DESCRIPTIVE STATISTICS

STATISTICS

Science of collecting,organizing,analyzing, interpreting and presenting data.

TYPES OF STATISTICS

DESCRIPTIVE

Helps us simplify and organise large amounts of data. This makes large amounts of data easier to understand.

INFERENTIAL

It uses smaller data to conclude a larger group. It helps us predict and draw conclusions about a population.

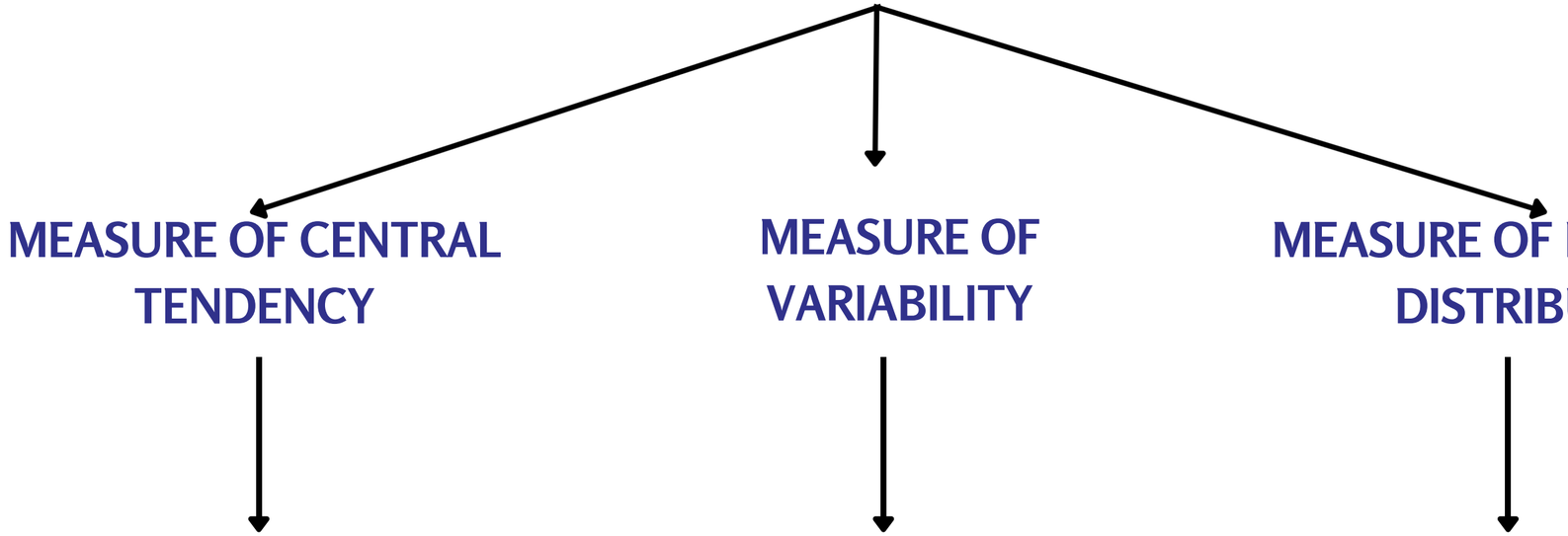
DESCRIPTIVE STATISTICS

- Central Tendency
- Dispersion
- Distribution Shape
- Visuals

INFERENTIAL STATISTICS

- Confidence Intervals
- Hypothesis Testing
- Regression Analysis
- P- Values & Z- score

DESCRIPTIVE STATISTICS



MEASURE OF CENTRAL TENDENCY

- Mean
- Median
- Mode

MEASURE OF VARIABILITY

- Range
- Variance
- Standard Deviation

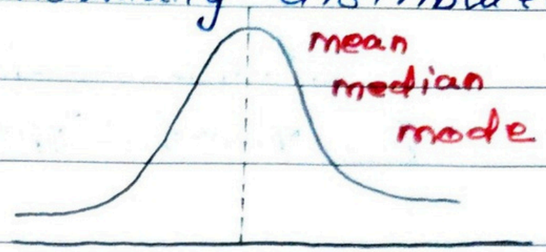
MEASURE OF FREQUENCY DISTRIBUTION

- Histogram
- Box Plot
- Pie Chart

MEASURE OF CENTRAL TENDENCY

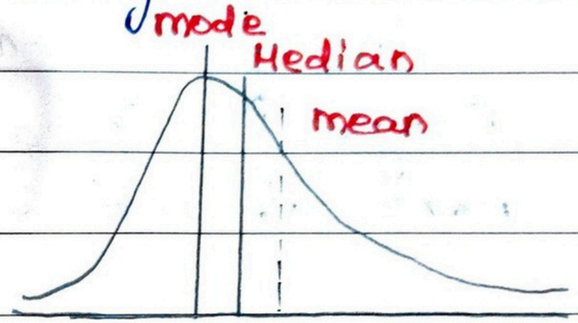
Mean ≈ Median ≈ Mode ⇒ normally distributed

(Heights of adults, IQ Scores)



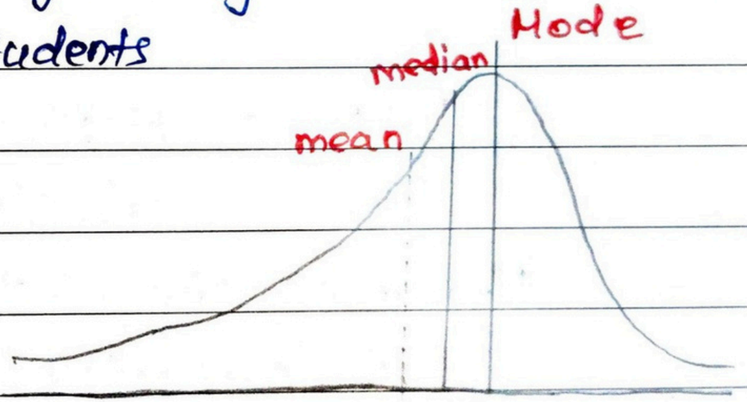
Mean > Median ⇒ Positively Skewed

(Income distribution -
a few rich people pull up
the mean)



Mean < Median ⇒ negatively Skewed

(Test Scores if most students
score high, but a few
score very low)



We can detect Outliers = Values that are very different from the rest.

Eg,[50, 51,52, 53,1000]

Mean = 241, Median = 52 → huge difference b\w mean and median.
In ML:

- Pre-processing data
- Detecting outliers feature
- scaling/normalisation
- understanding class balances

MEASURE OF VARIABILITY

Range:

Range=Largestvalue - smallestvalue

2. VARIANCE : defined as an average Squared deviation from the mean.
Measures the average squared distance of each point from the mean.

eg: Scores = [2, 4, 6, 8, 10]

$$\text{mean} = \frac{2+4+6+8+10}{5} = 6$$

Data	deviation (x − mean)	Squared
2	2 − 6 = −4	16
4	4 − 6 = −2	4
6	6 − 6 = 0	0
8	8 − 6 = 2	4
10	10 − 6 = 4	16
		40

$$\text{Var} = \frac{40}{5} = 8$$

Pop

$$\sigma = \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2$$

Sample

$$\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

--> Variance

Standard Deviation:

Tells how much the data values deviate from the mean, on average

→ Square root of variance.

→ same units as the original data.

Measures of Frequency distribution:

It is a way to organise and summarise data to show that, often, each value (or range of values) appears.
It is often presented in a table format or as a chart.

1. Histogram
2. Pie chart
3. cumulative line chart

THANK YOU