**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)

Code:

set.seed(6497)

Data = read.csv("~/Downloads/diabetes.csv")

View(Data)

dim(Data)

sample25 = Data[sample(768,25),]

View(sample25)

dim(sample25)

mean25 = mean(sample25$Glucose)

mean25

High\_gluco25 = max(sample25$Glucose)

High\_gluco25

mean\_all = mean(Data$Glucose)

mean\_all

High\_Gluco\_all = max(Data$Glucose)

High\_Gluco\_all

plot1 = barplot(c(mean25,mean\_all), col = c("blue", "red"), main = "Comparison of mean of Glucose", ylab = "Glucose", names.arg = c("mean25","mean\_all"))

plot2 = barplot(c(High\_gluco25,High\_Gluco\_all), col = c("grey", "green"), main = "Comparison of highest level of Glucose", ylab = "Glucose", names.arg = c("High\_gluco25","High\_Gluco\_all"))

Output Screenshots:

Graphical user interface, application

Description automatically generated

Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated

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

Code:

set.seed(6497)

Data = read.csv("~/Downloads/diabetes.csv")

View(Data)

dim(Data)

sample25 = Data[sample(768,25),]

View(sample25)

dim(sample25)

bmi\_25 = quantile(sample25$BMI,0.98)

bmi\_25

bmi\_all = quantile(Data$BMI,0.98)

bmi\_all

plot3 = barplot(c(bmi\_25,bmi\_all), col = c("orange", "white"), main = "Comparison of bmi of 98 percentile", ylab = "BMI", names.arg = c("bmi\_25","bmi\_all"))

Graphical user interface

Description automatically generated

Chart, square

Description automatically generated

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)

Code:

set.seed(6497)

Data = read.csv("~/Downloads/diabetes.csv")

View(Data)

dim(Data)

Bp = Data$BloodPressure

Bp\_mean = mean(Bp)

Bp\_sd = sd(Bp)

Bp\_perc = quantile(Bp,0.98)

boot\_strap <- function(Data, n\_boot, sample\_size) {

+ bs\_mean <- numeric(n\_boot)

+ bs\_sd <- numeric(n\_boot)

+ bs\_perc <- numeric(n\_boot)

+

+ for (i in 1:n\_boot) {

+ sample\_bs <- sample(Data, size = sample\_size, replace = TRUE)

+

+

+ bs\_mean[i] <- mean(sample\_bs)

+ bs\_sd[i] <- sd(sample\_bs)

+ bs\_perc[i] <- quantile(sample\_bs, 0.98)

+ }

+ list(mean = bs\_mean, sd = bs\_sd, perc\_98 = bs\_perc)

+ }

n\_boot <- 500

sample\_size <- 150

boot\_results <- boot\_strap(Bp, n\_boot, sample\_size)

boot\_mean <- mean(boot\_results$mean)

boot\_sd <- sd(boot\_results$mean)

boot\_perc <- perc\_98(boot\_results$mean)

plot3 = barplot(c(Bp\_mean,boot\_mean), col = c("black", "white"), main = "Comparison of mean of 500 and pop", ylab = "BP", names.arg = c("Bp\_mean","boot\_mean"))

plot4 = barplot(c(Bp\_sd,boot\_sd), col = c("red", "blue"), main = "Comparison of sd of 500 and pop", ylab = "BP", names.arg = c("Bp\_sd","boot\_sd"))

plot5 = barplot(c(Bp\_perc,boot\_perc), col = c("green", "yellow"), main = "Comparison of percentile of 500 and pop", ylab = "Bp", names.arg = c("Bp\_perc","boot\_perc"))

Graphical user interface, application

Description automatically generated

Chart

Description automatically generated

Graphical user interface, application

Description automatically generated

Chart, bar chart

Description automatically generated

Square

Description automatically generated

Chart, shape, bar chart, square

Description automatically generated

Findings:

The sample means and sample percentile is almost the same as the population mean and standard deviation, respectively, as shown by the plot. The standard deviation for the sample match is much more varied with those of the population. Overall, there is a strong correlation between the statistics from the samples and the statistics from the population, proving the validity of the sample data.