Probability

- Likelihood of occurrence of an event
- Expressed in 0-1 or in percentage
- If probability is 0.7, it indicates the event will occur 70% in total trial
- Calculated as $P(A) = \frac{\#\, Favorable\ Outcome}{Total\ Outcome}$

Statistics

- Science of collecting, analyzing, interpreting and presenting data
- Type:
  - Descriptive Statistics (Summarizing Data)
  - Inferential Statistics (Making prediction from data)

Variable

- Measurable characteristics that can change
- Example: Age, Height, Income, Color, Rating
- Building block of data analysis
- Types:
  - Numerical: Discrete or Continuous
  - Categorical: Nominal or Ordinal (with order)

Sampling

- Population → Entire group that we want to study
- Sample → Smaller group drawn from the population
- Sampling is practical, less expensive & saves time
- We will do stratified sampling later on ML

Central Tendency

These values denotes the center of data

- Mean
- Median
- Mode

Mean / Average

- Arithmetic mean & weighted mean
- Sum of observed value divided by total item
- Takes account of all observation
- Change in any data value affects the mean
- **Heavily influenced by outliers**
- **Not Suitable for asymmetric data**
- In weighted mean observation have some value with it
- $\overline{x} = \frac{\sum x}{N}$ / $\overline{x} = \frac{\sum x.f}{\sum f}$

Median

- Center of data
- Middle value in observation
- Calculated based on position of data rather than value though value is sorted in prior
- **Immune to outliers**
- Median = value of $\frac{n+1}{2}$ item / $\frac{value\ of\ \frac{n}{2} + value\ of\ (\frac{n}{2}+1)}{2}$
- Frequency distribution: calculate C.F & class > N/2

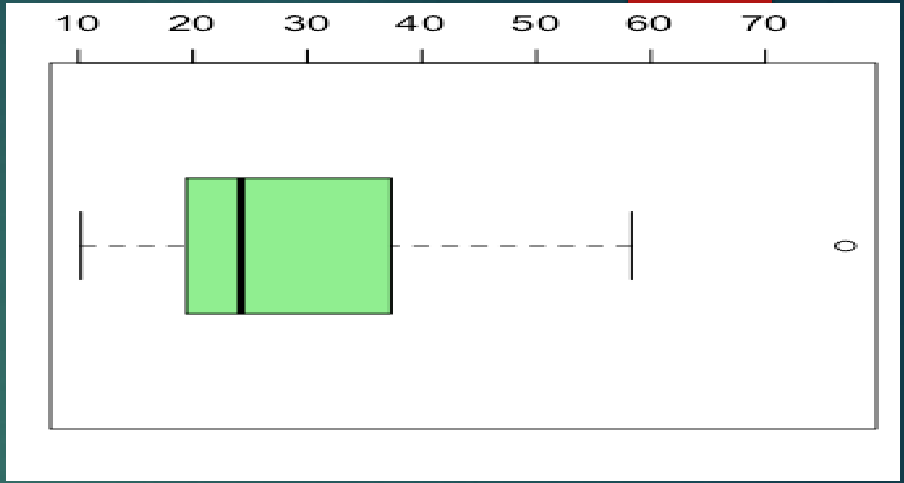
Mode

- Most repeated value
- Can apply to non-numeric data as well
- For frequency distribution table it's the value with highest frequency

Percentile

- Data divided into 100 parts
- Median in 50^{th} percentile
- Useful in interpreting how much bigger the data is compared to other data as single value mayn't give this information
- Example: If student scored 60 marks in exam we may assume he performed bad. But if he lies in 90^{th} percentile then he performed better than remaining 90% and can conclude question was tough or any other reason

Box Plot



- It gives 25th, 50th, and 75th percentile
- For effective visualization of percentile box plot is used
- Box contains 50% of data. i.e. IQR (Q3 – Q1)
- Whisker of plot is at either 1.5 * IQR value or min/max (lesser)
- Data below and above 1.5 * IQR are denoted by dot & considered outliers

Dispersion

- Variance
- Standard Deviation

Variance

- Measures spread of data from mean
- High ➔ Data is more spread, Less ➔ Clustered
- Mean squared distance of data from mean
- Has unit squared compared to data
- Formulae: $\sigma^2 = \frac{\sum(x_i - \bar{x})^2}{n}$ (For sample n-1)
- $\sigma^2 = \frac{\sum fx^2}{N} - (\frac{\sum fx}{N})^2$

Standard Deviation

- Square root of variance
- Has same unit as that of data point hence comparable

Covariance

- Measure of how two quantity varies together
- Measure of linear relationship between variable
- Used as dimension reduction technique
- Value can be any real number, so for standardization we use correlation
- Formulae: $Cov(x, y) = \frac{\sum(x-\bar{x})(y-\bar{y})}{n}$ (n-1 for sample)

Correlation

- Measure strength & direction of linear relationship between two variables (Continuous data)
- Its standardized (unit less) and always between +1 & -1
- Not a measure of causation but measure of association
- +1 ➔ Perfect positive linear relationship
- -1 ➔ Perfect negative linear relationship
- 0 ➔ No Linear relationship
- Pearson Correlation: $r = \frac{COV(x,y)}{\sigma_x \sigma_y}$

Spearman Rank Correlation

- Used when data is ordinal or not normally distributed
- Measure strength & direction of monotonic relationship
- Formulae: $r_s = 1 - \frac{6 \sum {d_i}^2}{n(n^2-1)}$
  - di ➔ Difference in ranks of each observation
  - n ➔ Number of observation

Example

Student	Math Score (X)	Physics Score (Y)
A	80	85
B	70	78
C	90	92
D	60	65
E	85	88

Example

Step 1: Rank the data

Math Scores (X):

Score	Rank
60	1
70	2
80	3
85	4
90	5

Physics Scores (Y):

Score	Rank
65	1
78	2
85	3
88	4
92	5

Example

Step 2: Calculate rank differences

Student	Rank X	Rank Y	$d_i = R_x - R_y$	d_i^2
A	3	3	0	0
B	2	2	0	0
C	5	5	0	0
D	1	1	0	0
E	4	4	0	0

$$\sum d_i^2 = 0$$

Example

Step 3: Apply Spearman's formula

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

$$r_s = 1 - \frac{6 \times 0}{5(25-1)} = 1$$

Answer: $r_s = 1$ → perfect positive monotonic correlation

✅ *This means students who score high in math also tend to score high in physics, in the exact same rank*

Probability Rule

- P(A or B) = P(A) + P(B) – P(A n B)
- P(A and B) = P(A) * P(B) [or P(B | A)]
- Permutation (order matters): $^nP_r = \frac{n!}{(n-r)!}$
- Combination: $^nC_r = \frac{n!}{r!(n-r)!}$