21AIC401T- Inferential Statistics and Predictive Analytics

Analyzing Health Indicators: A Statistical Exploration of the Pima Indians Diabetes Dataset

Krishna Palod – RA2212701010034

Abstract

This case study applies inferential statistics to the Pima Indians Diabetes dataset to test whether clinical variables differ significantly from medical guidelines and across patient subgroups. Three analyses were conducted: a one-sample t-test comparing mean glucose with the WHO diagnostic threshold of 126 mg/dL, a Welch two-sample t-test comparing BMI between diabetic and non-diabetic women, and a one-way ANOVA comparing glucose across age categories (≤30, 31–50, >50) with Tukey HSD post-hoc tests.

Results indicated that average glucose was significantly lower than the 126 mg/dL threshold ($p < 0.001$). BMI was significantly higher in diabetic patients compared to non-diabetic patients ($p < 0.001$). ANOVA revealed significant glucose differences across age groups ($p < 0.001$), with Tukey tests confirming significant differences between all pairs of groups. These findings highlight the importance of BMI and age as predictors of diabetes risk and demonstrate the application of inferential statistics in predictive health analytics.

Introduction

Diabetes is a chronic disorder marked by abnormal glucose metabolism and associated with obesity, age, and genetic risk. According to WHO, fasting glucose ≥126 mg/dL is diagnostic for diabetes. Body mass index (BMI) is strongly linked to diabetes risk, and age significantly influences metabolic function.

This study applies inferential statistics to the Pima Indians Diabetes dataset (768 records) to address three core questions:

1. Is the mean glucose level significantly different from the WHO threshold?
2. Does BMI differ significantly between diabetic and non-diabetic women?
3. Do glucose levels vary significantly across age groups

Dataset Description

- Source: Pima Indians Diabetes dataset (UCI Repository / Kaggle)
- Size: 768 women, 9 variables

Variables used:
- Glucose (mg/dL)
- BMI (kg/m^2)
- Age (years) – categorized into Young (≤30), Middle (31–50), Old (>50)
- Outcome (0 = Non-diabetic, 1 = Diabetic)

Descriptive Statistics:
Glucose: Mean = 120.89, SD = 31.97
BMI: Mean = 32.45, SD = 6.88
Age: Mean = 33.2, SD = 11.8

Hypotheses

One-Sample t-test (Glucose vs WHO cutoff)
- H_0: μ_glucose = 126 mg/dL
- H_1: μ_glucose ≠ 126 mg/dL

Two-Sample t-test (BMI by Outcome)
- H_0: μ_BMI_non-diabetic = μ_BMI_diabetic
- H_1: μ_BMI_non-diabetic ≠ μ_BMI_diabetic

One-Way ANOVA (Glucose by AgeGroup)
- H_0: μ_glucose equal across age groups
- H_1: At least one group differs significantly

Methods

- Tools: Python (pandas, numpy, scipy.stats, statsmodels)
- Tests: One-sample t-test, Welch's t-test, One-way ANOVA with Tukey HSD post-hoc
- Assumptions:
  - Normality: Approximate normal distribution of continuous variables

- Independence: Each patient is independent
- Equal variances: Welch's t-test used to relax equal variance assumption

Results

1. One-Sample t-test (Glucose vs WHO 126 mg/dL)

Problem Statement: Is the mean fasting glucose level in this population significantly different from the WHO diagnostic threshold of 126 mg/dL?

Result: t(767) = -4.425, p < 0.001. Mean glucose = 120.89 mg/dL, SD = 31.97. Conclusion: Mean glucose is significantly lower than 126 mg/dL.

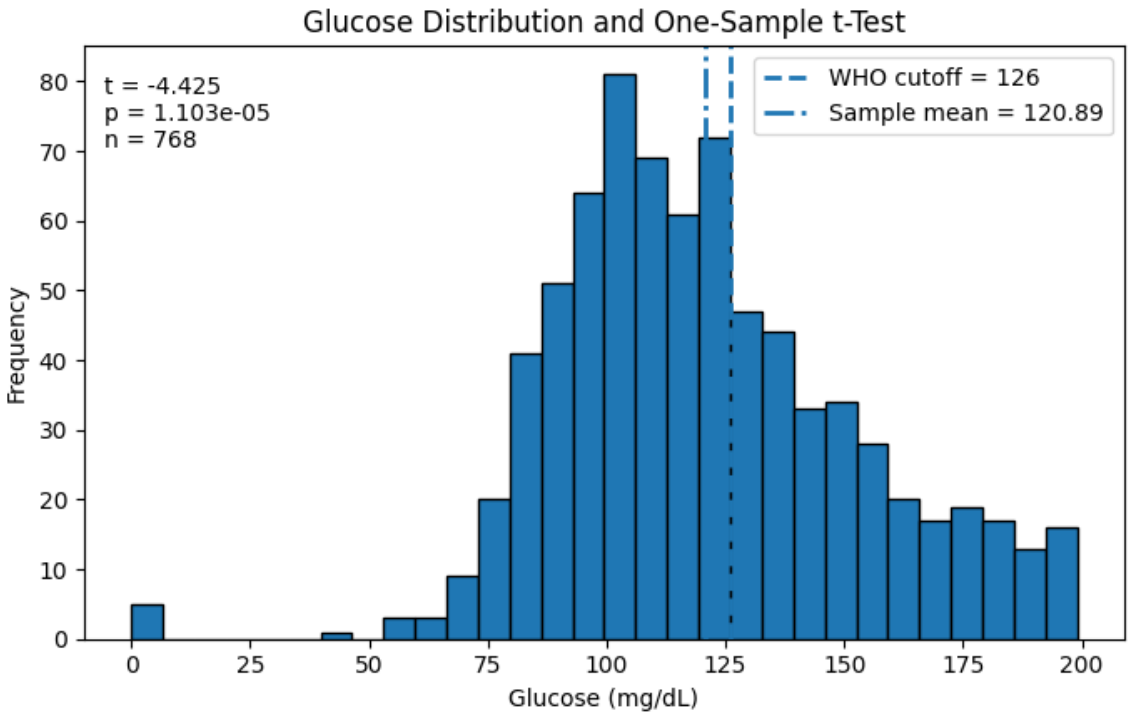

Fig 1.1: Glucose histogram with WHO cutoff (One-sample t-test)

2. Two-Sample t-test (BMI by Outcome)

Problem Statement: Is there a significant difference in BMI between diabetic and non-diabetic women?

Result: t = -8.619, p < 0.001. Mean BMI (non-diabetic) = 30.30, Mean BMI (diabetic) = 35.14. Conclusion: Diabetic women have significantly higher BMI than non-diabetic women.

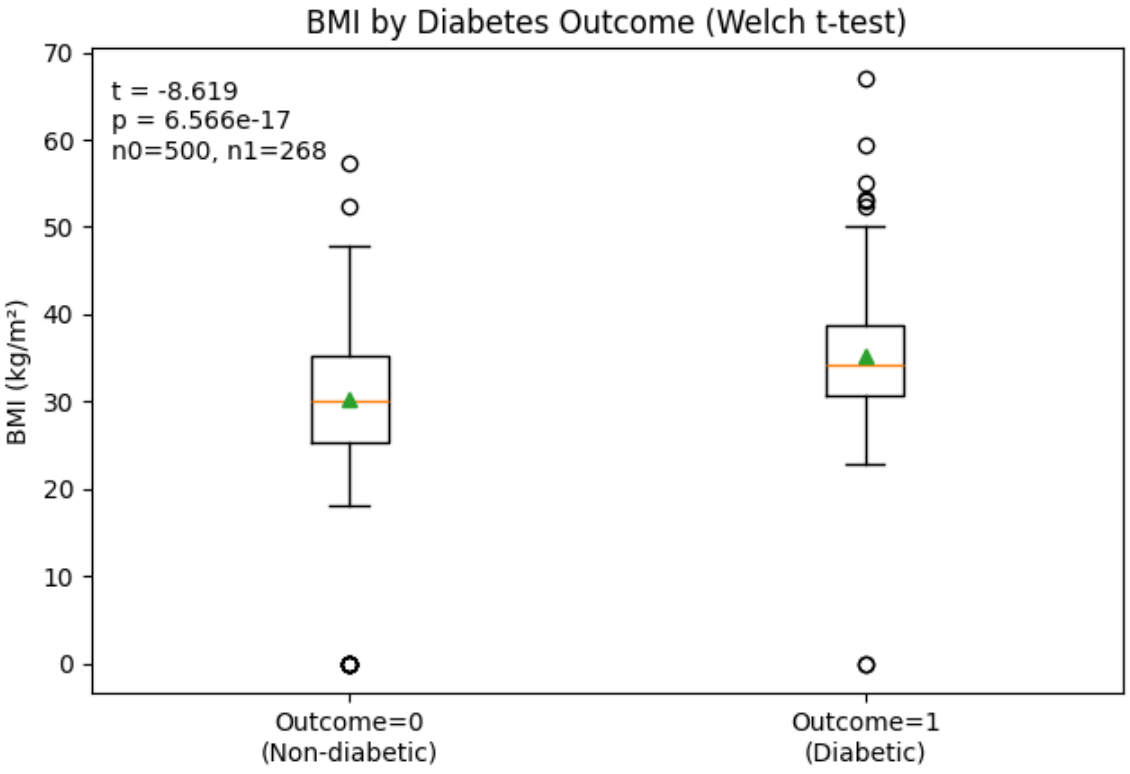

Fig 2.1: Boxplot of BMI by outcome (Two-sample t-test)

3. One-Way ANOVA (Glucose by Age Group)

Problem Statement: Do mean glucose levels differ across three age categories (≤30, 31–50, >50)?

Result: F(2,765) = 28.001, p < 0.001. Group means: Young = 114.18, Middle = 125.64, Old = 139.68. Tukey post-hoc confirmed significant differences between all groups.

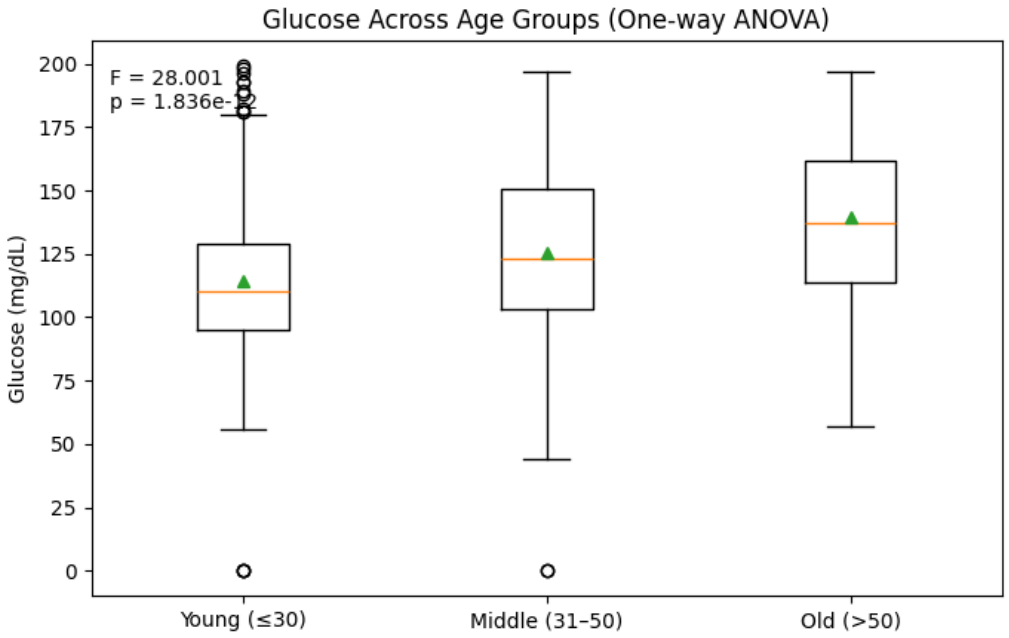

Fig 3.1: Glucose by age group (ANOVA)

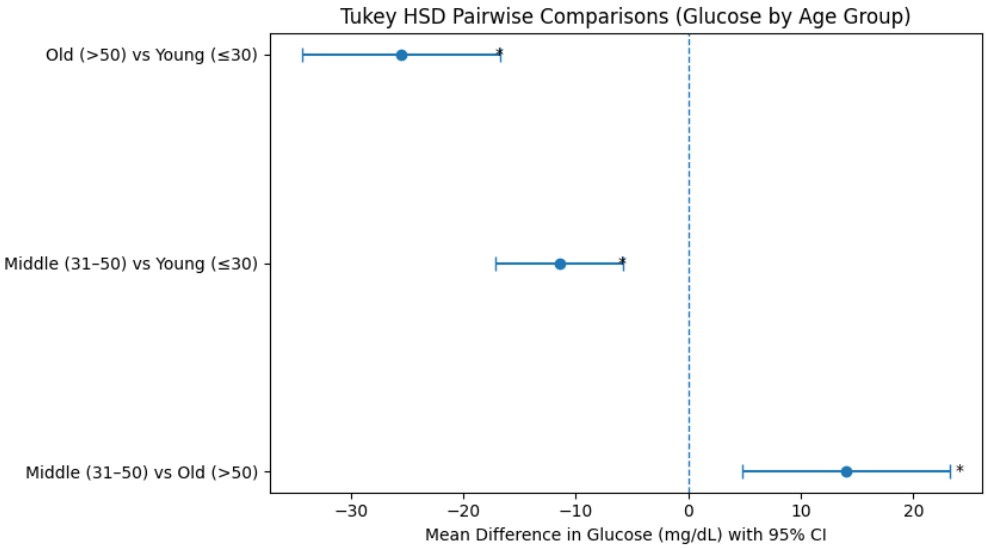

Fig 3.2: Tukey HSD forest plot

Discussion

- Glucose: Mean glucose is below the diagnostic threshold, but variability suggests subgroups at elevated risk.
- BMI and Diabetes: BMI strongly differentiates diabetic from non-diabetic patients, reinforcing obesity as a key predictor.
- Age and Glucose: Older age correlates with higher glucose levels, aligning with metabolic decline.

Clinical Implications:
- Age and BMI are crucial for risk stratification.
- Early lifestyle interventions in high-BMI groups may reduce diabetes onset.
- Inferential statistics provide a foundation for predictive healthcare analytics.

Limitations

- Dataset limited to Pima Indian women → limited generalizability.
- Missing values and zeros may bias estimates.
- Only univariate tests used; multivariate regression could yield deeper insights.

Conclusion

This study confirmed:
1. Mean glucose is significantly below the WHO threshold (with high individual variability).
2. BMI is significantly higher in diabetic patients.
3. Glucose significantly increases with age.

Inferential statistics applied to healthcare datasets can inform risk stratification and preventive strategies. Integrating such analyses into predictive modeling can enhance personalized medicine.