Tom Steinman
Assignment 3 - Diabetes
5/8/2025

Data Cleaning

The dataset had some impossible 0 values in the Glucose, BloodPressure, SkinThickness, BMI and Insulin columns that were removed. No null values were found.

1a)

There is very little difference in sample and population Glucose mean, with the population being 122.63 versus the sample mean of 120.88. The max has a larger gap with the population being 198.00 versus the sample at 181.00. There is an outlier within the population that did not get included in the sample. The distributions were similar, with less samples in the sample of the population at the outliers and center of the distribution.

1b)

The 98th percentile of BMI for the population and sample were also similar. The population had a 98th percentile of 47.00, while the sample was 49.07. The distribution between the two were also similar, with less outliers and higher observations in the sample compared to the population.

1c)

The bootstrap averages of BloodPressure mean, standard deviation and median (50th percentile) were all very similar. The 98th percentile bootstrapping average was pretty significant at 106.08 versus the population at 94.72. While the bootstrapping method had excellent accuracy in sampling the population of mean, standard deviation and median, it was not as accurate in estimating the 98th percentile, and likely high percentiles, of the population. The distribution of the bootstrapping method however is accurate outside of the higher percentile to demonstrate the observations of the population.