Statistical Analysis of Cardiovascular Indicators in the Heart Disease UCI Dataset

Name: Jacob Immanuel C
Reg. No.: RA2212701010008
Course: M.Tech Integrated AI
Subject: ISPA

Abstract

This report applies three inferential statistical tests—one-sample t-test, two-sample t-test, and one-way ANOVA with post-hoc analysis—to the Heart Disease UCI dataset to evaluate key cardiovascular parameters. Results indicate that (i) mean cholesterol levels differ significantly from the guideline value of 200 mg/dL, (ii) resting blood pressure differs significantly between male and female patients, and (iii) maximum heart rate varies significantly across chest pain types. These findings highlight the utility of hypothesis testing for uncovering differences between population subgroups in healthcare data.

1. Introduction

Heart disease, encompassing a range of conditions that affect the heart and blood vessels, remains one of the foremost causes of morbidity and mortality worldwide. According to the World Health Organization, cardiovascular diseases are responsible for millions of deaths each year, placing an enormous burden not only on health systems but also on individuals, families, and economies. Prevention and early detection are therefore critical components of global health strategies aimed at reducing the incidence and impact of these diseases.

To understand and manage cardiovascular risk, clinicians and researchers rely on a set of well-established indicators. Among the most important are serum cholesterol levels, resting blood pressure, and the heart's performance under physical stress or exertion. Elevated serum cholesterol is closely associated with the development of atherosclerosis and coronary artery disease. Resting blood pressure provides a direct measure of vascular health and resistance; sustained high blood pressure (hypertension) is a key risk factor for stroke and heart attack. Meanwhile, heart rate under stress or the maximum heart rate achieved during exertion offers insights into cardiac fitness and the presence of latent ischemic conditions. Together, these indicators form a multifaceted profile of cardiovascular health.

Yet simply measuring these indicators is not enough. To derive meaningful conclusions about whether a population or subgroup deviates from healthy norms, statistical analysis is essential. Hypothesis testing, a fundamental component of inferential statistics, allows researchers to move beyond descriptive summaries and determine whether observed differences or deviations are likely to be genuine or merely due to random chance. For example, a one-sample t-test can be used to compare a group's average cholesterol level against the recommended clinical guideline of 200 mg/dL. A two-sample t-test can determine whether men and women in a sample differ significantly in their average resting blood pressure. One-way analysis of variance (ANOVA) can assess whether maximum heart rate achieved differs among patients with different types of chest pain. In each case, the statistical test produces a

p-value, which quantifies the probability that the observed difference could occur if there were no true difference in the underlying population.

Applying these statistical techniques to a dataset such as the Heart Disease UCI dataset has multiple benefits. It helps to identify high-risk groups, understand gender- or symptom-related differences, and validate whether existing clinical guidelines are being met within a given cohort. This kind of evidence is invaluable for shaping public health policies, designing targeted interventions, and allocating healthcare resources more effectively. It also reinforces evidence-based medicine, where clinical decisions and preventive strategies are grounded in quantitative analysis rather than anecdote or assumption.

2. Dataset Description

The Heart Disease UCI dataset contains patient information including age, sex, chest pain type, resting blood pressure, serum cholesterol, maximum heart rate achieved, and target (presence of heart disease).

Key variables used in this study:

- **cholestoral (mg/dL)** – serum cholesterol level
- **resting_blood_pressure (mmHg)** – resting blood pressure
- **sex** – Male/Female
- **chest_pain_type** – Asymptomatic, Atypical angina, Non-anginal pain, Typical angina
- **Max_heart_rate (bpm)** – maximum heart rate achieved

3. Hypotheses and Methods

3.1 One-Sample t-Test

- *Null hypothesis (H_0):* Mean cholesterol = 200 mg/dL
- *Alternative hypothesis (H_1):* Mean cholesterol ≠ 200 mg/dL

3.2 Two-Sample t-Test

- *Null hypothesis (H_0):* Mean resting blood pressure (Male) = Mean resting blood pressure (Female)
- *Alternative hypothesis (H_1):* Means differ

3.3 One-Way ANOVA

- *Null hypothesis (H_0):* Mean Max_heart_rate is the same across all chest_pain_type groups
- *Alternative hypothesis (H_1):* At least one group differs

If ANOVA was significant, post-hoc Tukey HSD test was performed to identify specific group differences.

All analyses were conducted in Python using pandas, SciPy, and statsmodels.

4. Results

4.1 One-Sample t-Test (cholestoral vs 200 mg/dL)

- *T-statistic:* 28.545
- *p-value:* < 0.00001
- **Decision:** Reject H_0
- **Interpretation:** Mean cholesterol in this sample is significantly different from the guideline value of 200 mg/dL.

```
===== One-Sample T-Test (cholestoral vs 200 mg/dl) =====
T-statistic = 28.545, p-value = 0.00000
👉 Reject H0: Mean cholesterol is significantly different from 200 mg/dl.
```

4.2 Two-Sample t-Test (Resting BP: Male vs Female)

- *T-statistic:* –2.369
- *p-value:* 0.01822
- **Decision:** Reject H_0
- **Interpretation:** Resting blood pressure differs significantly between male and female patients.

```
===== Two-Sample T-Test (Resting BP: Male vs Female) =====
T-statistic = -2.369, p-value = 0.01822
👉 Reject H0: Significant difference in resting BP between males and females.
```

4.3 One-Way ANOVA (Max_heart_rate ~ chest_pain_type)

Source	sum_sq	df	F	p-value
Chest Pain Type	84,498.53	3	62.86	2.67×10^{-37}
Residual	457,467.12	1,021	–	–

- **Decision:** Reject H_0
- **Interpretation:** Maximum heart rate differs significantly across chest pain types.

```
===== One-Way ANOVA (Max_heart_rate ~ chest_pain_type) =====
                          sum_sq      df          F        PR(>F)
C(chest_pain_type)   84498.527469     3.0  62.862804  2.670722e-37
Residual            457467.117409  1021.0        NaN           NaN
👉 Reject H0: At least one chest_pain_type group differs in Max_heart_rate.

/tmp/ipython-input-3675392361.py:41: FutureWarning: Series.__getitem__ treating
  if anova_table['PR(>F)'][0] < 0.05:
===== Tukey HSD Post-Hoc Test =====
          Multiple Comparison of Means - Tukey HSD, FWER=0.05
=============================================================================
     group1           group2      meandiff p-adj   lower    upper   reject
-----------------------------------------------------------------------------
    Asymptomatic  Atypical angina   5.0498 0.3076  -2.4534  12.5529  False
    Asymptomatic Non-anginal pain  -2.0137 0.8807  -9.0121   4.9848  False
    Asymptomatic   Typical angina -17.0776    0.0 -23.7485 -10.4067   True
 Atypical angina Non-anginal pain  -7.0634 0.0036  -12.375  -1.7519   True
 Atypical angina   Typical angina -22.1273    0.0 -26.9992 -17.2554   True
Non-anginal pain   Typical angina -15.0639    0.0 -19.1156 -11.0122   True
-----------------------------------------------------------------------------
```

Tukey HSD Post-Hoc Test

- Typical angina vs all other groups: significantly lower Max_heart_rate.
- Atypical angina vs Non-anginal pain: significantly different.
- Asymptomatic vs others: mostly not significant except vs Typical angina.

These results indicate that the pattern of chest pain relates strongly to maximum heart rate achieved.

5. Discussion/Reports

The analysis demonstrates clear statistical differences in major cardiovascular indicators: cholesterol deviates from standard guidelines; men and women differ in resting blood pressure; and heart rate performance varies by chest pain presentation. These findings are consistent with clinical observations and support the utility of statistical hypothesis testing in exploratory health research.

6. Limitations

- The dataset may not represent all populations (sample bias).
- Some variables may have measurement error or missing data.
- Cross-sectional design limits causal inference.

7. Conclusion

Hypothesis testing provides an effective framework for identifying significant differences in health indicators across subgroups. The Heart Disease UCI dataset illustrates how clinical data can be interrogated to yield actionable insights about cardiovascular risk factors and patient characteristics.