UNIT 2

Data science & big data analytics

Statistical Interference

Basic Formulas :-

- Mean = $\frac{\sum x_i}{n}$

- Median =

For Datasets with Odd Numbers:

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

For Datasets with Even Numbers:

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

- Mode = the value with the highest frequency

- Variance

$$\sigma^2 = \frac{\sum (x-\mu)^2}{N} \quad \text{Population Variance}$$

$$s^2 = \frac{\sum (x-\bar{x})^2}{n-1} \quad \text{Sample Variance}$$

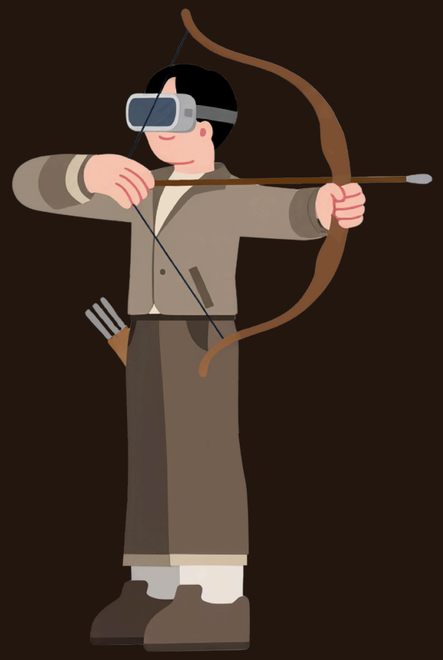
Basic Formulas :-

- Standard Deviation =

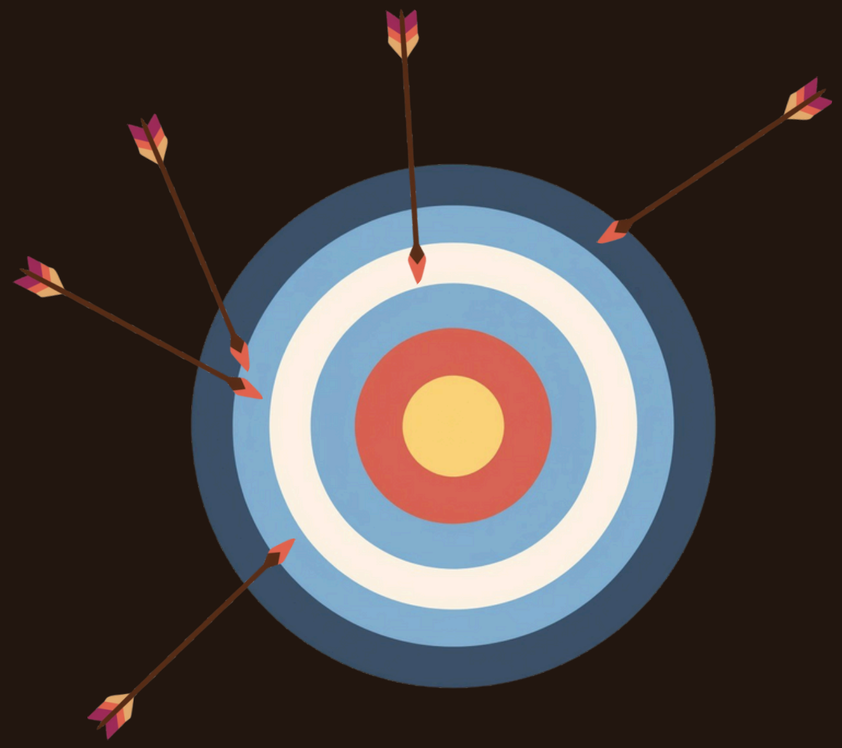
$$\text{Standard Deviation (population)} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \text{mean})^2}{n}}$$

$$\text{Standard Deviation (sample)} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \text{mean})^2}{n-1}}$$

Standard Deviation



- Spread Out = High SD



- Standard Deviation = How ‘SCATTERED’ Data is

Standard Deviation

- Standard Deviation is also a measure of dispersion in a dataset.
- It indicates how much data values deviate from the mean.
- Standard Deviation is the square root of variance.
- It helps determine the size and consistency of data spread.
- It is easier to interpret compared to variance because it is in the same unit as the data.

- Standard Deviation is symbolized by σ (Greek letter sigma).
- Formula of Standard Deviation:

$$\sigma = \sqrt{\frac{\sum (x - m)^2}{n}}$$

Variance

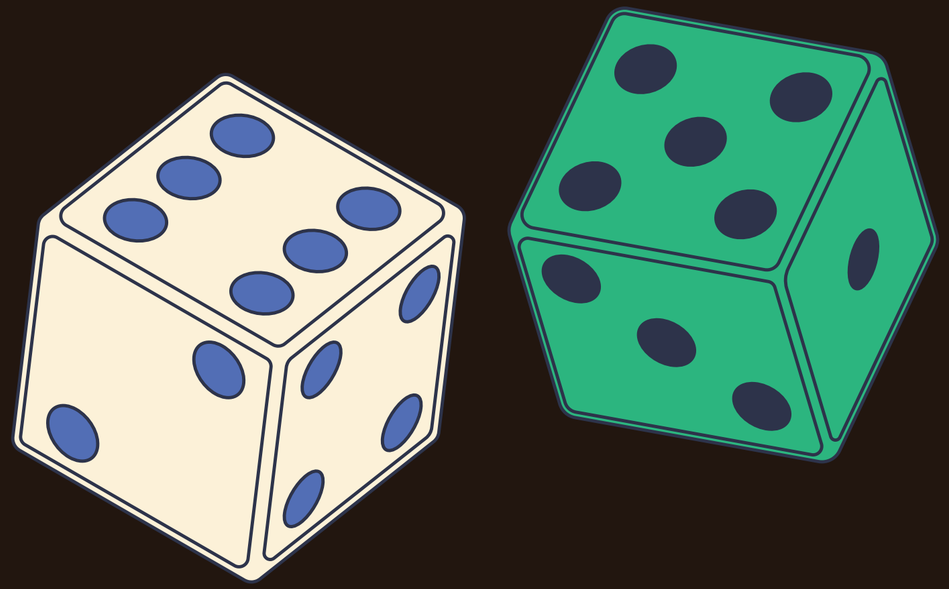
- Variance is a measure of dispersion in a dataset.
- It shows how far the values of a dataset are spread from their mean.
- Variance is a statistical measure used in probability and statistics.
- It helps in understanding the degree of variability in data.
- Variance measures the absolute variability of dispersion.
- Variance is denoted by σ^2 (sigma squared).

Formula of Variance:

$$\sigma^2 = \frac{\sum(x-m)^2}{n}$$

Baye's Theorem

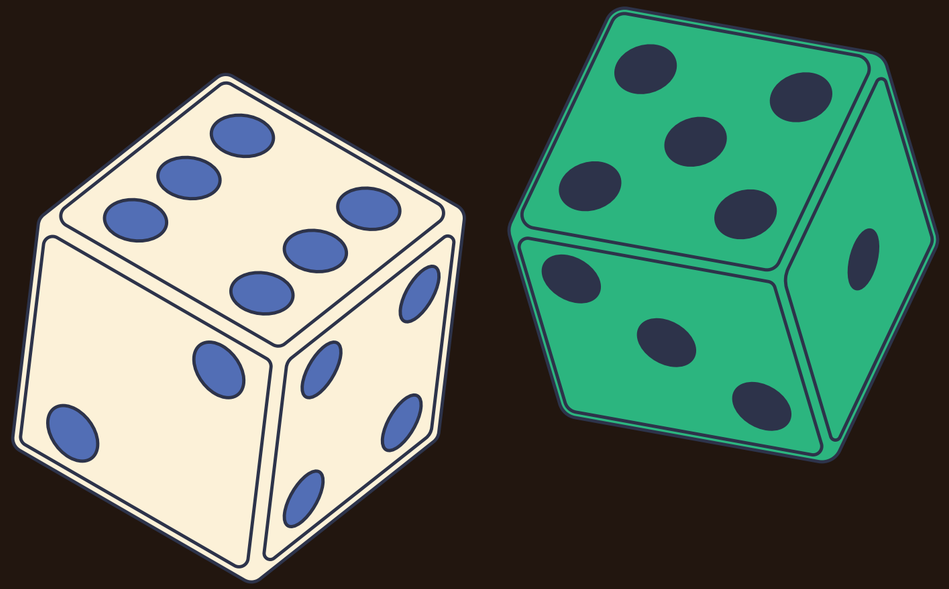
- Bayes' Theorem is a method to update or revise probability when new information is available.
- It is used to calculate conditional probability, also called posterior probability.
- If A and B are two events, P(A|B) denotes the probability of A occurring given that B has occurred.
- Similarly, P(B|A) denotes the probability of B occurring given that A has occurred.
- Bayes' Theorem establishes a relationship between P(A|B) and P(B|A).



Baye's Theorem

- Formula:

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

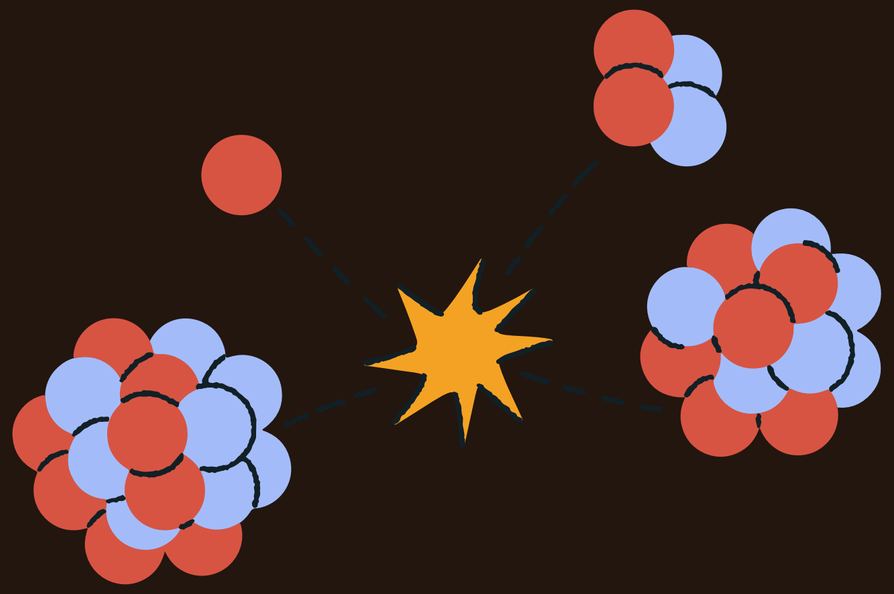


- Applications

- Used in statistics, machine learning, medical diagnosis, and decision making.

Hypothesis Testing

- Hypothesis Testing is a statistical test procedure used to choose between two possible statements about a population.
- It uses sample data to evaluate a hypothesis about a population.
- The main goal is to analyze a sample and decide which population characteristics are likely or unlikely to occur.

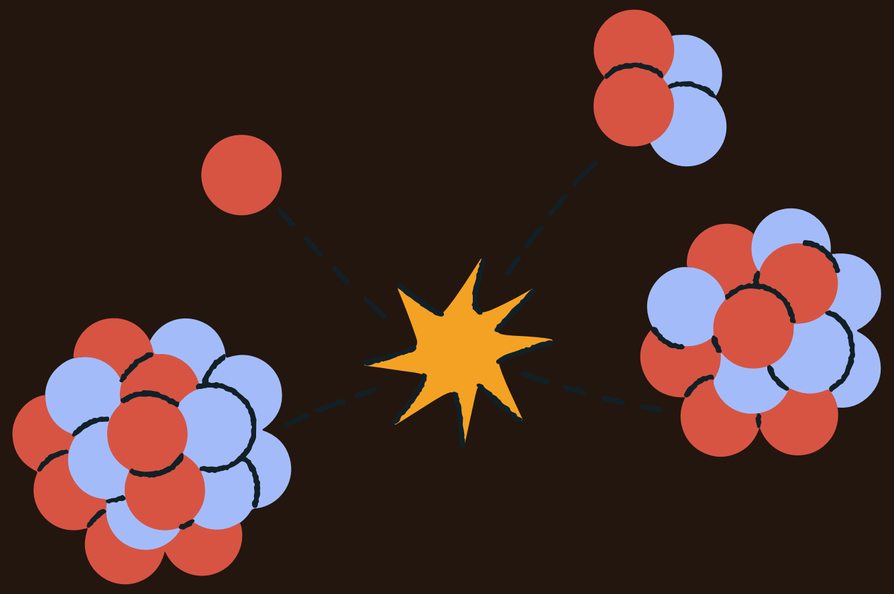


Steps in Hypothesis Testing

Hypothesis Testing

1) Formulate the Hypotheses

- Null Hypothesis (H_0):
- Represents the status quo; states no difference or no effect.
- Alternative Hypothesis (H_1):
- Indicates that some difference or effect exists.

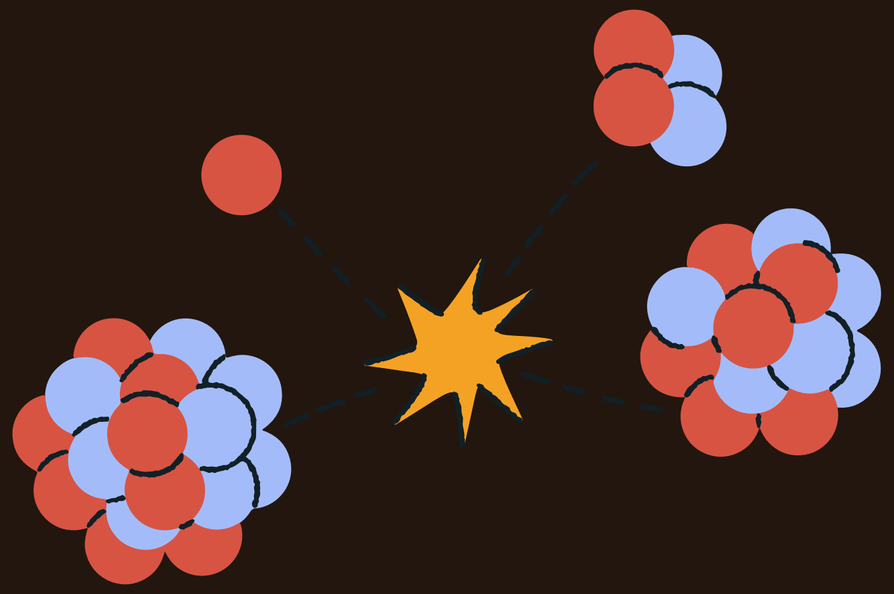


2) Select an Appropriate Test

- Choose a suitable statistical test (Z-test, t-test, etc.).

Hypothesis Testing

- The test statistic measures how close the sample result is to the null hypothesis.



3) Choose Level of Significance (α)

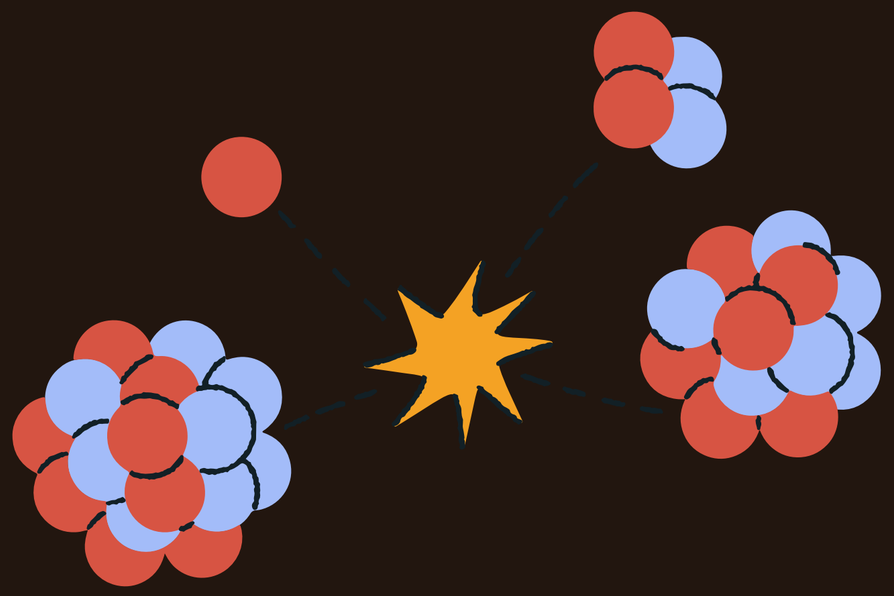
- The level of significance (α) is the probability of committing a Type I error.

Types of Errors

- Type I Error (α):

Hypothesis Testing

- Occurs when the null hypothesis is rejected even though it is true.

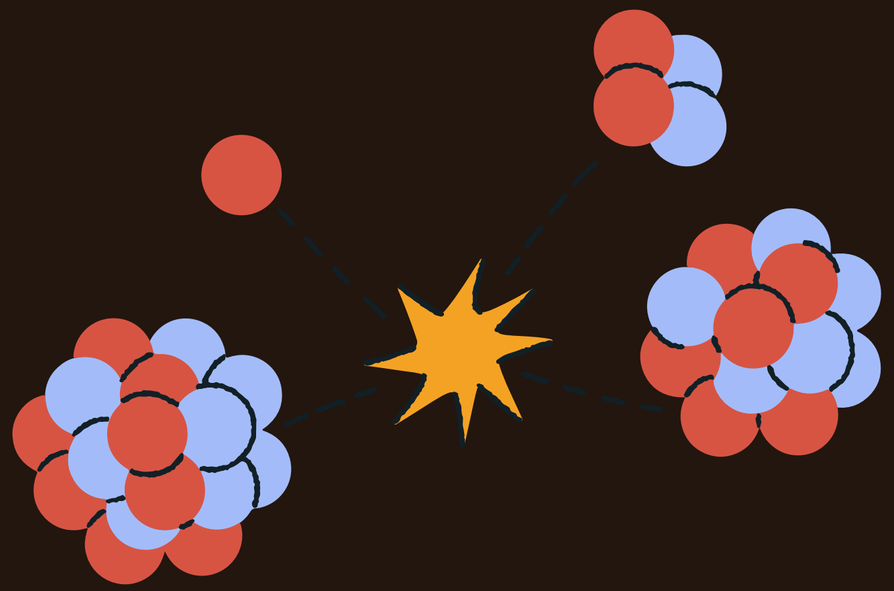


Type II Error (β):

- Occurs when the null hypothesis is not rejected even though it is false.
- The probability of Type II error depends on the actual value of the population parameter.
- It is necessary to balance both types of errors.

Hypothesis Testing

Final Steps



4) Collect Data and Calculate Test Statistic

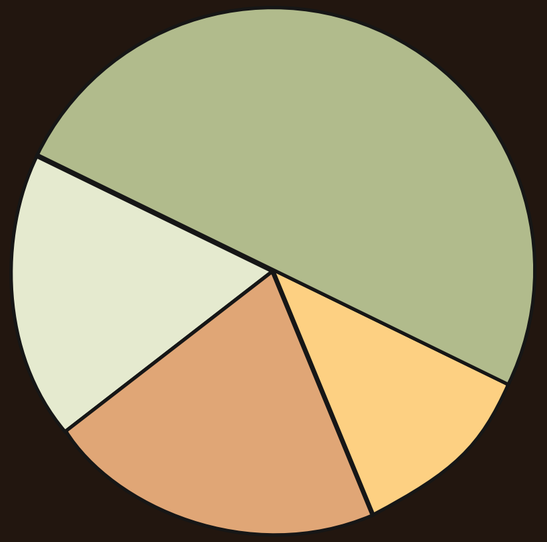
- The Z-test statistic is calculated using sample data.

5) Determine p-value or Critical Value

- Compare the calculated value with the critical value or p-value.
- Values are obtained using standard normal distribution tables.

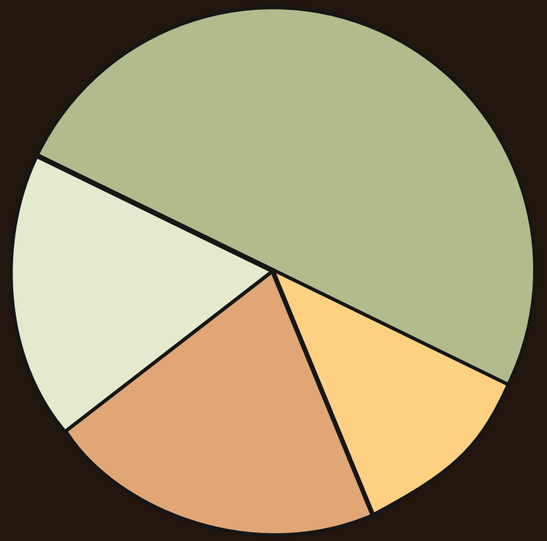
Wilcoxon Rank–Sum Test

- The Wilcoxon Rank–Sum Test is a non-parametric test used as an alternative to the two-sample t-test.
- It is applied when data does not follow a normal distribution.
- The test is based on the ranks of observations, not on the actual data values.
- It tests the null hypothesis that the medians of two populations are equal.



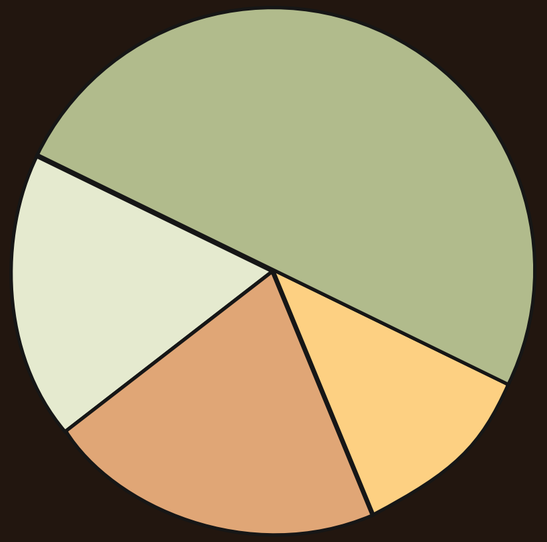
Wilcoxon Rank–Sum Test

- Working Principle
- All observations from both samples are combined and ranked.
- Two rank sums are calculated, one for each group.
- If a systematic difference exists between the two groups:
- Higher ranks tend to appear in one group.
- Lower ranks tend to appear in the other group.
- As a result, one rank sum becomes small.
- If the two groups are similar:



Wilcoxon Rank–Sum Test

- High and low ranks are evenly distributed.
- Both rank sums are relatively large.

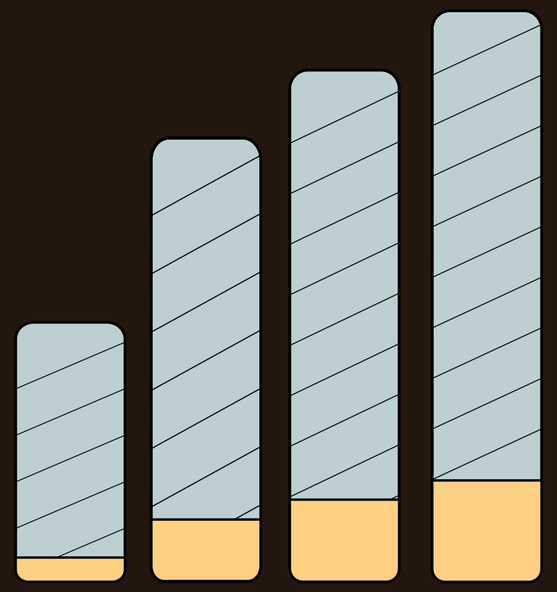


Test Statistic

- The Wilcoxon test statistic W is defined as the smaller of the two rank sums.
- The null hypothesis is rejected when W deviates significantly from its expected mean.
- This indicates that the two populations do not have identical distributions.

Descriptive Statistics

- Descriptive Statistics focuses on summarizing and organizing data.
- It involves calculating values that describe the main features of a dataset.
- Common measures include mean, median, mode, maximum, and error measures.
- It helps convert large amounts of raw data into meaningful information.
- Descriptive statistics are often used to present data using tables, charts, and graphs.



Inferential Statistics

- Inferential Statistics is used to draw conclusions about a population.
- It analyzes sample data to make predictions or decisions about the entire population.
- It helps in estimating parameters and testing hypotheses.

Chi-Square Test for Independence

- The Chi-Square Test for Independence is used to check whether there is a significant association between two categorical (nominal) variables.
- It examines how the frequency of one categorical variable changes across the categories of another variable.
- This test is also referred to as the Chi-Square Test of Association.
- It helps determine whether the variables are independent or related.

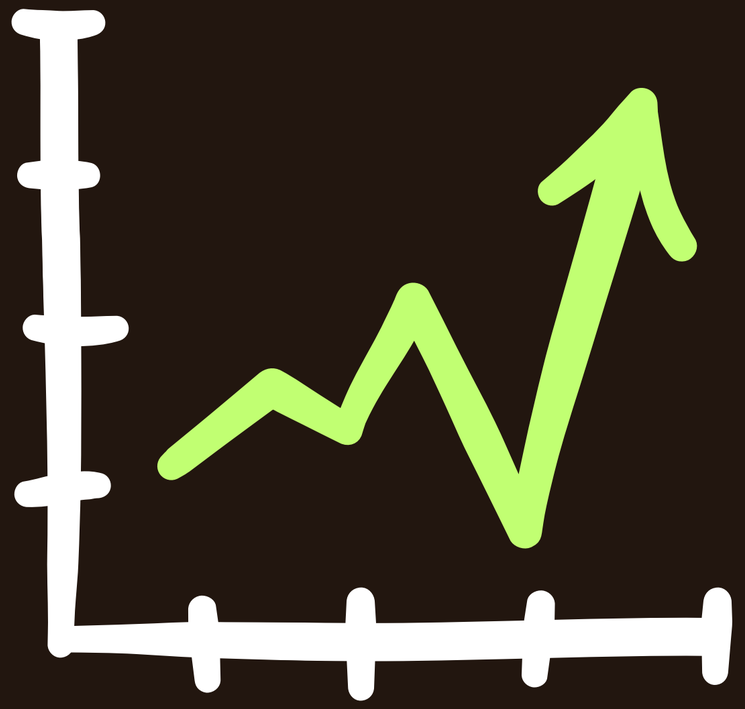
Chi-Square Test for Independence

Data Representation

- The test is performed using a contingency table.
- In the table:
  - Rows represent the categories of one variable.
  - Columns represent the categories of the other variable.
- The table displays the observed frequencies for each category combination.

Contingency Table

- A contingency table is a statistical table that shows the frequency distribution of two categorical variables.
- It is used to summarize categorical data in a clear and structured form.
- Contingency tables help in identifying relationships, patterns, or associations between variables.
- Each cell of the table represents the number of observations for a specific category combination.



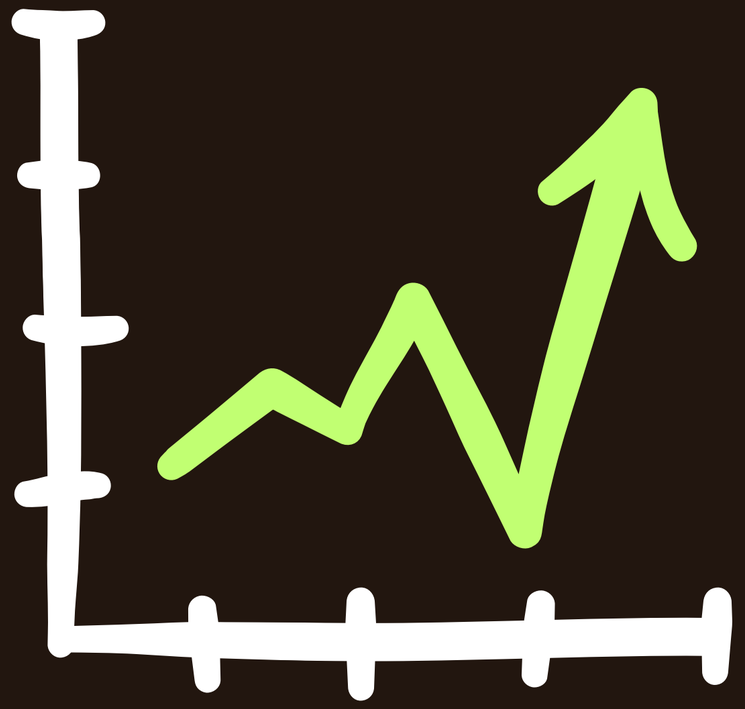
Contingency Table

Education Level	Employed	Unemployed	Total
School	120	80	200
Graduate	180	40	220
Postgraduate	70	10	80
Total	**370**	**130**	**500**

- Suppose data is collected on education level and employment status of individuals.

Contingency Table

- Rows represent categories of education level.
- Columns represent categories of employment status.
- Each cell shows the frequency of individuals in that category.
- This table helps analyze whether education level is related to employment status.
- It can be used as input for the Chi-Square Test of Independence.



One-Tailed and Two-Tailed t-Test

One-Tailed t-Test

- A one-tailed t-test is used when the hypothesis predicts a specific direction of difference.
- It tests whether a population mean is greater than or less than a given value or another mean.
- This test focuses on one side of the distribution.

- Example:
- Group A (Traditional method):

One-Tailed and Two-Tailed t-Test

- Mean = 75, Standard Deviation = 10, Sample Size = 30
- Group B (New method):
- Mean = 80, Standard Deviation = 12, Sample Size = 30
- A one-tailed t-test can be applied to check whether the mean score of Group B is significantly higher than Group A.

Two-Tailed t-Test

One-Tailed and Two-Tailed t-Test

- A two-tailed t-test is used when the hypothesis checks for any difference, without specifying direction.
- It tests whether the population mean is not equal to a specified value or another mean.
- This test considers both ends of the distribution.

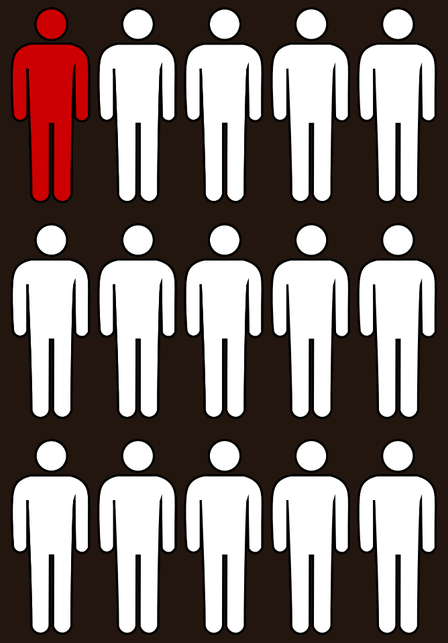
Example:

- Using the same data for Group A and Group B,
- A two-tailed t-test is conducted to determine whether there is any significant difference in mean scores between the two groups.

Population and Sample

Population

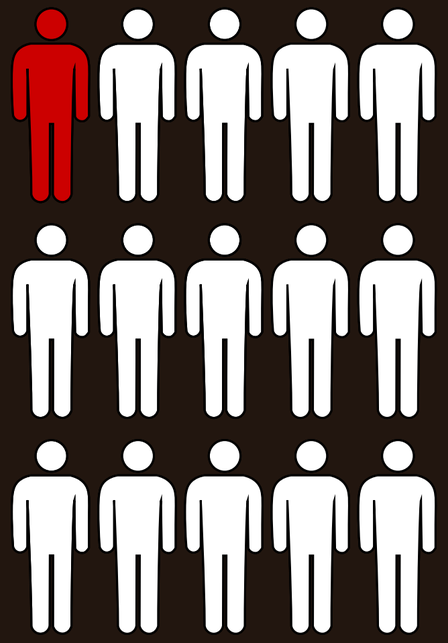
- Population includes the entire set of individuals or objects that are the focus of a study.
- It represents the complete group about which conclusions are to be drawn.
- Example:
- If a study examines the height of adult males in a country, the population consists of all adult males in that country.



Population and Sample

Sample

- A sample is a smaller portion selected from the population.
- It is chosen for observation, measurement, or analysis.
- The purpose of a sample is to estimate or infer characteristics of the whole population.
- Using samples makes studies practical, cost-effective, and time-efficient.



Finding Correlation

c) Is there a correlation between the variables in the following data set? [6]

Hours	9	15	25	14	10	18	19	16	20	18
Marks	39	56	93	61	50	75	42	70	66	32

Step 1 : - Prepare a tale

X (Hours)	Y (Marks)	X^2	Y^2	XY
9	39	81	1521	351
15	56	225	3136	840
25	93	625	8649	2325
14	61	196	3721	854
10	50	100	2500	500
18	75	324	5625	1350
19	42	361	1764	798
16	70	256	4900	1120
20	66	400	4356	1320
18	32	324	1024	576
Σ		**2892**	**37196**	**10034**

Finding Correlation

Step 2 : - Calculation of Mean Values

$\bar{X}$ (Mean of X):

$$\bar{X} = \frac{\Sigma X}{n} = \frac{164}{10} = \mathbf{16.4}$$

$\bar{Y}$ (Mean of Y):

$$\bar{Y} = \frac{\Sigma Y}{n} = \frac{584}{10} = \mathbf{58.4}$$

Step 3: Correlation Coefficient (r)

Finding Correlation

Formula:

$$r = \frac{n\sum XY - \sum X \sum Y}{\sqrt{(n\sum X^2 - (\sum X)^2)\,(n\sum Y^2 - (\sum Y)^2)}}$$

Given Values:

- $n = 10$
- $\sum X = 164$
- $\sum Y = 584$
- $\sum X^2 = 2892$
- $\sum Y^2 = 37196$
- $\sum XY = 10034$

Finding Correlation

Substitution:

Numerator

$$10(10034) - (164)(584) = 100340 - 95776 = 4564$$

Denominator

$$\sqrt{(10 \times 2892 - 164^2)(10 \times 37196 - 584^2)}$$
$$= \sqrt{(28920 - 26896)(371960 - 341056)}$$
$$= \sqrt{(2024)(30904)}$$
$$= \sqrt{62549696} \approx 7909$$

Correlation Coefficient:

$$r = \frac{4564}{7909} \approx \mathbf{0.58}$$

- The value of r ≈ 0.58
- Hence, there is a moderate positive correlation between Hours studied and Marks obtained

r ki value	Meaning
r > 0	**Positive correlation**
r < 0	**Negative correlation**
r = 0	**No correlation**

c) Assume that a patient X took a lab test for a certain disease and tested positive. The lab test returns a positive result in 95% of the cases in which the disease is actually present and it falsely returns a positive result in 6% of the cases in which the disease is not present. Further more only 1% of the entire population has this disease. What is the probability that X actually has the disease given that he is tested positive. **[5]**

EGP013091
20/03/2024 10:45:49

Let:

- D = patient has the disease
- $\bar{D}$ = patient does NOT have the disease
- + = test result is positive

Given:

- $P(D) = 0.01$ (1% population has disease)
- $P(\bar{D}) = 0.99$
- $P(+|D) = 0.95$ (true positive rate)
- $P(+|\bar{D}) = 0.06$ (false positive rate)

Required:

$$P(D|+) = ?$$

Bayes' Theorem Formula

$$P(D|+) = \frac{P(+|D)\,P(D)}{P(+|D)\,P(D) + P(+|\bar{D})\,P(\bar{D})}$$

Substitution

$$\begin{aligned} P(D|+) &= \frac{0.95 \times 0.01}{(0.95 \times 0.01) + (0.06 \times 0.99)} \\ &= \frac{0.0095}{0.0095 + 0.0594} \\ &= \frac{0.0095}{0.0689} \\ &\approx 0.138 \end{aligned}$$

Interpretation (very important for exams)

- Even though the test is highly accurate, the probability that the patient actually has the disease after testing positive is only about 13.8%, because the disease is very rare in the population.

Thank You