STATISTICS - I

Statistics is the art of learning from data. It is concerned with collection of data, their subsequent description and analysis which leads to drawing of conclusions.

1. Major Branches.

(i) Descriptive Statistics.

Part of statistics concerned with description and summarization of data.

(ii) Inferential Statistics.

Drawing of conclusions from data.

Note!- If analysis is completely done on given data only, then it is descriptive.
but if we draw conclusion, then it is inferential.

2. Data

Facts and figures collected, analyzed and summerized for presentation and interpretation.

(ii) Unstructured and Structured Data.

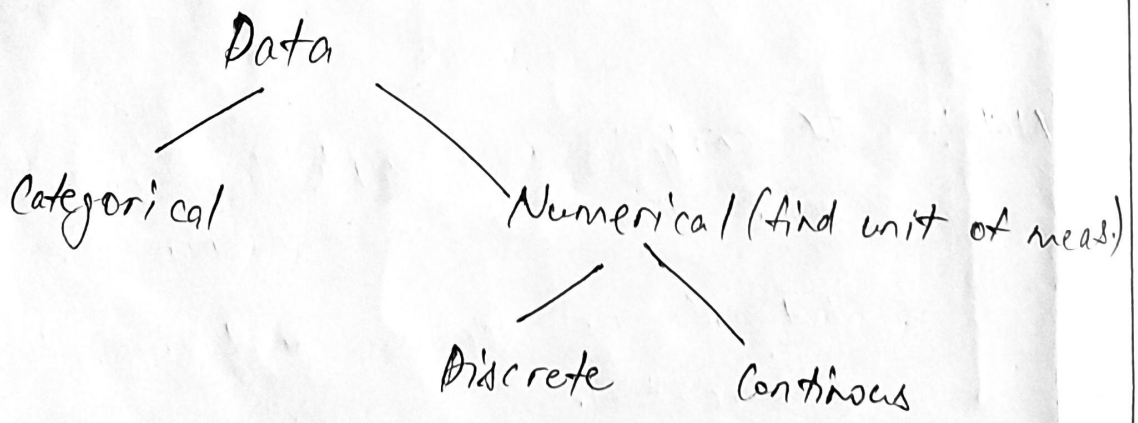
1 Maggi, Cold dr. - - - - - Unstr.

| 1 | Maggi | 30+ASt - - - St.

There is a diff. b/w 0 & ----

3. Categorical and Numerical Data

Qualitative ~~Data~~ Var.	Quantitative Var.
Identify group membership.	Desc. Num. prop. of cases.
	Have meas. units.



4. Cross Sectional And Time-Series Data.

Time Series Data - Data recorded over time

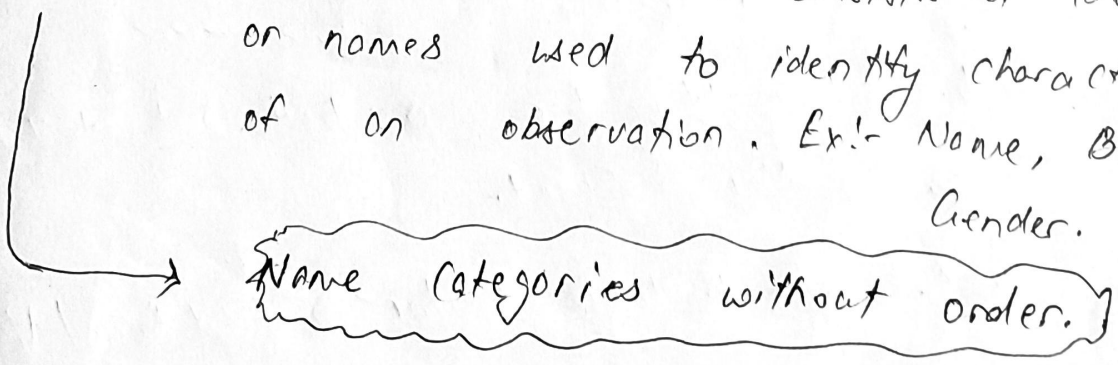
Time plot - Graph of time series showing values in chronological order.

Cross-Section - Data observed at some time

5. Scales of Measurement

(i) Data collection requires one of the following scales of measurement: nominal, ordinal, interval, ratio.

Nominal :- Data for a variable consists of labels or names used to identify characteristics of an observation. Ex:- Name, Board, Gender.



Ordinal:- Prop. of Nominal Data + Order or Rank in meaningful way, but no fixed interval.

Ex:- Army Pos.

Interval:- Ordinal + Fixed Interval.

Ex:- No. System. (1.2.3.4.5...)

a. Always Numeric.

b. Ratio of values have no meaning.

Ratio:- Interval + Ratio is meaningful.

a. True Zero Exists.

6. Categorical Data (Detailed)

(i) Frequency Dist. (Google Sheet)

a. Select the cells having data you want to vizualize

b. Formatting Bar → Data. → Pivot Table → New Sheet.

c. Pivot Table Editor → Add Rows/Values.

(ii) Relative Frequency.

Ratio of frequency to total no. of observations.

Total Sum should be '1'.

iii) Graphs.

a. To show proportions pie chart is good.

b. In pareto chart categories are sorted by frequency.

c. If ordinal, order of categories is preserved.

d. Bar Graphs are used to compare things b/w diff. groups

e. Follow Area Principle.

f. Round offs create errors.

7. Median and Mode.

Descriptive measures that indicate where the centre or most typical value of data set lies is measure of central tendency.

ii) Mode

Mode of categorical variable is most common category, the category with highest freq.

(a) Long Bar, Widest slice,

(b) Binodal and Multinodal do exist.

(ii) Median (Ordinal)

Median is category of middle observation of sorted values.

8. Numerical Data.

(i) Organising Num. Data.

Discrete Variable involves a count of someth.

Continous Variable involves measurement of something.

a. First group observations into classes (categories) and then treat classes as distinct value of qualitative data.

(ii) Organising Discrete Values (single).

a. Represent the value as a category and then create freq. table.

b. This is for small data.

(iii) Organising Continous Data.

a. No. of class - Suggested are 5 to 20.

b. Each obs. should belong to a class and no obs. should belong to mult. class.

c. Suggested to start with equal class interval.

Lower Class Limit :- Smallest value that could go in class
Upper Class Limit :- Largest value for a class
Class Width :- Diff b/w lower limit of class and lower limit of next class
Class Mark :- Average of upper and lower limit.

d. Class interval contains it's left-end but not it's right end.

(iv) Constructing Histogram (Continous).

Don't leave space b/w histogram bars.

(v) Stem and Leaf Diagram

Stem ↓ 7 5 ↑ Leaf

for 75, 78

Stem	Leaf
7	5, 8

for multiple values always sort stem in ascending order and remove comma.

(vi) Descriptive Measures

a. Measure of Central Tendency.
b. Measure of Dispersion.

a. 1. mean (average)

for discrete values, $\bar{x} = \frac{\sum f_i x_i}{\sum f_i}$

for continous Data,

Class Interval.	Frequency (f_i)	Mid Point (m_i)	$f_i m_i$	$\bar{x} = \frac{\Sigma f_i m_i}{n}$

Above value will only be approximation.

For exact, use discrete method.

a.1.a. Adding const. to all will increase mean to mean + const.

a.1.b. Multiplying const to all will convert mean to mean * const.

a.2. Compute Median

For ordered data, median $= \frac{n+1}{2}$ (odd)

$$\text{median} = \frac{\left(\left(\frac{n}{2}\right) + \left(\frac{n}{2}+1\right)\right)}{2}$$

a.3. Mode. (Highest Frequency)

b.1. Measure of Dispersion.

It shows amount of variation, or speed or dispersion.

b.1.a. Range is diff b/w highest & lowest value.

Range is sensitive to outliers.

b.1.b Variance takes into account all values.

Consider deviation of data values from central tendency.

Pop. variance $= \sigma^2 = \left(\frac{(x_i - \bar{x})^2 + \cdots}{N}\right)$ where, $\bar{x}$ = mean.

Sample Variance $= s^2 = \frac{(x_i - \bar{x})^2}{n-1} \cdots$

b.1.c. Variance + const = Variance

Variance * c = Variance * c^2

b.1.d. Standard Deviation = $\sqrt{\text{Deviation}}$

It is done to restore original unit which has been doubled.

Same as b.1.c

(vi) Percentile.

Value below which a percent of data falls.

a. Rank = Percentile × (n-1) + 1

Ex:- Rank = 0.25 × (10-1) + 1 = 3.25 (25th percentile)

b. Split it into fractional and int. part.
= 3 + 0.25

c. Find ordered data value corresponding to int. part. (x_i)

Percentile = x_i + fractional × $[x_{(i+1)} - x_{(i)}]$

(vii) Quartiles.

Sample 25th percentile - ~~Min~~. First (Lower)
50th - Median
75th - Third (Upper)

Five Number Summary includes First, Med. Third, Min, Max.

(viii) Inter Quartile Range

$$IQR = Q_3 - Q_1$$

9. Association b/w two categorical variables.

Ex:- Market research about ownership of mobile phone by gender.

Categorical variables in this example.

Gender :- Male or Female - Nominal

Own a Smartphone :- Yes or No. - Nominal

(ii) Row - Relative frequency

Divide each cell freq. in row by its row total.

(iii) Column - Relative frequency

Divide each cell freq in col by its column total.

(iii) Association b/w two variables.

If row or column related freq. are same for all row or column then we say that the two variables are not associated. with each other.

Stacked Bar chart is used to represent this.

10. Association b/w two numerical values.

(i) Scatter Plot. (x-y graph)

A graph that displays pairs of values as points on a 2D plane.

(ii) Describing Association.

a. Direction - up or down.
b. Curvature - Linear or Curved.
c. Variation - Tightly clustered or not.
d. Outliers -

(iii) Measure of Association.

a. Covariance

Covariance quantifies the strength of linear association b/w two numerical values.