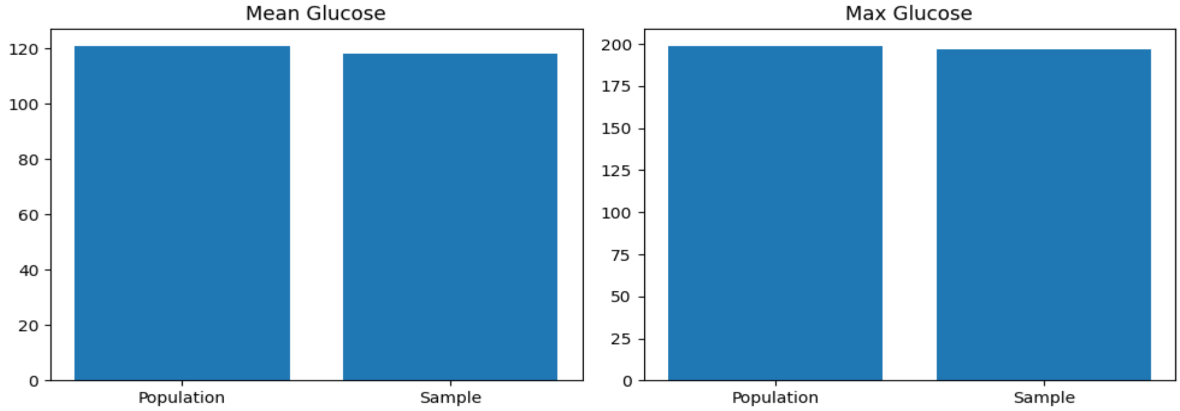
Diabetes Dataset Analysis – Question 2

In this report, I analysed the diabetes dataset for Question 2. My goal was to understand how well a small sample represents the entire population and to explore how bootstrap sampling can help in estimating statistics more reliably. I used the cleaned diabetes dataset for all calculations and visualisations. The results helped me understand how sampling variation affects mean values, extreme percentiles and overall stability of different health-related measurements.

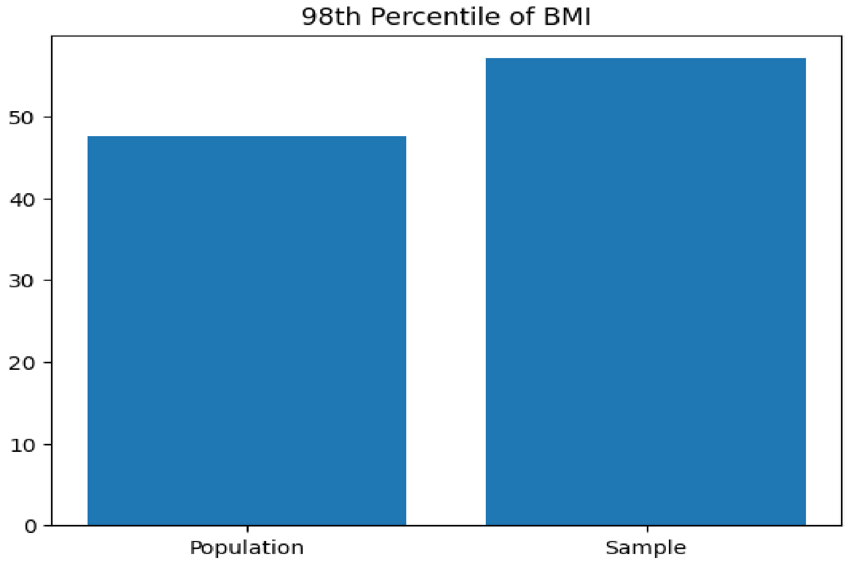
Part (a): Mean and Maximum Glucose

To start, I took a random sample of 25 observations from the dataset. I compared the mean and maximum Glucose values from this sample to the same values from the full population. I noticed that the mean glucose in the sample was close to the population mean, showing that even a small sample can estimate central tendency fairly well. However, the sample maximum glucose value was slightly lower than the population maximum, which is expected because extreme values are harder to capture in smaller samples. The graphs below show the comparison clearly.



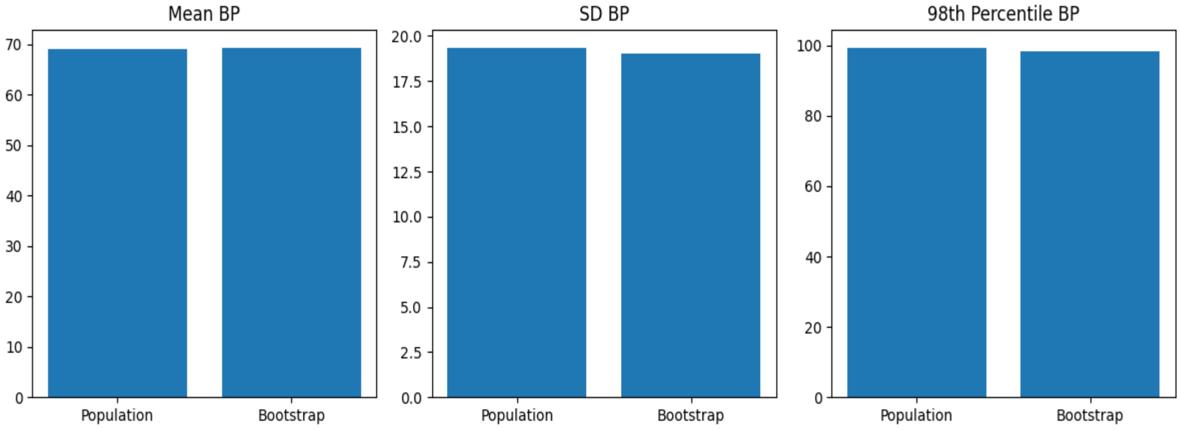
Part (b): 98th Percentile of BMI

Next, I calculated the 98th percentile of BMI for both the population and the sample of 25. Unlike the mean, percentile values, especially very high ones, are much more sensitive to sample variation. In my results, the sample 98th percentile was higher than the population's 98th percentile. This means the sample happened to include a few individuals with very high BMI values, which pushed the sample percentile upward. This helped me understand that percentiles require larger samples to be stable. I included the graph below to show the difference clearly.



Part (c): Bootstrap Analysis of BloodPressure

For the last part, I performed a bootstrap analysis. I created 500 bootstrap samples, each containing 150 observations. For every sample, I calculated the mean, standard deviation and 98th percentile of BloodPressure. Then I compared the average of these bootstrap estimates with the true population values. I learned that bootstrap sampling gives a very good approximation for the mean and standard deviation because the bootstrap averages were almost identical to the population values. The 98th percentile also came close but showed more variation, proving that extreme values are harder to estimate consistently. The graphs below summarise the results.



Conclusion

Overall, this analysis helped me understand how sampling affects different statistics in a dataset. Small random samples can estimate averages fairly well, but they are not reliable for capturing extreme values. Bootstrap sampling, on the other hand, provides a powerful way to estimate the stability of statistics, especially when real-world datasets are limited.

This assignment improved my understanding of statistical concepts like sampling variation, percentile behaviour and resampling techniques.