Page No.
Date

Session-38

Descriptive Statistics - 1

- Statistics Definition ✓

Statistics

↓ Descriptive
↳ Summarize the data
→ mean, median etc

↓ Inferential
↳ Drawing conclusion and making predictions
→ predicting election result

- Population V/S Sample ✓

- Inferential Stats -
  - Hypothesis testing
  - Confidence interval
  - ANOVA
  - Regression analysis
  - Chi-Square tests
  - Sampling Techniques

- Types of Data

↓ categorical / qualitative
↓ nominal
↓ ordinal

↓ numerical / quantitative
↓ Discrete
↓ continuous

- Central tendency -
  - Single value that represent center of data

mean | median | mode

1) mean = $\frac{\text{Sum of All values}}{\#\text{ values}}$

Populat^n mean = $\mu = \frac{\sum_{i=1}^{N} x_i}{N}$

Sample mean = $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$

$N \rightarrow$ No. of items in populat^n
$n \rightarrow$ " " " " sample

2) ~~Median~~ · If outlier, then we may get present wrong mean.

2) median -
  - middle value in the dataset when arranged in order

- 5, 1, 2, 3, 8 ⇒ 1, 2, [3], 5, 8

- 5, 7, 8, 9, 2, 4 ⇒ 2, 4, (5, 6), 7, 8

$\frac{5+5}{2} = 5.5$

Page No.
Date

3) Mode - most frequent value in Dataset

1, 1, 5, 7, 2, 3, 2, 5 ∴ mode = 1

4) weighted mean -

	1	2	3
weight	0.2	0.3	0.8

$$\text{weighted mean} = \frac{1\times0.2 + 2\times0.3 + 3\times0.8}{0.2 + 0.3 + 0.8}$$

5) Trimmed mean

1, 2, 3 | 8, 10, 12 | 20, 22

Trim it out — calculate it's mean — trim it out

- measure of Dispersion -
  - How much spread data is

1) Range → (min_value, max_value)
   - It is affected by outlier

2) Variance -

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

- Variance is proportional to spread.

Page No.
Date

Can't be used in inferential step.

- Mean Absolute Deviation

$$= \frac{\sum |x_i - \bar{x}|}{n}$$

↳ less prone to outlier

- Variance is highly prone to outlier.
- Population variance $\sigma^2 = \frac{\sum (x - \mu)^2}{N}$
- Sample variance $s^2 = \frac{\sum (x - \bar{x})^2}{n-1}$

3) Standard deviation -

$\sqrt{\text{variance}}$

- Its unit is same as Data.

4) Coefficient of variation (CV)

$$CV = \left(\frac{\sigma}{M}\right) * 100$$

- Salary and age can't be compared. By calculating CV - we can tell which of Salary & age are more closer to there mean.

- Graphs for univariate Analysis
  - ↳ Depends whether data is categorical or numerical

1) Categorical column –

i) Frequency distribution table

category	frequency

=> using this we can build Bar graph

ii) Relative frequency

↳ we ~~get~~ get % of frequency

↳ we can build pie chart

iii) cumulative frequency –

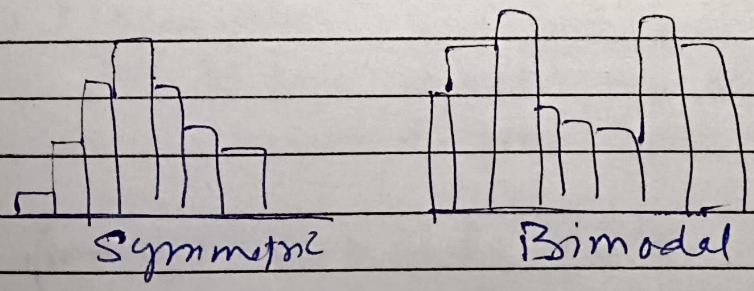
A	fq	c.freq
x	20	20
y	30	50
z	40	90

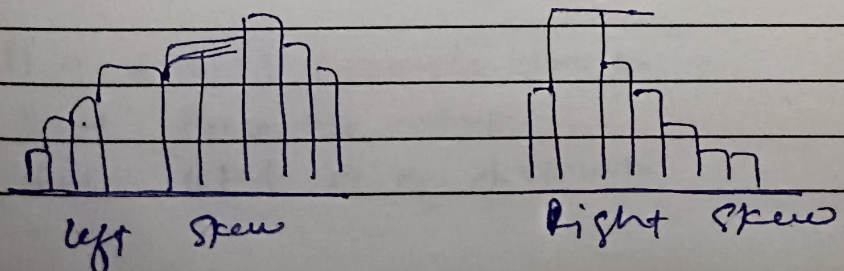
2) Numerical column

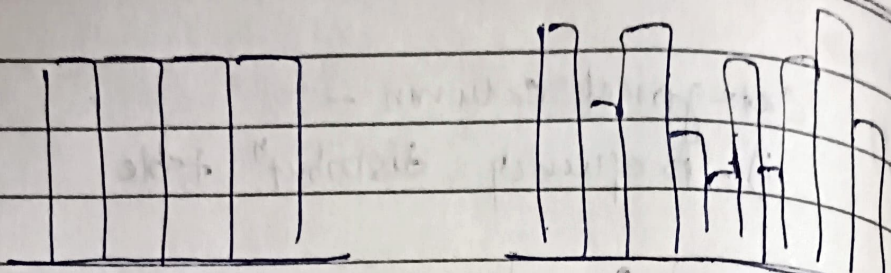
↳ frequency Distribution ✓

↳ Histogram ✓

↳ we use bins







uniform

No pattern.

- Graphs for Bivariate analysis -
  - ↳ cat - cat
  - ↳ cat - Numerical
  - ↳ Num - Num

1) Categorical - categorical
   - ↳ we make contengency / crosstab table

2) Numerical - Numerical
   - ↳ Scatter plot

3) categorical - Numerical
   - ↳ a lot of things can be done
   - ↳ can make contengency.