Sardar Patel Institute of Technology,Mumbai

Department of Electronics and Telecommunication Engineering

B.E. Sem-VII- PE-IV (2024-2025)

**IT 24 - AI in Healthcare**

**Experiment 3: Hypothesis Testing**

**Name: Adwait Purao UID : 2021300101** **Date: 9/9/24**

**Objective:**

* **Write a program for Hypothesis tests such as Chi square test and ANOVA test.**

**Outcomes:**

* **Appropriately interpret results of chi-square tests**
* **Identify the appropriate hypothesis testing procedure based on type of outcome variable and number of samples**

**System Requirements:**

Linux OS with Python and libraries or R or windows with MATLAB

**Theory:**

**What is Hypothesis?**

**Hypothesis:** A formal statement about a population parameter or a probability distribution. It serves as the basis for statistical testing, where we aim to determine if there's enough evidence in a sample to infer a certain condition for the entire population.

### **Null Hypothesis (H₀)**

* A statement of "no effect" or "no difference."
* Represents the status quo or a default position that indicates no association between variables or no change from the norm.

### **Alternative Hypothesis (H₁ or Hₐ)**

* A statement that contradicts the null hypothesis.
* Suggests that there is an effect, a difference, or an association.

### **Hypothesis Testing Procedure**

1. **State the Hypotheses:** Formulate H₀ and H₁.
2. **Choose the Significance Level (α):** Typically set at 0.05 or 0.01.
3. **Select the Appropriate Test Statistic:** Depends on the data type and sample size.
4. **Compute the Test Statistic:** Use sample data to calculate the value.
5. **Determine the Critical Value or P-value:** Based on the test statistic's distribution.
6. **Make a Decision:**
   * If the test statistic exceeds the critical value or if the P-value is less than α, reject H₀.
   * Otherwise, fail to reject H₀.

**Chi Square Test with mathematical approach**

The chi-square (χ²) test is used to determine if there's a significant association between categorical variables. It compares observed frequencies in each category to expected frequencies under the assumption of no association.

### **Types of Chi-Square Tests**

1. **Chi-Square Goodness-of-Fit Test:** Checks if a sample data matches a population.
2. **Chi-Square Test for Independence:** Determines if there's an association between two categorical variables.

### **1. Chi-Square Goodness-of-Fit Test**

**Purpose:** Tests if the observed frequency distribution of a categorical variable differs from an expected distribution.

**Formula:**

χ² = Σ((Oᵢ - Eᵢ)² / Eᵢ)

* Oᵢ: Observed frequency for category i
* Eᵢ: Expected frequency for category i
* k: Number of categories

**Steps:**

1. **State the Hypotheses:**
   * H₀: The observed frequencies match the expected frequencies.
   * H₁: The observed frequencies do not match the expected frequencies.
2. **Calculate Expected Frequencies (Eᵢ):** Based on the hypothesized distribution.
3. **Compute the Chi-Square Statistic (χ²):** Using the formula above.
4. **Determine Degrees of Freedom (df):** df = k - 1.
5. **Find the Critical Value or P-value:** From the Chi-square distribution table.
6. **Make a Decision:** Compare χ² to the critical value.

**Example:**

Suppose we roll a die 60 times, and the outcomes are as follows:

| Face | Observed (Oᵢ) |
| --- | --- |
| 1 | 8 |
| 2 | 10 |
| 3 | 9 |
| 4 | 12 |
| 5 | 11 |
| 6 | 10 |

Expected frequency for each face if the die is fair: Eᵢ = 60/6 = 10.

Calculate χ²:

χ² = Σ((Oᵢ - 10)² / 10) = (8-10)²/10 + ... + (10-10)²/10

Compute and compare χ² to the critical value with df = 5.

### **2. Chi-Square Test for Independence**

**Purpose:** Determines if there's a significant association between two categorical variables.

**Formula:**

χ² = ΣΣ((Oᵢⱼ - Eᵢⱼ)² / Eᵢⱼ)

* Oᵢⱼ: Observed frequency in cell (i, j)
* Eᵢⱼ: Expected frequency in cell (i, j), calculated as:  
  Eᵢⱼ = (Row Totalᵢ \* Column Totalⱼ) / Grand Total
* r: Number of rows
* c: Number of columns

**Steps:**

1. **State the Hypotheses:**
   * H₀: The variables are independent.
   * H₁: The variables are dependent.
2. **Create a Contingency Table:** Organize the observed frequencies.
3. **Calculate Expected Frequencies (Eᵢⱼ):**
4. **Compute the Chi-Square Statistic (χ²):**
5. **Determine Degrees of Freedom (df):** df = (r - 1)(c - 1).
6. **Find the Critical Value or P-value:**
7. **Make a Decision:**

**Example:**

Suppose we have the following data on smoking habits by gender:

|  | Smoker (S) | Non-Smoker (NS) | Total |
| --- | --- | --- | --- |
| Male (M) | 40 | 60 | 100 |
| Female (F) | 30 | 70 | 100 |
| Total | 70 | 130 | 200 |

Calculate expected frequencies:

E\_MS = (Row Total\_M \* Column Total\_S) / Grand Total = (100 \* 70) / 200 = 35

Repeat for each cell, compute χ², determine df = (2 - 1)(2 - 1) = 1, and make a decision.

**ANOVA test with mathematical approach**

**Analysis of Variance (ANOVA)** is used to compare the means of three or more samples to determine if at least one sample mean is significantly different from the others.

### **One-Way ANOVA**

**Purpose:** Tests for significant differences among group means when there is one independent variable.

**Model:**

Yᵢⱼ = μ + τᵢ + εᵢⱼ

* Yᵢⱼ: Observation from group i, subject j
* μ: Overall mean
* τᵢ: Effect of treatment (group i)
* εᵢⱼ: Random error term, assumed εᵢⱼ ~ N(0, σ²)

**Hypotheses:**

* H₀: μ₁ = μ₂ = ... = μₖ
* H₁: At least one μᵢ differs

**ANOVA Table Components:**

1. **Total Sum of Squares (SST):**  
   SST = Σᵢ Σⱼ (Yᵢⱼ - Ȳ..)²  
   * Ȳ..: Grand mean of all observations
   * nᵢ: Number of observations in group i
2. **Between-Groups Sum of Squares (SSB):**  
   SSB = Σᵢ nᵢ (Ȳᵢ. - Ȳ..)²  
   * Ȳᵢ.: Mean of group i
3. **Within-Groups Sum of Squares (SSW):**  
   SSW = Σᵢ Σⱼ (Yᵢⱼ - Ȳᵢ.)²  
   * Note: SST = SSB + SSW

**Degrees of Freedom:**

* df\_Total = N - 1
* df\_Between = k - 1
* df\_Within = N - k
* N: Total number of observations

**Mean Squares:**

* MSB = SSB / df\_Between
* MSW = SSW / df\_Within

**F-Statistic:**

F = MSB / MSW

**Steps:**

1. **Calculate Group Means (Ȳᵢ.) and Grand Mean (Ȳ..)**.
2. **Compute SSB, SSW, and SST**.
3. **Calculate Degrees of Freedom**.
4. **Compute Mean Squares (MSB and MSW)**.
5. **Calculate the F-Statistic**.
6. **Determine the Critical F-Value** from the F-distribution table with df\_Between and df\_Within.
7. **Make a Decision:**
   * If F exceeds the critical value, reject H₀.
   * Otherwise, fail to reject H₀.

**Example:**

Suppose we have test scores from three different teaching methods:

* Method A: n₁ = 10, Ȳ₁. = 85
* Method B: n₂ = 10, Ȳ₂. = 80
* Method C: n₃ = 10, Ȳ₃. = 75
* Grand Mean (Ȳ..) = 80

Calculate SSB:

SSB = 10(85 - 80)² + 10(80 - 80)² + 10(75 - 80)² = 10(25) + 0 + 10(25) = 500

Calculate SSW using individual data (not provided here), compute MSB and MSW, calculate F, and make a decision.

**Dataset Description :**

**Link :** <https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset>

### **Context**

This data set dates from 1988 and consists of four databases: Cleveland, Hungary, Switzerland, and Long Beach V. It contains 76 attributes, including the predicted attribute, but all published experiments refer to using a subset of 14 of them. The "target" field refers to the presence of heart disease in the patient. It is integer valued 0 = no disease and 1 = disease.

### **Content**

Attribute Information:

1. age
2. sex
3. chest pain type (4 values)
4. resting blood pressure
5. serum cholestoral in mg/dl
6. fasting blood sugar > 120 mg/dl
7. resting electrocardiographic results (values 0,1,2)
8. maximum heart rate achieved
9. exercise induced angina
10. oldpeak = ST depression induced by exercise relative to rest
11. the slope of the peak exercise ST segment
12. number of major vessels (0-3) colored by flourosopy
13. thal: 0 = normal; 1 = fixed defect; 2 = reversable defect  
    The names and social security numbers of the patients were recently removed from the database, replaced with dummy values.

**ALGORITHM STEPS:**

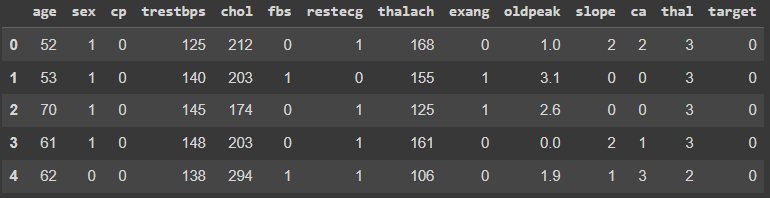
**Code:**

**import pandas as pd**

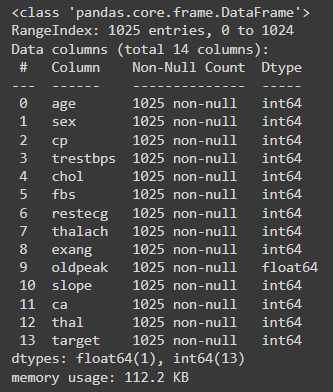
**import numpy as np**

**df = pd.read\_csv('/content/heart.csv')**

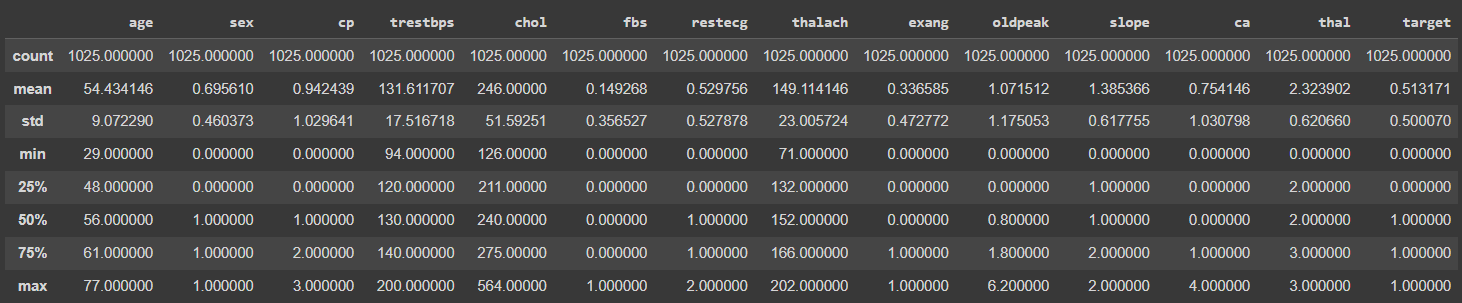
**df.head()**

****

**df.info()**

****

**df.describe()**

****

## 2 Sample T test

**Columns Taken:**

**trestbps (resting blood pressure)**

**thalach (maximum heart rate achieved)**

### Hypotheses

Null Hypothesis (H0): The means of trestbps for two independent groups (e.g., target = 0 and target = 1) are equal.

Alternative Hypothesis (H1): The means of trestbps for two independent groups are not equal.

**from scipy import stats**

**import pandas as pd**

**# Assuming df is your DataFrame**

**# Replace these lines with your actual data selection**

**group1 = df[df['target'] == 0]['trestbps']**

**group2 = df[df['target'] == 1]['trestbps']**

**# Perform the t-test**

**t\_stat, p\_val = stats.ttest\_ind(group1, group2, equal\_var=False)**

**# Set significance level**

**alpha = 0.05**

**# Print results**

**print("T-statistic:", t\_stat)**

**print("P-value:", p\_val)**

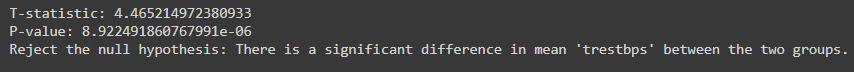
**# Make a decision**

**if p\_val <= alpha:**

**print("Reject the null hypothesis: There is a significant difference in mean 'trestbps' between the two groups.")**

**else:**

**print("Fail to reject the null hypothesis: There is no significant difference in mean 'trestbps' between the two groups.")**

****

## Chi Square Test

**Columns Taken:**

cp - chest pain type

target

import pandas as pd

from scipy import stats

# Assuming df is your DataFrame

# Create a contingency table

contingency\_table = pd.crosstab(df['cp'], df['target'])

# Perform the Chi-Square test

chi2\_stat, p\_val, dof, expected = stats.chi2\_contingency(contingency\_table)

# Set significance level

alpha = 0.05

# Print results

print("Chi-Square Statistic:", chi2\_stat)

print("P-value:", p\_val)

print("Degrees of Freedom:", dof)

print("Expected Frequencies:\n", expected)

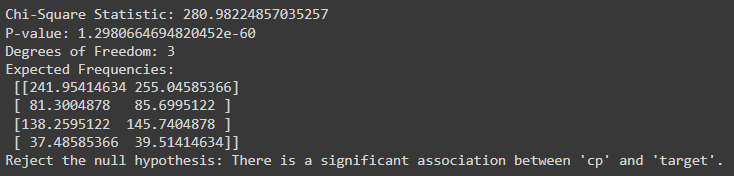
# Make a decision

if p\_val <= alpha:

print("Reject the null hypothesis: There is a significant association between 'cp' and 'target'.")

else:

print("Fail to reject the null hypothesis: There is no significant association between 'cp' and 'target'.")

****

**Conclusion:**

**Chi-Square Test Conclusion**: With a Chi-Square statistic of 280.98 and a p-value of 1.30e-60 (well below the 0.05 significance level), we reject the null hypothesis. This suggests a strong association between chest pain type (cp) and the presence of heart disease (target).

**T-Test Conclusion**: The t-test results, showing a t-statistic of 4.465 and a p-value of 8.92e-06 (significantly below 0.05), lead us to reject the null hypothesis. This indicates a meaningful difference in mean resting blood pressure (trestbps) between patients with and without heart disease.