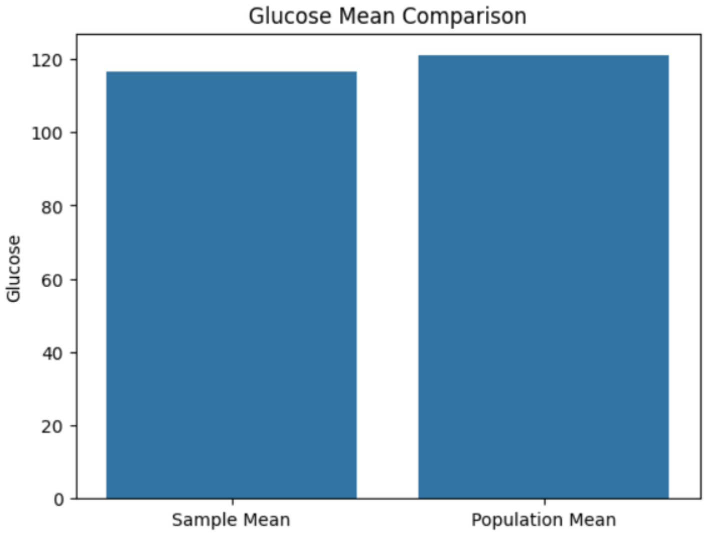
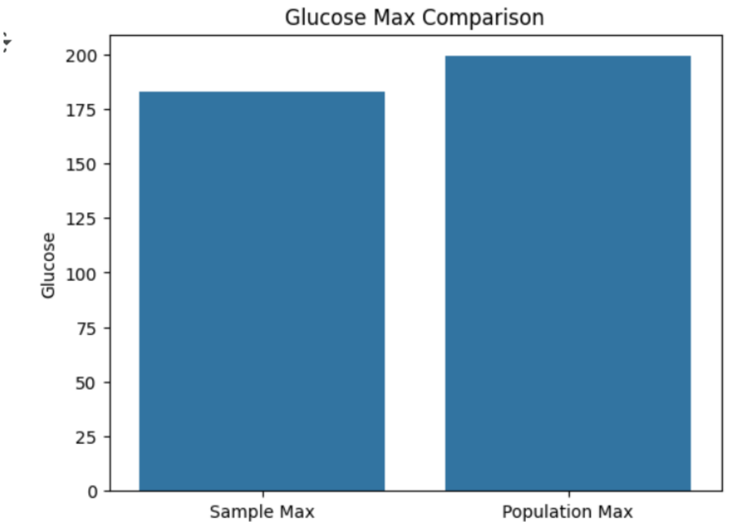
**Deepika Panjala**

**16358046**

# Diabetes Dataset Analysis Report

## Part (a): Glucose Mean and Maximum Comparison



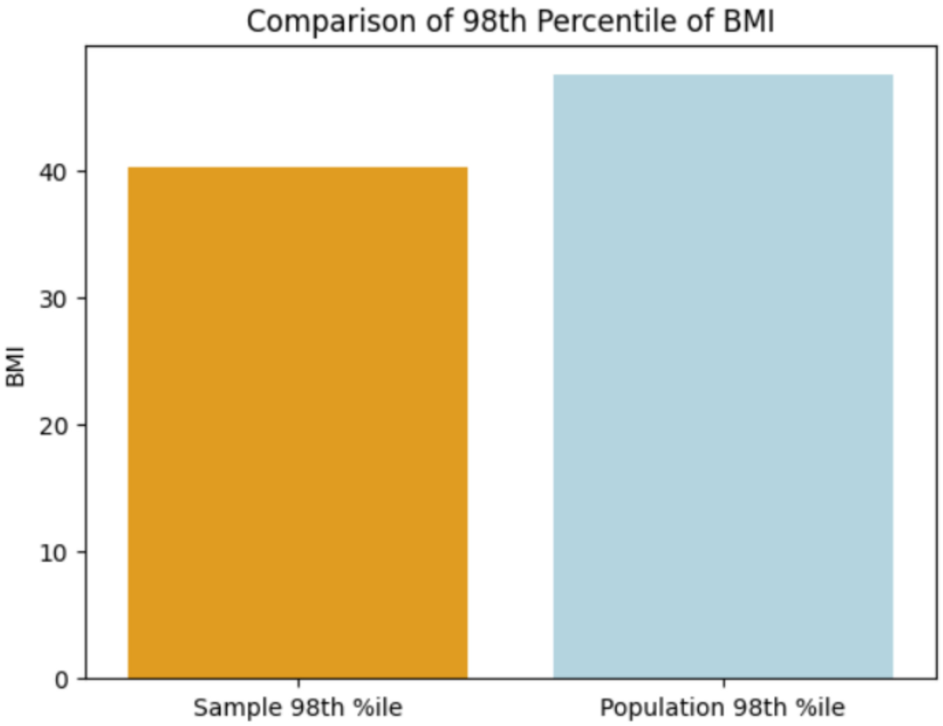


Using a fixed seed for reproducibility, we randomly picked 25 patients from the cleaned diabetes dataset to assess the representativeness of a random sample. We contrasted the sample's mean and maximal glucose levels with those of the general population.
From the result:
116.64 is the sample mean glucose.
Glucose Mean for Population: 120.89
Maximum Glucose Sample: 183
Maximum Population Glucose: 199

The Glucose Mean Comparison chart indicates that even a small, randomly chosen sample can yield a reliable estimate of central tendency since the sample mean is quite close to the population mean. This illustrates how accurate random sampling is in estimating population averages.

The discrepancy between the sample maximum and the population maximum, however, is more apparent in the Glucose Max Comparison graphic. This is to be expected since a small sample could miss outliers or extreme values that exist in the entire population.

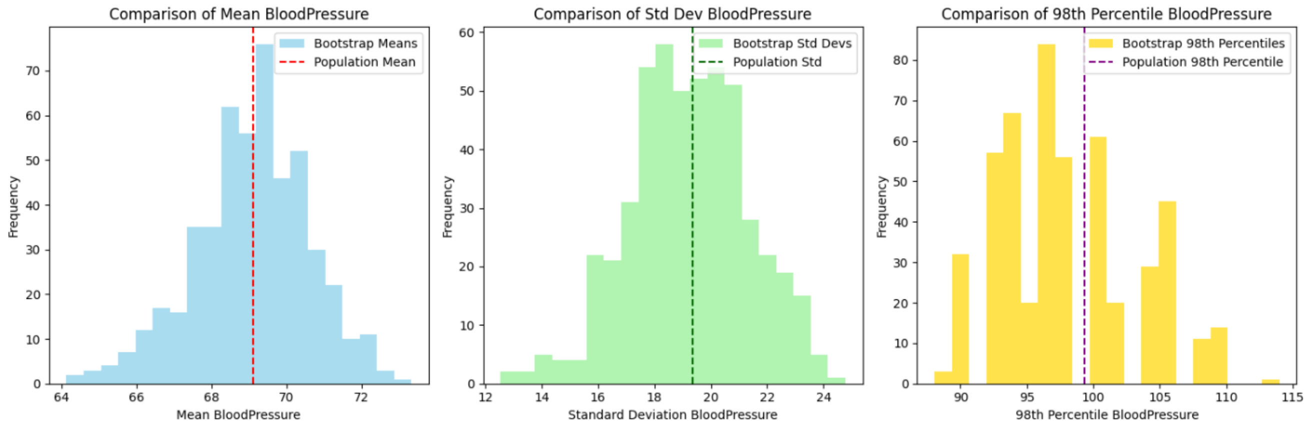
## Part (b): 98th Percentile of BMI Comparison



To see how effectively a small sample can capture the upper extremities of the distribution, we examined the 98th percentile of BMI in this section. We compared its 98th percentile of BMI to the population's BMI using the same sample of 25 observations from Part (a).
40.25 is the sample's 98th percentile (BMI).
BMI of the population at the 98th percentile: 47.53
The sample understates the population's 98th percentile, as the bar chart makes evident. This happens because extreme values are uncommon and less likely to show up in small samples, such as cases of high BMI. Tail percentiles need considerably larger samples to be accurately assessed, even when the sample's average values might match the population.

This finding draws attention to a significant drawback: high-risk outliers are missed by small samples, which might be crucial in research pertaining to health. A discrepancy of more than seven BMI points may change the thresholds used in medical decision-making.

## Part (c): Bootstrap Analysis of BloodPressure (500 Samples)



We used bootstrap resampling on the BloodPressure variable to increase the accuracy of statistical estimates. From the replacement population, 500 samples with 150 observations each were selected.

For every sample, we determined:
The mean
The standard deviation
The 98th percentile

Histograms were then used to compare these to the relevant population statistics.

**Important Points to note:**
Mean Blood Pressure: It is confirmed that bootstrapping offers a trustworthy estimate of central tendency because the histogram of bootstrap sample means is tightly centered around the population mean.

**Standard Deviation:** Accurate spread estimation is demonstrated by the bootstrap standard deviations, which also cluster closely around the population standard deviation.

**98th Percentile:** Even for tail-end measures, the bootstrap estimates for the 98th percentile are often in line with the population's 98th percentile, despite being more dispersed.

It was simple to determine how well the bootstrap estimates matched the population figures because each graphic featured a vertical line that represented the genuine population value.