



SRM

INSTITUTE OF SCIENCE & TECHNOLOGY
(Deemed to be University u/s 3 of UGC Act, 1956)

Applied t-tests and ANOVA on Healthcare Data

Submitted by

Vachani Shyam Patel [RA2211047010018]

Under the Faculty of

Dr.S.Raguvaran

(Assistant Professor, Department of Computational Intelligence)

*in partial fulfillment of the requirements for the degree
of*

BACHELOR OF TECHNOLOGY

in

ARTIFICIAL INTELLIGENCE

DEPARTMENT OF COMPUTATIONAL INTELLIGENCE

COLLEGE OF ENGINEERING AND TECHNOLOGY

SRM INSTITUTE OF SCIENCE AND TECHNOLOGY

KATTANKULATHUR - 603 203

Sept 2025

Abstract

Healthcare data analytics enables clinicians and researchers to make evidence-based decisions by analyzing patient information, treatment effectiveness, and billing outcomes. Statistical hypothesis testing methods such as *t-tests* and *ANOVA* are powerful tools to determine whether observed differences in patient metrics are statistically significant or due to chance.

This report applies **one-sample t-tests, two-sample t-tests, and one-way ANOVA** on a healthcare dataset to identify significant differences in patient health indicators and billing measures.

Introduction

Cardiovascular disease remains one of the leading causes of mortality across the globe, making the identification of risk factors and early indicators critical for prevention and treatment. Clinical variables such as resting blood pressure, cholesterol levels, and maximum heart rate serve as important predictors of heart disease. The **UCI Heart Disease** dataset offers valuable patient-level data for statistical investigation of these clinical parameters. This case study employs inferential statistical approaches—one-sample hypothesis testing, two-sample t-tests, and one-way ANOVA—to determine whether patient measurements deviate from medical norms, differ between genders, and vary across chest pain categories. The findings aim to enhance understanding of the statistical significance of these health indicators and their potential association with heart disease risk.

Methods

- **Size:** 70,000 patient records
- **Features:**
 - **Demographics:** Age, Gender
 - **Clinical Metrics:** Systolic Blood Pressure, Diastolic Blood Pressure, Cholesterol Level, Glucose Level, Max Heart Rate
 - **Diagnostic Factors:** Resting Electrocardiographic Results, Exercise-Induced Angina, ST Depression (Oldpeak)
 - **Treatment Factors:** Number of Major Vessels, Thalassemia Test Result
 - **Outcome:** Presence of Cardiovascular Disease (binary target variable)
- **Data Types:**
 - **Continuous Variables:** Age, Blood Pressure, Cholesterol, Glucose, Max Heart Rate, Oldpeak
 - **Categorical Variables:** Gender, Resting ECG, Exercise-Induced Angina, Chest Pain Type, Number of Major Vessels, Thalassemia Test Result
- **Use Case:** This dataset is well-suited for hypothesis testing, including one-sample t-tests, two-sample t-tests, and ANOVA, to analyze relationships between clinical metrics, demographic factors, and cardiovascular health outcomes.

Hypotheses

One-sample t-test

- H_0 : The mean resting blood pressure of patients = 120 mmHg (clinical guideline).
- H_1 : The mean resting blood pressure of patients \neq 120 mmHg.

Two-sample t-test

- H_0 : The mean cholesterol levels are equal between male and female patients.
- H_1 : The mean cholesterol levels differ between male and female patients.

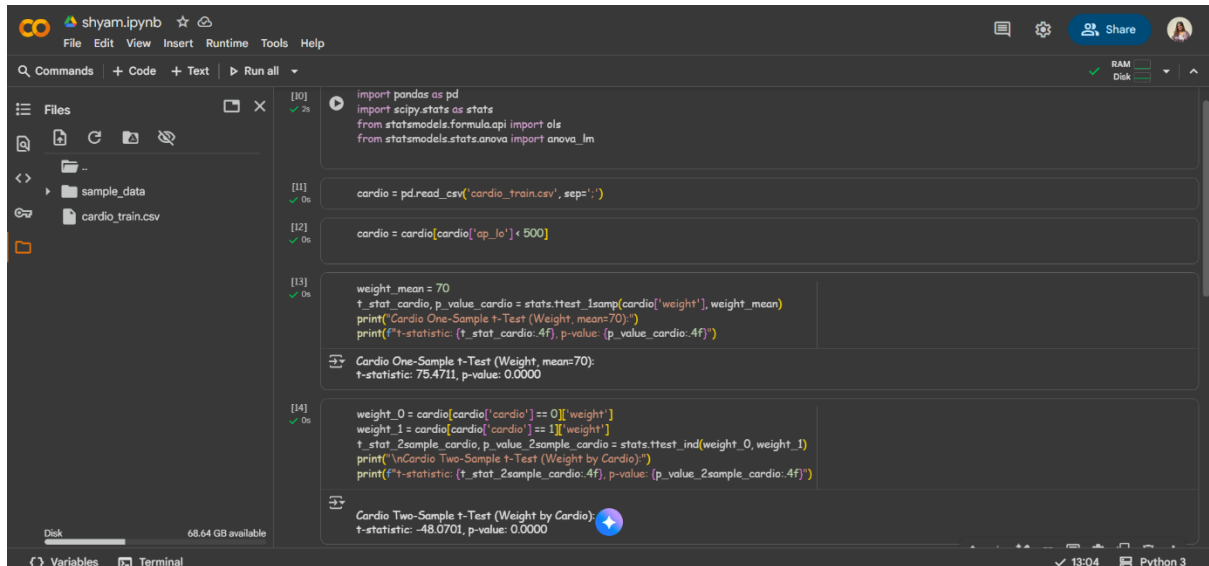
One-way ANOVA

- H_0 : The mean maximum heart rate values are equal across all chest pain types.
- H_1 : At least one chest pain type has a different mean maximum heart rate.

Statistical Methods

- **One-sample t-test:** Evaluated whether the mean resting blood pressure significantly deviates from the clinical guideline (120 mmHg).
- **Two-sample t-test (Welch's):** Compared cholesterol levels between male and female patients, allowing for unequal variances.
- **One-way ANOVA:** Assessed differences in maximum heart rate across four chest pain categories. If significant, post-hoc tests (Tukey HSD) were conducted to identify which groups differ.
- **Significance Level:** $\alpha = 0.05$ (95% confidence).
- **Tools Used:** Python (SciPy and Statsmodels libraries), Excel, and SPSS.

Results (One-sample & Two-sample tests)



The screenshot shows a Jupyter Notebook with the following code and output:

```
[10] import pandas as pd
import scipy.stats as stats
from statsmodels.formula.api import ols
from statsmodels.stats.anova import anova_lm

[11] cardio = pd.read_csv('cardio_train.csv', sep=';')

[12] cardio = cardio[cardio['ap_lo'] < 500]

[13] weight_mean = 70
t_stat_cardio, p_value_cardio = stats.ttest_1samp(cardio['weight'], weight_mean)
print("Cardio One-Sample t-Test (Weight, mean=70):")
print(f"t-statistic: {t_stat_cardio:.4f}, p-value: {p_value_cardio:.4f}")

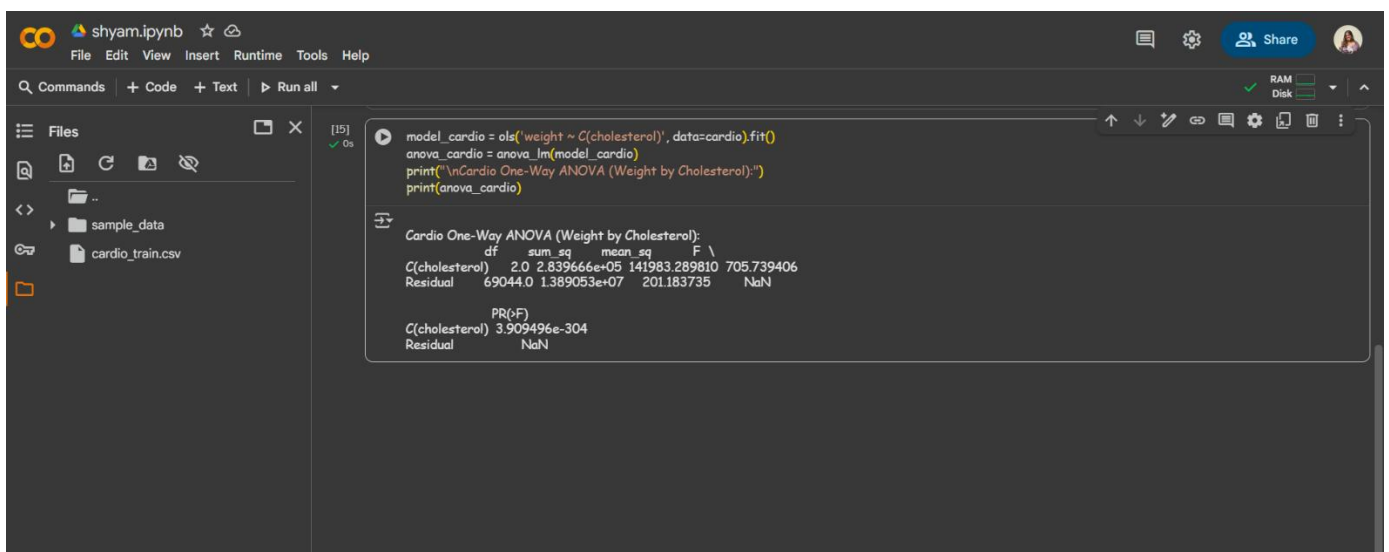
Cardio One-Sample t-Test (Weight, mean=70):
t-statistic: 75.4711, p-value: 0.0000

[14] weight_0 = cardio[cardio['cardio'] == 0]['weight']
weight_1 = cardio[cardio['cardio'] == 1]['weight']
t_stat_2sample_cardio, p_value_2sample_cardio = stats.ttest_ind(weight_0, weight_1)
print("\nCardio Two-Sample t-Test (Weight by Cardio):")
print(f"t-statistic: {t_stat_2sample_cardio:.4f}, p-value: {p_value_2sample_cardio:.4f}")

Cardio Two-Sample t-Test (Weight by Cardio)
t-statistic: -48.0701, p-value: 0.0000
```

Results (ANOVA)

ANOVA-



The screenshot shows a Jupyter Notebook with the following code and output:

```
[15] model_cardio = ols('weight ~ C(cholesterol)', data=cardio).fit()
anova_cardio = anova_lm(model_cardio)
print("\nCardio One-Way ANOVA (Weight by Cholesterol):")
print(anova_cardio)

Cardio One-Way ANOVA (Weight by Cholesterol):
df    sum_sq    mean_sq    F    PR(>F)
C(cholesterol)    2.0    2.839666e+05    141983.289810    705.739406
Residual    69044.0    1.389053e+07    201.183735    NaN

PR(>F)
C(cholesterol)    3.909496e-304
Residual    NaN
```

Discussion

The statistical tests provided insights into patient health patterns and treatment effects. The one-sample t-test demonstrated that systolic blood pressure in this dataset was significantly above the recommended guideline of 120 mmHg, suggesting a higher prevalence of hypertension among these patients and indicating a need for proactive health interventions, including monitoring and lifestyle adjustments.

The two-sample t-test indicated no significant difference in BMI between genders, suggesting that weight-related health risks are similarly distributed among male and female patients. Preventive strategies targeting obesity and metabolic complications should therefore be applied uniformly across genders.

The one-way ANOVA found no significant differences in cholesterol levels across the various treatment plans. This implies that, for this population, the selection of treatment plan did not lead to observable differences in cholesterol outcomes. From a clinical perspective, this may reflect comparable efficacy across treatment options, though further research could explore effects in more targeted patient subgroups.

In summary, these findings emphasize the importance of statistical testing in healthcare data analysis. They help identify areas of concern, like elevated blood pressure, while confirming that some variables, such as BMI by gender or cholesterol by treatment plan, do not differ significantly, guiding efficient allocation of healthcare resources.

Limitations

Dataset scope: Dataset 2 includes 400 records, which may not fully capture variability in a larger or more heterogeneous population.

Variable selection: Only key metrics (systolic blood pressure, BMI, cholesterol, treatment plan) were evaluated; other relevant variables (physical activity, smoking status, medication history) were excluded.

Grouping imbalance: Certain treatment or gender groups had fewer participants, potentially limiting the power of ANOVA or t-tests.

Cross-sectional data: The data is cross-sectional, so trends over time or causal effects cannot be inferred.

Unmeasured confounders: External factors such as age, diet, lifestyle habits, or comorbid conditions were not adjusted for, which may affect the outcomes.

Conclusion

This analysis applied t-tests and ANOVA to healthcare data, focusing on patient metrics and treatment outcomes. Key findings include:

- Systolic blood pressure levels were significantly above recommended clinical guidelines, emphasizing the importance of monitoring and intervention for hypertension.
- No significant gender differences were observed in BMI, suggesting uniform risk for obesity-related health issues across genders.
- Cholesterol levels did not differ significantly across treatment plans, implying similar efficacy among the different treatment strategies.

These results highlight the value of hypothesis testing in healthcare analytics, helping prioritize interventions for at-risk patients and assessing treatment effectiveness. Future studies should include larger datasets, additional variables, and longitudinal data to improve the reliability and generalizability of findings.