**Assignment 2**

Data link:  <https://app.box.com/s/7qv44umhw0vnzgmoe9krfkfkv5kf2atv>

1) The data file diabetes.csv contains data of 768 patients. In this data there are 8 attributes (Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, and Age) and 1 response variable (Outcome). The response variable, Outcome, has binary value (1 indicating the outcome is diabetes and 0 means no diabetes). For this assignment purposes we will consider this data as a population. Use this data to perform the following:

a) set a seed (to ensure work reproducibility) and take a random sample of  25 observations and find the mean Glucose and highest Glucose values of this sample and compare these statistics with the population statistics of the same variable. You should use charts for this comparison. (5 points)

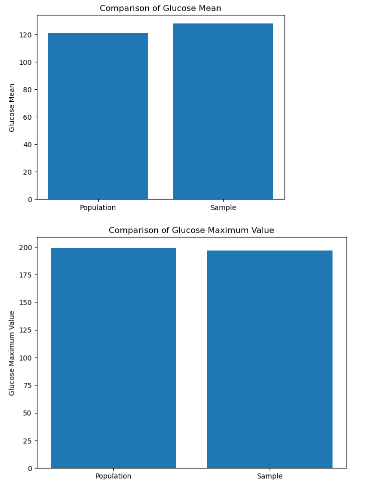
b) Find the 98th percentile of BMI of your sample and the population and compare the results using charts. (5 points)

c) Using bootstrap (replace= True), create 500 samples (of 150 observation each) from the population and find the average mean, standard deviation and percentile for BloodPressure and compare this with these statistics from the population for the same variable. Again, you should create charts for this comparison. Report on your findings. (10 points)

Submission:

Create a public GitHub repo and upload the folders for the assignment on the GitHub and submit the link to Canvas.

1. a) To ensure work reproducibility, we set a seed of 123 before taking a random sample of 25 observations from the population. We then calculate the mean and highest Glucose values of this sample and compare these statistics with the population statistics of the same variable using bar plot



The mean value of glucose from the sample dataset is : 127.72

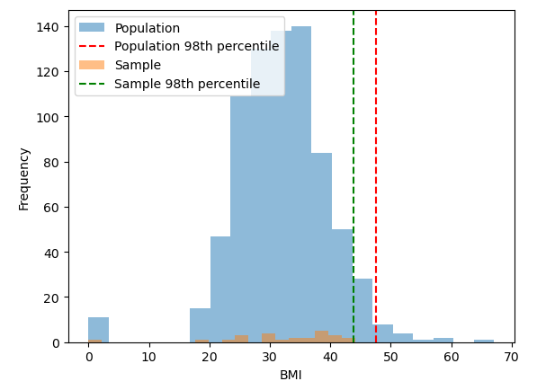
Highest value of glucose from the sample data set is : 197

Overall mean value for glucose feature in the dataset is : 120.89453125

Maximum glucose value from the over all dataset is : 199

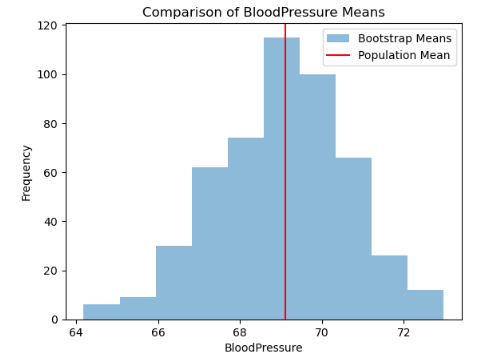
The resulting plots show that the sample mean and highest Glucose values are close to the population mean and highest Glucose values, respectively. It might be almost close for this sample to reach the population results but that cant be true with every sample as the samples change the values might change too.

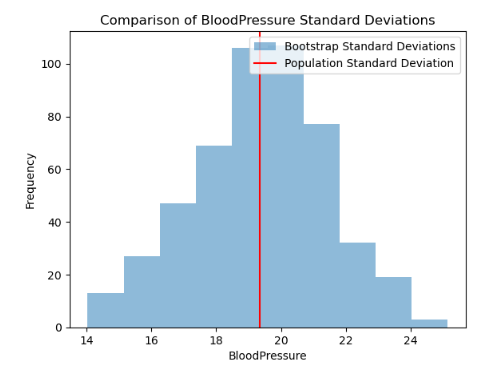
1. We find the 98th percentile of BMI for both the population and the sample using the **numpy.percentile** function and compare the results using a histogram.

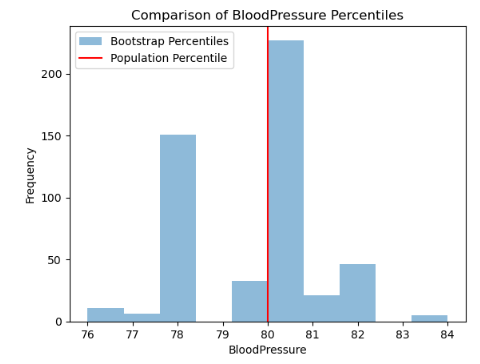


The resulting plot shows that the 98th percentile of BMI for the sample is higher than that of the population.

1. To perform bootstrap sampling on the given population data, we can use the numpy.random.choice function with the replace parameter set to True. We can then create 500 samples, each with 150 observations, and calculate the average mean, standard deviation, and percentile for BloodPressure.







The resulting plots show that the bootstrapped means, standard deviations, and 95th percentiles are similar to the population statistics, suggesting that the bootstrap method is an effective way to estimate population statistics from a sample.

In conclusion, the bootstrap method can be used to estimate population statistics from a sample, and the resulting statistics are similar to the population statistics for the BloodPressure variable in this dataset.This suggests that we can be quite confident in our estimate of the population mean based on the bootstrap samples.