**PDS ASSIGNMENT 2**

**SOURCE CODE**

**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 theoutcome is diabetes and 0 means no diabetes). For this assignment purposes we will consider this dataas 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 andfind the mean Glucose and highest Glucose values of this sample and compare these statisticswith the population statistics of the same variable. You should use charts for this comparison.**

import pandas as pd

import numpy as np

import random

import matplotlib.pyplot as plt

# Set random seed

np.random.seed(123)

# Read in data

data = pd.read\_csv("/babulu/diabetes.csv")

# Take a random sample of 25 observations

sample = data.sample(n=25)

# Calculate mean and max glucose for sample and population

sample\_mean\_glucose = np.mean(sample["Glucose"])

sample\_max\_glucose = np.max(sample["Glucose"])

pop\_mean\_glucose = np.mean(data["Glucose"])

pop\_max\_glucose = np.max(data["Glucose"])

# Create bar chart comparing mean and max glucose between sample and population

fig, ax = plt.subplots()

x = np.arange(2)

width = 0.35

ax.bar(x, [pop\_mean\_glucose, sample\_mean\_glucose], width, label='Mean Glucose')

ax.bar(x + width, [pop\_max\_glucose, sample\_max\_glucose], width, label='Max Glucose')

ax.set\_ylabel('Glucose')

ax.set\_title('Comparison of Glucose between Population and Sample')

ax.set\_xticks(x + width / 2)

ax.set\_xticklabels(['Population', 'Sample'])

ax.legend()

plt.show()

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

sample\_98\_percentile = np.percentile(data["BMI"].sample(n=25), 98)

pop\_98\_percentile = np.percentile(data["BMI"], 98)

fig, ax = plt.subplots()

ax.bar(["Population", "Sample"], [pop\_98\_percentile, sample\_98\_percentile], label="98th Percentile BMI")

ax.set\_ylabel("BMI")

ax.set\_title("Comparison of 98th Percentile BMI between Population and Sample")

ax.legend()

plt.show()

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

n\_samples = 500

sample\_size = 150

sample\_means = np.zeros(n\_samples)

sample\_stds = np.zeros(n\_samples)

sample\_percentiles = np.zeros(n\_samples)

def bootstrap\_sample(data):

return data.sample(n=len(data), replace=True)

for i in range(n\_samples):

sample = bootstrap\_sample(data["BloodPressure"]).sample(n=sample\_size, replace=True)

sample\_means[i] = sample.mean()

sample\_stds[i] = sample.std()

sample\_percentiles[i] = np.percentile(sample, 75)

pop\_mean\_bp = data["BloodPressure"].mean()

pop\_std\_bp = data["BloodPressure"].std()

pop\_percentile\_bp = np.percentile(data["BloodPressure"], 75)

fig, axes = plt.subplots(3, 1, figsize=(8, 12))

axes[0].hist(sample\_means, bins=20)

axes[0].axvline(pop\_mean\_bp, color="red", linestyle="--", label="Population Mean")

axes[0].set\_xlabel("Mean BloodPressure")

axes[0].set\_ylabel("Frequency")

axes[0].legend()

axes[1].hist(sample\_stds, bins=20)

axes[1].axvline(pop\_std\_bp, color="red", linestyle="--", label="Population Std")

axes[1].set\_xlabel("Std BloodPressure")

axes[1].set\_ylabel("Frequency")

axes[1].legend()

axes[2].hist(sample\_percentiles, bins=20)

axes[2].axvline(pop\_percentile\_bp, color="red", linestyle="--", label="Population 75th Percentile")

axes[2].set\_xlabel("75th Percentile BloodPressure")

axes[2].set\_ylabel("Frequency")

axes[2].legend()

plt.tight\_layout()

plt.show()