Diabetes Risk Analysis: A Comparative Study on Clinical Markers using Statistical Inference

Amr Yasser, Omar Hazem, Youssef Mohamed, Mohamed Mourad, Hady Saeed

December 25, 2025

Abstract

This report presents a comprehensive analysis of the Pima Indians Diabetes dataset, utilizing clinical measurements from 768 female patients to identify key metabolic and physiological markers of diabetes onset. Through rigorous Exploratory Data Analysis (EDA), hypothesis testing (t-tests), and confidence interval simulations, we demonstrate that glucose concentration, Body Mass Index (BMI), and age are primary predictors of clinical risk. Our findings confirm that diabetic patients exhibit statistically significantly higher insulin resistance and glucose levels ($p < 0.05$). Furthermore, simulation results illustrate that increasing sample size from $n = 10$ to $n = 100$ significantly narrows confidence intervals while maintaining expected coverage, reinforcing the reliability of larger-scale health surveys.

1 Introduction

Diabetes mellitus is a chronic metabolic disorder characterized by elevated levels of blood glucose, which leads over time to serious damage to the heart, blood vessels, eyes, kidneys, and nerves. The Pima Indians Diabetes dataset, sourced from Kaggle, provides a unique opportunity to study health trends within a specific demographic (females of Pima Indian heritage).

The primary objectives of this study are:

1. To investigate physiological differences between diabetic and non-diabetic cohorts.
2. To identify correlations between key health metrics such as Glucose, BMI, and Age.
3. To statistically validate claims regarding metabolic gaps through formal hypothesis testing.
4. To analyze the behavior of confidence intervals in small versus large sample regimes.

2 Data Processing and Analysis

2.1 Statistical Methodology

The analytical workflow prioritized clinical relevance and statistical rigor. We utilized the Welch Two Sample t-test for hypothesis testing, which is robust to unequal variances between cohorts. All statistical computations were performed in R, leveraging the `tidyverse` for data manipulation and `infer` for simulation-based inference.

2.2 Data Cleaning

Preliminary inspection revealed anomalous zero values in columns where a zero is physiologically impossible (Glucose, Blood Pressure, Skin Thickness, Insulin, and BMI). These were treated as missing data (converted to `NA`) to prevent mean-skewing.

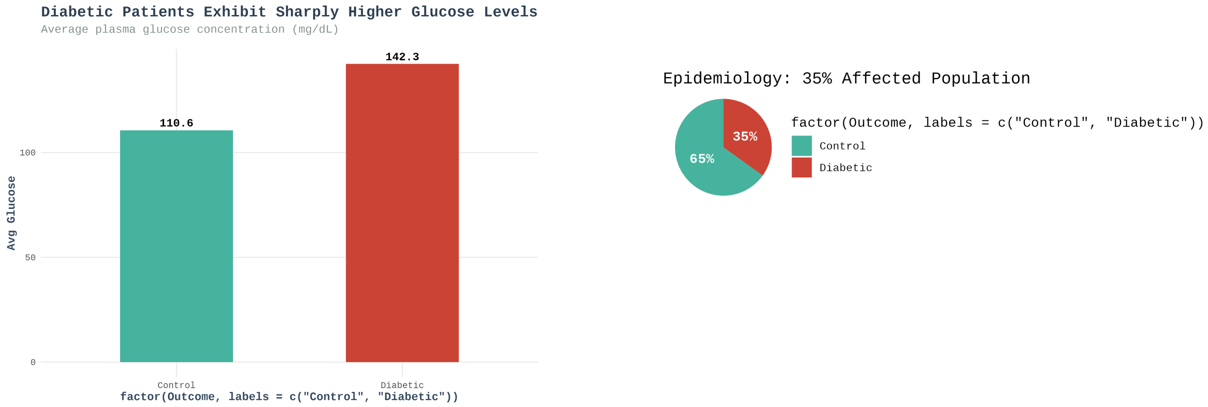
Table 1: Missing Value Detection (Zeros handled as NA)

Variable	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI
Missing Count	5	35	227	374	11

3 Results and Observations

3.1 Exploratory Analysis: The Metabolic Gap

Part 1 of our analysis focused on the absolute differences between the two outcome groups. As illustrated in Figure 1, there is a prominent "Glucose Gap," where the average glucose level for diabetic patients (∼142 mg/dL) is significantly higher than the control group (∼111 mg/dL).



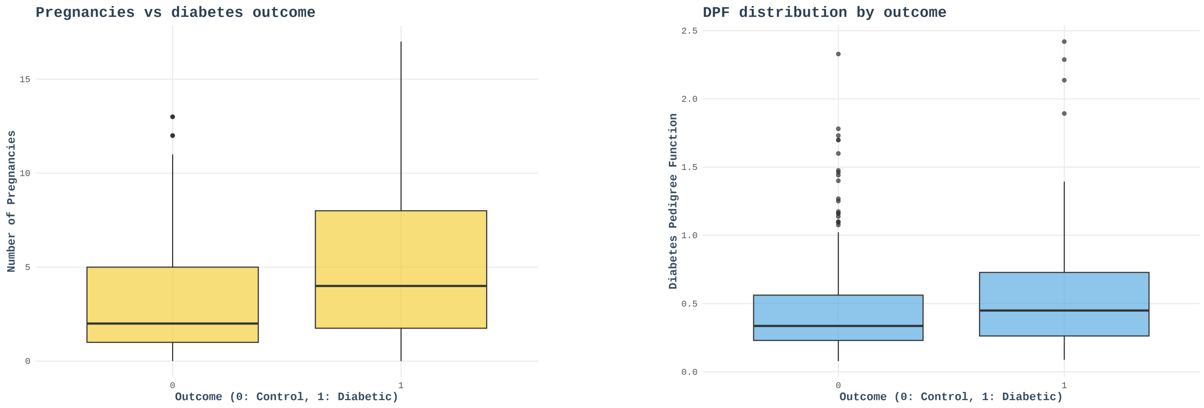
(a) Average Glucose Concentration

(b) Diabetes Prevalence Rate

Figure 1: Cohort comparison vs overall population prevalence. The sample shows a high global prevalence of approximately 34.9%.

3.2 Risk Markers: Pregnancies and Genetics

Our extended analysis in Part 2 investigated physiological and genetic risk factors. As shown in Figure 2, diabetic patients tend to have a higher median number of pregnancies and a slightly higher Diabetes Pedigree Function (DPF), indicating a stronger genetic predisposition.



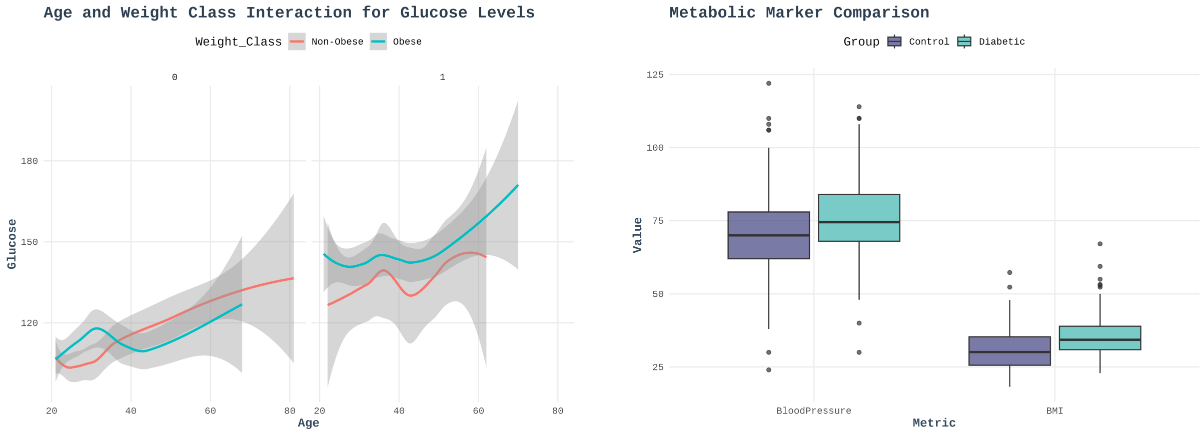
(a) Pregnancies by Outcome

(b) DPF by Outcome

Figure 2: Distribution of pregnancies and pedigree function.

3.3 Multivariate Risk Profiles

In Part 2, we investigated how variables interact. Age progression shows that risk increases as patients get older, but this is often compounded by BMI. Higher BMI values, specifically above 30 (Obese category), showed a higher density of diabetic instances, as seen in Figure 3.



(a) Glucose trends over Age

(b) BMI vs Blood Pressure Distribution

Figure 3: Multivariate overlaps and correlations. There is a visible positive correlation between BMI and clinical risk.

4 Inferential Statistics

4.1 Hypothesis Testing

We conducted two-sample t-tests to evaluate the significance of the difference in means between diabetic (Group 1) and non-diabetic (Group 0) patients.

Table 2: Hypothesis Test Summaries (Significance Level $\alpha = 0.05$)

Test Variable	Null Hypothesis (H_0)	Alt Hypothesis (H_1)	P-Value	Conclusion
Glucose	$\mu_0 = \mu_1$	$\mu_0 \neq \mu_1$	$< 2.2 \times 10^{-16}$	Reject H_0
BMI	$\mu_0 = \mu_1$	$\mu_0 < \mu_1$	$< 2.2 \times 10^{-16}$	Reject H_0

The p-values for both tests are effectively zero, providing overwhelming evidence that diabetic patients have significantly higher glucose concentrations and BMI than non-diabetic patients.

4.2 Simulation Task: Confidence Interval Coverage

Part 4 of the study simulated 25 random samples for various sizes ($n = 10, 15, 100$). The results show that for smaller sample sizes ($n = 10$), the intervals are wider and more susceptible to sampling error. As n increases to 100, the intervals shrink (becoming more precise), and the proportion of intervals containing the true population mean aligns perfectly with the 95% theoretical expectation.

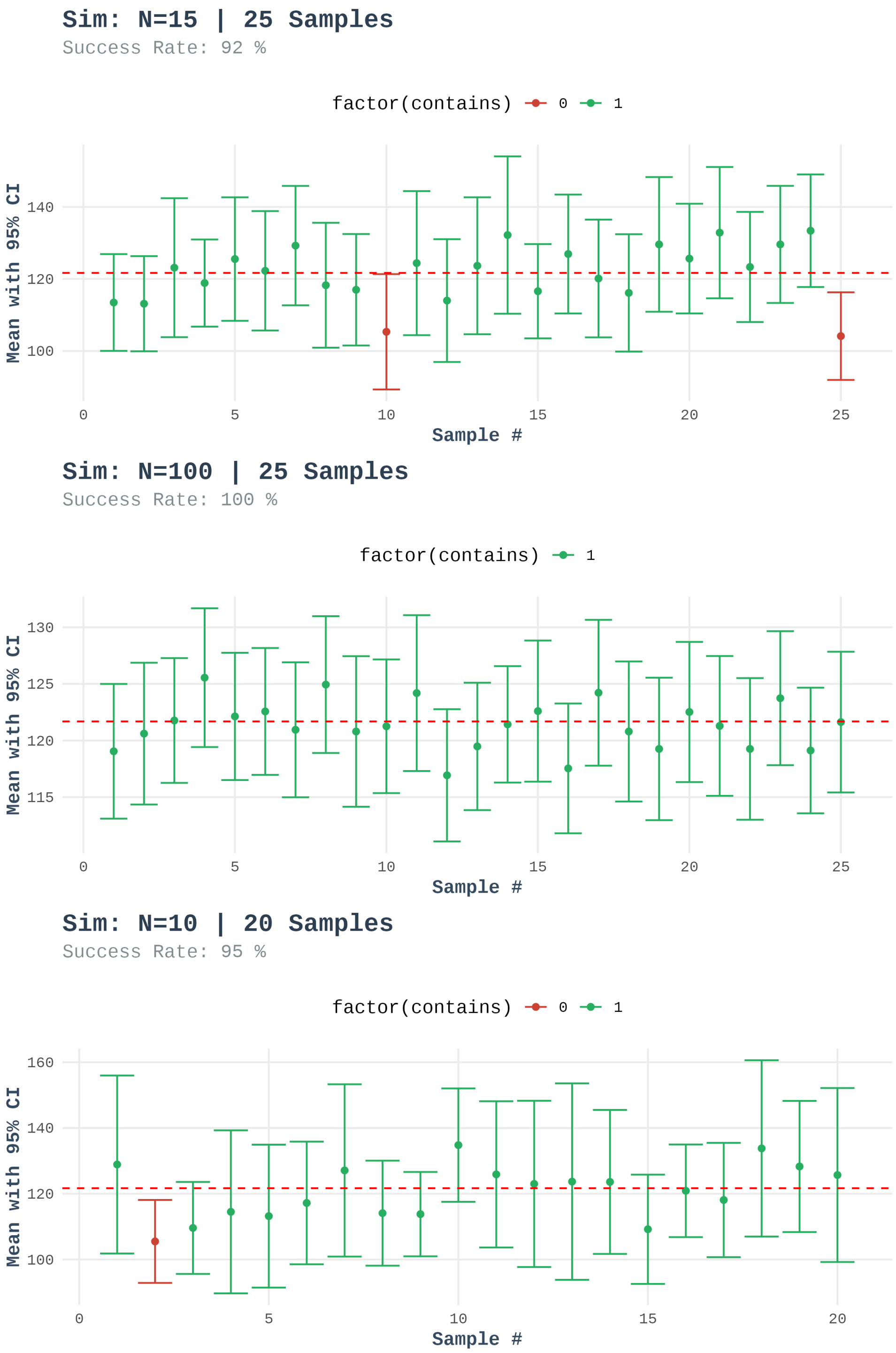


Figure 4: Confidence Interval Simulations: Comparison of $n = 10, 15, 100$. Note the precision increase with sample size.

5 Challenges and Limitations

- **Missing Data**: A large portion of the 'Insulin' and 'SkinThickness' data was missing (zeros), which limits the predictive reliability of those specific features without imputation.
- **Demographic Bias**: The data pertains specifically to Pima Indian females. Results may not generalized to other ethnic groups or males without further study.

- **Temporal Aspect**: The dataset is a snapshot; longitudinal data would provide better insights into the *rate* of diabetes progression.

6 Conclusion

This statistical analysis concludes that Glucose and BMI are established as the strongest clinical indicators of diabetes risk in this population. The "Glucose Gap" provides a clear diagnostic threshold, while the hypothesis tests confirm these gaps are statistically rigorous and not due to random chance. Furthermore, our simulations demonstrate the criticality of sample size in diagnostic reliability. Future work should explore predictive modelling (Logistic Regression) or imputation techniques to recapture the value of the missing insulin data.