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 (H_0)

- 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 (H_1 or H_a)

- 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 H_0 and H_1.
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 H_0.
   - Otherwise, fail to reject H_0.

Chi Square Test with mathematical approach

The chi-square (χ^2) 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:

$$\chi^2 = \Sigma((O_i - E_i)^2 / E_i)$$

- O_i: Observed frequency for category i
- E_i: Expected frequency for category i
- k: Number of categories

Steps:

1. **State the Hypotheses:**
   - H_0: The observed frequencies match the expected frequencies.
   - H_1: The observed frequencies do not match the expected frequencies.
2. **Calculate Expected Frequencies (E_i):** Based on the hypothesized distribution.
3. **Compute the Chi-Square Statistic (χ^2):** 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 χ^2 to the critical value.

Example:

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

Face	Observed (O_i)
1	8
2	10
3	9
4	12
5	11
6	10

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

Calculate χ^2:

$\chi^2 = \Sigma((O_i - 10)^2 / 10) = (8-10)^2/10 + ... + (10-10)^2/10$

Compute and compare χ^2 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:

$\chi^2 = \Sigma\Sigma((O_{ij} - E_{ij})^2 / E_{ij})$

- O_{ij}: Observed frequency in cell (i, j)
- E_{ij}: Expected frequency in cell (i, j), calculated as: $E_{ij} = (\text{Row Total}_i * \text{Column Total}_j) / \text{Grand Total}$

- r: Number of rows
- c: Number of columns

Steps:

1. **State the Hypotheses:**
   - H_0: The variables are independent.
   - H_1: The variables are dependent.
2. **Create a Contingency Table:** Organize the observed frequencies.
3. **Calculate Expected Frequencies (E_{ij}):**
4. **Compute the Chi-Square Statistic (χ^2):**
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 χ^2, 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:

$Y_{ij} = \mu + \tau_i + \varepsilon_{ij}$

- Y_{ij}: Observation from group i, subject j
- μ: Overall mean
- τ_i: Effect of treatment (group i)
- ε_{ij}: Random error term, assumed $\varepsilon_{ij} \sim N(0, \sigma^2)$

Hypotheses:

- H_0: $\mu_1 = \mu_2 = ... = \mu_k$
- H_1: At least one μ_i differs

ANOVA Table Components:

1. **Total Sum of Squares (SST):**
   $SST = \Sigma_i \Sigma_j (Y_{ij} - \bar{Y}..)^2$
   - $\bar{Y}..$: Grand mean of all observations
   - n_i: Number of observations in group i
2. **Between-Groups Sum of Squares (SSB):**
   $SSB = \Sigma_i n_i (\bar{Y}_i. - \bar{Y}..)^2$
   - $\bar{Y}_i.$: Mean of group i
3. **Within-Groups Sum of Squares (SSW):**
   $SSW = \Sigma_i \Sigma_j (Y_{ij} - \bar{Y}_i.)^2$
   - 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 ($\bar{Y}_i.$) and Grand Mean ($\bar{Y}..$)**.
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 H_0.
   - Otherwise, fail to reject H_0.

Example:

Suppose we have test scores from three different teaching methods:

- Method A: $n_1 = 10$, $\bar{Y}_1. = 85$
- Method B: $n_2 = 10$, $\bar{Y}_2. = 80$
- Method C: $n_3 = 10$, $\bar{Y}_3. = 75$
- Grand Mean ($\bar{Y}..$) = 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()
```

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	52	1	0	125	212	0	1	168	0	1.0	2	2	3	0
1	53	1	0	140	203	1	0	155	1	3.1	0	0	3	0
2	70	1	0	145	174	0	1	125	1	2.6	0	0	3	0
3	61	1	0	148	203	0	1	161	0	0.0	2	1	3	0
4	62	0	0	138	294	1	1	106	0	1.9	1	3	2	0

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1025 entries, 0 to 1024
Data columns (total 14 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   age       1025 non-null   int64
 1   sex       1025 non-null   int64
 2   cp        1025 non-null   int64
 3   trestbps  1025 non-null   int64
 4   chol      1025 non-null   int64
 5   fbs       1025 non-null   int64
 6   restecg   1025 non-null   int64
 7   thalach   1025 non-null   int64
 8   exang     1025 non-null   int64
 9   oldpeak   1025 non-null   float64
 10  slope     1025 non-null   int64
 11  ca        1025 non-null   int64
 12  thal      1025 non-null   int64
 13  target    1025 non-null   int64
dtypes: float64(1), int64(13)
memory usage: 112.2 KB
```

```
df.describe()
```

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
count	1025.000000	1025.000000	1025.000000	1025.000000	1025.00000	1025.000000	1025.000000	1025.000000	1025.000000	1025.000000	1025.000000	1025.000000	1025.000000	1025.000000
mean	54.434146	0.695610	0.942439	131.611707	246.00000	0.149268	0.529756	149.114146	0.336585	1.071512	1.385366	0.754146	2.323902	0.513171
std	9.072290	0.460373	1.029641	17.516718	51.59251	0.356527	0.527878	23.005724	0.472772	1.175053	0.617755	1.030798	0.620660	0.500070
min	29.000000	0.000000	0.000000	94.000000	126.00000	0.000000	0.000000	71.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	48.000000	0.000000	0.000000	120.000000	211.00000	0.000000	0.000000	132.000000	0.000000	0.000000	1.000000	0.000000	2.000000	0.000000
50%	56.000000	1.000000	1.000000	130.000000	240.00000	0.000000	1.000000	152.000000	0.000000	0.800000	1.000000	0.000000	2.000000	1.000000
75%	61.000000	1.000000	2.000000	140.000000	275.00000	0.000000	1.000000	166.000000	1.000000	1.800000	2.000000	1.000000	3.000000	1.000000
max	77.000000	1.000000	3.000000	200.000000	564.00000	1.000000	2.000000	202.000000	1.000000	6.200000	2.000000	4.000000	3.000000	1.000000

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.

```python
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.")
```

```
T-statistic: 4.465214972380933
P-value: 8.922491860767991e-06
Reject the null hypothesis: There is a 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'.")
```

```
Chi-Square Statistic: 280.98224857035257
P-value: 1.2980664694820452e-60
Degrees of Freedom: 3
Expected Frequencies:
 [[241.95414634 255.04585366]
 [ 81.3004878   85.6995122 ]
 [138.2595122  145.7404878 ]
 [ 37.48585366  39.51414634]]
Reject the null hypothesis: There is a 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.