**Statistics Final Project**

1. Pick two separate corporations and from www.nasdaq.com, pick the “historical quote” option and randomly select 30 opening stock prices for the last two years. Place the data into an Excel spreadsheet and outline your sampling method in the main Word document for the project.

I have picked General Electric (GE) and Bank of America (BAC) for the last 02 years from [www.nasdaq.com](http://www.nasdaq.com). Then randomly selected 30 opening stock prices based on **Simple Random Sampling (SRS)**, are attached below.

**R Code:**

*#Set Directory*

*setwd("C:/WQU/Final\_Project")*

*#Get data for General Electric (GE) & Bank of America (BAC)*

*GE <- read.csv("HistoricalQuotes\_GE.csv", header=TRUE, sep=",")[,c(1,4)]*

*BAC <- read.csv("HistoricalQuotes\_BAC.csv", header=TRUE, sep=",")[,c(1,4)]*

*#Creating a random sample of 30 opening prices for the last 02 years*

*set.seed(123)*

*GE.Sample <- GE[sample(nrow(GE), size = 30, replace = FALSE, prob = NULL),2]*

*BAC.Sample <- BAC[sample(nrow(BAC), size = 30, replace = FALSE, prob = NULL),2]*

*#Export Sample Data*

*write.csv(GE.Sample,"GE.Sample.csv")*

*write.csv(BAC.Sample,"BAC.Sample.csv")*

2. Determine a claim (prior to analyzing the descriptive statistics) based on two population means/samples.

A shortened version of claim: The opening stock prices for GE and BAC are different.

Hypothesis

Ho (Null Hypothesis): The Opening Stock Prices for GE and BAC are same. Price(GE) = Price(BAC) OR-- Price(GE) - Price(BAC) = 0

Ha (Alternate Hypothesis): The Opening Stock Prices for GE and BAC are not same. Price(GE) not equal to Price(BAC) OR-- Price(GE) - Price(BAC) not equal to 0.

**R Code (has only the comments section):**

*#Claim: The opening stock prices for GE and BAC are different.*

*#Hypothesis*

*#Ho (Null Hypothesis): The Opening Stock Prices for GE and BAC are same. Price(GE) = Price(BAC)*

*#OR-- Price(GE) - Price(BAC) = 0*

*#Ha (Alternate Hypothesis): The Opening Stock Prices for GE and BAC are not same.*

*#Price(GE) not equal to Price(BAC) OR-- Price(GE) - Price(BAC) not equal to 0.*

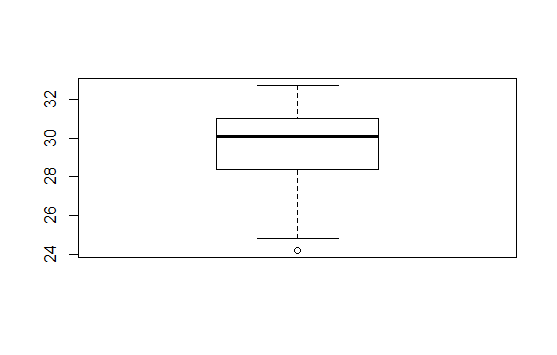
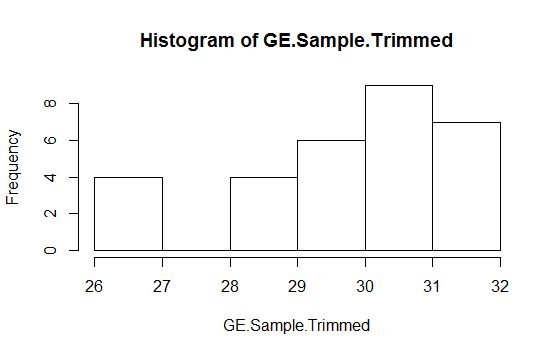
3. Calculate the descriptive statistics for the two separate corporations. The distribution shape should be analyzed to determine the appropriate descriptive statistics to use (mean/standard deviation versus median/IQR). The graph should be included in the main Word document for the project. Find the appropriate descriptive statistics and place this in the Word document noting any outliers or any irregularities discovered.

**Summary Statistics for GE Sample:**

Min. 1st Qu. Median Mean 3rd Qu. Max.

24.20 28.54 30.08 29.56 31.00 32.73

The range of price is 8.53 USD with the minimum price of 24.20, and the highest price of 32.73. The mean price is 29.56 USD; the median is 30.08. As the mean is slightly lesser than median, the distribution is slightly left skewed. The first quartile is 28.54 USD; the third quartile is 31.00 USD; the IQR is 2.46 USD.

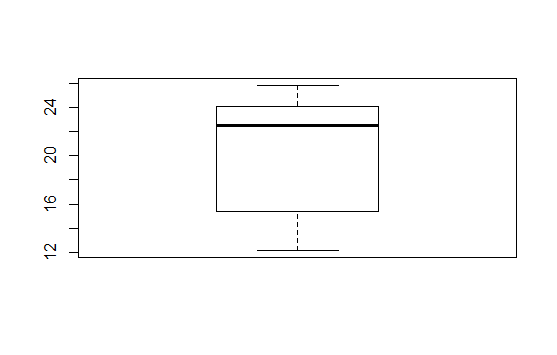
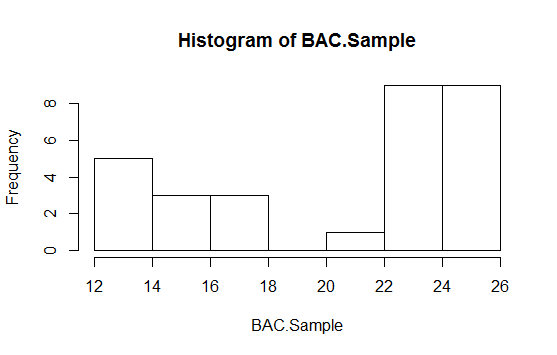
There is one outlier in GE Sample data. Its value is 24.2, which was much lesser than mean and median of the sample data. After flooring the values at 10th percentile, the mean value is slightly being increased from 29.56133 to 29.64633. Based on that, I used the capped sample for hypothesis testing. The histogram shows distribution after outlier treatment.

**Summary Statistics for BAC Sample:**

Min. 1st Qu. Median Mean 3rd Qu. Max.

12.14 15.76 22.54 20.34 24.10 25.86

The range of price is 13.72 USD with the minimum price of 12.14, and the highest price of 25.86. The range is much higher than GE sample data. The mean price is 20.34 USD; the median is 22.54, much lower than GE sample data. As the mean is lesser than median, the distribution is left skewed. The first quartile is 15.76 USD; the third quartile is 24.86 USD; the IQR is 8.34 USD. Much more than the GE sample data. The data in BAC sample has more spread than GE sample data.

There is no outlier in GE Sample data. Therefore, no outlier treatment is needed.

**R Code:**

*#Descriptive Statistics for GE Sample*

*summary(GE.Sample)*

*#Box Plot for GE Sample*

*boxplot(GE.Sample)*

*#Outlier Detection and Treatment*

*GE.Outliers <- boxplot.stats(GE.Sample)$out #05 Outliers Detected for GE*

*GE.Outliers*

*#Mean of outliers*

*mean(GE.Outliers)*

*#Comparison of means with and without outliers*

*GE.Non.Outliers <- ifelse(GE.Sample %in% GE.Outliers, NA, GE.Sample) #One Outlier Detected for GE*

*#Box Plot of GE Sample without outliers*

*boxplot(GE.Non.Outliers)*

*#GE Sample mean without outliers*

*mean(GE.Non.Outliers, na.rm = T)*

*#GE Sample mean with outliers*

*mean(GE.Sample)*

*#Outlier Treatment*

*#capping values at 10th and 90th percentiles*

*#Function to cap or floor extreme values*

*fun.capping.flooring <- function(x){*

*quantiles <- quantile( x, c(.10, .90 ) )*

*x[ x < quantiles[1] ] <- quantiles[1]*

*x[ x > quantiles[2] ] <- quantiles[2]*

*x*

*}*

*GE.Sample.Trimmed <- fun.capping.flooring(GE.Sample)*

*mean(GE.Sample.Trimmed)*

*mean(GE.Sample)*

*# After capping/flooring the outlier, the mean value has slightly been increase from 29.56133*

*#to 29.64633. Therefore we will use the capped sample for hypothesis testing*

*#Descriptive Statistics for GE Sample*

*summary(BAC.Sample)*

*#Box Plot for GE Sample*

*boxplot(BAC.Sample)*

*#Outlier Detection*

*BAC.Outliers <- boxplot.stats(BAC.Sample)$out #No Outlier Detected for BAC*

*#Will use all the obs without any outlier treatments (e.g.Mean/Median Imputation, Capping/Flooring)*

*#Histogram to look at the shape of distribution of samples*

*hist(GE.Sample.Trimmed)*

*hist(BAC.Sample)*

4. Conduct the hypothesis test based on the claim from item 2. Outline all the specifics in the Word document describing your conclusion. Note any possible reasons for the results. This might include some research on the corporations chose in item 1.

Hypothesis Testing: Mean with Sigma Not Known

**Requirements:**

1. The sample is obtained using simple random sampling or from a randomized experiment.
2. The sample has no outliers, and the population from which the sample is drawn is normally distributed or the sample size, n, is large (n > 30).
3. The sampled values are independent of each other.

Based on these assumption, we can assume the samples follow T-Distribution.

**Claim:** The opening stock prices for GE and BAC are different.

**Hypothesis:**

Ho (Null Hypothesis): The Opening Stock Prices for GE and BAC are same. Price(GE) = Price(BAC) or, Price(GE) - Price(BAC) = 0

Ha (Alternate Hypothesis): The Opening Stock Prices for GE and BAC are not same. Price(GE) not equal to Price(BAC) or, Price(GE) - Price(BAC) not equal to 0.

**Determine the claim:**

As p-value (1.166751e-10) is very less, and much lesser than alpha (05%), which provides very strong evidence to reject Null Hypothesis. Based on p-value we can reject Null Hypothesis (Ho): The Opening Stock Prices for GE and BAC are same.

We accept the Alternate Hypothesis, i.e. The Opening Stock Prices for GE and BAC are not same.

**R Code:**

*#Hypothesis Testing: Mean with Sigma Not Known*

*# The sample is obtained using simple random sampling or from a randomized experiment.*

*# The sample has no outliers, and the population from which the sample is drawn*

*#is normally distributed or the sample size, n, is large (n > 30).*

*# The sampled values are independent of each other.*

*# On these assumption, we can assume the samples follow T-Distribution*

*#Claim: The opening stock prices for GE and BAC are different.*

*#Hypothesis:*

*#Ho (Null Hypothesis): The Opening Stock Prices for GE and BAC are same. Price(GE) = Price(BAC)*

*#OR-- Price(GE) - Price(BAC) = 0*

*#Ha (Alternate Hypothesis): The Opening Stock Prices for GE and BAC are not same.*

*#Price(GE) not equal to Price(BAC) OR-- Price(GE) - Price(BAC) not equal to 0.*

*#Two Tail t-Distribution test*

*#Get Difference of Prices*

*Diff.Prices <- GE.Sample.Trimmed - BAC.Sample*

*Diff.Prices.Mean <- mean(Diff.Prices)*

*Diff.Prices.SD <- sd(Diff.Prices)*

*n <- length(Diff.Prices)*

*#Find test statistic: ts = (Xbar - Muo)/(sd/sqrt(n))*

*ts <- ((Diff.Prices.Mean-0)/(Diff.Prices.SD/sqrt(n)))*

*#Calculate p=value*

*p.value = 2 \* (pt(abs(ts),df=n-1,lower.tail=FALSE))*

*p.value*

*#Determine the claim*

*#As p-value (1.166751e-10) is very less, much lesser than alpha (05%),*

*#which provides very strong evidence to reject Null Hypothesis. 𝛼*

*#Based on p-value we can reject Null Hypothesis*

*#(Ho): The Opening Stock Prices for GE and BAC are same.*

*# We accept the Alternate Hypothesis, i.e.*

*#The Opening Stock Prices for GE and BAC are not same.*