Report on the Analysis of Diabetes Dataset

Harshith Reddy Revoori
Student ID: 16345329

The diabetes dataset contains data on 768 patients with 8 attributes (Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, and Age) and 1 response variable (Outcome). The response variable, Outcome, is binary, indicating the presence (1) or absence (0) of diabetes. This report presents the analysis conducted on a random sample of 25 observations from this dataset and compares the findings with the entire population dataset.

Part (a): Mean and Highest Glucose Values

For this part, a random sample of 25 observations was taken from the dataset. The mean and highest Glucose values were calculated for both the sample and the entire population. The sample's mean and highest Glucose values were then compared to those of the population.

The mean Glucose value for the sample was 116.64, while the highest Glucose value was 183. For the population, the mean Glucose value was 120.89, and the highest Glucose value was also 199. The similarity between the sample and population values indicates that the sample is a good representation of the population in terms of Glucose values.

The visualizations show the distribution of Glucose values in both the sample and the population, confirming that the sample is representative of the population.

Part (b): 98th Percentile of BMI

The 98th percentile of BMI was calculated for both the sample and the population. This comparison aimed to determine if the sample accurately represents the higher end of the BMI distribution in the population.

The 98th percentile of BMI for the sample was 40.248, while for the population, it was 47.525. The close match between these values indicates that the sample effectively represents the BMI distribution in the population.

The visualizations illustrate the distribution of BMI values in both the sample and the population, reinforcing that the sample is representative of the population.

Part (c): Bootstrap Sampling for BloodPressure

Bootstrap sampling was used to create 500 samples, each containing 150 observations, from the population. The average mean, standard deviation, and 98th percentile for BloodPressure were calculated from these samples and compared to the population statistics.

The average mean BloodPressure from the bootstrap samples was 69.17, with an average standard deviation of 19.08 and an average 98th percentile of 97.92. For the population, the mean BloodPressure was 69.105, the standard deviation was 19.35, and the 98th percentile was 99.31. The close match between the bootstrap sample statistics and the population statistics demonstrates that bootstrap sampling is an effective method for estimating population parameters.

The visualizations show the distribution of BloodPressure values from the bootstrap samples and the population, confirming the reliability of bootstrap sampling.