Analytical Methods for Business
Pablo Zumba

Preprocessing

1. Load the data file associated with this assignment into R. This file contains NCUA data on 2262 credit unions operating in the United States. This will be your master data set. Variables are:

   charter.number: The number of the governmentally-issued charter or founding document of the credit union.

   name: The name of the credit union.

   city: The city in which the credit union's headquarters is located.

   state: The US state in which the credit union's headquarters is located.

   members: The number of members of the credit union.

   total.assets: The total assets of the credit union in millions of US dollars.

   total.loans: The total loans held as assets by the credit union in millions of US dollars.

   total.deposits: The total funds deposited with the credit union in millions of US dollars.

```
#Pablo Zumba
#U54252888
rm(list=ls())
library(rio)
masterDataset = import("6304 Module 3 Assignment Data.xlsx")
attach(masterDataset)
str(masterDataset)
> str(masterDataset)
'data.frame':    2262 obs. of  8 variables:
 $ charter.number: num  6 12 13 22 28 42 47 48 53 60 ...
 $ name          : chr  "THE NEW ORLEANS FIREMEN'S" "FRANKLIN TRUST" "EFCU
FINANCIAL" "WATERBURY CONNECTICUT TEACHERS" ...
 $ city          : chr  "METAIRIE" "HARTFORD" "BATON ROUGE" "WATERBURY" ...
 $ state         : chr  "LA" "CT" "LA" "CT" ...
 $ members       : num  26573 9311 53556 19760 117191 ...
 $ total.assets  : num  226.3 61.6 687.5 302.9 989.7 ...
 $ total.loans   : num  146.1 21.7 558.3 199.5 779 ...
 $ total.deposits: num  204.5 59.2 601.4 262.9 847.5 ...
```

2. Use R to create a) a data frame with California credit unions, b) a data frame with Florida credit unions, and c) a data frame with all credit unions based in New York or New Jersey. These are your intermediate data frames.

```
#2 Processing intermediate datasets for CA, FL, NY-NJ:
ca_CU = subset(masterDataset, state == "CA")
fl_CU = subset(masterDataset, state == "FL")
nynj_CU = subset(masterDataset, state == "NY" | state == "NJ")
```

3. Use the method demonstrated in class to take a random sample of n=20 from each of your intermediate data frames. These are your primary data frames. In total you will have created have 7 data frames.

```
#Processing 3: Taking 20 random samples from each intermediate dataframe
set.seed(54252888)
ca_20samples = ca_CU[sample(1:nrow(ca_CU),20),]
fl_20samples = fl_CU[sample(1:nrow(fl_CU),20),]
nyny_20samples = nynj_CU[sample(1:nrow(nynj_CU),20),]
```

Analysis

Using R and your primary data sets report the following:

1. The R code and resulting 90% confidence interval on the members variable for Florida credit unions. Give a clear written interpretation of your confidence interval.

```
#Analysis 1
results_fl = t.test(fl_20samples$members, conf.level = 0.9)
results_fl$conf.int
width_fl = results_fl$conf.int[2]-results_fl$conf.int[1]
width_fl
hist(fl_20samples$members,col="brown",main="Histogram of FL members")
abline(v=results_fl$conf.int[1],lwd=3,col="blue")
abline(v=results_fl$conf.int[2],lwd=3,col="blue")
abline(v=mean(fl_20samples$members),lwd=3,col="green")
> results_fl$conf.int
[1] 16714.91 39614.09
attr(,"conf.level")
[1] 0.9
> width_fl = results_fl$conf.int[2]-results_fl$conf.int[1]
> width_fl
[1] 22899.17
```

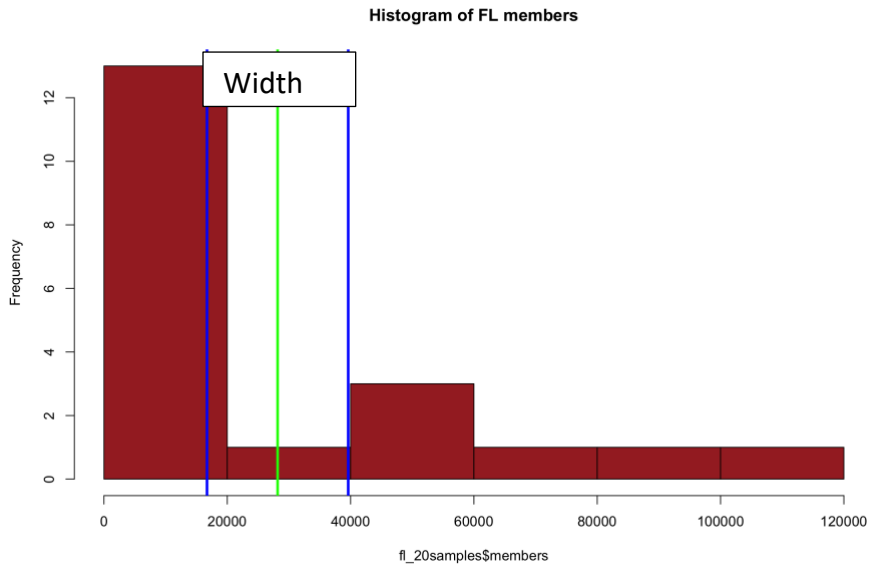


Figure 1 Histogram of FL members at conf.int=0.9

Interpretation: We can be 90% sure that the "Population Mean" on the members variable for Florida credit unions is between 16714.91 and 39614.09. In this case, with 20 random samples, the t-test gives us a sample mean of 28164.5 and width = 22899.17 which is represented in Figure 1.

2. The R code and resulting 99% confidence interval on the members variable for Florida credit unions. State the difference in width for the two confidence intervals.

```
#Analysis 2
results_fl2 = t.test(fl_20samples$members, conf.level = 0.99)
results_fl2$conf.int
width_fl2 = results_fl2$conf.int[2]-results_fl2$conf.int[1]
width_fl2
hist(fl_20samples$members,col="brown",main="Histogram of FL members at
conf.int=0.99")
abline(v=results_fl2$conf.int[1],lwd=3,col="blue")
abline(v=results_fl2$conf.int[2],lwd=3,col="blue")
abline(v=mean(fl_20samples$members),lwd=3,col="green")
> results_fl2 = t.test(fl_20samples$members, conf.level = 0.99)
> results_fl2$conf.int
[1]  9220.603 47108.397
attr(,"conf.level")
[1] 0.99
> width_fl2 = results_fl2$conf.int[2]-results_fl2$conf.int[1]
> width_fl2
[1] 37887.79
```

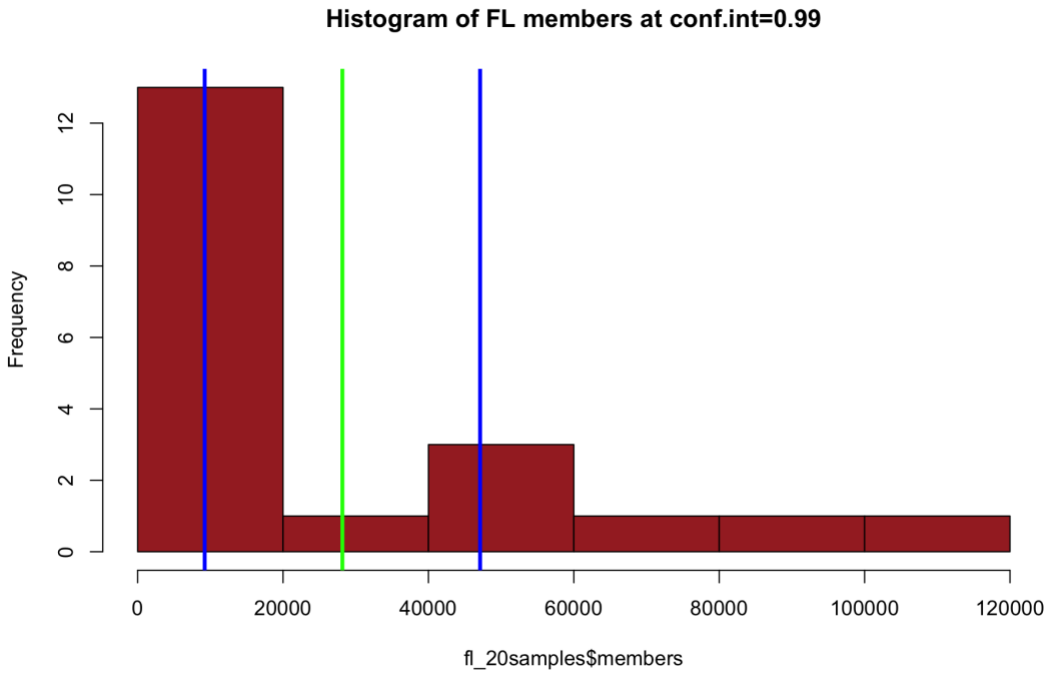


Figure 2 Histogram of FL members at conf.int=0.99

Interpretation: We can be 99% sure that the "Population Mean" on the members variable for Florida credit unions is between 9220.603 and 47108.397. Having these results (width at 90% = 22899.17) and (width at 99% = 37887.79), we can state that the confidence interval is directly proportional to the width. In other words, as the confidence interval increases, the width also increases. This can be noticed if we compare the width in Figure 1 against Figure 2.

3. The R code and results of a hypothesis test on the total assets of California credit unions, testing whether the population mean is greater than $170 million. Report the null and alternate hypotheses and the p-value resulting from the hypothesis test. Give a brief but clear written interpretation of the results of the test.

```
hist(ca_20samples$total.assets,col="red",main="20 samples of total assets of
CA")
abline(v=170,lwd=3,col="blue")
abline(v=mean(ca_20samples$total.assets),lwd=3,col="green")
summary(ca_20samples$total.assets)
mean(ca_20samples$total.assets)
sd(ca_20samples$total.assets)
moments::skewness(ca_20samples$total.assets)
moments::kurtosis((ca_20samples$total.assets))
t.test(ca_20samples$total.assets,mu=170,alternative = "greater")
> summary(ca_20samples$total.assets)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  48.93   71.88  117.61  214.37  379.55  643.38
> mean(ca_20samples$total.assets)
[1] 214.3667
```

```
> sd(ca_20samples$total.assets)
[1] 181.4441
> moments::skewness(ca_20samples$total.assets)
[1] 0.9010709
> moments::kurtosis((ca_20samples$total.assets))
[1] 2.568731
> t.test(ca_20samples$total.assets,mu=170,alternative = "greater")

      One Sample t-test

data:  ca_20samples$total.assets
t = 1.0935, df = 19, p-value = 0.1439
alternative hypothesis: true mean is greater than 170
95 percent confidence interval:
 144.2121      Inf
sample estimates:
mean of x
 214.3667
```

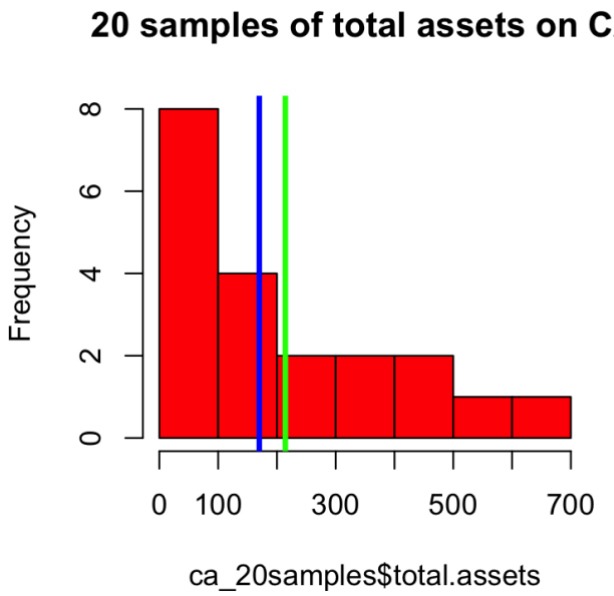


Figure 3 Histogram of the total assets in CA

Interpretation: The t-test was set up for alternative = "greater" I.e. $\boldsymbol{H_0 : u_{CA} \leq 170}$ and $\boldsymbol{H_a : u_{CA} > 170}$. Since the p-value=0.1439 is greater than 5%, we failed to reject the null, that is, the true sample mean of the total assets in CA credit unions is below $170 million despite the histogram in Figure 3 showing the mean is greater than 170$.

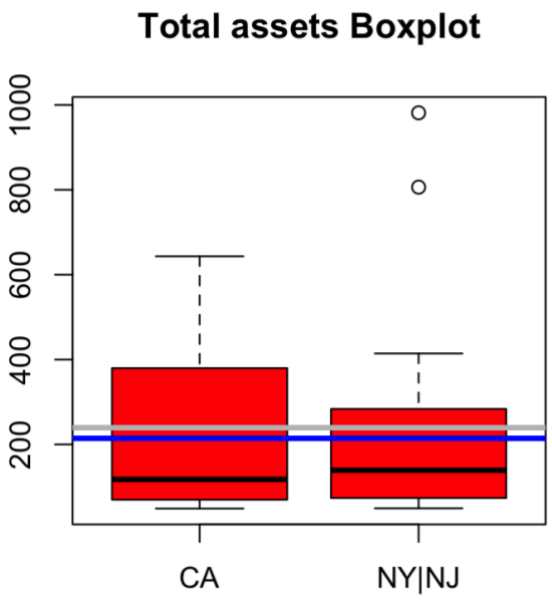
4. The R code and results of a hypothesis test determining if the population mean total.assets differs between California and New York/New Jersey. Report the null and alternate hypotheses and the p-value resulting from the hypothesis test. Give a brief but clear written interpretation of the results of the test.

```
#Analysis 4 Independent Sampling or Paired comparisons
pairedComparison=t.test(ca_20samples$total.assets,nyny_20samples$total.asset
s,mu=0, alternative = c("two.sided"))
pairedComparison
boxplot(ca_20samples$total.assets,nyny_20samples$total.assets,notch=FALSE,co
l="red",
        main="Total assets Notched Boxplot",
        names=c("CA","NY|NJ"))
abline(h=mean(ca_20samples$total.assets),col="blue",lwd=3)
abline(h=mean(nyny_20samples$total.assets),col="grey",lwd=3)
> pairedComparison

      Welch Two Sample t-test

data:  ca_20samples$total.assets and nyny_20samples$total.assets
t = -0.36287, df = 34.66, p-value = 0.7189
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -165.3934  115.2481
sample estimates:
mean of x mean of y
 214.3667  239.4394

>
boxplot(ca_20samples$total.assets,nyny_20samples$total.assets,notch=FALSE,co
l="red",
+         main="Total assets Notched Boxplot",
+         names=c("CA","NY|NJ"))
> abline(h=mean(ca_20samples$total.assets),col="blue",lwd=3)
> abline(h=mean(nyny_20samples$total.assets),col="grey",lwd=3)
```



Interpretation: The Null Hypothesis states that there is no difference in the population mean between CA and NY|NJ; I.e. $\boldsymbol{H_0}: \boldsymbol{u_{CA}} - \boldsymbol{u_{NY|NJ}} = 0$ and $\boldsymbol{H_a}: \boldsymbol{u_{CA}} - \boldsymbol{u_{NY|NJ}} \neq 0$. Since the p-value=0.7189 is

greater than 5% we failed to reject the null, that is, the population means of the total.assets between California and New York/New Jersey do not differ too much from each other (they are very similar) and it can be corroborated on the Total assets boxplot where the blue line corresponds to the sample mean of CA and the grey line to the sample mean of NY|NJ.

5. The R code and results to extract and report only the a) name, b) city, c) members, and d) total assets of the largest Florida credit union in your sample.

```
#Analysis 5 | The largest FL Credit Union
#Assuming the "largest" means the one that has more members
attach(fl_CU)
fl_CU[which.max(members), c("name", "city", "members", "total.assets")]
> fl_CU[which.max(members), c("name", "city", "members", "total.assets")]
        name        city members total.assets
1352 FLORIDA GAINESVILLE  126483     1656.334
```