Statistics

Week 1 to Week 3

- Descriptive Statistics : → done on complete population
  → used while summarizing data
  → to describe the data

- Inferential Statistics : → done on sample data
  → used to infer something from data
  → to draw conclusions about population from sample data

- Population : complete data
- Sample : subset of population

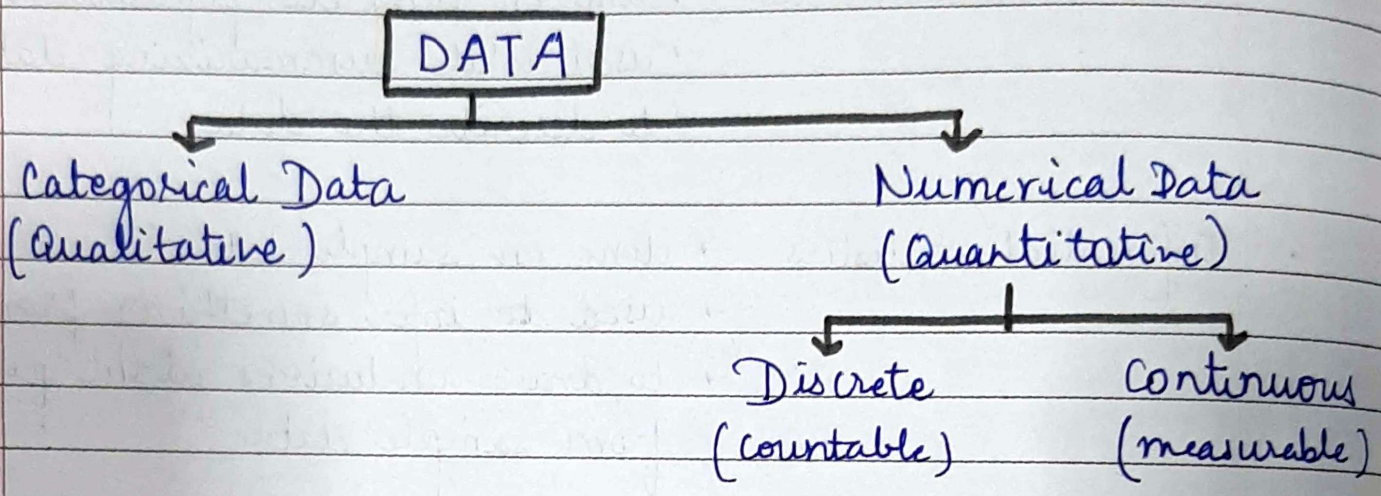
- Structured Data : organised in a table in a systematic manner

- Unstructured Data : Raw form of data, scattered like comments on youtube.

- Dataset : Structured Collection of values - could be numbers, names, tell numbers.

- Variables : → Something that varies
  → In a table, variables are the columns
  → Variables must have same unit throughout!

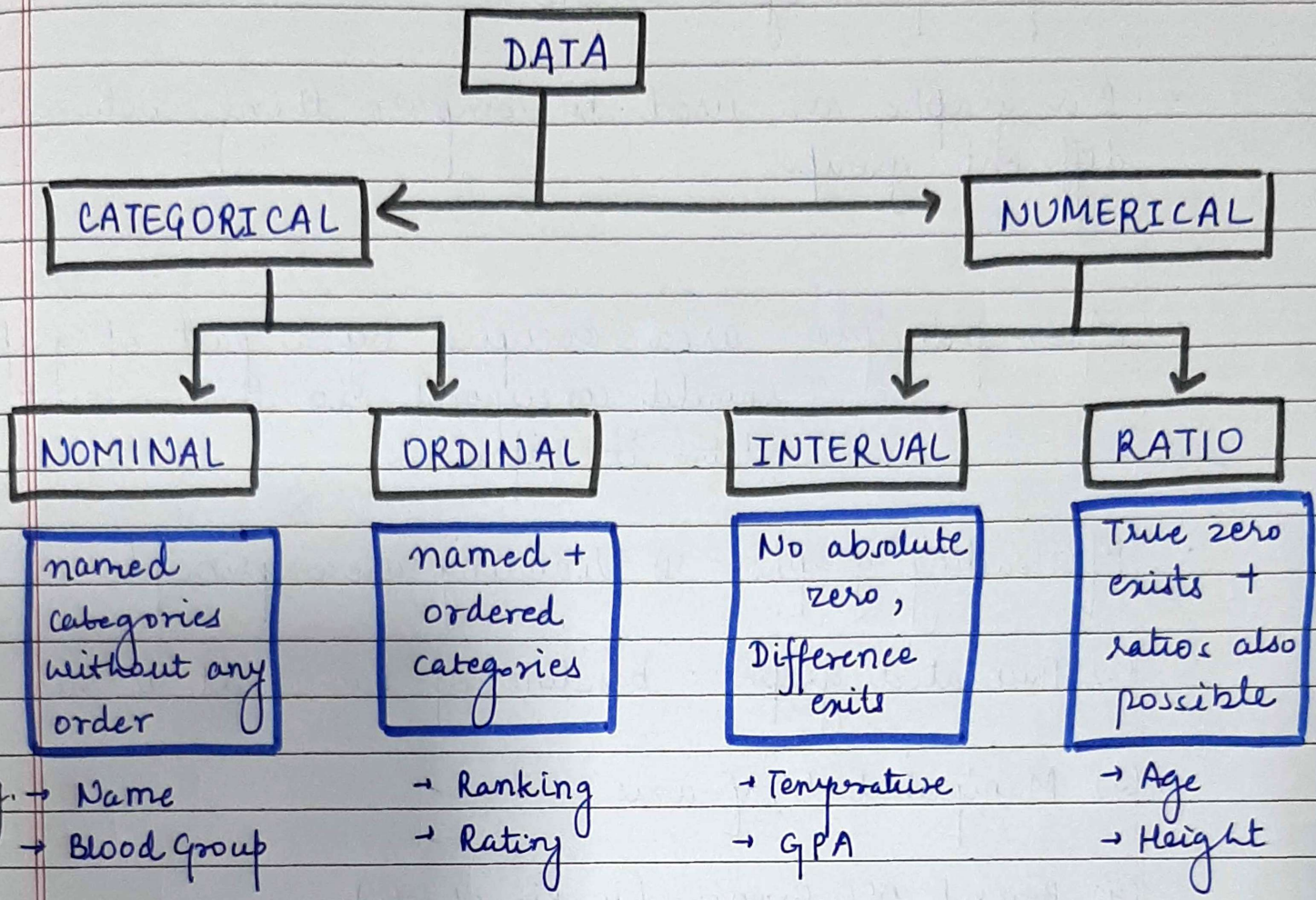
Case: → observation / record
→ in a table, a case is a row
→ for example, each student in student's dataset is a case.
→ for each case, same attribute is recorded



- Time-Series Data: data recorded over time, on a particular variable
- Cross-Sectional Data: data observed at the same time
- Timeplot: graph of a time series showing values in chronological order

→ Check whether data is collected at a point of time (cross-sectional) or over-time (time series)

- Scales Of Measurement



- Representing Data with Graphical Summaries

PIE CHART → used to show the proportion

BAR CHART → used to show the frequencies / relative frequencies
→ If ordinal, the order of categories is preserved

PARETO CHART → it is kind of bar chart where categories are sorted by frequency, either ascending or descending

→ Pie charts are best to use when you are trying to compare parts of a whole

→ Bar graphs are used to compare things between different groups

- Area principle : area occupied by a part of graph should correspond to the amount of data it represents

- Misleading graphs : 1) Violating area principle

(2) Truncated graphs : baseline of bar chart is not at zero

(3) Manipulated Y-axis

(4) Round Off Errors (in Pie charts)

DESCRIBING CATEGORICAL DATA

→ Mode : → highest frequency
→ The longest bar in bar chart
→ The widest slice in a pie chart

→ Multimodal Data : If two or more categories tie for highest frequencies, the data is said to be multimodal.

→ Bimodal for two frequencies.

→ Median :→ We can find median for only ordinal data in categorical data.

→ Median : middle observation

→ for even obs : Median = $\left(\frac{n}{2}\right)^{th}$ & $\left(\frac{n}{2}+1\right)^{th}$ obs

→ for odd obs : Median = $\left(\frac{n+1}{2}\right)^{th}$ obs.

⇒ DESCRIBING NUMERICAL DATA

GRAPHICAL SUMMARIES : 1. Histogram
2. Stem & Leaf Diagram

NUMERICAL SUMMARIES: 1. Measures of Central Tendency
2. Measures of Dispersion

→ Measures of Central Tendency

(1) Mean : Average of given data

$$\text{Sample Mean} = \bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n}$$

$$\text{Population Mean} = \bar{\mu} = \frac{x_1 + x_2 + \dots + x_N}{N}$$

→ for grouped data, i.e, with frequencies,

$$\bar{x} = \frac{f_1 x_1 + f_2 x_2 + \dots + f_n x_n}{f_1 + f_2 + \dots + f_n}$$

$$\bar{x} = \sum_{i=1}^{n} \frac{f_i x_i}{f_i} \qquad [f_i \rightarrow \text{frequency}]$$

→ for grouped data, i.e, class intervals

$$\bar{x} = \sum_{i=1}^{n} \frac{f_i m_i}{f_i}$$ [m_i - mid point of class interval]

But with this mid point we get the approx mean and not the exact mean.

(2) Median : mid value of dataset

→ for odd no. of observations : median = $\left(\frac{n+1}{2}\right)^{th}$ observation

→ for even no. of obs : median = avg. of $\left(\frac{n}{2}\right)^{th}$ & $\left[\left(\frac{n}{2}\right)+1\right]^{th}$ obs.

(3) Mode : most frequent value

→ Measures Of Dispersion

(1) Range : Max - Min

(2) Variance : (1) $\sigma^2 = \frac{(x_1-\mu)^2 + (x_2-\mu)^2 + \ldots + (x_n-\mu)^2}{N}$

(2) $s^2 = \frac{(x_1-\bar{x})^2 + (x_2-\bar{x})^2 + \ldots + (x_n-\bar{x})^2}{n-1}$

(3) Standard Deviation : $s = \sqrt{\text{Variance}}$

(4) Interquartile Range : Q3 - Q1

'PERCENTILES'

→ Computing percentiles : n = Total observation

$p = \frac{\text{percentile}}{100}$

∴ For example, we want to find 25th percentile of a data having 20 observations,

→ $n = 20$, $p = 0.25$ ∴ $np = 20 \times 0.25 = 5$

If 'np' is integer then, mean of $(np)^{th}$ & $(np+1)^{th}$ is the given percentile.

If 'np' is not an integer then, $(np+1)^{th}$ obs is given percentile.

→ Five Number Summary

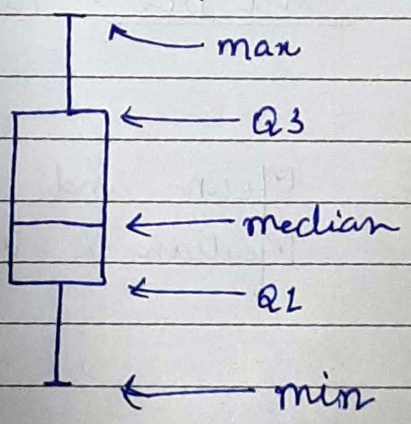
- Minimum
- Q1 - quartile 1, 25th percentile, lower quartile
- Q2 - quartile 2, 50th percentile, median
- Q3 - quartile 3, 75th percentile, upper quartile
- Maximum

→ Outliers

$Q3 + 1.5\,IQR < \text{Outlier} < Q1 - 1.5\,IQR$

3rd quartile ↑ interquartile range ↑ first quartile

→ Box plot



Effects on Numerical summaries After Making Changes In Data

(1) Adding a constant to each data value

Mean : new mean = old mean + constant

Median : new median = old median + constant

Mode : new mode = old mode + constant

Variance : no change

Standard Deviation : no change

(2) Multiplying a constant to each data value

Mean : new mean = constant x old mean

Median : new median = constant x old median

Mode : new mode = constant x old mode

Variance : new variance = $(\text{constant})^2$ x old variance

Std. Dev : new std. dev = constant x old dev

Note : Mean and Range are sensitive to outliers.
Median & Mode are least affected by outliers.