Statistics & Statistical Analysis

Module-4

Part-2

Hypothesis Testing

Marks: 50

All questions must be answered in a .ipynb file.
Use the necessary Python libraries to perform the tasks, and include text cells to provide interpretations or explanations where required.

Question 1

Scenario:

A common guideline suggests that a healthy resting blood pressure is around 130 mm Hg. You want to investigate whether the average resting blood pressure (trestbps) of patients in this Heart Disease dataset is significantly different from this standard value.

Dataset Link: https://www.kaggle.com/datasets/redwankarimsony/heart-disease-data

You are required to:

- Clean the trestbps column by removing null values.
- Take a random sample of 30 patients (use random_state = 42).

(a) Type of Test (0.5 marks)

What type of hypothesis test will you use for this scenario?
Explain why this test is appropriate.

(b) Hypotheses Formulation (2 marks)

Formulate the hypotheses clearly:

- Null hypothesis (H_0)
- Alternative hypothesis (H_1)

(c) Assumptions & Distribution Check (2.5 marks)

Check the normality assumption of your sample data.

- Which specific test will you use to check normality?
- Perform and show the result of the test.
- Interpret whether the data meet the assumption of normality.

(d) Interpretation (5 marks)

Using the appropriate test, determine whether the average resting blood pressure (trestbps) of patients in this dataset is significantly different from the population mean of 130 mm Hg.

- Report the test statistic and p-value.
- State whether you reject or fail to reject H_0 at a 5% significance level.
- Write a short on the hypothesis test's conclusion in plain language

Question 2

Scenario:
You suspect that the maximum heart rate achieved during exercise (thalach) might differ between male and female patients. You decide to compare the thalach values for these two independent groups to determine whether there is a significant difference between them.

Dataset Link: https://www.kaggle.com/datasets/redwankarimsony/heart-disease-data

You are required to:

- Clean the thalach column by removing null values.
- Split the dataset into two groups based on gender (Male and Female).
- Take a random sample of 30 patients from each group (use random_state = 42).

(a) Type of Test (0.5 marks)
What type of hypothesis test will you use for this scenario?
Explain why this test is appropriate.

(b) Hypotheses Formulation (2 marks)
Formulate the hypotheses clearly:

- Null hypothesis (H_0)
- Alternative hypothesis (H_1)

(c) Assumptions & Distribution Check (2.5 marks)
Check the normality assumption for both male and female samples.

- Which specific test will you use to check normality?
- Perform and show the result of the test for both groups.
- Interpret whether the data for both groups meet the assumption of normality.

(d) Interpretation (5 marks)
Using the appropriate test, determine whether the mean maximum heart rate (thalach) of male and female patients differs significantly.

- Report the test statistic and p-value.
- State whether you reject or fail to reject H_0 at a 5% significance level.

- Write a short on the hypothesis test's conclusion in plain language.

Question 3

Scenario:
The dataset consists of monthly revenue figures (in thousands of dollars) for 100 stores. The revenue was recorded **before** and **after** a 4-week marketing campaign. The purpose of this dataset is to evaluate whether the marketing campaign was effective in increasing store revenue.

Dataset: Monthly Revenue (in thousands) - Sheet1.csv

You are required to:

- Clean the dataset by removing any null values in the Revenue_Before and Revenue_After columns.

- Select a random sample of 30 stores (use random_state = 42).

(a) Type of Test (0.5 marks)
What type of hypothesis test will you use for this scenario?
Explain why this test is appropriate.

(b) Hypotheses Formulation (2 marks)
Formulate the hypotheses clearly:

- Null hypothesis (H_0)

- Alternative hypothesis (H_1):

(c) Assumptions & Distribution Check (2.5 marks)
Check the normality assumption for the difference between Revenue_After and Revenue_Before.

- Which specific test will you use to check normality?

- Perform and show the result of the test.

- Interpret whether the difference data meets the assumption of normality.

(d) Interpretation (5 marks)
Using the appropriate test, determine whether the average revenue after the marketing campaign is significantly higher than the revenue before the campaign.

- Report the test statistic and p-value.
- State whether you reject or fail to reject H_0 at a 5% significance level.
- Write a short on the hypothesis test's conclusion in plain language, explaining whether the marketing campaign had a statistically significant effect on store revenue.

Question-4

Scenario:
A botanist wants to know if the average petal length differs across three Iris species (setosa, versicolor, virginica).

You are required to:

- Use the species and petal_length columns from the Iris dataset (available via seaborn).
- Take a random sample of 30 observations from each species (use random_state = 42).

(a) Type of Test (0.5 marks)
What type of hypothesis test will you use for this scenario?
Explain why this test is appropriate.

(b) Hypotheses Formulation (2 marks)

Formulate the hypotheses clearly:

- Null hypothesis (H_0)
- Alternative hypothesis (H_1)

(c) Assumptions & Distribution Check (2.5 marks)
Check the assumptions for One-Way ANOVA:

- Normality of petal lengths in each species group

- Perform and show the result of the test(s)

(d) Interpretation (5 marks)

- Conduct the One-Way ANOVA and report the F-statistic and p-value.
- If the ANOVA is significant, perform **Post Hoc Tukey HSD test** and interpret which species differ from each other.

- Write a short on the hypothesis test's conclusion in plain language about petal length differences among species.

Question-5

Scenario:
A health researcher wants to investigate whether smoking status is associated with the presence of cardiovascular disease in patients.

Dataset: cardio_data_processed.csv

The dataset contains the following columns:

- smoke (0 = Non-Smoker, 1 = Smoker)

- cardio (0 = No Disease, 1 = Disease)

You are required to:

- Load the dataset from 'cardio_data_processed.csv'.

- Select the relevant columns (smoke and cardio) and remove any missing values.

- Optionally, map the numeric values to readable labels: 0 = Non-Smoker/No Disease, 1 = Smoker/Disease.

- Create a **contingency table** showing counts of patients by smoking status and cardiovascular disease.

(a) Type of Test (2.5 marks)
What type of hypothesis test will you use for this scenario?
Explain why this test is appropriate.

(b) Hypotheses Formulation (2.5 marks)
Formulate the hypotheses clearly:

- Null hypothesis (H_0)
- Alternative hypothesis (H_1)

(c) Test Execution (2.5 marks)

- Report the **test statistic**, **p-value**, **degrees of freedom**, and **expected frequencies**.

(d) Interpretation (2.5 marks)

- State whether you reject or fail to reject H_0.
- Determine whether there is a significant association between smoking and cardiovascular disease at a 5% significance level.