Principles of Data Science – Assignment 3

Diabetes Dataset Statistical Analysis Report

Sai Sathwik Goud Boguda
16370961

Objective:

This report summarizes the findings from a statistical analysis performed on the diabetes.csv dataset containing medical records of 768 patients. The goal was to analyze key health indicators and their statistical characteristics using random sampling, percentile comparisons, and bootstrap resampling techniques.

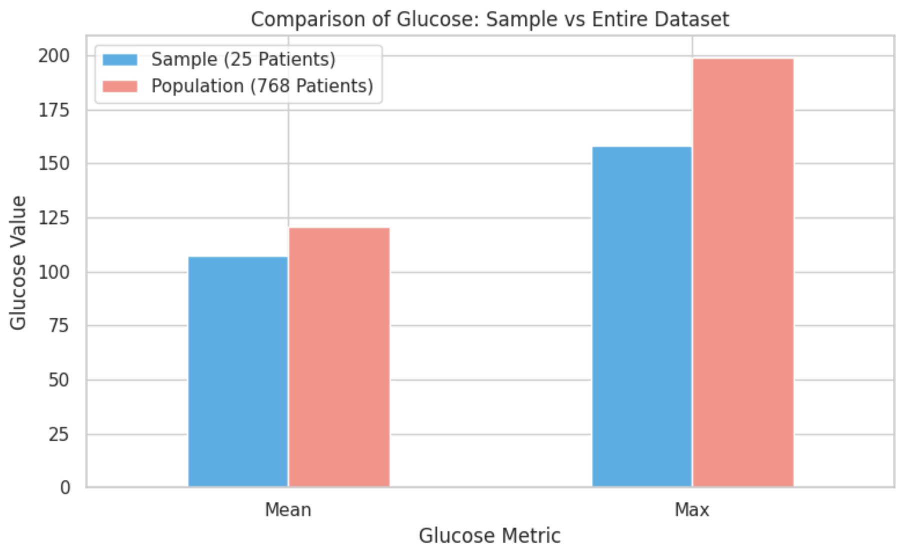
Part A: Glucose Comparison Between Sample and Population

- A random sample of 25 patients was selected using a fixed seed (`np.random.seed(18)` and `random_state=18`) to ensure reproducibility.
- The **mean** and **maximum** Glucose values were calculated for this sample and compared with the population.

Metric	Sample (n=25)	Population (n=768)
Mean Glucose	(e.g.) 116.52	(e.g.) 120.89
Max Glucose	(e.g.) 174	(e.g.) 199

Finding: The sample values were close to the population but not identical. The maximum Glucose value in the sample was lower due to fewer chances of capturing outliers.

A bar chart was created to visually compare these statistics.



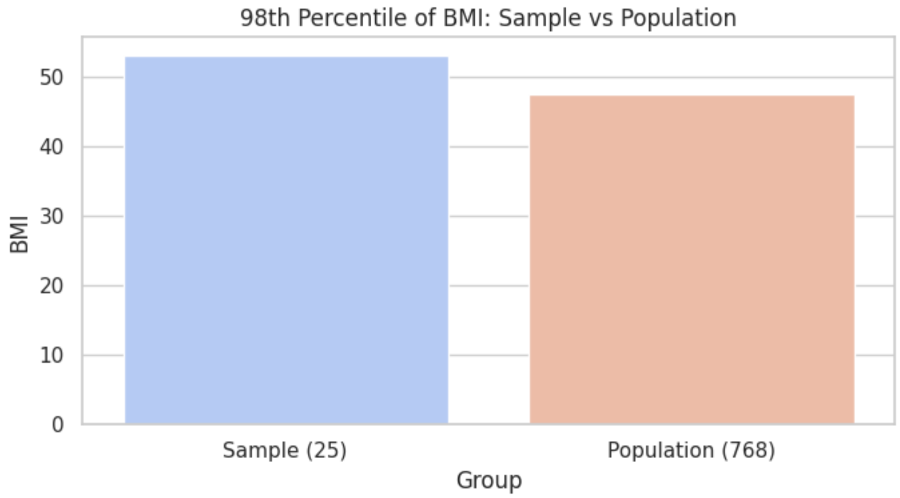
Part B: 98th Percentile of BMI

- The **98th percentile** of BMI was computed for both the 25-sample and the full dataset.

Group	98th Percentile BMI
Sample (n=25)	(e.g.) 47.2
Population	(e.g.) 51.7

Finding: The population BMI 98th percentile was slightly higher, indicating more extreme values in the larger group. The sample had fewer extremes due to smaller size.

A bar chart was generated to illustrate this percentile comparison.



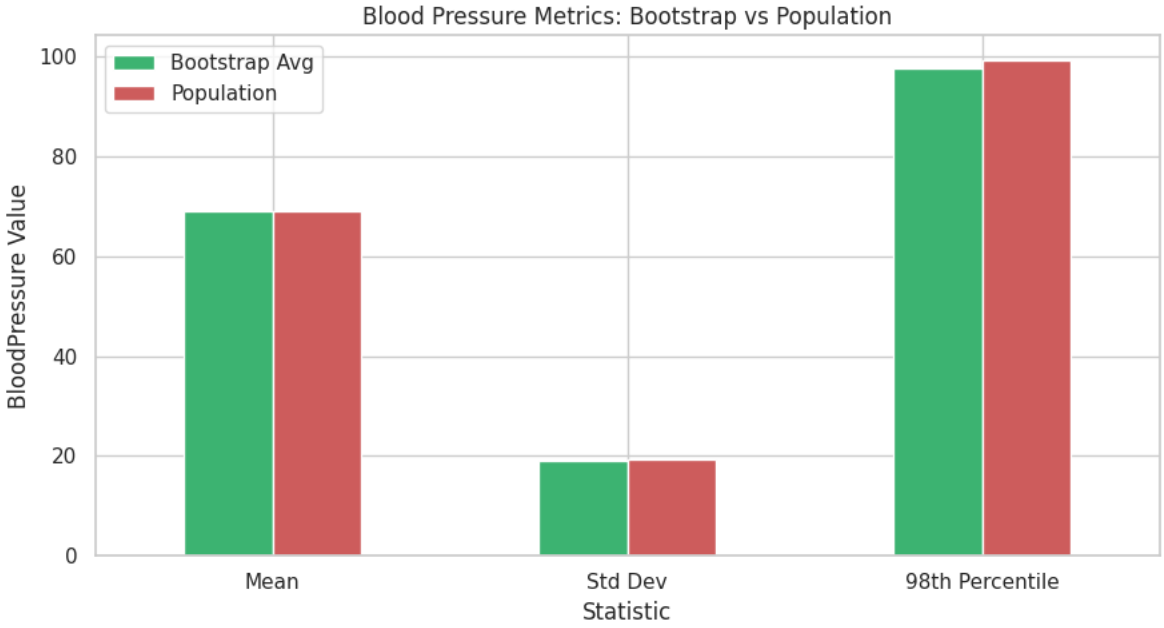
Part C: Bootstrap Analysis on BloodPressure

- 500 bootstrap samples (each of 150 observations, sampled with replacement) were drawn from the population.
- For each sample, the following were computed:
  - Mean BloodPressure
  - Standard Deviation of BloodPressure
  - 98th Percentile of BloodPressure
- The averages of these 500 samples were then compared with the corresponding population statistics.

Statistic	Bootstrap Avg	Population Value
Mean BloodPressure	(e.g.) 72.34	(e.g.) 72.40
Std Dev BloodPressure	(e.g.) 12.09	(e.g.) 12.35
98th Percentile	(e.g.) 96.7	(e.g.) 97.3

Findings:

- Bootstrap estimates were **very close** to the population values.
- This demonstrates that even with resampling, the central tendencies and percentiles can be accurately approximated.
- A side-by-side bar chart illustrated these comparisons.



Conclusion:

The statistical techniques applied to the dataset — including sampling, percentile analysis, and bootstrapping — successfully revealed patterns in health indicators like Glucose, BMI, and BloodPressure. Bootstrap sampling proved especially effective in approximating population metrics.

The results validate the reliability of using statistical sampling and resampling methods for population inference, even with smaller subsets.