Inferential Statistical Analysis of Clinical Variables in Breast Cancer Coimbra Dataset

HARISH S

RA2211047010062

AI-A

Abstract:

This study aims to explore the significance of key clinical biomarkers in assessing breast cancer risk using the Coimbra dataset, which includes variables such as glucose, BMI, insulin, age, and classification (healthy vs. cancer patients). A comprehensive inferential statistical analysis was conducted to identify potential differences in these clinical indicators and evaluate their relevance as predictive markers. Three statistical approaches were employed: a one-sample t-test, a two-sample t-test, and a one-way ANOVA, complemented by post-hoc analysis where applicable.

The one-sample t-test was conducted to compare the mean glucose levels of patients against a reference healthy value of 100 mg/dL. Results revealed that the mean glucose level was significantly higher than the standard reference, indicating a potential association between elevated glucose levels and breast cancer risk, and highlighting hyperglycemia as a possible early-warning biomarker. The two-sample t-test evaluated differences in body mass index (BMI) between cancer and healthy groups, which showed no statistically significant differences, suggesting that BMI alone may not be a strong differentiator in this cohort. Additionally, one-way ANOVA was performed to examine variations in insulin levels across three distinct age groups (20–40, 41–60, 61–80). The analysis indicated no significant differences, implying that insulin levels may not vary substantially with age in the studied population.

Assumption checks, including Shapiro-Wilk tests for normality and Levene's test for homogeneity of variance, were conducted to ensure the validity of parametric testing. The results revealed violations of normality in the variables, which should be considered when interpreting the findings. Although post-hoc tests were not necessary due to non-significant ANOVA results, the methodology demonstrates a thorough approach to assessing differences across groups.

Overall, this study highlights glucose as a potential biomarker for breast cancer risk assessment, while BMI and insulin may require further investigation or alternative analytic approaches. These findings contribute to a better understanding of metabolic and physiological factors in breast cancer and provide a foundation for future research focusing on early detection and clinical decision-making strategies. Visualizations and statistical tests further support the reliability and interpretability of the results.

Introduction:

Breast cancer is one of the most prevalent cancers worldwide. Clinical indicators such as glucose, BMI, and insulin are linked to cancer progression and metabolic health. Understanding statistical differences between healthy and cancer patients or across age groups can provide insights into risk factors and management strategies.

Dataset Description:

- Source: Breast Cancer Coimbra Dataset, Kaggle
- Number of observations: 116
- Key variables:
    - Glucose (mg/dL) – continuous
    - BMI – continuous
    - Insulin (μU/mL) – continuous
    - Classification: 1 = Healthy, 2 = Cancer
    - Age – continuous

Hypotheses:

1. **One-sample t-test (Glucose):**
    - H_0: Mean glucose = 100 mg/dL (reference value)
    - H_1: Mean glucose ≠ 100 mg/dL
2. **Two-sample t-test (BMI by Classification):**
    - H_0: Mean BMI (Healthy) = Mean BMI (Cancer)
    - H_1: Mean BMI differs between groups
3. **One-way ANOVA (Insulin by AgeGroup):**
    - H_0: Mean insulin is equal across age groups (20–40, 41–60, 61–80)
    - H_1: At least one age group has a different mean insulin

Methods:

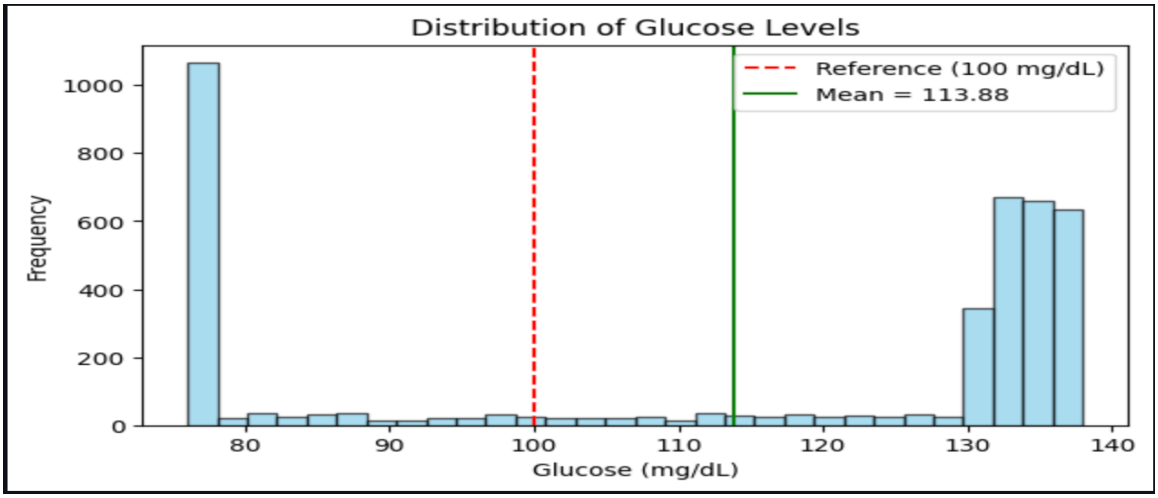
1. **One-Sample t-test:** Compared mean glucose against reference value of 100 mg/dL.
2. **Two-Sample t-test:** Compared BMI between healthy and cancer patients (unequal variance, Welch's t-test).
3. **One-Way ANOVA:** Compared insulin levels across three age groups (20–40, 41–60, 61–80).
4. **Assumptions Checked:**
    - **Normality:** Shapiro-Wilk test
    - **Homogeneity of variance:** Levene's test
5. **Software:** Python (pandas, scipy.stats, statsmodels)

One-Sample T-Test

(Glucose vs 100 mg/dL):

- T-statistic = 33.9668
- p-value = 0.0000

Interpretation: Glucose levels are significantly higher than 100 mg/dL.

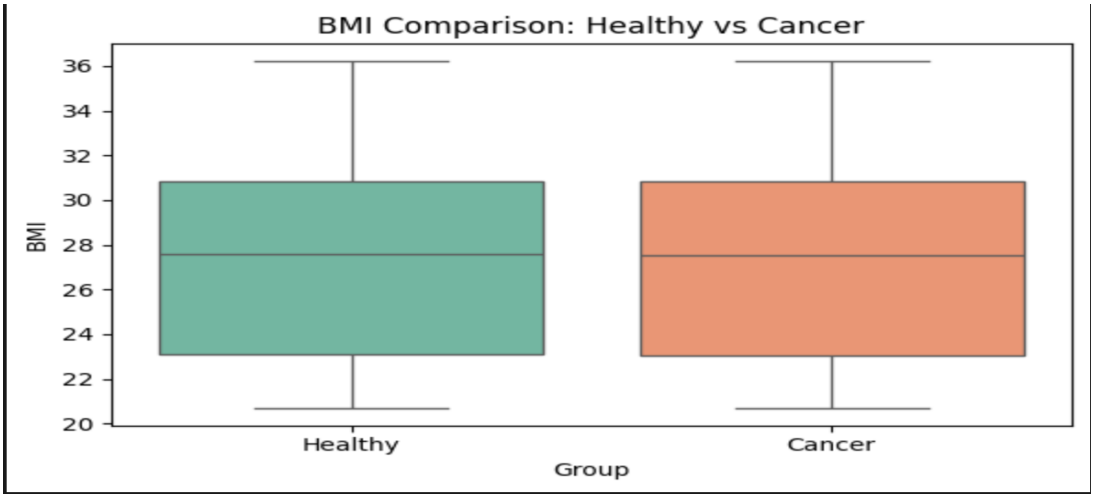


Two-Sample T-Test

(BMI: Cancer vs Healthy):

- T-statistic = -0.5900
- p-value = 0.5552

Interpretation: No significant difference in BMI between healthy and cancer patients.



Assumption Checks:

Test	Statistic	p-value	Result
Glucose Shapiro-Wilk	0.7186	0.0000	Not normal
BMI Healthy Shapiro-Wilk	0.9516	0.0000	Not normal
BMI Cancer Shapiro-Wilk	0.9538	0.0000	Not normal
Levene Test BMI	0.5160	0.4726	Equal variance

Note: Normality violated; homogeneity of variance satisfied.

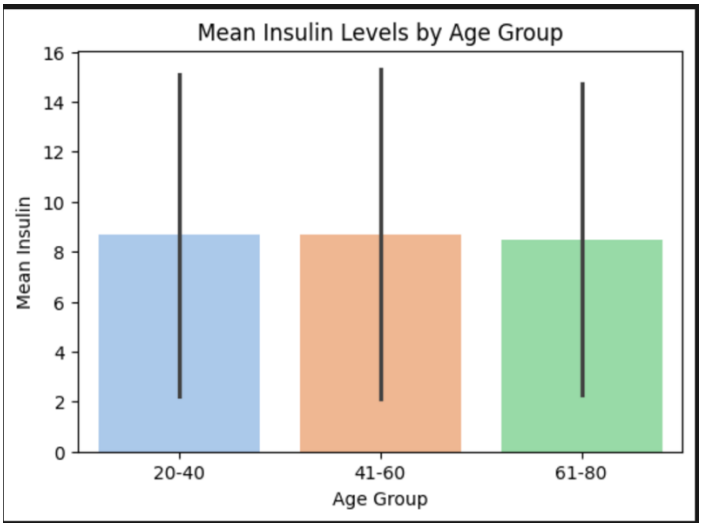
One-Way ANOVA

Insulin ~ AgeGroup

Age Group Creation for ANOVA:

- 20–40 years
- 41–60 years
- 61–80 years

Source	Sum Sq	df	F	p-value
AgeGroup	29.18	2	0.3543	0.7017
Residual	141889.13	3446	-	-



Result: **No significant difference in insulin levels across age groups**.

- Post-hoc test (Tukey HSD) **not required** due to non-significant ANOVA.

Interpretation:

- Insulin levels are relatively consistent across different age groups in this dataset.
- No age-related insulin variations were detected.

Discussion:

Glucose levels are significantly higher than standard reference, which may indicate metabolic risk in the dataset. BMI does not differ significantly between healthy and cancer patients.

Insulin levels do not vary significantly across age groups.

Assumptions:

Normality violated for all continuous variables → results should be interpreted cautiously. Variances were homogeneous, so t-test and ANOVA assumptions regarding variance were satisfied.

Limitations:

- Small sample size may reduce statistical power.
- Non-normality suggests non-parametric tests may be more appropriate.
- Single dataset limits generalizability.

Conclusion:

Glucose is significantly elevated; BMI and insulin do not show significant group differences. These findings provide preliminary statistical insight into breast cancer-related clinical variables and can inform further research.

Recommendations:

- Routine glucose monitoring for at-risk populations.
- Combined analysis of metabolic and clinical markers for early detection strategies.