Statistics

Statistics

- **It is the Science of Collection , Organizing, Analysis, Interpretation & Presentation of Data**
- **Examples: Population Density, Literacy Rate, Life Expectancy, Sales Growth**

- Statistics can be divided into two types
  - **Descriptive: Summarization** of Data in a meaningful way (A way to Describe Data)
  - **Inferential:** Inference associated with Data-set, conclusion drawn about population from the sample

Descriptive Statistics include creation of Graphs, Charts, Tables and include summarizing measures such as average, percentile etc. However, limited to data we are analyzing does not help in analyzing something that is beyond data

Let us look at the Sample Data Collected for a set of people....

Dataset

Data is a collection of set of Values of Numerical or Categorical Variables

Here Data Includes:

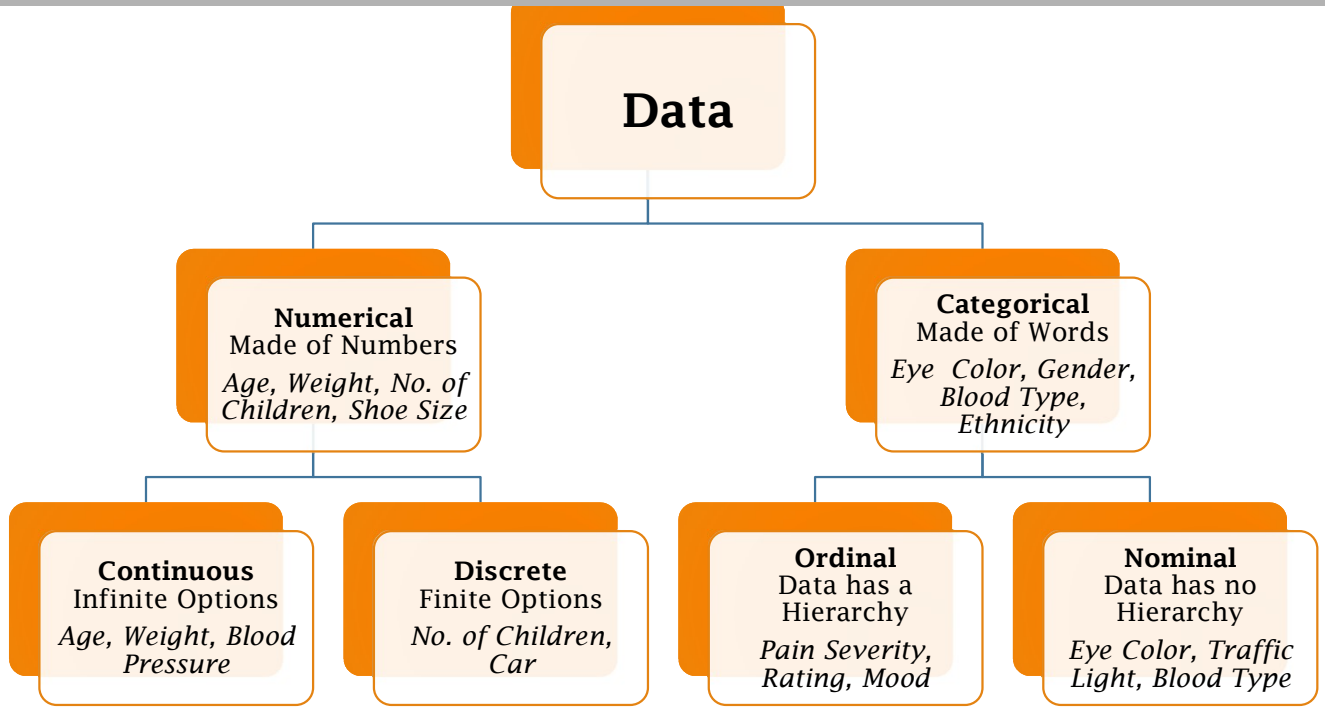
- Age of Dwellers in a particular area
- Their Height
- Their Gender
- Their Lung Capacity (LungCap)
- Their Smoking Habit

Snapshot of Data Collected from the City

Age	Height	Smoke	Gender	LungCap
6	62.1	No	Male	6.48
18	74.7	Yes	Female	10.13
16	69.7	No	Female	9.55
14	71	No	Male	11.13
5	56.9	No	Male	4.80
11	58.7	No	Female	6.23
8	63.3	No	Male	4.95
11	70.4	No	Male	7.33
15	70.5	No	Male	8.88
11	59.2	No	Male	6.80
19	76.4	No	Male	11.50
17	71.7	No	Male	10.93

Type of Data

Understanding type of Data



Ratio variables → never fall below zero. E.g., Height and weight measure

Interval variable → indicates difference between 2 entities, No True Zero Value E.g., Temperature

Dependent / Independent Variables

Here LungCap is taken as Dependent Variable / Response Variable / Target or Output Variable and rest are Independent Variables....

Age	Height	Smoke	Gender	LungCap
6	62.1	No	Male	6.48
18	74.7	Yes	Female	10.13
16	69.7	No	Female	9.55
14	71	No	Male	11.13
5	56.9	No	Male	4.80
11	58.7	No	Female	6.23
8	63.3	No	Male	4.95
11	70.4	No	Male	7.33
15	70.5	No	Male	8.88
11	59.2	No	Male	6.80
19	76.4	No	Male	11.50
17	71.7	No	Male	10.93

Type of Analysis

C

Types of Analysis

1

Univariate Analysis

Focusing & Analyzing One Variable at a time

2

Bivariate Analysis

Comparing two variables at a time which may be Categorical or Numeric. 1) Numeric – Numeric 2) Categorical – Numeric 3) Numeric – Categorical 4) Categorical - Categorical

3

Multivariate Analysis

Comparing More than two variables at a time which may be Categorical or Numerical

Univariate Analysis focuses & Analyses One Variable at a time …

Univariate Analysis

Mathematical Measurements

Measures of Central Tendency

- **Mean:** Totalling all dataset values & dividing by number of Values
- **Median:** Central Value in Dataset
- **Mode:** Most Frequently occurring value in Dataset

Measures of Dispersion

- Range
- Inter-Quartile Range
- **Variance:** Dispersion around mean
- **Standard Deviation:** Square Root of Variance

Measures of Shape

- **Skewness:** Symmetry in Distribution
- **Kurtosis:** Shape of Peaks in Distribution of Data

Visual Analysis

- **Histogram**
- **Bar Chart**
- Box Plot
- Pie Chart

Measures of Central Tendency / Location

- Estimate Central Point of Sample. Different ways of estimating central point of dataset are as follows:

  - Mean → Most representative value in the data. Used with only numerical data
  - Median → Midpoint of a ranked distribution (sorted data in increasing order)
  - Mode → Most Common data value or highest frequency

Above All three help us in summarizing data

Drawback of Mean: Very sensitive to outlier. Outlier leads to wrong summary.

Arithmetic Mean

- Arithmetic mean is a mathematical average, and it is the most popular measures of central Tendency. It is frequently referred to as 'Mean'. It is denoted by $\bar{x}$.
- "It is obtained by dividing sum of all the values by the total number of observations"
- Say we measure the height of 10 students in class and calculate the average
- Mean is affected by extreme values

Median

- Median is a middlemost value of the distribution, or the value which divides the distribution in equal parts, when the values are arranged in order of magnitude
- Median for Raw data:
  - Arrange data in Ascending order.
  - Apply the formula.
- Median is not affected by extreme values
- Quartiles, Deciles, Percentiles

Mode

- Mode is the most frequent (most frequent) value in the distribution
- Mode is not affected by extreme values
- It is denoted by Z

Types of Means

- Arithmetic Mean
- Geometric Mean
- Harmonic Mean

Geometric Mean (Average of Growth)

- Geometric Mean used in case of average of consecutive growth rates or Shrinkage rates (Return)

  - 100 Rs becomes 120 in first year, it becomes 90 in 2nd year, 110 in 3rd year and 120 in 4th year

+20% -10% +10% +20%

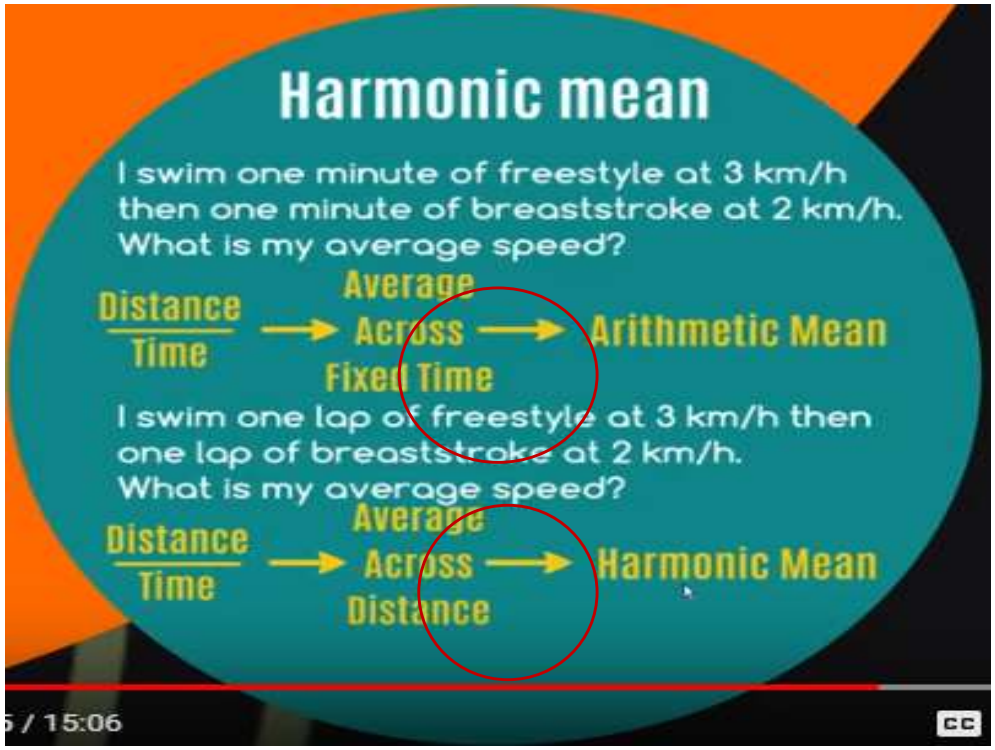
GROWTH FACTORS: 1.2 0.9 1.1 1.2

$$\overline{x}_g = \sqrt[4]{1.2 \cdot 0.9 \cdot 1.1 \cdot 1.2} \approx 1.093$$

Average Growth 9.3% per year

Geometric Analogy – A cube of 1.2, 1.1, 09 can give a side same side length of a cube, volume remaining same

Harmonic Mean (Average of Rate Such as Speed)



Speed = Distance / Time

Fixed Time (say 1 min) → AM
Fixed Distance (say 1 lap) → HM

Measures of Dispersion

- Range → Does not tell how the data is distributed between the max and min boundary. Limitation of Range is that it is sensitive to outliers
- Variance / Standard Deviation
- Interquartile Range → (Quartile contains 25% of data) ie. Mini Range that excludes Outliers. Advantage of IQR is that it does not take outlier into account. It uses central 50% of data
  - Upper Limit = Q1-1.5(IQR)
  - Lower Limit =Q3+1.5(IQR)
- **Dispersion**
  - More Similar the data points → Less Dispersion
  - Less Similar the data points → More Dispersion (More Spread-Out distribution = Larger Dispersion)

Measures of Shape

- Distribution of Data provides the shape of the data. Distribution of Data is visually represented by Histogram.
- Histogram has two properties
  - Degree of Skewness → Deviation from Symmetry → Degree of Symmetry
    - Gives → Direction and Amount of Skewness
  - Kurtosis → How tall / sharp central peak is → Degree of Peakedness

Testing of Normality → In stats the data needs to be normal

Normal Distribution = Skewness =0 , Kurtosis =0

Bivariate Analysis compares two variables at a time which may be Categorical or Numerical…

Bivariate Analysis

Mathematical Measurements

Correlation & Covariance

- Compare two Numerical Variables

Contingency Table

- Table for 2 Categorical Variables

T-Test & F-Test (ANOVA)

- Compare One Numerical (Dependent / Response Variable) & One Categorical Variable

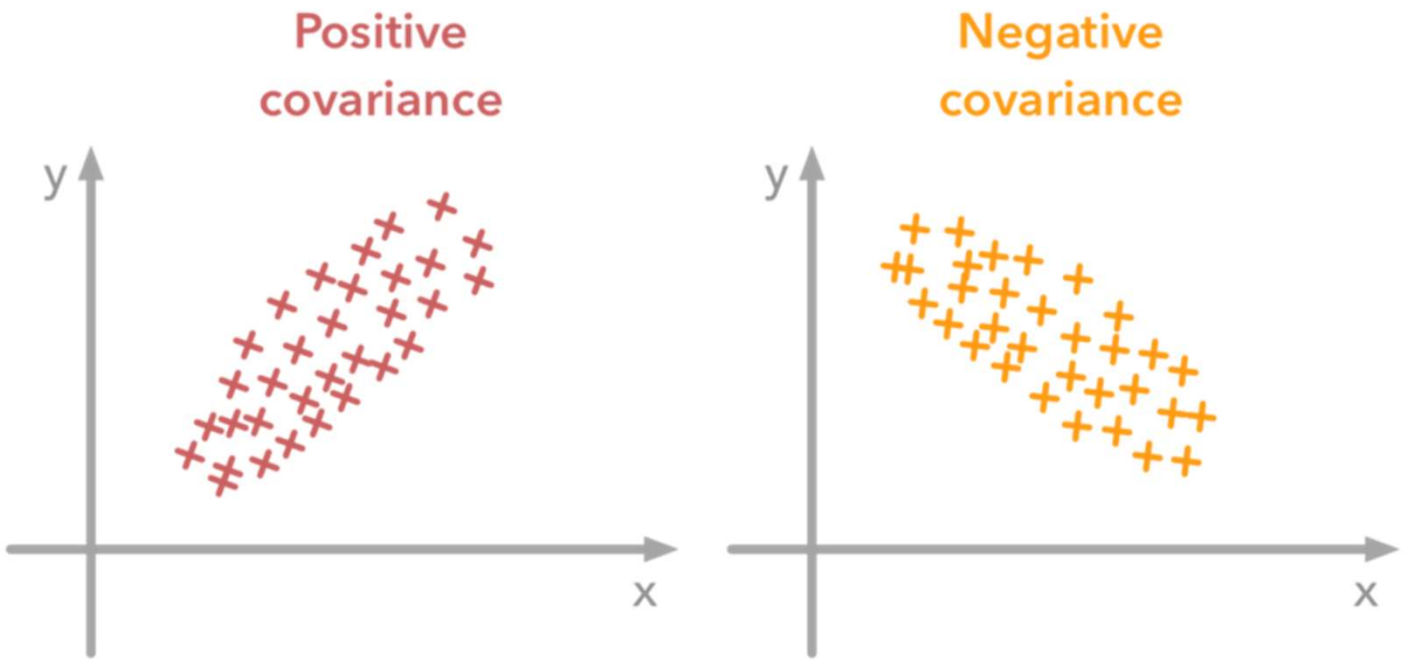
Chi Square Test

- Compare 2 Categorical Variables

Visual Analysis

- Correlation Plot
- Scatter Plot
- Stacked Box Plot
- Stacked Bar Plot

Covariance



Correlation