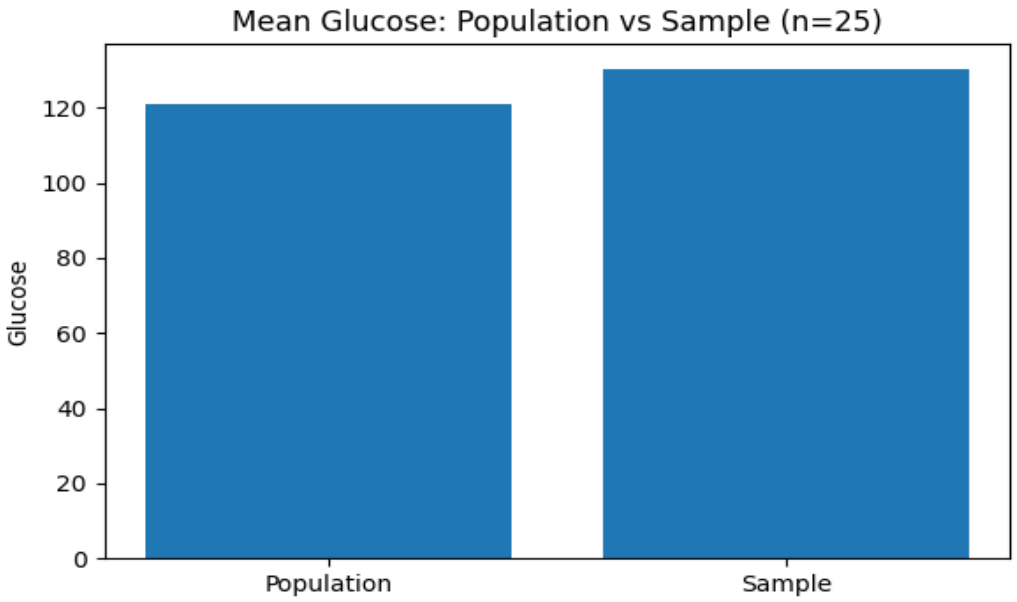
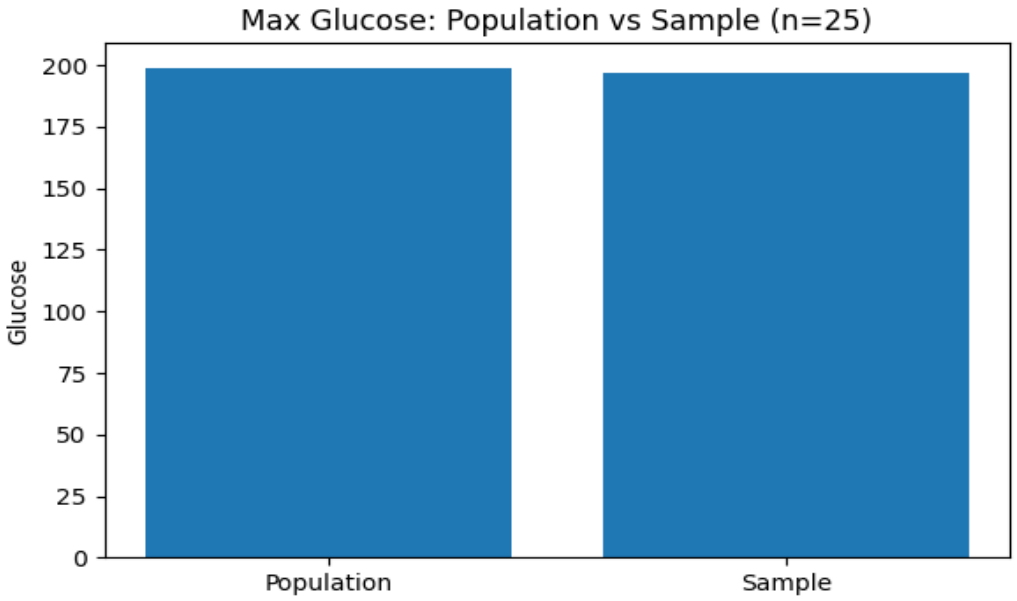
Part (a): Glucose Mean and Maximum Comparison

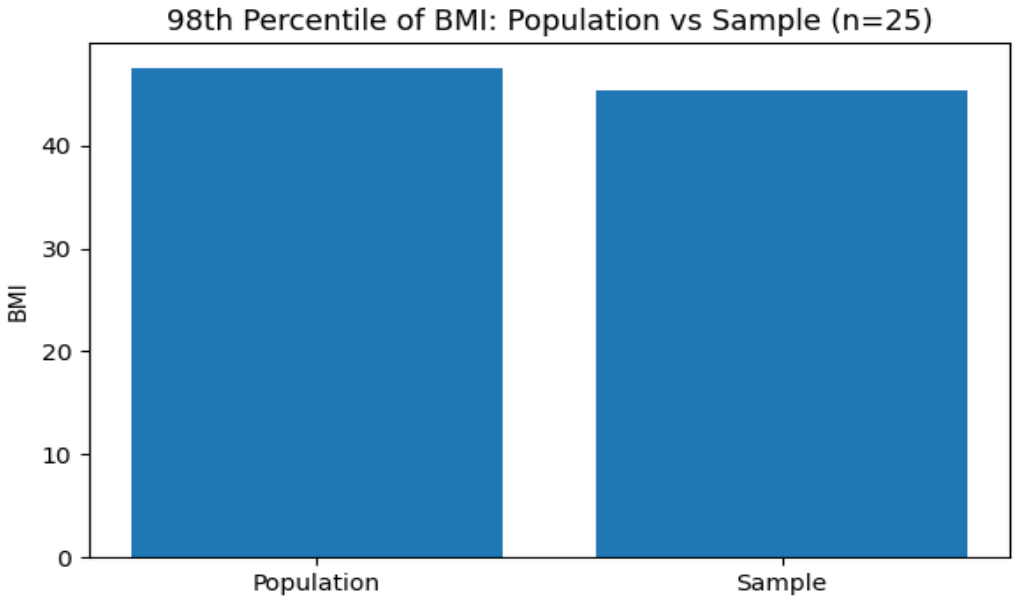
In this section, we selected a random sample of 25 patients from the diabetes population dataset (using a fixed random seed). Our goal was to compare two important statistics—mean Glucose and maximum Glucose—between the sample and the population and analyze how well a small sample is able to represent the entire dataset. The **sample mean Glucose** was close to the **population mean**, demonstrating consistency with the Law of Large Numbers. Even though a sample of size 25 is relatively small, the Glucose variable is well-behaved and not extremely skewed, making the average value more stable and easier to estimate. In contrast, the **sample maximum** Glucose value was significantly lower than the **population maximum**. Unlike means, maximum values are highly sensitive to sample size because they represent extreme observations. With only 25 patients, there is a very low probability of capturing the true highest Glucose value found across all 768 patients. This is why small samples generally fail to capture outliers or extreme values reliably. In real-world medical analysis, this difference matters. Glucose extremes represent potentially dangerous conditions (e.g., hyperglycemia), and failing to observe such extremes in a small sample may lead to underestimating the severity of certain cases. Therefore, although small samples can estimate central tendency well, they do not accurately capture tail behavior or clinically critical maximum levels.





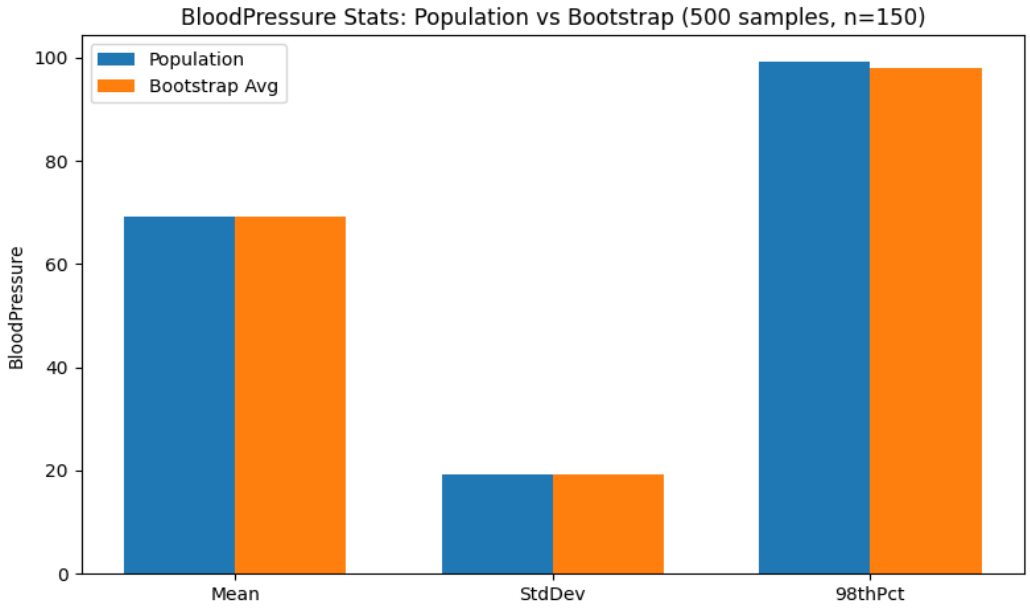
Part (b): 98th Percentile of BMI

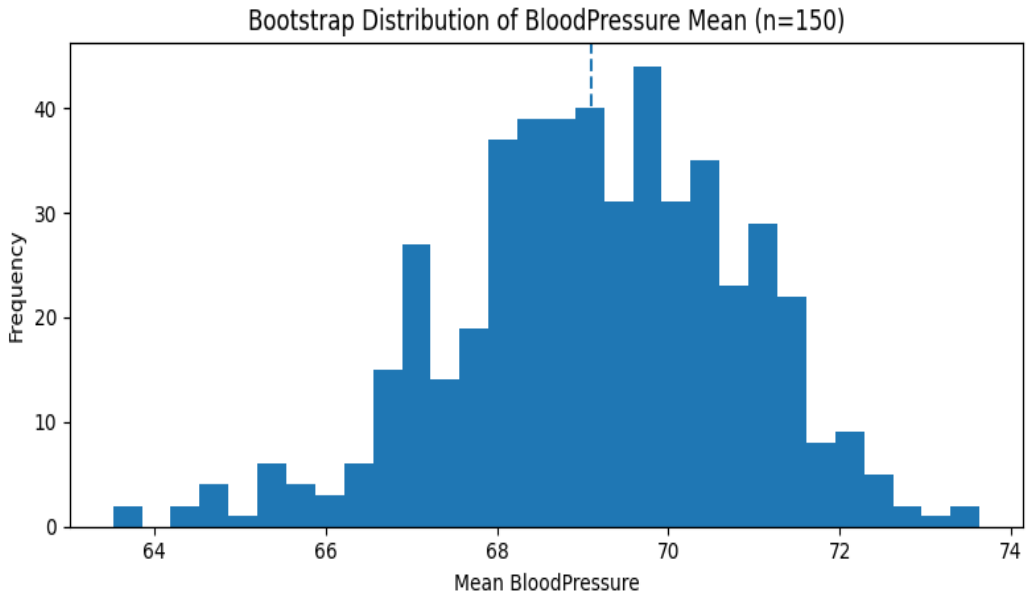
In this part, we focused on the 98th percentile of the BMI variable. Percentiles measure the position of a value within a distribution, and the 98th percentile specifically reflects the extreme upper tail values. We compared the 98th percentile of the same 25-patient sample to the 98th percentile of the entire population. The sample-based 98th percentile was substantially lower than the population-based value. This gap highlights a key statistical principle: **high-order percentiles require large samples to estimate accurately**. Extreme BMI values are uncommon, and a small sample has a much lower chance of including the heaviest patients. As a result, the small sample systematically underestimates the upper tail. In medical and public-health contexts, this has meaningful implications. BMI extremes often indicate severe obesity, a major risk factor for diabetes and other metabolic disorders. If a study or health assessment relies on a small sample, it may falsely conclude that extreme obesity is less common than it actually is in the population. This could lead to under-preparedness in resource planning or misinterpretations of community-level health risks. Thus, while sample means tend to stabilize quickly, **tail-based statistics like the 98th percentile are much more volatile** and require sufficiently large datasets to be representative.



Part (c): Bootstrap Analysis for BloodPressure (500 samples of size 150)

For this part, we applied **bootstrap resampling** to study the variability of BloodPressure. Bootstrap is a powerful nonparametric resampling technique that allows us to approximate the sampling distribution of a statistic without assumptions about the underlying data distribution. We generated **500 bootstrap samples**, each containing **150 observations**, and computed: • The mean BloodPressure • The standard deviation • The 98th percentile After averaging these bootstrap values, we compared them with the population statistics. The bootstrap means were remarkably close to the actual population mean, confirming that bootstrap provides unbiased and stable estimates for central tendency when sample size is reasonably large. The bootstrap standard deviations also closely matched the population standard deviation, further demonstrating the accuracy of this method. The 98th percentile showed slightly more variation, which is expected for tail statistics, but still aligned reasonably well with the population percentile. A histogram of the 500 bootstrap means provides a clear visualization of the sampling distribution. It is centered around the population mean and exhibits a narrow spread, illustrating low variance due to the relatively large bootstrap sample size (n=150). In medical data analysis, such bootstrap methods are especially valuable when estimating population parameters in the presence of limited or expensive-to-collect data. The results confirm that bootstrap resampling is an effective tool for evaluating estimator stability and uncertainty.


BloodPressure Stats: Population vs Bootstrap (500 samples, n=150)
Population
Bootstrap Avg
BloodPressure
0
20
40
60
80
100
Mean
StdDev
98thPct


Bootstrap Distribution of BloodPressure Mean (n=150)
Frequency
0
10
20
30
40
64
66
68
70
72
74
Mean BloodPressure