## **ASSIGNMENT-3**

## **Name: ENUKONDA SAI DEEPTHI STUDENT ID:16343756**

**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**

The code calculates mean and maximum 'Glucose' values for both a random sample of 25 observations and the entire population. It visually compares these statistics using side-by-side bar charts, employing distinct colors for sample and population data. The lightseagreen and lightcoral bars in the first subplot represent mean glucose values, while the sandybrown and lightgreen bars in the second subplot represent maximum glucose values. This comparison aims to provide a quick visual assessment of whether the small sample is representative of the population in terms of 'Glucose' characteristics.

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

The code determines the 98th percentile of 'BMI' for both a sample of 25 and the entire population. A bar chart, using different colors to distinguish between sample and population percentiles, visually contrasts the results. This graphical comparison facilitates a swift evaluation of how effectively the sample represents the 98th percentile 'BMI' of the overall population. If the bars align closely, it indicates a close match between the sample and population. The visual representation streamlines the assessment process, offering insights into the sample's accuracy in capturing this specific percentile characteristic.

**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)**

The provided Python code uses bootstrap sampling to create 500 samples, each with 150 observations, from the 'BloodPressure' variable in the population dataset. This resampling is conducted with replacement to simulate diverse subsets of the population. For each of these bootstrap samples, the code computes the mean, standard deviation, and 95th percentile of the 'BloodPressure' variable. These metrics offer insights into the central tendency, variability, and higher-end threshold of blood pressure values.

The comparison process involves generating visualizations, including histograms for means and standard deviations, and boxplots for percentiles. These charts facilitate a quick assessment of how well the characteristics of the bootstrap samples align with the overall population. A close match between the distributions of bootstrap and population statistics indicates the reliability of using bootstrap samples for estimating 'BloodPressure' metrics.

This analysis goes beyond numerical comparisons by providing visual aids that enhance the understanding of the representativeness and robustness of the bootstrap sampling method. Overall, it offers a comprehensive evaluation of how effectively the 500 bootstrap samples capture the key statistical properties of 'BloodPressure' in comparison to the entire population.

Top of Form

Bottom of Form

**Submission Link:**

<https://github.com/deepthi978/PDS-Assignment-3>