate AQI

Output → Dependent feature