Math 332 Water Quality Final Report

Sean Tocci

4/21/2018

Water Quality Across the World

Abstract

Water makes up about 70% of our planet but only 2% of that is freshwater. Our freshwater is becoming more and more in demand with our growing population on the planet and changing climate. The United Nations provides water quality data from 1990 to 2015 for over 200 countries. The data set is used to generate questions about trends or patterns one might find. A big assumption implemented on the data is the omitting NA values but this gives a more accurate picture of the data.

Introduction

Water is one of the most abundant resources on the planet and is used for anyting from growing our food to generating electricity. Today our freshwater is under more stress than ever due to our growing population and changing climate. Looking at water quality in our past can help tell a story on how we got here and where we are going. The goal is better understand the data and get possible questions about it.

United Nations Data Set

The United Nations provides a comma separated value or CSV file of all countries from the year 1990 to 2015. The code below show the data set which contains 3 variables. These variables are columns in our are data set. Our columns are: Country, Year, and Value. Country is a string that holds the name of the country. Year is an integer. Value is a numeric score from 0 to 100 with 100 being perfect water quality to 0 being very unhealthy and likely dangerous.

```
##       Country Year Value
## 1 Afghanistan 2015  55.3
## 2 Afghanistan 2014  55.2
## 3 Afghanistan 2013  53.4
## 4 Afghanistan 2012  51.6
## 5 Afghanistan 2011  49.8
## 6 Afghanistan 2010  48.0
```

Methods

Some countries may be missing data for a year and that is represented by an NA value. To deal with this we simply removed all years that have an NA value. Some countries who have no data at all such as Bermuda, they will be completely removed from the data set. There is a 4th column called other which contains no information and is removed by setting it to NULL then omitting NA values.

```
UnitedNationsData <- read.csv('UNdata.csv', header=TRUE,sep=",",col.names = c('Country','Year','Value',
UnitedNationsData$Other <- NULL
UnitedNationsData<-na.omit(UnitedNationsData)
```

Results

After we have removed all our NA we can now start to get an idea on how the data is starting to look such as how many countries we will be working with.

```
AllCountriesMultipleValues<-UnitedNationsData$Country
OnlyUniqueCountries<-unique(AllCountriesMultipleValues)
result<- length(OnlyUniqueCountries)
paste0('The number of unique countries in our dataset is: ',result)
```

```
## [1] "The number of unique countries in our dataset is: 212"
```

This also raises many other questions about the data set and how we can use it further understand the water dilemma. Some of these being:

- How might the water quality be different between neighboring countries?
- Which country has the best water? Which has the worst?
- How might a natural disaster affect the water quality?
- Is there an particular region with bad water?

There are many more but these are just a few you might think of investigating.

Discussion

A big assumption that might not have been obvious is omitting all NA values. There are two choices: replace all NA with a value or remove them all. Choosing to replace them with a value, usually 0, would have thrown off our results since 0 in this case would mean their water for that year was the worst it could be. This would be detrimentally to finding trends in the data from year to year. So by choosing to omit all NA values we lose some data but keeps the validity of the data.

References

United Nations Water Data: http://data.un.org/Data.aspx?q=water&d=MDG&f=seriesRowID%3a665

Competing for Clean Water Has Led to a Crisis. (2017, January 27). Retrieved from https://www.nationalgeographic.com/environment/freshwater/freshwater-crisis/

United Nations Water Quality of the World: Descriptive Statistics

Abstract

Descriptive statistics are helpful for getting an understanding of the data and how it is distributed. There are many techniques to test to see if the data we have is irregular or strange. When looking at water quality from the years 2000 and 2015 we can see trends between them. Through these techniques we discovered that for both of the data sets they are negatively skewed meaning they have a negative tail. This is useful since now we know there are some low water quality values that are affecting our averages or mean and to be careful knowing this going forward.

Introduction

The first step of any data analysis to examine the descriptive statistics of the data set. This provides insight on how the data is distributed and shaped. More in particular they are used to help understand or describe a

feature in our data set. We begin here since starting by looking at all the data can become very confusing quickly. So by starting by looking at one feature such as year or country, we can learn the data set and know how to go to more complex figures.

Methods

For our descriptive statistics we will only be using uni-variate variables. This means we won't be looking at all our columns, country, year, and values, but just one. For this we will be looking at all the values for the year 2015 since it is the most recent. The techniques applied on the year 2015 are: mean, median, quantiles, standard deviation, histogram, skewness, kurtosis, and quantile plot. These techniques are shown by the R command window below:

```
year2015<- subset(UnitedNationsData, Year==2015)
hist(year2015$Value)
summary(year2015$Value)
sd(year2015$Value)
kurtosis(year2015$Value)
skewness(year2015$Value)
qqnorm(year2015$Value)
qqline(year2015$Value)
```

We now create another histogram and quantile plot for the year 2000. This will be used to compare results and see a trend over the 15 years.

Results

To get our mean, median, max, and quantiles we used a summary method built in R. To get the kurtosis and skewness the library called moments was used. The results are shown below:

```
## [1] 100
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   40.00   86.65   96.30   89.77   99.70  100.00
```

```
## [1] "The mode for year 2015 is:"
```

```
## [1] 100
```

```
## [1] "The skewness for year 2015 is:"
```

```
## [1] -1.580149
```

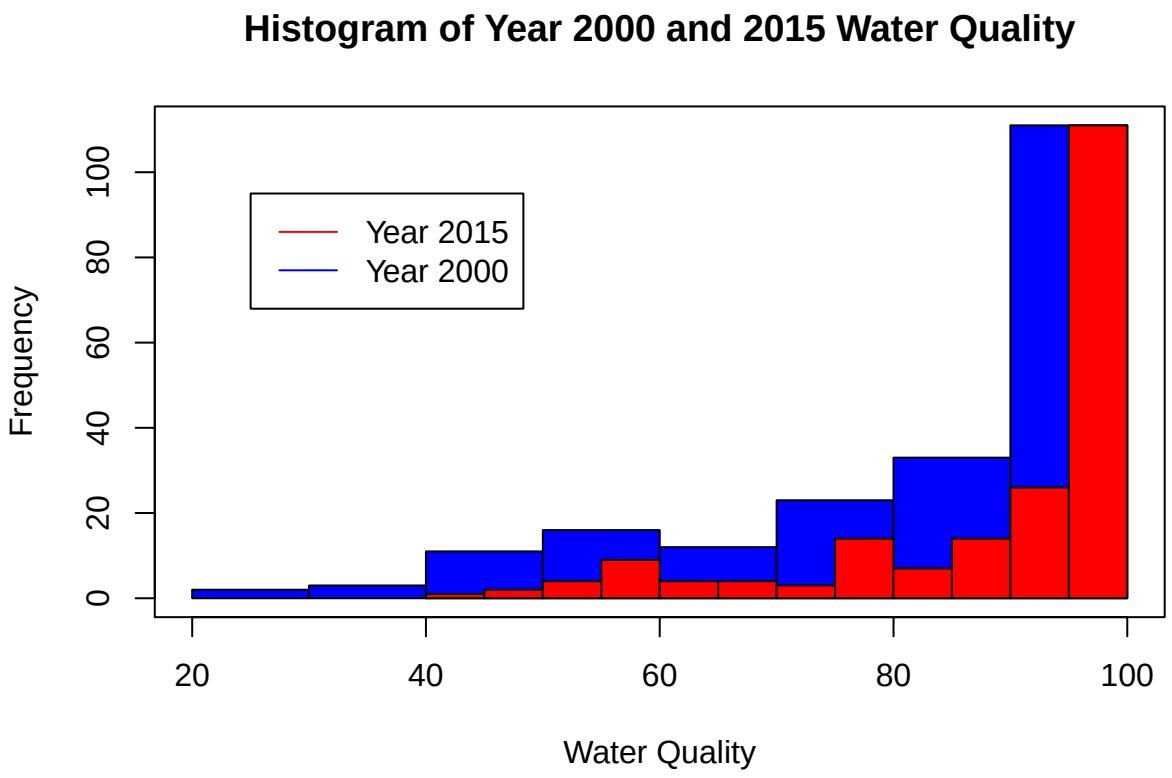
```
## [1] "The kurtosis for year 2015 is:"
```

```
## [1] 4.513122
```

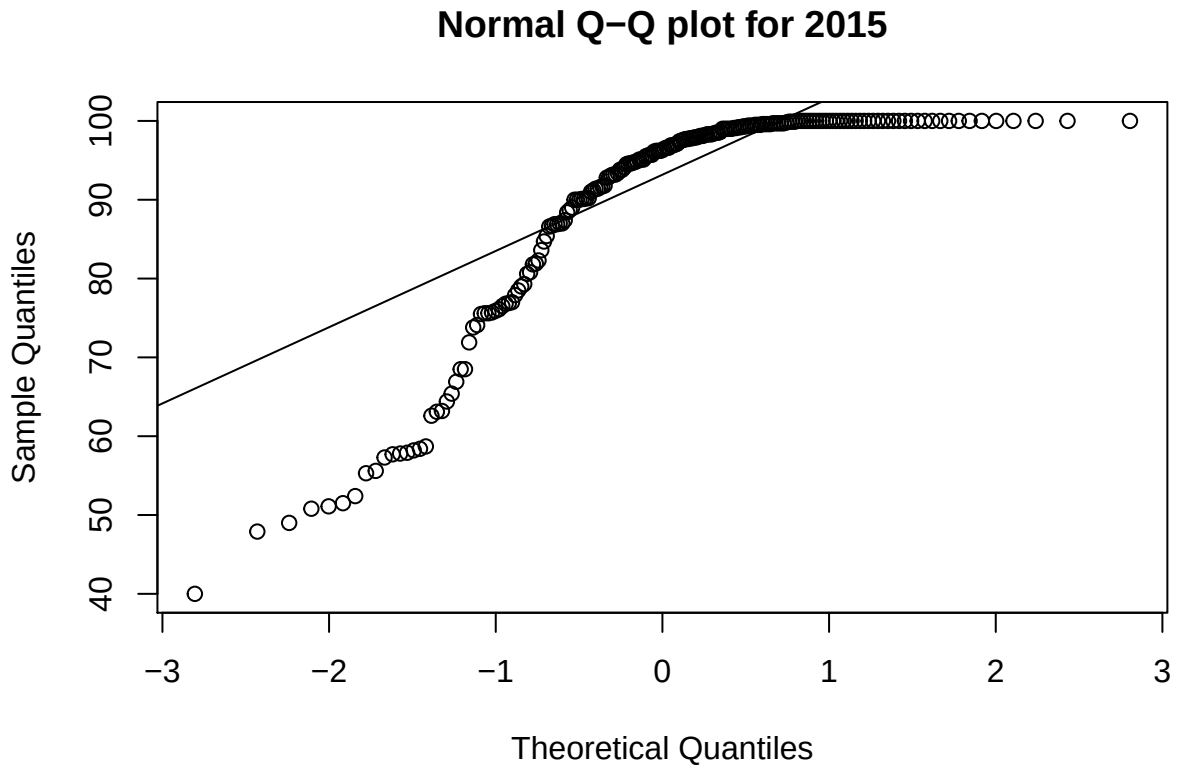
```
## [1] "The Standard Deviation for year 2015 is:"
```

```
## [1] 14.08849
```

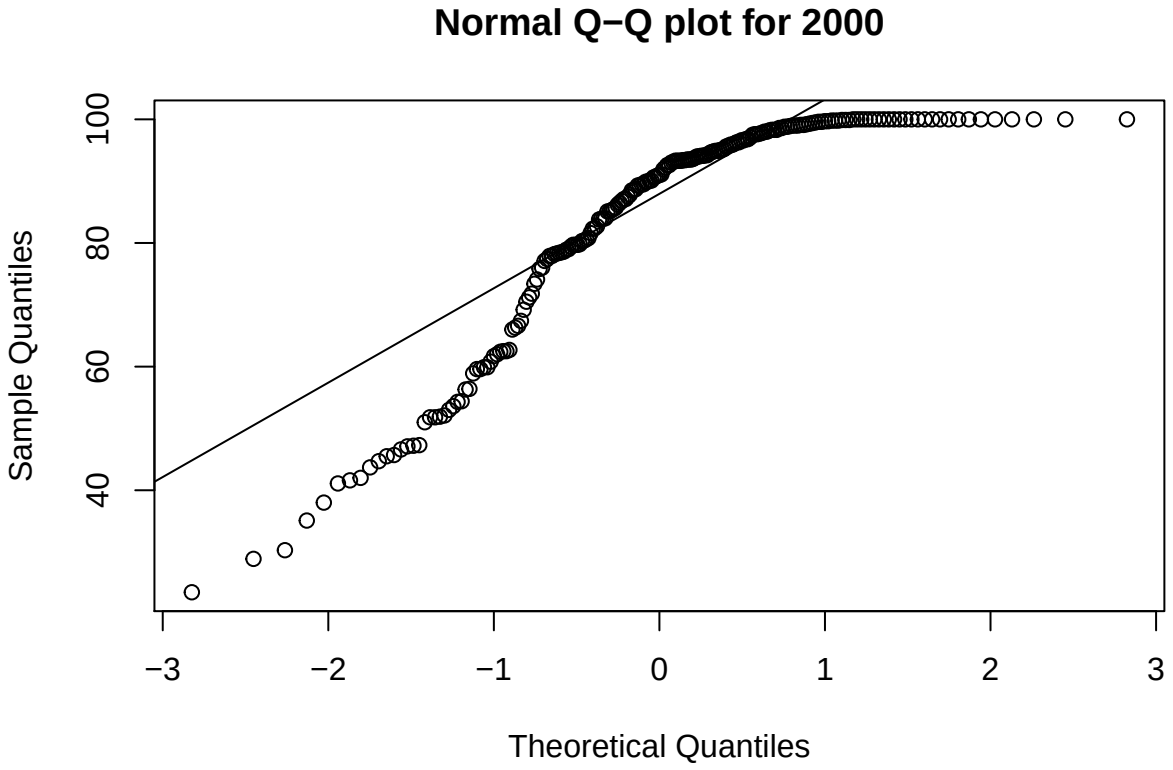
Our histogram shows year 2000 shown in blue and year 2015 in red. In year 2000 there are 12 more values taken. Year 2000 has 211 values used where as 2015 only has 199.



Our Q-Q plot for the year 2015 is shown below. Only 199 values were used for this plot.



Our Q-Q plot for the year 2015 is shown below. This plot uses 211 values which is 12 more than year 2015.



Discussion

The goal of this section is to break down each statistically technique and interpret what the value means. For the histogram and Q-Q plots the goal is to show what is being shown and how it useful in understanding how the data is distributed.

Mean, Median, Mode, and Quantiles

Mean is the total water quality data added up then divided by the number of observations. This will tell you the average water quality for the world for that year. In 2015 the mean was 89.77 where as the mean for 2000 is 83.69. This is saying that over 15 years overall water quality of the world was raised by 6.08. Now if you divide this by 15 years you can get a rough estimate of the yearly growth which .4053. This isn't 100% accurate since NA omit may have removed different countries for for these years if they missed it. Median is the middle value of the data set. This is useful for a set that may have very large or very small outliers that could skew the mean making not be 100% true to the true mean. The median value for 2015 is 96.3 and 91.0 for 2000. This is saying that the mean is could be pulled down more by smaller values. This is expected since the max you can get is 100 so there are no large outliers but a country who could have been hit a natural disaster would have a worse score and thus bring down the mean. The mode for both year is 100 this is saying the most most frequent value seen in the water quality data is 100. This makes sense as it the max score and many countries have great water quality for years and only gets better. This is also shown clearly in the histogram later. Quantiles are helpful for seeing the distribution of the data. Quartiles tell you the value at which that percent of the data falls below. Quartiles are broken down into the minimum, 25%, Median, 75%, and the max. In 2015 you can see the minimum is 40, the 25% quartile says that 25% of the data is below 86.65, the median is 96.3, 75% quartile is saying that 75% of the data falls below 99.7, and the max is 100. In 2000 you can see the minimum is 23.5, the 25% is below 77.65, the median is 91, 75% quartile is saying that 75% of the data falls below 98.25, and the max is 100. The interesting part of this data that is

telling you that for both 2015 and 2000 the data falls heavily towards the upper end of the scale. This is interpreted as many countries having good water quality.

Skewness

Skewness is showing how the data is distributed. This means that if we have a negative value for our skewness our tail is below our central value. In 2015 and in 2000 both have a negative skewness so what does this mean? This means that more values fall above our central value which is saying there are more countries with good water quality than bad. The closer our value is to 0 the closer it is to being normally distributed.

Kurtosis

Kurtosis is showing how sharp or extreme the tails are. If our kurtosis was 0 this would mean are data is uniform and are tails aren't extreme. If we get a kurtosis near 5 and up live in year 2015 which has a kurtosis of 4.5 this is saying we have a sharp tail. This we already knew since there are a lot of high scores but some low score which creates the sharp tails. In year 2000 our kurtosis is 3.5 which is good since a normal distribution curve has a kurtosis of 3. This is saying our tails are extreme but also not flat.

Histograms

A histogram is a binned bar graph. This means it uses a bar to show an interval of values rather than a single value like in a bar graph. In our water quality data for the year 2000 we used 8 bins for example a bin for the year 2000 is [20,30] this means any value that falls between 20 and 30 we will add a point to this bin. This is usually for seeing which intervals are popular. For the year 2015 we used 12 bins since our data is more compacted in the higher score range. This allows us see a better trend of the data since if we used 8 bins our red bars would be very big at the end and wouldn't really show up for the lower scores. There is no correct number of bins to choose but the number of bins allows you to see how many values fall into that interval.

Q-Q Plots

A Quantile-quantile plot is similar histogram but now puts into comparison with a normal distribution. The line on the graph is representative of what the normal distribution would be but since we found out earlier our water quality data isn't normally distributed we would expect some irregularities. By looking at this we can see that many of the points between -3 and 0 theoretical quantiles are far away from the line. This is due to the fact that we are negatively skewwed. You can also see at the top of both Q-Q plots they are a lot of points clumped together which is due to many countries having high water quality but also messes up our normal distribution since many are repeating the same value.

References

Q–Q plot. (2018, April 10). Retrieved from https://en.wikipedia.org/wiki/Q\T1\textendashQ_plot

Measuring Skewness and Kurtosis. (n.d.). Retrieved April 20, 2018, from https://www.itl.nist.gov/div898/handbook/eda/section3/eda35b.htm

United Nations Water Quality of the World: Linear Regression

Abstract

Finding patterns and trends in data can be helpful for predicting future values. Using a simple linear model can be helpful to predict multivariate data. This gets messy when using many points such as the world water quality data for 25 years since your model overfits and won't be a good estimator. When used only one country, Afghanistan, the model was able to get a much higher score to predict into the future. Althought it might be accurate for a few years the model can't predict the growth of technology and use it as a factor so the model will have to updated after a few years.

Introduction

Trying to find a pattern or predict a future value can very difficult if you have alot of points. Using a linear regression you can see the trend of the data in a very simple fashion. A simple linear regression model will try to graph a line through the data to show what a future trend might be. When applying this to water quality data you could tell the trend of the water quality and when you can predict when it will reach a certain quality. This is useful to see how quickly a nation is improving their water.

Methods

For linear regression we are going to have to use multivariate data. This means we are using two or more features which in the water quality data are year, country, and value. First we will create a plot all the values for all countries from 1990 to 2015. Next we will create a linear model to fit to this data. After we will add this model to our other plot and look at the summary features of the linear model. This is shown below in code:

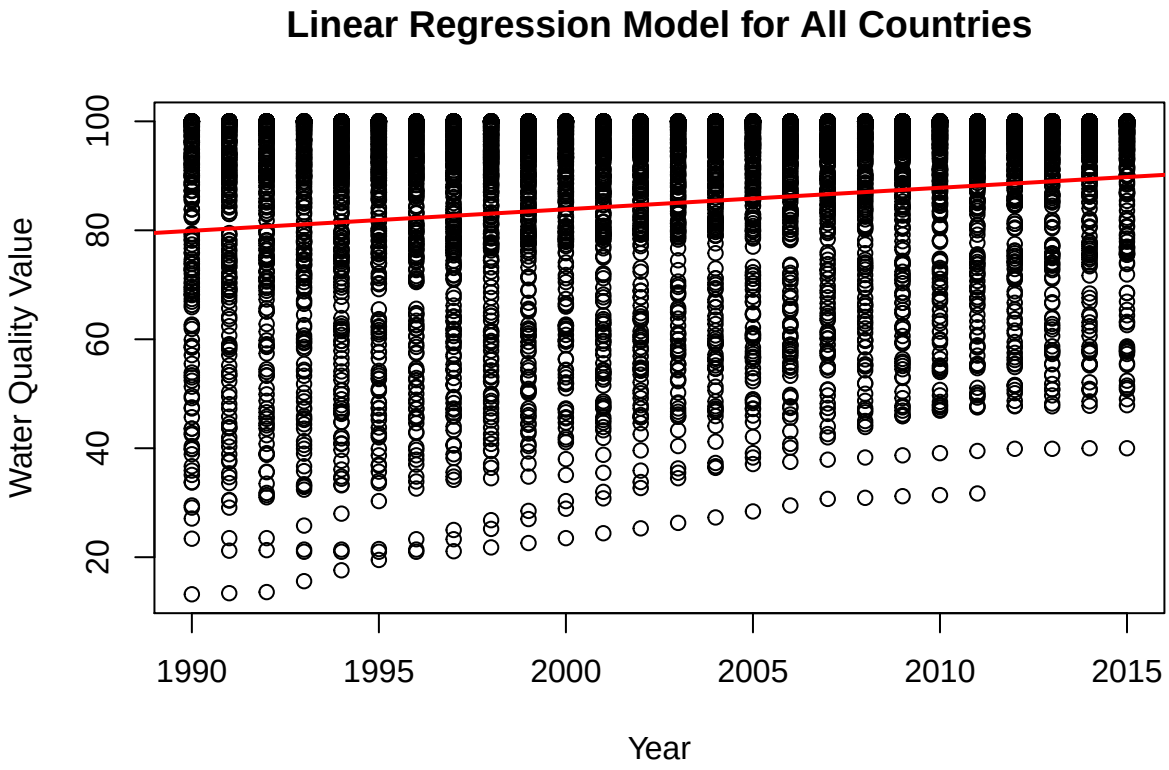
```
plot(UnitedNationsData$Year,UnitedNationsData$Value)
linModel<- lm(UnitedNationsData$Value~UnitedNationsData$Year)
abline(linModel,col="red",lwd=2)
summary(linModel)
```

Now we will do the same thing but for Afghanistan. After we will compare the linear models and see which model best fits the data.

```
plot(Afghanistan$Year,Afghanistan$Value)
linModel<- lm(Afghanistan$Value~Afghanistan$Year)
abline(linModel,col="red",lwd=2)
summary(linModel)
```

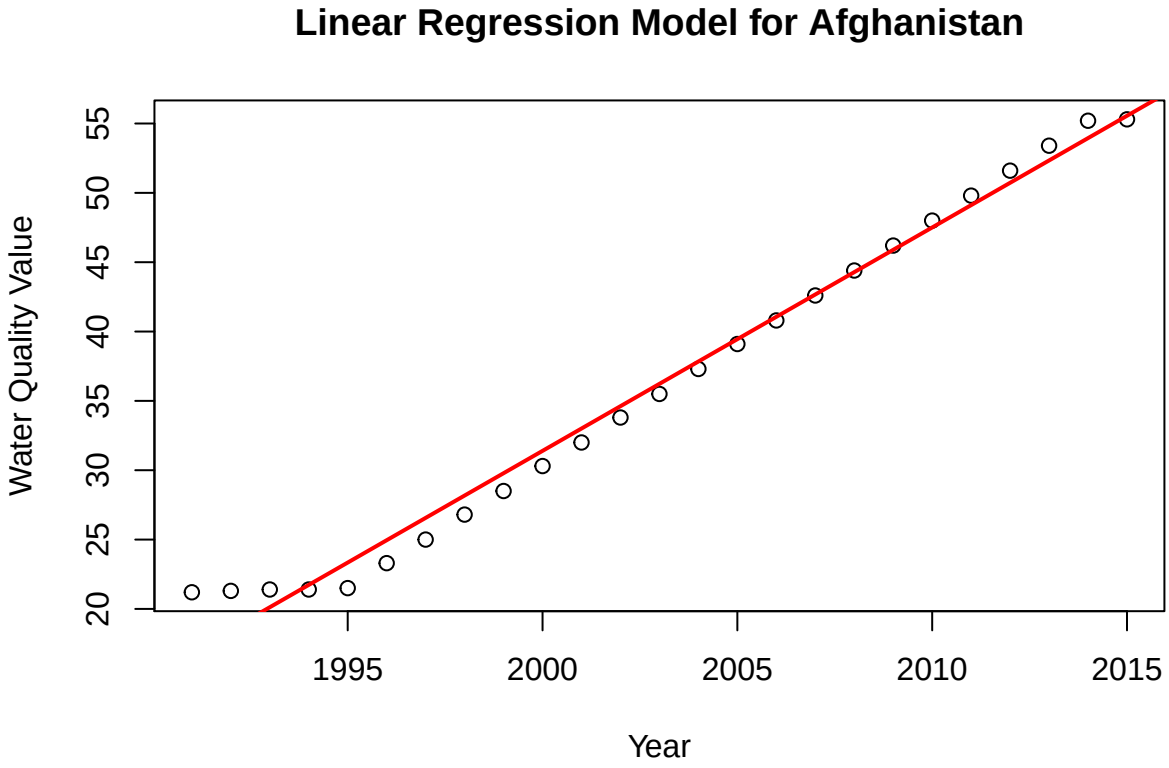
Results

The graph below shows all values plotted as a white circle with a black border. The linear model is shown as a red line.




```
##
## Call:
## lm(formula = UnitedNationsData$Value ~ UnitedNationsData$Year)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -67.101  -7.007   7.541  12.281  20.089
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)            -705.88647   66.38796  -10.63   <2e-16 ***
## UnitedNationsData$Year    0.39487    0.03315   11.91   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.76 on 5274 degrees of freedom
## Multiple R-squared:  0.0262, Adjusted R-squared:  0.02601
## F-statistic: 141.9 on 1 and 5274 DF,  p-value: < 2.2e-16
```

The graph below shows Afghanistan plotted as white circles with a black border. The linear model is still shown as a red line.




```
##
## Call:
## lm(formula = Afghanistan$Value ~ Afghanistan$Year)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.8418 -1.0065 -0.2603  0.6858  4.3012
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -3.190e+03  8.059e+01  -39.59   <2e-16 ***
## Afghanistan$Year  1.611e+00  4.023e-02   40.04   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.451 on 23 degrees of freedom
## Multiple R-squared:  0.9859, Adjusted R-squared:  0.9852
## F-statistic:  1603 on 1 and 23 DF,  p-value: < 2.2e-16
```

Discussion

World Linear Regression Model

The world regression model is very clustered and can be hard to tell what is going on. The red line is the predicted trend based on the world water quality data. By looking at you can tell its not a good fit for good reason. There are three types of countries in this data, one being those who have had perfect water quality for years, another being those whose water has been bad but got a little better, and then there are countries who have made a lot of progress in water quality. This makes trying to fit a model to this almost impossible

with so much data. This is called over-fitting and gives you a bad model and should not be used to predict future values.

Afghanistand Linear Regression Model

Looking at Afghanistan linear model you can tell that the red line captures the overall trend of the water quality data improving. From the summary you can see that under the Estimate column and Afghanistan$Year you see 1.611 this is your slope. The model is saying that for every year going by Afghanistan's water quality is going up by 1.611 which is relatively slow. How do we know if this is a good model though? By looking at our residual standard error which is also known as our mean square error. This tell us the difference between the actual value and what the model predicted. Our residual standard error is 1.451 which is a good indictation that our model is well fit and not over-fitted like the world linear regression model. Although it could be argued that this model won't be good for predicting far into the future since the model can't predict information being exchanged but that predicted on the points given. With technology always advancing making it hard to predict future water quality after a few years.

References

Mean squared error. (2018, April 10). Retrieved from https://en.wikipedia.org/wiki/Mean_squared_error

Linear Regression. (n.d.). Retrieved from http://www.stat.yale.edu/Courses/1997-98/101/linreg.htm

New York Harbor Data: Water Quality

Abstract

Being able to use sample data to make inferences about poplulation data is a powerful statistic techinique. Using sample data from New York harbor data we can create a 95% confidence interval for the population. The confidence interval for the top is between 516.3 and 536.19 and for the bottom is between 481.39 and 499.41. This is a safe range for the wildlife in New York harbor.

Introduction

Imagine having heights of 10 trees in a forest of 100 trees, could we use this to figure out the height of the other trees in the forest? With what level of confiedence? Using samples to predict the a feature of the population is a very powerful skill in statistics. Using a confidence interval allows us to find a range based on a sample that we can use on the population with a confidence level that is numeric since just having a good inclination that it will be isn't good enough. We will look at samples of disolved oxygen in the New York harbor water quality and generate confidence intervals for the whole harbor with 95% certainy.

Methods

First we have to import the data for New York Harbor data and change the type of DO.Top and DO.Bot to numeric. DO.Top is the disolved oxygen at the top of the harbor and DO.Bot is the disolved oxygen at the bottom of the harbor. Now we change the type of these two columns to numeric which will remove all NS or NA values.

```
harborData<-read.csv('betterHarbor.csv',header=TRUE)
harborData$DO.Top<-as.numeric(harborData$DO.Top)
harborData$Do.Bot<-as.numeric(harborData$Do.Bot)
head(harborData)
```

Next we get will get min, 1st quartile, median, mean, 3rd quartile, and max. We will also plot a histogram to see the distrubtion of the two layers compared to each other.

```
harborData<-read.csv('betterHarbor.csv',header=TRUE)
harborData$DO.Top<-as.numeric(harborData$DO.Top)
harborData$Do.Bot<-as.numeric(harborData$Do.Bot)
summary(harborData$DO.Top)
summary(harborData$Do.Bot)
hist(harborData$DO.Top, col = 'Blue', xlab = 'Oxygen in Water (mg/L)', main ='Histogram of Disolved Oxy

hist(harborData$Do.Bot, col = 'Red',add=TRUE)
legend(25, 190, legend=c("DO.Bot", "DO.Top"),
       col=c("red", "blue"), lty=1)
box()
```

Now that we know the distribution and summary statistics we can construct a confidence interval by using a package in R called Rmisc. This is a function call that you pass your DO.Top with your confidence level which for this case is .95 and it will return an interval corresponding to that. Then we will do the same with DO.Bot.

```
library(Rmisc)
CI(harborData$DO.Top,ci=.95)
CI(harborData$Do.Bot,ci=.95)
```

Results

First we will look at our disolved oxygen at the top statistics

```
## Loading required package: lattice
```

```
## Loading required package: plyr
```

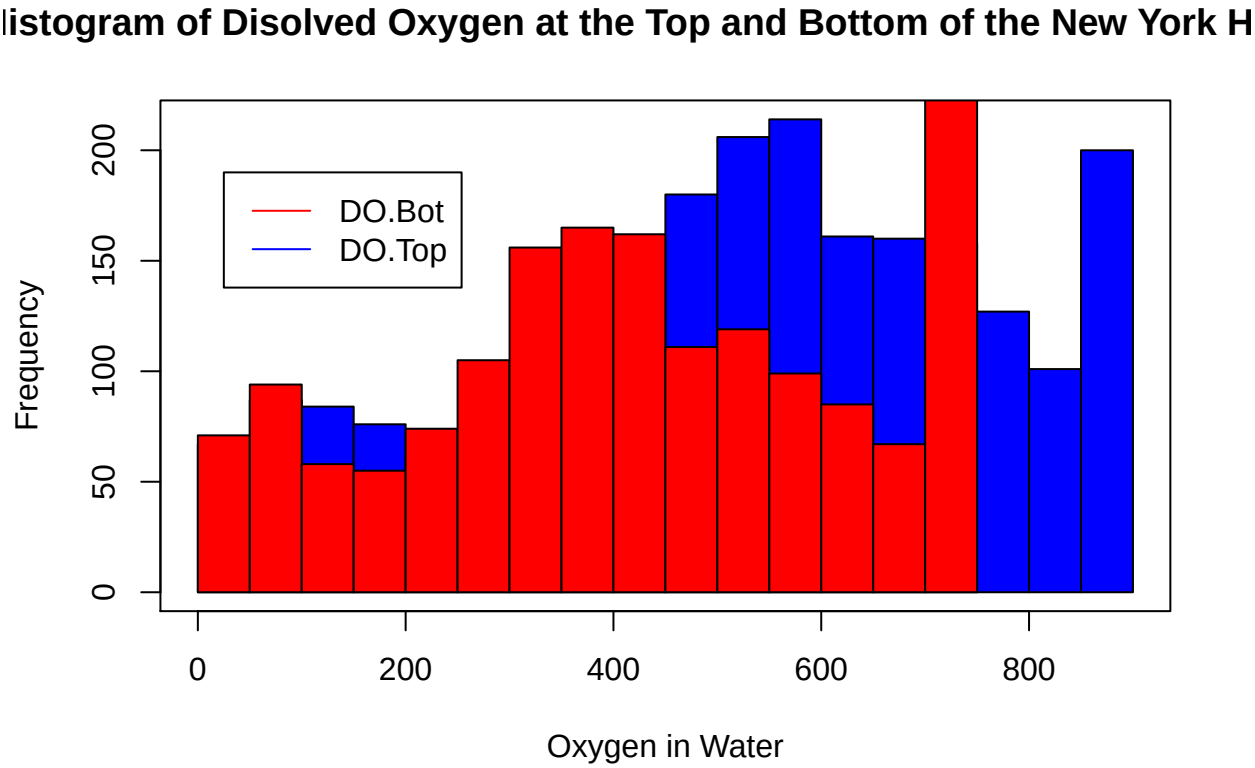
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     1.0   381.2   552.0   526.3   710.0   884.0
```

```
##    upper     mean    lower
## 536.1995 526.2520 516.3046
```

Now we will look at the disolved oxygen at the bottom statistics and the histogram with both of them on it.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     1.0   335.0   523.0   490.4   707.0   707.0
```

```
##    upper     mean    lower
## 499.4163 490.4062 481.3961
```



You can see in the histogram that there is more oxygen at the top of the harbor than the bottom. This is also shown in the confidence intervals by the ranges not overlapping.

Discussion

The confidence interval for the disolved oxygen at the top is saying that if another sample were to be test you could say with a 95% certainty that it will fall between 516.3 and 536.19. This is the same with the disolved oxygen at the bottom but between 481.39 and 499.41. We can conclude that there is more oxygen to be found at the top of the harbor than at the bottom. This matters because many fish can be harmed or killed if there is too much oxygen in water. Fish can also die from a lack of oxygen in the water. With a low disolved oxygen it will also kill of plants and slow down decompostion at the bottom of the harbor which could have other unintended effects. This infomation is usefully for mantaining health level for fish and other wild life in the harbor. The confidence intervals show safe ranges for the wild life in harbor and that there is nothing to worry about.

References

What are Confidence Intervals? (n.d.). Retrieved from https://www.itl.nist.gov/div898/handbook/prc/section1/prc14.htm

Global Water Quality: Comparing Means

Abstract

Water quality varies drastically across the globe but how can we prove it statistically? Through the use of two techniques, t-test and bootstrapping, we are able to prove if our data sets are significantly different or not.

Introduction

Water quality throughout the world has a wide range. How can we compare two or more countries and prove if their water is significantly different or not? This is done through comparing the means. There are two main techniques that are important for comparing means, a t-test and bootstrapping. Both of these use the difference in means to prove statistically with a level of confidence if the data sets are the same or not.

Methods

First we will load in the United Nations Water quality data. Next we will removed the extra Other column in the data set since it contains no information. Now we will omit all values that have an NA value.

```
UnitedNationsData <- read.csv('UNdata.csv', header=TRUE,sep=",",col.names = c('Country','Year','Value',
UnitedNationsData$Other <- NULL
UnitedNationsData<-na.omit(UnitedNationsData)
```

Now we will get the water quality for three neighboring countries, the United States, Canada, and Mexico. These are the countries will be preforming t-test and boot strapping methods on.

```
Mexico <- subset(UnitedNationsData, Country=='Mexico')
Canada <- subset(UnitedNationsData, Country=='Canada')
USA <- subset(UnitedNationsData, Country=='United States')
```

For the t-test we will first take Mexico and compare it with Canada. Next we will preform a t-test on Canada and USA. Finally we will compare USA and Mexico.

```
t.test(Canada$Value,Mexico$Value,paired = TRUE)
t.test(Canada$Value,USA$Value,paired = TRUE)
t.test(USA$Value,Mexico$Value,paired = TRUE)
```

For our bootstrapping method we will take two countries and their absolute difference and compare them 100,000 with our psuedo A and psuedo b. First we will compare Canada and Mexico.

```
difference=abs(mean(Canada$Value)-mean(Mexico$Value))
pooleddata<-c(Canada$Value,Mexico$Value)
dmeans <- numeric(100000) # vector to store means
for(i in 1:100000){
  g<-sample(52,26)
  dmeans[i]<-abs(mean(pooleddata[g])-mean(pooleddata[-g]))
}
tails <- which(dmeans >abs(mean(Canada$Value)-mean(Mexico$Value)))
length(tails)/100000

hist(dmeans, col = "gray")
abline(v=difference,col="red",lwd=2)
```

These steps are repeated for USA and Canada. Then again for USA and Mexico.

Results

For our t-test score between Canada and Mexico we got:

```
##
##  Paired t-test
##
## data:  Canada$Value and Mexico$Value
```

```
## t = 11.775, df = 25, p-value = 1.077e-11
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   8.244547 11.740068
## sample estimates:
## mean of the differences
##                9.992308
```

For our t-test score between Canada and Mexico we got:

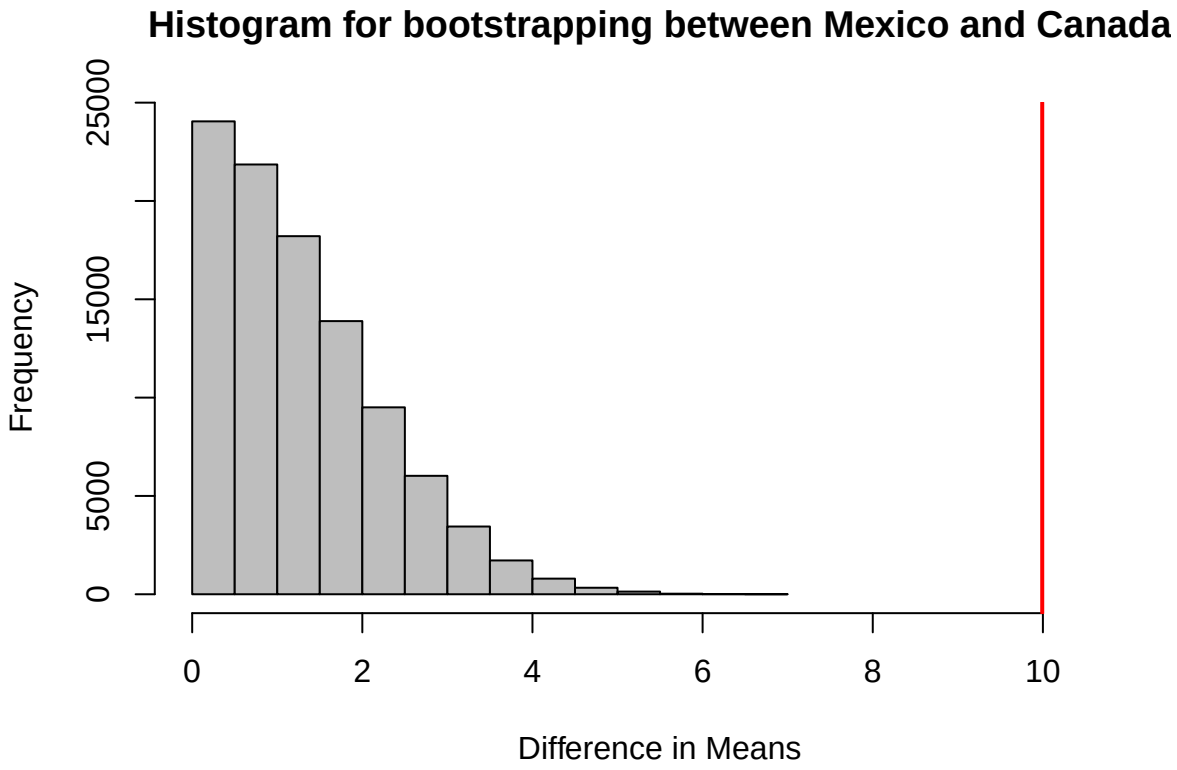
```
##
##  Paired t-test
##
## data:  Canada$Value and USA$Value
## t = 19.285, df = 25, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.8451083 1.0471994
## sample estimates:
## mean of the differences
##               0.9461538
```

For our t-test score between Canada and Mexico we got:

```
##
##  Paired t-test
##
## data:  USA$Value and Mexico$Value
## t = 11.308, df = 25, p-value = 2.54e-11
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   7.398514 10.693794
## sample estimates:
## mean of the differences
##                9.046154
```

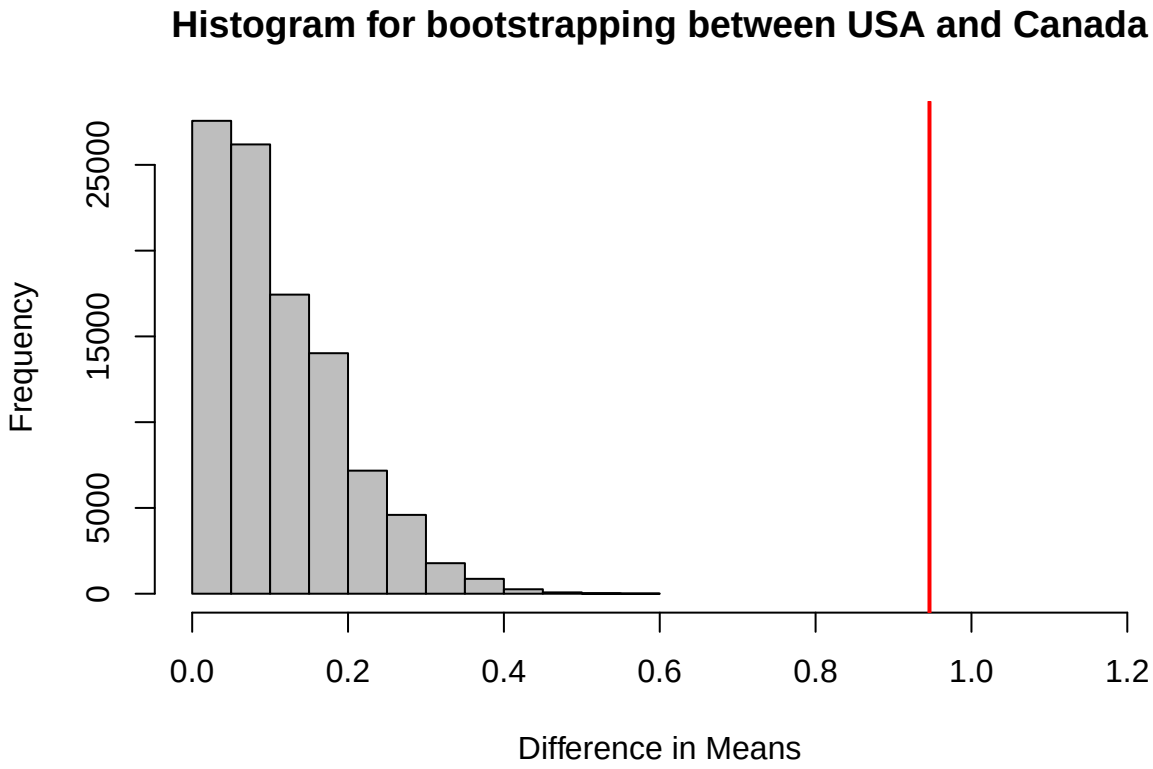
For our bootstrapping between Mexico and Canada we got:

```
## [1] 0
```



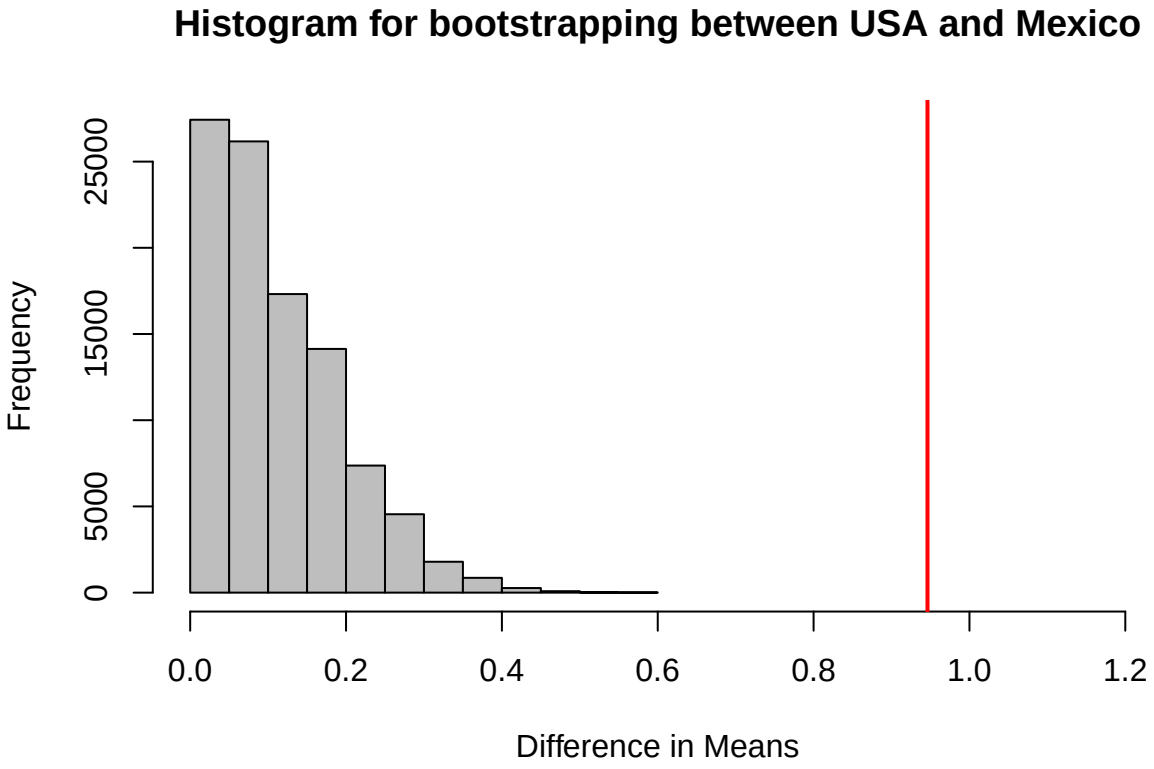
For our bootstrapping between USA and Canada we got:

```
## [1] 0
```



For our bootstrapping between USA and Canada we got:

```
## [1] 0
```



Discussion

T-Test

For all of the t-test between the 3 countries the p-value returned is very small. This means the null hypothesis has a very small chance of being true. For our water quality data our null hypothesis is that there is a difference of means is equal to 0. Our p value is the chance or likelihood that the null hypothesis is true and since our p-values are very small we can say with confidence that for all the sets are significantly different. By doing a 3 paired t-test were are allowing for more error since each test already has a 5% for each test. This doesn't play much of role in the water quality in the 3 countries since our p-value is so low but it is something to be careful of.

Bootstrapping

Bootstrapping is when you take the difference between means of two data sets so for this example 2 countries and you compare it with the differnce in your psuedo A and psuedo B. This techinique is general used when you don't have many data points and need to generate more. Your two psuedo columns come from pooling your data sets together or countries in this case and randomly sampling half into psuedo A and the others into psuedo B. Then you compare the true difference with the 100,000 psuedo As and psuedo Bs differences and count how many are larger than your true difference. In each of the histograms the red line represents the true differnce and as you can see the true difference is much larger than all the psuedo A and B differences. This means that our data isn't very similiar because your true difference should be closer to your psuedo A and B difference the closer your data sets are to each other.

References

Students t-test. (2018, April 23). Retrieved from https://en.wikipedia.org/wiki/Students_t-test

The T-Test. (n.d.). Retrieved from https://socialresearchmethods.net/kb/stat_t.php

Permutation Tests. (n.d.). Retrieved from https://thomasleeper.com/Rcourse/Tutorials/permutationtests.html

What is the Bootstrap? (n.d.). Retrieved from https://www.methodsconsultants.com/tutorial/what-is-the-bootstrap/