4/18/18

ISYE 2028 Project: Crime in the United States

Akwule, Preston & Yoyin, Ayotunde

Table of Contents

[Module 1: Introduction and Data Collection 4](#_Toc511750138)

[Introduction 4](#_Toc511750139)

[Data Collection and the Cleaning Process 4](#_Toc511750140)

[Variable List 6](#_Toc511750141)

[Preview of the Project 9](#_Toc511750142)

[Module 2: Modeling Data with Probability Distributions 10](#_Toc511750143)

[Module Introduction 10](#_Toc511750144)

[Continuous Variable Analysis: Median Income 10](#_Toc511750145)

[Graphical and Numerical Summaries 10](#_Toc511750146)

[Determining Probabilty Model and Parameter Estimation 12](#_Toc511750147)

[Goodness of Fit Test 15](#_Toc511750148)

[Confidence Interval 15](#_Toc511750149)

[Categorical Variable Analysis: Dominiant Political Party 16](#_Toc511750150)

[Graphical and Numerical Summaries/Probability Model and Parameter 16](#_Toc511750151)

[Binary Variable Analysis: Does State Allow Open Carry? 19](#_Toc511750152)

[Graphical and Numerical Summaries/ Probability Model and Parameter 19](#_Toc511750153)

[Ordinal Variable Analysis: Economic Brackets 21](#_Toc511750154)

[Graphical and Numerical Summaries/ Probability Model and Parameter Estimation 22](#_Toc511750155)

[Module Conclusion 25](#_Toc511750156)

[Module 3: Single sample analysis 26](#_Toc511750157)

[Module Introduction 26](#_Toc511750158)

[How does the Population Mean Unemployment Rate of the United States in 2014 Differ from the Population Mean Unemployment Rate in 2010? 26](#_Toc511750159)

[Power Calculation 31](#_Toc511750160)

[What Proportion of States had more than $60,000 Median Income? (Test on Population Proportion) 32](#_Toc511750161)

[Foundation for Further Exploration: Is the True Population Mean for Prisoners Per 1000 People Between Different Between the Years 2010 and 2014? 35](#_Toc511750162)

[Conclusion 37](#_Toc511750163)

[Module 4: Two Sample Analysis 38](#_Toc511750164)

[Is Median Income Bracket Independent of Dominant Political Party? (Independence of Variables) 38](#_Toc511750165)

[Test of Independence 38](#_Toc511750166)

[Process Done via Code 40](#_Toc511750167)

[Does the Incarceration Rate of GOP Controlled States Greater than Democratic Controlled Ones? (Two Sample Hypothesis Test on Population Means) 41](#_Toc511750168)

[Is There a Higher Proportion of Violent Crimes in GOP States than that of Democratic Ones? (Difference in Proportions) 45](#_Toc511750169)

[Conclusion 49](#_Toc511750170)

[Module 5: Prediction 50](#_Toc511750171)

[Introduction 50](#_Toc511750172)

[Which Variable is Explanatory and Which is Response?/Linear Regression Model Motivation 50](#_Toc511750173)

[Scatterplot Production, and Analysis on the Direction, Strength, and Linearity of the Association Between Incarceration Rate and Median Income. Analysis of Potential Influential Points 51](#_Toc511750174)

[Determining the Least Squares Line for Predicting Incarceration Rate from Median Income/Interpretation of the Value of the Slope Coefficient and Intercept. 52](#_Toc511750175)

[Reporting and Interpretation of the Coefficient of Determination. 53](#_Toc511750176)

[What Incarceration Rate Would the Line Predict for the 2014 Median Income of Washington D.C. 53](#_Toc511750177)

[Examination of Residual Plots to Assess Whether the Technical Conditions for Inference Appear to be Satisfied 54](#_Toc511750178)

[Hypothesis Test 56](#_Toc511750179)

[Constructing a 95% Confidence Interval for the Population Slope Coefficient and Interpretation of this Interval. 57](#_Toc511750180)

[Steps for Developing a More Complex Model 57](#_Toc511750181)

[Conclusion 57](#_Toc511750182)

[Module 6: Summary and Conclusion 58](#_Toc511750183)

[Appendicies 64](#_Toc511750183)

ISYE 2028 Project: Crime in the United States

Preston Akwule and Ayotunde Yoyin

4/18/18

## Module 1: Introduction and Data Collection

### Introduction

The United States of America is one of the most developed countries in the world. Each state within this country has distinguishing characteristics such as population, economics, and political status. Crime in the United States has changed since the past, as it was a significant problem in the 1970’s. Now, in 2018 it is relatively controlled in most parts of the country. However, some states exhibit significant deviance. Thus, we wanted to analyze what factors correlate (or don’t correlate) with the heightened crime rates across each of the 50 states.

This project focused on data collected in the year 2014. Some data from the year 2010 was collected as well to give a relative perspective on how crime and other statistics have changed within a five year period. The goal of this analysis was to determine correlations (if any) between variables and crime rates and to have a better understanding of what phenomena occur in tandem within each state. Thus, we wanted to answer the question: “What are the common factors that are prevalent in states with heightened occurrences of crime?” so we could then recommend which variables should be further explored in order to decrease crime rates.

### Data Collection and the Cleaning Process

Multiple sources were used in the data collection process for this project. The majority of the data was retrieved from government provided data sources including the Federal Bureau of Intelligence (FBI), United States Census, and the Bureau of Justice Statistics (BJS). These government based sources release yearly reports containing population based data including the data used in this analysis. The data was compiled from different sources and consolidated into one data table. All sources are listed in the appendix of this project. Many relevant variables were meant to be included, however we faced the problem that most statisticians tend to face when performing statistical studies. Initial variables such as the number of “the total amount of crimes per state” had to be adjusted to ones such as “the total number of violent crimes in each state” to accommodate what public data sources had already published for the 50 states. Thus, we were able to compile data that would produce meaningful analysis while still having reliable data sources.

The data from the 50 states are the “sample” that is analyzed in this project, it is also considered to be random because of the times in which data was compiled at. There is no way that the samples were fixed in choice because of how state data is organized and compiled. Thus, we can view this sample as a random sample because it was only collected from one point in time along a timeline. For example, 2013’s census data and 2014’s census data are two different random samples. This was important because it allowed us to successfully complete other tests done in the rest of this project that require a random sample as an assumption to conduct. The observational unit of this consisted of states which were the sample from the whole counts as the population which would include the 50 states and other U.S territories and districts such as (Puerto Rico, Guam, and Washington D.C.)

Our data cleaning process consisted of the discovery of errors in a data record to remove or correct these mistakes. Data cleaning was not necessary for this project. All of the data for each variable in 2014 fell within the reasonable range of values. No outliers were found, therefore the initial data was sufficient, as the government data was most likely cleaned before being released to the public.

The data was collected for each of the 50 states in the United States of America. As aforementioned, data was collected for variables that were found for each state in America.

### 

### Variable List

**Expounding on Variables:** The following are details on certain variables that require further explanation on the data collection process and method of measurement.

Number of Prisoners: This variable measures the correctional population in the given state. Correctional population is defined in this case to be individuals on probation or parole, and incarcerated individuals in state prisons, local prisons, and local jails.

Median Income: This variable was compiled from the 2014 data provided from census.gov and its value was calculated in terms of the 2014 value of the USD. Therefore, if this were to be compared with 2017 values, it would need to have inflation accounted for.

Does State Allow Open Carry?: Unrestricted states are defined as those that allow permit-less carry in their state (coded as 1) and unrestricted states that require citizens to go through some sort of process to earn the right of open carry (coded as a 0).

Dominant Political Party: This variable pertains to which political party has more seats in congress within the respective state. The possible values were either “Republican (GOP)”, “Democrat”, or “Split”.

Poverty Level: The United States Census defines a poverty stricken family as any household with a yearly household income less than 12,060 USD + 4,180 USD(per person). For example, a two person household would need to have above 16,240 USD as a combined income to be above the poverty line.

Economic Bracket: Each state’s median income was taken and coded with a 1, 2, 3, 4 or 5. 30,000-39,999(USD) corresponds to a 1, 40,000-49,999(USD) corresponds to a 2, 50,000-59,999(USD) corresponds to a 3, 60,000-69,999(USD) corresponds to a 4, and 70,000-79,999(USD) corresponds to a 5.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable Name (Project) | Concept | Var Name (coding in R) | Type of Var | Reasonable Range | Unit of measurement |
| Population | The amount of people in the state. (Recorded April 14) | pop | discrete numerical | 600,000-40,000,000 | people |
| Prisoners | The number of prisoners in state and local jails | prisoners | discrete numerical | 140 - 320 | people |
| Number of Violent Crimes | The number of violent crimes committed in the state | vc | discrete numerical | 80 - 300 | N/A |
| Unemployment Rate | The rate of unemployment by state | ur | continuous numerical | 3-10 | percent |
| Median Income | The median income of the state collected by the census | mi | continuous numerical | 30,000-80,000 | USD |
| Does State Have Open Carry? | Whether or not state law allows for open carry | oc | binary | Restricted = state allows for open carry = 0 Unrestricted = state does not allow for open carry = 1 | N/A |
| Dominant Political Party | The Political party that holds most seats in state | dpp | nominal | D = Dem G = GOP Split | N/A |
| Income Bracket | Grouping based upon state median income | eb | ordinal | 1 = $30,000s, 2 = $40,000s, 3 = $50,000s 4 = $60,000s 7 = $70,000s | USD |
| Poverty Stricken Families | The number of families that are below the poverty line defined by the U.S census | psf | continuous | 5%-30% | N/A |
| Number of Violent Crimes Per 1000 Persons | Variable calculated using the number of violent crimes and population | vc | continuous | 0-10 | N/A |
| Prisoners Per 1000 People | Variable calculated using the number of prisoner crimes and population | pp1k | continuous | 0-10 | People |

### 

### Preview of the Project

The project will give a comprehensive overview of statistical concepts and methods of analysis including, one sample and two sample hypothesis testing, confidence intervals, distribution analysis and linear regression to better understand the multifaceted makeup of crime in the United States of America. Once we had executed and analyzed these concepts, we could then be able to draw connections and make conclusions about the different kinds of variables which we chose to study. The main goal of this project was to make a structured claim pertaining to crime in the United States based on our findings for further data analysis. This goal was revisited in Module 6 upon completion of hypothesis tests and linear regression analysis.

## Module 2: Modeling Data with Probability Distributions

### Module Introduction

The goal of this module was to summarize the data associated with four variables from our project. These variables are of the following types: continuous, binary, categorical, and ordinal. The four variables chosen for this section were “Median Income”, “Does state allow open carry?”, “Dominant Political Party”, and “Economic Bracket”. We chose these variables because we thought that these variables were pertinent to our overall understanding of crime related statistics. These variables were summarized both graphically and numerically. Then, they were modeled with a probability distribution which seemed to fit best. After that, parameters were estimated using various methods learned in ISYE 2028 class. Procedures were used only for those of the appropriate data types. These models gave us a better understanding of variables that we explored alongside crime data.

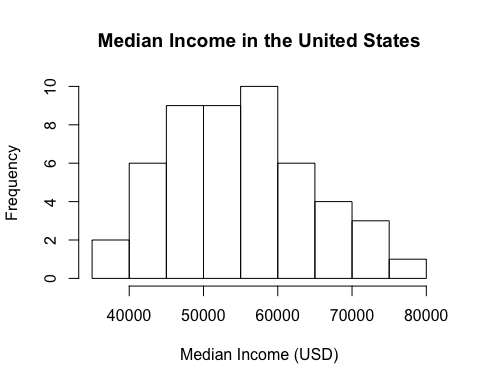
library(readxl)  
project\_data <- read\_excel("CrimeProjData.xlsx")  
attach(project\_data)

### Continuous Variable Analysis: Median Income

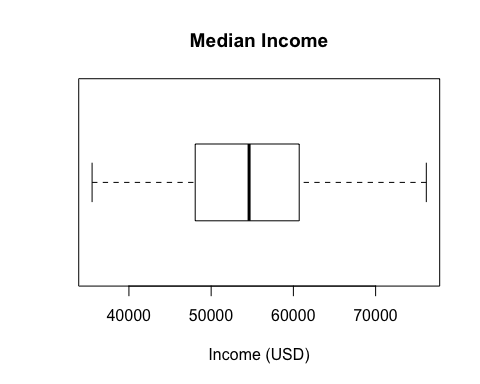
As stated in the expounding variables section above, the Median Income of each state in the United States was determined by the Census in 2014 with the 2014 USD amount. Dr. A states that money can be considered a continuous variable so we treated this variable as such. First, a histogram and box plot were drawn for visualization purposes.

### Graphical and Numerical Summaries

mi=project\_data$`Median Income` #Access column in a excel  
len\_of\_mi = length(mi)  
mi\_break\_len = sqrt(len\_of\_mi) #To use for breaks  
hist(mi, breaks = mi\_break\_len+1, xlab = "Median Income (USD)",main = "Median Income in the United States")



boxplot(mi, horizontal =TRUE, main="Median Income", xlab="Income (USD)")



summary(mi)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 35521 48358 54613 54963 60653 76165

sd(mi)

## [1] 9109.319

The visual interpretations of the “Median Income” variable show that the distribution for median income:

1.Is fairly symmetric when analyzing the shape of the histogram

2. Has a unimodal shape

3. has a mean of the median income variable as well as a median of the median income variable that are similar in value.

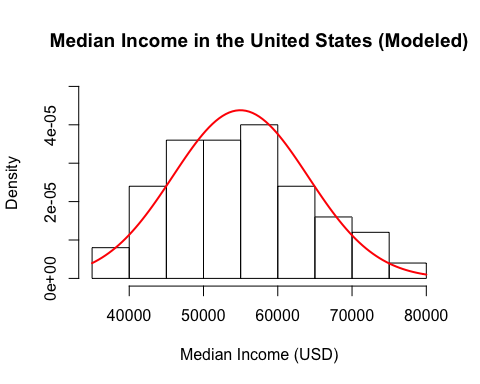
The percent difference between the mean median income in the 50 states (54,963 USD) and median of the median income in the 50 states (54,613 USD) was computed as 0.636%, and 4. Both boxes on the boxplot are of similar size. Therefore, the data moderately models a Normal Distribution from this perspective. The boxplot produced highlights this view of data as well, as there are no outliers and both tails of the image are of the same size.

The numerical summary of the median income across the 50 states shows that the mean of “median income” (measurement of central tendency) is 54,963 USD and the standard deviation of “median income” is 9109.319. The three quartiles from the given dataset are 48,358 USD, 54613 USD, and 60653 USD representing Q1, Q2, and Q3 respectively. This numerical summary demonstrates that the distribution of median incomes is normal because it was mainly symmetric about the mean of 54,963 USD. (Q2-Q1 ~ Q3-Q2, 54,613 USD -48,358 USD = 6255 USD ~ 60,653 USD -54,613 USD = 6040 USD).

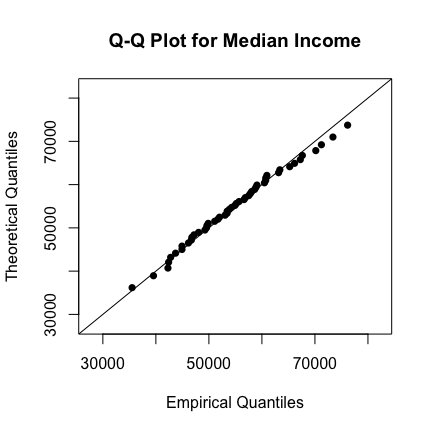
### Determining Probability Model and Parameter Estimation

To further explore the idea of the median income data following a Normal Distribution, we chose to use the sample mean of median income and sample variance of median income to check how well they match with a histogram (with model overlaid). Thus, n(54,963,9,109.319^2) was our probability model. These parameters were chosen using the Wisdom of the Ages method, making a decision based off of citation authority. Wisdom of the ages typically accepts sample values as the estimate, thus we did that by that by basing our estimates off of the sample mean and variance.

hist(mi, freq = FALSE, breaks = mi\_break\_len+1, xlab = "Median Income (USD)",main = "Median Income in the United States (Modeled)", ylim= c(0,5e-05))  
curve(dnorm(x, mean = mean(mi), sd = sd(mi)), add=TRUE, lwd=2, col = "red")



# set the number of data points  
n = length(mi)  
  
m = 54963  
s = 9109.31  
limits = c(m-3\*s,m+3\*s)  
probs = (1:n)/(n+1)  
norm.quants = qnorm(probs,m,s)  
plot(sort(mi),sort(norm.quants),ylab="Theoretical Quantiles",  
 xlab="Empirical Quantiles", main="Q-Q Plot for Median Income",  
 xlim=limits, ylim=limits, pch=16)  
abline(0,1)



The density curve fits the model moderately well. A large majority of the data lies in the middle, symmetrically about the mean of the graph of the median income variable, and there are less and less data points as the tails are approached. Lastly, the curve captures most of the data.

Thus, for further confirmation of this potential model, a QQ Plot was generated to test whether or not this model fits well. A QQ plot is a model which compares the empirical Cumulative Density Function (CDF) of a data set with the theoretical CDF of a particular distribution. It can also be used to compare two empirical CDFs.

The QQ plot shows that the distribution fits the model well. The middle portion contains the majority of the data and most of the data points fall on the line (as expected with a normal distribution). Aside from the fact that the model partially underestimates the data to the right of the median, all signs show that this is a good normal distribution.

### Goodness of Fit Test

#alpha = .05  
shapiro.test(mi)

##   
## Shapiro-Wilk normality test  
##   
## data: mi  
## W = 0.9901, p-value = 0.9485

The Shapiro-Wilk Normality Test was the Hypothesis Test used to determine the goodness of fit of the model. The null hypothesis of this test is that the median income data is normally distributed. The alternative hypothesis is that the data is not normally distributed. Therefore, “if the p-value is less than the chosen alpha level, then the null hypothesis is rejected and there is evidence that the data tested are not from a normally distributed population; in other words, the data are not normal. On the contrary, if the p-value is greater than the chosen alpha level, then the null hypothesis that the data came from a normally distributed population cannot be rejected.” Thus, if we apply this logic to this case, we conclude that our p-value of 0.9485 is greater than our standard alpha value of 0.05. Therefore, the null hypothesis cannot be rejected and our sample mean and variance are good parameters. The data can be modeled with a N(54963,9109.319^2).

### Confidence Interval

Lastly, the confidence interval was determined using a 95% confidence level. The mean and standard deviation chosen via the Wisdom of the Ages method were used in the calculation of this confidence interval. Lower limit = Mean - & Upper limit = Mean +

a <- 54963  
s <- 9109.319  
n <- 50  
total\_error <- qnorm(0.975)\*s/sqrt(n)  
left\_error <- a-total\_error  
right\_error <- a+total\_error  
left\_error

## [1] 52438.07

right\_error

## [1] 57487.93

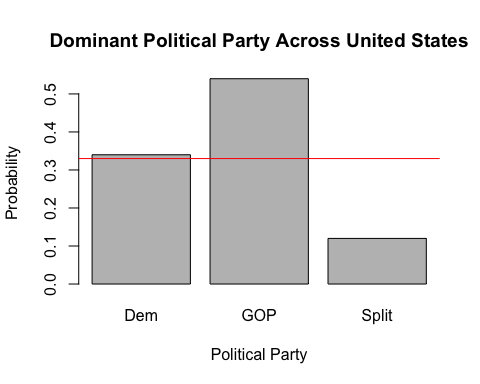
The lower boundary is 52438.07 USD and the upper boundary is 57487.93 USD. Thus, our interpretation is if we repeated this sample collection, drawing the random sample of size n thousands of times, created a confidence interval for each data set, and knew the true mean, the confidence intervals would contain the true mean 95% of the time. The 95% confidence interval for the mean of median income is [52438.07, 57487.93].

### Categorical Variable Analysis: Dominant Political Party

In 2014, the number of congress representatives by political party was taken in each state by political party. Each state was either categorized as GOP (Republican majority), Dem (Democratic majority), or Split (Even). For this categorical variable, we compared it with a discrete uniform distribution. The dominant political party in each state was an interesting factor which we wanted to analyze when we analyzed crime rates

### Graphical and Numerical Summaries/Probability Model and Parameter

dpp=project\_data$`Dominant Political Party` #Access column in a excel  
dpp=table(dpp)  
dpp=as.data.frame(dpp)  
names(dpp)[1]="Dominant Political Party"  
height\_dpp=dpp$Freq/50  
barplot(height\_dpp, names.arg=dpp$`Dominant Political Party`, xlab = "Political Party", ylab = "Probability", main = "Dominant Political Party Across United States")  
abline(.33,0,col="red")



dpp

## Dominant Political Party Freq  
## 1 Dem 17  
## 2 GOP 27  
## 3 Split 6

Both the bar plot and frequency table of the dominant political parties out of the 50 states in 2014 convey the fact that GOP controlled most states at a frequency of 27 controlled states, while Democrats controlled 17, and 6 were under split control.

We used a discrete uniform distribution to model this data, as Dr. A states, that is the proper procedure for categorical data. This is because the categories are unordered and the probability value of each category would be unchanged upon rearrangement. The selected parameters values were 1 and 3 because they were the indexes of the variables (minimum and maximum values of our political party frequency). Since we implemented the uniform model, the probability of any category being selected was determined to be 0.333. This fact is conveyed by the red line overlaid on this graph.

We conducted A Chi Square Goodness of Fit Test to determine whether or not the Uniform Distribution was a good model for our data pertaining to dominant political party in the United States. α = 0.05 , the standard chosen probability of Type I Error. In a Hypothesis Test, we determine a p-value. If it was lower than the chosen alpha value, then we would have rejected the null hypothesis in favor of the alternative hypothesis.

**Null Hypothesis, :** The Discrete Uniform (1,3) is an adequate model for our data pertaining to the dominant political party of a state in the United States.

**Alternative Hypothesis, :** The Discrete Uniform (1,3) is not an adequate model for our data pertaining to the dominant political party of a state in the United States.

Chi Squared Table:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Political Party | Observed Frequency | Probability of X Being selected | Expected Frequency | (Oi-Ej)^2/Ej |
| GOP | 27 | 0.3333 | 16.665 | 6.409 |
| Dem | 17 | 0.3333 | 16.665 | .00673 |
| Split | 6 | 0.3334 | 16.665 | 6.825 |
| Total | 50 | 1 | 50 | 13.241 |

Our test statistic: = = = 13.241

This test statistic was conducted with = 5.991 = Critical Value

chisq.test(dpp$Freq)

##   
## Chi-squared test for given probabilities  
##   
## data: dpp$Freq  
## X-squared = 13.24, df = 2, p-value = 0.001333

Our P-Value was determined to be 0.001333. When we compared this to our alpha value of 0.05, we confirmed that we could reject our null hypothesis of the Discrete Uniform (1,3) being an adequate model for our data pertaining to dominant political party in the United States in favor of the alternative one (The Discrete Uniform (1,3) is not an adequate model for our data pertaining to dominant political party in the United States).

### Binary Variable Analysis: Does State Allow Open Carry?

Unrestricted states were defined as those that allow permit-less carry in their state (r = restricted) and unrestricted states that require citizens to go through some sort of process to earn the right of open carry (u as unrestricted).

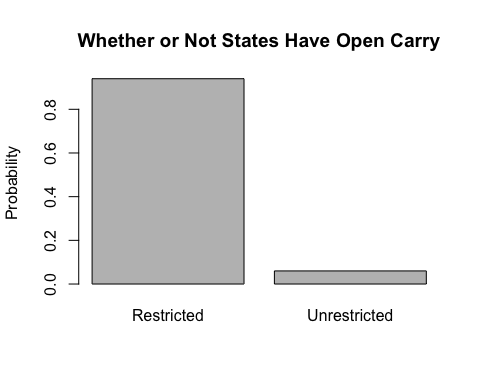
### Graphical and Numerical Summaries/ Probability Model and Parameter

oc=project\_data$`Does State Have Open Carry? R = Restricted; U = Unrestricted` #Access column in a excel  
oc=table(oc)  
oc=as.data.frame(oc)  
names(oc)[1]="Does State Have Open Carry? R = Restricted; U = Unrestricted"  
oc

## Does State Have Open Carry? R = Restricted; U = Unrestricted Freq  
## 1 r 47  
## 2 u 3

The frequencies of each of the two observations were determined and shown in the numerical summary. Restricted open carry state observations were coded as a 0 and unrestricted open carry states were a 1. Thus, the mean of a binary variable can be coded as the fraction of the sample which was coded as a 1 (unrestricted open carry states). The mean for this distribution of restricted vs. unrestricted open carry was 0.06. The standard deviation of this distribution was not calculated in this numerical summary. It was not necessary since there were only two variables involved.

proboc=oc$Freq/50  
barplot(proboc,main="Whether or Not States Have Open Carry", names.arg = c("Restricted", "Unrestricted"),ylab="Probability")



For graphical purposes, a bar plot with the corresponding observation probability was made. The graph also conveyed the fact that the frequency of restricted open carry states was significantly greater than the frequency of unrestricted states. There were only two possible options for our observations of this variable. The only probability model that satisfies this is the Bernoulli Random Variable. Thus, no Goodness of Fit test was necessary for this random variable. The parameter for a Bernoulli Random Variable is the probability of a “success” being observed. A success in this instance is represented by a state that has unrestricted open carry. (Not to give a connotation of preference, this is solely because the unrestricted observations were coded as 1s). Therefore, the parameter value that came from the restricted and unrestricted variables corresponded with 0.94 (47/50) and 0.06 (3/50) respectively.

### Ordinal Variable Analysis: Economic Brackets

As previously mentioned, each state’s median income was taken and coded with a 1,2,3,4 or 5. 30,000s(USD) corresponds to a 1, 40,000s(USD) corresponds to a 2, 50,000s(USD) corresponds to a 3, 60,000s(USD) corresponds to a 4, 70,000s(USD) corresponds to a 5. Economic brackets were viewed as an important factor to study when it came down to analyzing crime rates.

### 

### Graphical and Numerical Summaries/ Probability Model and Parameter Estimation

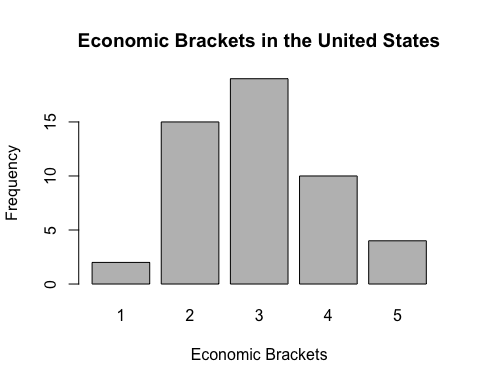
eb=project\_data$`Economic Bracket` #Access column in a excel  
eb

## [1] 2 4 2 2 4 4 5 3 2 2 5 3 3 2 3 3 2 2 3 5 4 3 4 1 3 3 3 2 5 4 2 3 2 4 2  
## [36] 2 3 3 3 2 3 2 3 4 4 4 3 1 3 3

summary(eb)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.00 2.00 3.00 2.98 4.00 5.00

len\_of\_eb = length(eb)  
eb\_break\_len = sqrt(len\_of\_eb)  
barplot(table(eb), , main="Economic Brackets in the United States", ylab="Frequency", xlab="Economic Brackets")



#curve(dpois(x, 2.98), add=TRUE, lwd=2, col = "red") Model for poisson dist. cannot be overlaid on barplot

Both the graphical and numerical summaries support that the data pertaining to economic brackets in the United States follows a Poisson distribution. The following three factors of the data supported the claim that this data follows a Poisson distribution: it was discrete, unimodal, and is centered roughly on our mean of 2.98. Next, to confirm whether or not this distribution could be represented with a Poisson distribution, we used a Goodness of Fit hypothesis test and the Method of Moments to estimate the parameters for this distribution.

When estimating the parameter for an exponential distribution, Chapter 7.4 of *Applied Statistics and Probability for Engineers by Douglas C. Montgomery and George C. Runger* states, “Let , ,…, be a random sample from the probability distribution f (x) where f(x) can be a discrete probability mass function or a continuous probability density function. The kth population moment (or distribution moment) is E( ), k = 1, 2, …” When given Poisson distribution with mass function, P(X = x) = , x = 0, 1, 2, . . . , where λ is an unknown parameter. E(X) = λ. So, µ1 = E(X) = λ = X¯ = ˆµ. Hence, the method of moments estimator of λ is the sample mean. Thus, we chose the sample mean of 2.98 as our parameter.

We conducted A Chi Square Goodness of Fit Test to determine whether ot not the Uniform Distribution was a good model for our data pertaining to dominant political party in the United States. α = 0.05 is the standard chosen probability of a Type I Error. In a Hypothesis Test, we determine a p-value. If it was lower than the chosen alpha value, then we would have rejected the null hypothesis in favor of the alternative one.

**Null Hypothesis, :** The Poisson (2.98) is an adequate model for our our data pertaining to economic brackets in the United States.

**Alternative Hypothesis, :** The Poisson (2.98) is not an adequate model for our our data pertaining to economic brackets in the United States.

α = 0.05 Standard Alpha Value = Probability of Type I Error n = 50 (Therefore DF = 49) The test statistic is =

Distribution: ~

Table of Observed Values:

|  |  |
| --- | --- |
| Economic Bracket | Observed Frequency |
| 1 | 2 |
| 2 | 15 |
| 3 | 19 |
| 4 | 10 |
| 5 | 4 |

Table of Expected Frequencies:

|  |  |  |
| --- | --- | --- |
| Economic Bracket | Probability | Expected Frequency |
| 1 | 0.1513 | 7.57 |
| 2 | 0.2255 | 11.28 |
| 3 | 0.2240 | 11.20 |
| 4 | 0.1667 | 8.34 |
| 5 or more | .2321 | 11.62 |

Calculation of Chi Squared Test Statistic:

|  |  |  |
| --- | --- | --- |
| Economic Bracket | Observed Frequency | Expected Frequency |
| 1 | 2 | 7.57 |
| 2 | 15 | 11.28 |
| 3 | 19 | 11.20 |
| 4 | 10 | 8.34 |
| 5 or more | 4 | 11.62 |

The test statistic is = = 16.08

compared to ~ 67.05 (Critical Value)

**Conclusion:** Since the test statistic of 16.08 is less than the critical value of 67.05, we failed to reject the null hypothesis that the Poisson (2.98) was an adequate model for our data pertaining to economic brackets in the United States.

### Module Conclusion

The purpose of this module was to visually demonstrate the process of choosing suitable distribution models for a variety of different kinds of data and to determine how well each chosen model fit its respective data set. Extreme care was taken to ensure that the values which came from parameter estimations and confidence intervals accurately and properly represented our data sets. Overall, the probability models chosen were an accurate representation of the data collected. In the next Module, we continue our data analysis using Hypothesis Tests to justify the use of our estimated parameters.

## 

## Module 3: Single sample analysis

### Module Introduction

In this module, we aimed to demonstrate the analyses we have learned that involve only one sample from one population. This is the point in the project in which we planned to tell a story by posing questions, doing the analysis, and interpreting the results pertaining to Crime Data in the United States during the year 2014. Multiple forms of hypothesis tests were conducted in this module to explore different phenomena. In this module, we analyzed the variables Median Income, Number of Prisoners, and Unemployment Rate.

First, we analyzed the mean Unemployment Rate across the United States in the year of 2014 and compared it to the values recorded for 2010. Unemployment Rate data was chosen from this year for two reasons: 1. Census data taken each decade historically has the least margin of error and 2. 2010 was not far away from the beginning of the recession that occurred. Thus, a significant difference between the years 2010 and 2014 was expected going into the test. Although Unemployment Rate is not a direct crime statistic, it was included in this project to be further analyzed later to determine its correlation with the Prisoners Per 1000 People in a State. Thus, it was still deemed important enough to analyze overall.

### How does the Population Mean Unemployment Rate of the United States in 2014 Differ from the Population Mean Unemployment Rate in 2010?

The United States Census provided data on Unemployment Rates across the 50 states in 2010, this value was 8.748%, which is what we used as the null hypothesis in our analysis. We aimed to investigate if the 2014 Unemployment Rate differs in value when compared to the historical data.

**Parameter of interest:** For the purpose of this analysis, we declared U to be Unemployment Rate in the United States for 2014. Thus, the true population mean for Unemployment Rates is defined as .

**Null Hypothesis vs. Alternative Hypothesis**: Null Hypothesis: : The true population mean of Unemployment Rate is equal to 8.748%.

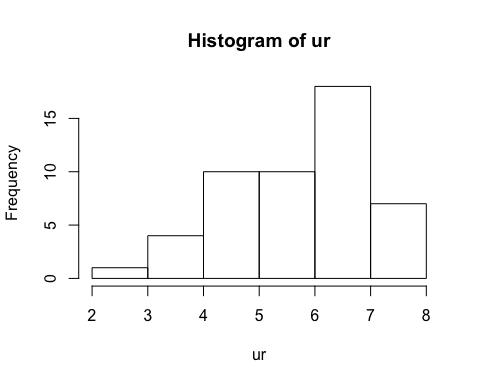
Alternative Hypothesis: : The true population mean of Unemployment Rate is not equal to 8.748%.

**Choosing Alpha/Test Setup**: To test these mutually exclusive, spanning claims, we conducted a one-sample two-sided hypothesis test on the mean for a normally distributed population with an unknown variance. We used the standard alpha-value of 0.05 for this test. (Because there was no reason to choose another). Thus, α = 0.05 Standard chosen probability of Type I Error.

**Assumptions**: When conducting hypothesis tests, critical assumptions are to be made before computations are to be made.

The data is normally distributed: Chapter 9 of *Applied Statistics and Probability for Engineers by Douglas C. Montgomery and George C. Runger*, states that “The development of the t-test assumes that the population from which the random sample is drawn is normal. This assumption is required to formally derive the t-distribution as the reference distribution for the test statistic…Because it can be difficult to identify the form of a distribution based on a small sample, a logical question to ask is, how important this assumption is? Studies have investigated this. Fortunately, studies have found that the t-test is relatively insensitive to the normality assumption. If the underlying population is reasonably symmetric and unimodal, the t-test will work satisfactory.”

ur=project\_data$`Unemployment Rate` #Access column in a excel  
hist(ur, breaks = 7)



Doing analysis on the shape of the graph, the distribution of Unemployment Rates in the United States is indeed reasonably symmetric and unimodal. There is no facet of this distribution that made us wary of continuing the test.

The sample is a random sample: Due to the nature of this analysis, it was important to note that the data came from all 50 states. However, the nature of this process was indeed random, therefore it satisfied that assumption. Because the sample was random and independent because independence is one of the criteria for randomness.

m\_ur = mean(ur)  
sd\_ur = sd(ur)  
m\_ur

## [1] 5.742

sd\_ur

## [1] 1.24787

**Test statistic, and Distribution:** Our test statistic is as follows:

Note: is equivalent to the sample standard deviation and n is the sample size. has a t-distribution and had a degrees of freedom of n-1. In this case n = 50, therefore we had a distribution.

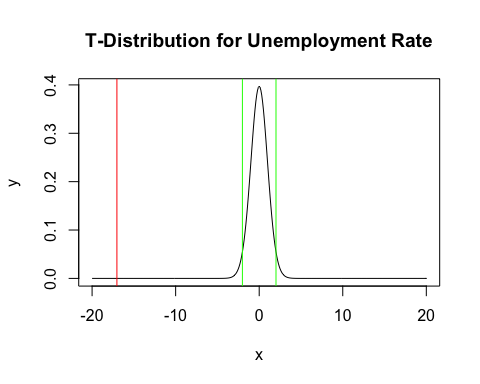
= -17.03 = Test Statistic

This value is compared with a distribution.

In this hypothesis test on mean Unemployment Rates, we calculated critical value and compared it to the test statistic. Due to the nature of a two-sided test, the critical values are the points where an area of 0.025, or half of the alpha value, falls to the right and left side of the values. Thus, was the lower critical value and was the upper critical value for this particular test.

P{}= 0.025 = -2.01 This value was calculated using the invT() function on the TI-84 Calculator. Since the t-distribution was determined to be symmetrical, the upper boundary was 2.01. = 2.01

x=seq(from=-20,to=20,by=0.1)  
y=dt(x,df=49)  
plot(x,y,type= "l", main="T-Distribution for Unemployment Rate")  
abline(v=-17.03,col="red")  
abline(v=2.01,col="green")  
abline(v=-2.01,col="green")



The distribution for unemployment rate was plotted with the red line being the test statistic and green lines being the critical values.

**Judgement/Recommendation:** The region to the left of the test statistic bounded by the model is half of the p-value. Thus, the p value for this test is approximately double the value that was determined (Approximately 0 Calculated by using Tcdf() on calculator). This value is compared to our α/2 (0.025). This value was significantly lower than half of our alpha value, so we rejected the hypothesis that the population mean of Unemployment Rate in 2014 in the United States was the same as the population mean of the Unemployment Rate in the United States in 2010. In context, one would expect for this to occur given that this was a five-year period of time. In 2010, the United States was still recovering from a recession, thus extreme differences were expected when comparing 2014 to 2010.

To provide further evidence on the accuracy for the hypothesis test conducted, we computed this in R for confirmation. The results aligned with the hypothesis test done by hand.

t.test(ur,y = NULL, alternative = "two.sided",8.748,conf.level = 0.95)

##   
## One Sample t-test  
##   
## data: ur  
## t = -17.034, df = 49, p-value < 2.2e-16  
## alternative hypothesis: true mean is not equal to 8.748  
## 95 percent confidence interval:  
## 5.387359 6.096641  
## sample estimates:  
## mean of x   
## 5.742

The results of the R computations were consistent with the hand calculations for the two-sided, one sample hypothesis test on the mean.

### Power Calculation

The power of a statistical test is the probability of rejecting the null hypothesis when the alternative hypothesis is true. The power is computed as 1−β, and power can be interpreted as the probability of correctly rejecting a false null hypothesis. The specific alternative value we used for the population mean of the Unemployment Rate was provided by the Bureau of Labor Statistics for 2014, 6.3. So, our specific alternative hypothesis was: : = 6.3

The power of the test was computed using the power.t.test function. The delta value is defined as the difference between the null and the new specific alternative, which was equal to 0.558 in this instance. The effect size was defined as delta divided by the sample standard deviation, and was equal to 0.447. This was considered to be a moderately small effect.

power.t.test(n=50,delta=0.558,sd=1.24787,sig.level=.05,type="two.sample",alternative="two.sided")

##   
## Two-sample t test power calculation   
##   
## n = 50  
## delta = 0.558  
## sd = 1.24787  
## sig.level = 0.05  
## power = 0.6002128  
## alternative = two.sided  
##   
## NOTE: n is number in \*each\* group

### 

### What Proportion of States had more than $60,000 Median Income? (Test on Population Proportion)

The United States Census provided data pertaining to median income in 2010. This data was used for a comparison against the data collected for 2014. The 2010 data showed that only 14% of the states had a median household income that was above 60,000 (USD). As mentioned in the module introduction, we aimed to determine whether or not the change between 2010 statistics and 2014 statistics were noticeable given the fact that 2010 was closer to the beginning of a recession. Thus, we aimed to find out whether or not the proportion of states who had a median income of above 60,000 (USD) had increased or not.

Let be p the parameter of population proportion. A test of proportion assesses whether or not a sample from a population represents a true proportion of the entire population. Thus, we let p represent the number of the states that had a median household income that was above 60,000 (USD). Our sample proportion was 14/50. We formed the following hypothesis to further explore this question.

Null Hypothesis: : p 0.14

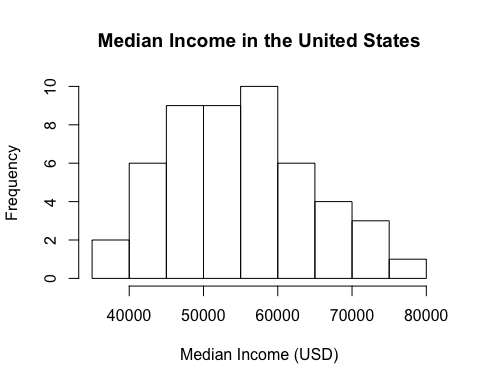
Alternative Hypothesis: p > 0.14

We conducted a one-sided single sample hypothesis test on the population proportion to test this claim. We compared the p-value generated from a Z approximation on population proportion to the standard alpha value (α = 0.05).

**Assumptions** :

The data is normally distributed:

mi=project\_data$`Median Income` #Access column in a excel  
len\_of\_mi = length(mi)  
mi\_break\_len = sqrt(len\_of\_mi) #To use for breaks  
hist(mi, breaks = mi\_break\_len+1, xlab = "Median Income (USD)",main = "Median Income in the United States")



The visual interpretations of the “Median Income” variable show that the distribution for median income: 1.Is fairly symmetric when analyzing the shape of the histogram, 2. Has a unimodal shape, 3. has a mean median income and median of the median income that are similar in value. The percent difference between the mean median income in the 50 states (54,963 USD) and median of the median income in the 50 states (54,613 USD) was computed as 0.636%, and 4.Both boxes on the boxplot are of similar size. Therefore, the data moderately models a Normal Distribution from this perspective.

The other assumptions of the sample being random and independently distributed were previously expounded upon in Module 2.

prop.test(14,50,0.14,alternative="greater",conf.level = 0.95)

##   
## 1-sample proportions test with continuity correction  
##   
## data: 14 out of 50, null probability 0.14  
## X-squared = 7.0183, df = 1, p-value = 0.004034  
## alternative hypothesis: true p is greater than 0.14  
## 95 percent confidence interval:  
## 0.180529 1.000000  
## sample estimates:  
## p   
## 0.28

The prop.test can be used for testing the null that the proportions (probabilities of success) in several groups are the same, or that they equal certain given values. The Z-Approximation on the population proportion produced a p-value of 0.004034. This value was less than our alpha value of 0.05. Therefore, we rejected the null hypothesis which stated that the proportion of the states with a median income of over 60,000 (USD) in 2014 was less than those in 2010 in favor of the alternative. In context, this matched up with the expectations.

### Foundation for Further Exploration: Is the True Population Mean for Prisoners Per 1000 People Between Different Between the Years 2010 and 2014?

In the year of 2010 the population mean of prisoners per 1000 people in each state was determined to be 4.361. As previously mentioned at the beginning of this module, this was the point in the project where we aimed to answer some questions and establish a foundation for the further analysis of correlation to be done in Module 5. Thus, we aimed to determine whether there was difference in amount of prisoners per 1000 people when analyzing population means.

To test our hypotheses and answer the question, we conducted a one-sample, one-sided hypothesis test on the mean of the population with an unknown variance. We also compared the generated p-value to the standard alpha value of 0.05 (α = 0.05).

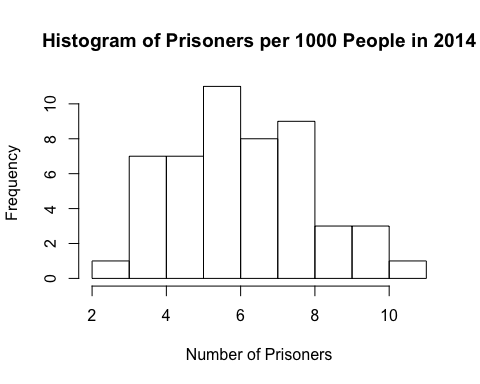
Null Hypothesis: : 4.361

Alternative Hypothesis: : 4.361

**Assumptions**:

The data is normally distributed:

pp1k = project\_data$"Prisoners Per 1000 People"  
hist(pp1k, xlab ="Number of Prisoners", main = "Histogram of Prisoners per 1000 People in 2014")



The histogram shows that the population of prisoners in the United States was normally distributed. This was a unimodal distribution with few outliers which also has a bell-like shape. Therefore, there were no concerns when proceeding with the test.

The entire data set was proven to be random and independently distributed in Module 1.

t.test(pp1k,alternative="greater",mu=4.361,conf.level = 0.95)

##   
## One Sample t-test  
##   
## data: pp1k  
## t = 6.2426, df = 49, p-value = 4.941e-08  
## alternative hypothesis: true mean is greater than 4.361  
## 95 percent confidence interval:  
## 5.57086 Inf  
## sample estimates:  
## mean of x   
## 6.015092

When the test was run with the use of R, a P-Value of 4.941e-08 was produced. This was much smaller than our standard alpha value of 0.05. Therefore, we rejected our null hypothesis that stated that our true population mean of prisoners per 100 people was less than 4.361 in favor of the alternative hypothesis. Therefore, we are accepting the alternative hypothesis that the mean prisoners per 1000 people in the United States were greater in 2014.

### Conclusion

In this module, we showed that the proportion of states with a median income of 60,000 USD was greater in 2014 than it was using the 2010 data as a comparison. The population mean for Unemployment Rates differed as expected as well. We predicted that this would occur because in 2010, the United States was coming off of a recession. Another possible reason was because Silent Generation and Baby Boomers were retiring at high rates at this time. However, the results for the population mean of prisoners per 1000 people was what we expected. This led us to execute a linear regression comparing the two variables: median income and prisoners per 1000 persons which was further explored in Module 5.

## Module 4: Two Sample Analysis

In this module, we continued along the lines of our work done in Module 3, now doing an analysis on two samples from the same population. Two sample hypothesis tests were done to distinguish different factors between independent and matched pairs between two groups. We asked questions, analyzed these questions, and discussed the results. These analyses better helped us to understand the relations (if any) between these variables. We then discussed how they may or may not be related when it comes to crime data in the United States to make a structured judgement and recommendation to further explore correlation in Module 5.

### Is Median Income Bracket Independent of Dominant Political Party? (Independence of Variables)

Two random variables x and y are independent if the probability distribution of one variable is not affected by the presence of another. We performed a Chi-Squared Hypothesis Test on Independence to determine independence using the seven key steps highlighted in *Applied Statistics and Probability for Engineers by Douglas C. Montgomery and George C. Runger*.

Note: Although we tested for independence, the observations within each group must be independent of each other. This was a key assumption necessary for conducting this test and it was met sufficiently.

### Test of Independence

**1. Parameter of interest:** X = Dominant Political Party in the United States, defined by either Democratic (Dem), Republican (GOP), or split control (Split). Y = Economic Brackets in the United States. Each state’s median income was taken and coded with a 1,2,3,4 or 5

|  |  |
| --- | --- |
| Bracket | Range |
| 1 | $30,000-$39,999 |
| 2 | $40,000-$49,999 |
| 3 | $50,000-$59,999 |
| 4 | $60,000-$69,999 |
| 5 | $70,000-$79,999 |

**2. Null Hypothesis vs. Alternative Hypothesis**: : The dominant political party in a state is independent from the median income in a state. Alternative Hypothesis: : The dominant political party in a state is dependent on state median income.

**3. Choosing Alpha** α = 0.05 Standard chosen probability of Type I Error

**4. Table of Counts, Test statistic, and Distribution:**

Table of Observed Counts

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Party/Economic Bracket | 1 | 2 | 3 | 4 | 5 | Total |
| Dem | 1 | 2 | 5 | 6 | 3 | 17 |
| Gop | 1 | 12 | 10 | 4 | 0 | 27 |
| Split | 0 | 1 | 4 | 0 | 1 | 6 |
| Total | 2 | 15 | 19 | 10 | 4 | 50 |

The test statistic was =

Because r = 3 and c = 5, the degrees of freedom for chi-square are (r – 1)(c – 1) = (2)(4) = 8 Distribution: ~

**5. Reject if:** We used a fixed-significance level test with α = 0.05. Therefore, because r = 3 and c = 5, the degrees of freedom for chi-square are (r – 1)(c – 1) = (2)(4) = 8, and we would reject if = = 15.507 = Critical Value

**6. Computations:** From the table of expected counts, we determined our values to use in the following formula for our test statistic.

Table of Expected Counts

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Party/Economic Bracket | 1 | 2 | 3 | 4 | 5 | Total |
| Dem | 0.68 | 5.1 | 6.46 | 3.4 | 1.36 | 17 |
| Gop | 1.08 | 8.1 | 10.26 | 5.4 | 2.16 | 27 |
| Split | .24 | 1.8 | 2.28 | 1.2 | .48 | 6 |
| Total | 2 | 15 | 19 | 10 | 4 | 50 |

= = 14.4

**7. Conclusions:** As the Chi Squared value of 14.4 was less than our critical value of 15.507, we failed to reject the null hypothesis that the dominant political party in the United States was independent of the median income brackets in the United States.

### Process Done via Code

Next, we checked the accuracy of our hand calculations against the output from R. The “chisq.test” function in R allowed us to conduct a contingency table test by inputting table data.

tbl = table(project\_data$`Dominant Political Party`, project\_data$`Economic Bracket`) #Acessing data from project  
tbl

##   
## 1 2 3 4 5  
## Dem 1 2 5 6 3  
## GOP 1 12 10 4 0  
## Split 0 1 4 0 1

chisq.test(tbl)

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

##   
## Pearson's Chi-squared test  
##   
## data: tbl  
## X-squared = 14.4, df = 8, p-value = 0.07191

Note: P Value = P { }

As the Chi Squared value of 14.4 was less than our critical value of 15.507, we failed to reject the null hypothesis that the dominant political party in the United States is independent of the median income brackets in the United States. Also, our P-Value of 0.07191 was greater than alpha (0.05). It was important to note that the approximation may be inaccurate. The expected number of cells in the contingency table should be no bigger than three for this case. Thus, they are improper for this interpretation.

## Does the Incarceration Rate of GOP Controlled States Greater than Democratic Controlled Ones? (Two Sample Hypothesis Test on Population Means)

We created two subpopulations of our data set to be analyzed in this section. Using our categorical variable “Dominant Political Party”, we created a binary subpopulation consisting of only GOP and Democratic states. We chose to neglect Split states because there were so few of them it would not be pertinent to this section. Our two main subpopulations to be studied now became states that are were controlled by the GOP (Republican) Party and states that were controlled by the Democratic Party. Combining this with our variable of incarceration rate, the comparison was done between the incarceration rate of GOP states and the incarceration rate of Democratic states. We theorized that the strictness pertaining to the incarceration rate in a state may have a tendency to change in tandem with its state policymakers, so, we wanted to explore whether or not the GOP controlled subpopulation had a greater mean incarceration rate per 1000 people than the Democratic one.

inc\_gop = project\_data$`Prisoners Per 1000 People`[project\_data$`Dominant Political Party`== "GOP"] #n gop = 27  
inc\_dem = project\_data$`Prisoners Per 1000 People`[project\_data$`Dominant Political Party`== "Dem"] #n dem = 17  
mean\_gop = mean(inc\_gop)  
mean\_dem = mean(inc\_dem)  
mean\_gop

## [1] 7.06803

mean\_dem

## [1] 4.807566

sd\_gop = sd(inc\_gop)  
sd\_dem = sd(inc\_dem)  
sd\_gop

## [1] 1.547504

sd\_dem

## [1] 1.460627

|  |  |
| --- | --- |
| GOP (Sample 1) | Dem (Sample 2) |
| N = 27 | N= 17 |
| Mean = 7.06803 | Mean = 4.807566 |
| Standard Deviation = 1.547504 | Standard Deviation = 1.460627 |

Population 1: All states that are under GOP control Population 2: All states that are under Democratic control, Null Hypothesis: 0 Alternative Hypothesis: 0

**Assumptions**:

Independence: The two samples are independent. The incarceration rates of each state do not impact the ones of another.

Large Sample: Neither of the samples are large, but we continued on with the test to show our understanding of this test.

hist(inc\_gop,xlab = "Prisoners Per 1000 People")



hist(inc\_dem,breaks = 4, xlab = "Prisoners Per 1000 People")



The histograms for both populations showed that the respective distributions are Normal. They were both unimodal with little skew, and fairly symmetric. Thus, this assumption was met.

The standard deviation of population 1 was 1.547504 and the variance of population 2 was 1.460627 (approximately equal). Therefore, this was indeed a pooled 2 sample T-Test on population means. (One sided upper)

Null Hypothesis: 0 Alternative Hypothesis: 0 α = 0.05

t.test(inc\_gop, inc\_dem, alternative = "greater",var.equal = TRUE, conf.level = 0.95)

##   
## Two Sample t-test  
##   
## data: inc\_gop and inc\_dem  
## t = 4.8191, df = 42, p-value = 9.556e-06  
## alternative hypothesis: true difference in means is greater than 0  
## 95 percent confidence interval:  
## 1.471522 Inf  
## sample estimates:  
## mean of x mean of y   
## 7.068030 4.807566

Our P-Value was determined to be 9.55e^-6. This was lower than our alpha value of 0.05. Therefore, we rejected the null hypothesis in favor of the alternative hypothesis. In context, this means that the mean incarceration rate for GOP controlled states is greater than that of the Democratic states.

### Is There a Higher Proportion of Violent Crimes in GOP States than that of Democratic Ones? (Difference in Proportions)

The last test of this module was similar to the previous one. Like incarceration rate, we wanted to determine whether or not the proportion of violent crimes was greater in GOP states than in Democratic states. We particularly chose to analyze the number of states with a violent crime rate of above 5 per 1000 people. As for this analysis, more than 5 violent crimes per 1000 people was deemed to be a large. To test this, we conducted a two sample hypothesis test on the difference in proportions (one sided lower).

vc\_gop = project\_data$`Number of Violent Crimes Per 1000 Persons`[project\_data$`Dominant Political Party`== "GOP"] #n gop = 27  
vc\_dem = project\_data$`Number of Violent Crimes Per 1000 Persons`[project\_data$`Dominant Political Party`== "Dem"]#n dem = 17  
meanvc\_gop = mean(vc\_gop)  
meanvc\_dem = mean(vc\_dem)  
vc\_gop

## [1] 4.274157 6.357807 3.998524 4.801493 5.404886 3.772973 2.121797  
## [8] 3.653104 3.485856 5.147455 4.273312 2.784830 4.429060 3.236682  
## [15] 3.295165 2.650504 2.848847 4.059771 3.140951 4.977153 3.265450  
## [22] 6.084266 4.058841 2.156375 1.962459 2.902964 1.954967

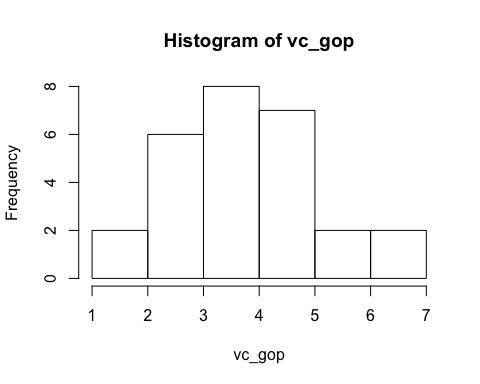
vc\_dem

## [1] 3.961317 3.090817 2.369409 4.890906 2.592351 3.700377 1.278110  
## [8] 4.461042 3.913625 2.291479 6.355890 2.611943 5.973901 2.323286  
## [15] 2.192058 0.992719 3.020008

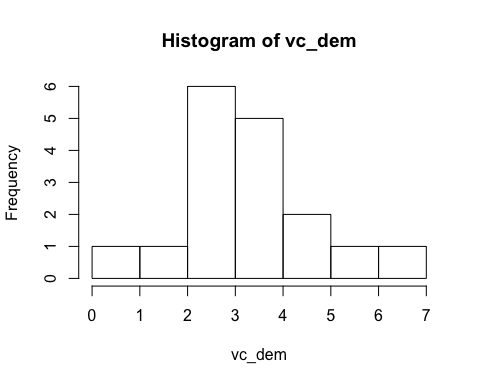
vc\_gop\_big = vc\_gop[vc\_gop>5]  
vc\_dem\_big = vc\_dem[vc\_dem>5]  
count\_gop\_big = length(vc\_gop\_big)  
count\_dem\_big = length(vc\_dem\_big)

There were 27 GOP states recorded and out of those states 4/27 had more than five violent crimes per 1000 people recorded. There were 17 Democratic states recorded and out of those states 2/17 had more than five violent crimes per 1000 people. Population 1: All states that are under GOP control Population 2: All states that are under Democratic control Let be the proportion GOP states that have over 5 violent crimes per 1000 people reported Let be the proportion Democratic states that have over 5 violent crimes per 1000 people reported Null Hypothesis: 0 Alternative Hypothesis: 0 **Assumptions**: The assumptions of randomness and independence were expounded upon in the first module. Therefore, the two subpopulations met these criteria.

hist(vc\_gop)



hist(vc\_dem)



The histograms for both populations showed that the respective distributions were Normal. They were both unimodal with little skew, and fairly symmetric.

prop.test(c(count\_gop\_big,count\_dem\_big),c(27,17),alternative="less",conf.level=0.95,correct=TRUE)

## Warning in prop.test(c(count\_gop\_big, count\_dem\_big), c(27, 17),  
## alternative = "less", : Chi-squared approximation may be incorrect

##   
## 2-sample test for equality of proportions with continuity  
## correction  
##   
## data: c(count\_gop\_big, count\_dem\_big) out of c(27, 17)  
## X-squared = 2.0928e-31, df = 1, p-value = 0.5  
## alternative hypothesis: less  
## 95 percent confidence interval:  
## -1.0000000 0.2317847  
## sample estimates:  
## prop 1 prop 2   
## 0.1481481 0.1176471

The proportion test conducted in R produced a p-value of 0.5. This value was larger than our alpha value of 0.05. Thus, we were unable to reject the null hypothesis that there is a higher proportion of GOP States with a number of violent crimes committed per 1000 people than that Democratic states. In context, there was no evidence to support the idea presented earlier in the module.

### Conclusion

The hypothesis tests conducted further peaked an interest in further understanding the relationships and facets of the data set that one might not have highlighted at first glance. The median income variable was the one out of this module that peaked our interest in terms of further analysis. After we failed to reject that it was independent of the political parties that control states, we wondered what significance median income plays within crime, which was our main topic to touch on. Thus, in Module 5 we planned to determine whether or not there was a correlation between a states median income and incarceration rate.

## Module 5: Prediction

### Introduction

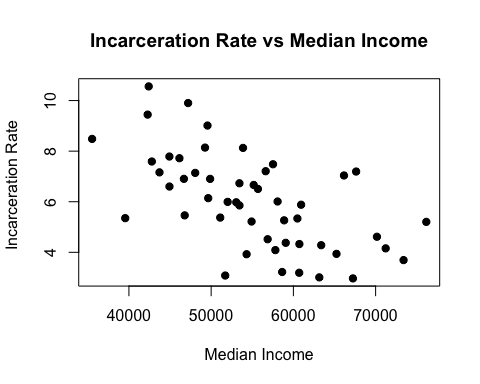
The purpose of Linear Regression is to predict the value of a response variable as a linear function of predictor variables. In this module, we demonstrated the analyses we have learned to predict the value of one variable using the values of one (or more) other variable(s). In context, we planned to explore the relationship between incarceration rates and median income by doing a linear regression and analyzing the results. All computations were done using R. We also illustrated the use of the model to predict the value of the incarceration rate when median income of a state is known. Lastly, we planned to include an analysis of the residuals and a discussion of how reasonable the assumptions were.

### Which Variable is Explanatory and Which is Response?/Linear Regression Model Motivation

In this case, our explanatory variable was median income and our response variable was incarceration rate per 1000 people. In context, this meant that we believed that knowing the median income of a state was a reasonable way to estimate it’s incarceration rate. This was motivated by everyday depictions of states that may have low median incomes and high incarceration rates such as Alabama and Arizona. These suspicions could not be further explored until the relationships between certain variables such as median income and incarceration were analyzed with a linear regression model.

### Scatterplot Production, and Analysis on the Direction, Strength, and Linearity of the Association Between Incarceration Rate and Median Income. Analysis of Potential Influential Points

pp1k = project\_data$"Prisoners Per 1000 People"  
plot(mi, pp1k, main="Incarceration Rate vs Median Income",   
 xlab="Median Income", ylab="Incarceration Rate ", pch=19)



The association between median income and incarceration rate in a state was negative because an increase in the median income of a state generally resulted in a decrease of the incarceration rate. We decided that the model was linear because the points plotted on the graph generally tended to follow a straight line of best fit. It was moderately strong because the majority of the points lied very close to the line of best fit we imagined while analyzing the plot. ###Calculation and Comment on the Correlation Coefficient.

cor(mi,pp1k)

## [1] -0.6082134

Our scatterplot analysis and comments were supported numerically by the correlation coefficient of -0.6082134. This represented a moderately strong negative association between the median income and incarceration rate of a state.

### Determining the Least Squares Line for Predicting Incarceration Rate from Median Income/Interpretation of the Value of the Slope Coefficient and Intercept.

model1 = lm(pp1k ~ mi)  
s=summary(model1)  
s

##   
## Call:  
## lm(formula = pp1k ~ mi)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.3395 -0.9256 -0.0013 0.6112 2.9740   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.289e+01 1.313e+00 9.821 4.55e-13 \*\*\*  
## mi -1.251e-04 2.357e-05 -5.309 2.80e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.503 on 48 degrees of freedom  
## Multiple R-squared: 0.3699, Adjusted R-squared: 0.3568   
## F-statistic: 28.18 on 1 and 48 DF, p-value: 2.797e-06

Our model using least squares was = 12.89 + -1.251e-04 \* Dollars. The slope coefficient for this model indicated that for every one unit decrease in dollars for a states median income, the predicted number of prisoners per 1000 people will decrease by -1.251e-04 people (a very small number, but makes sense in context as median income differs by thousands of dollars between states). The intercept shows us that if a state’s median income was 0 dollars then the number of prisoners per 1000 people would be 12.89 (approximately 13 people per 1000 in context).

### Reporting and Interpretation of the Coefficient of Determination.

Our coefficient of determination, otherwise known as adjusted R-squared, was 0.3568. This means that the model we produced explains 35.68% of the observed variation in incarceration rates of people across the United States.

### What Incarceration Rate Would the Line Predict for the 2014 Median Income of Washington D.C.

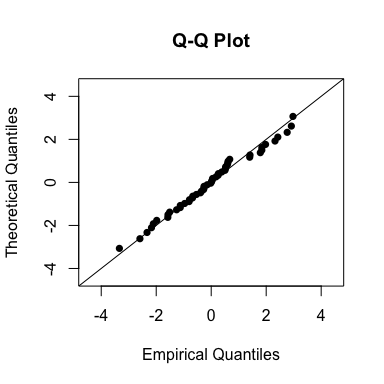
Washington D.C is the capital of the United States of America and technically, it is not a state. Therefore, its metrics were left out of this analysis and project as a whole. However, this allowed us to test the accuracy of our model by testing how closely it would predict the number of prisoners per 1000 people in Washington D.C. The United States reported the Washington D.C median income to be 68277 USD in 2014. The Federal Bureau of Investigation released their yearly Uniform Crime Report and the incarceration rate for Washington D.C. was 2.4283.

Our model produced the following: Predicted Incarceration Rate = 12.89 + -1.251e-04 \* Dollars Predicted Incarceration Rate = 12.89 + -1.251e-04 \* (Dollars)68277 = 4.3485 prisoners per 1000 people = − is called the residual. The residual describes the error in the fit of the model to the ith observation yi. In this case, the residual would be -1.92. This means that our line overpredicted the value of the incarceration rate for the median income of 68277 which was associated with Washington D.C.

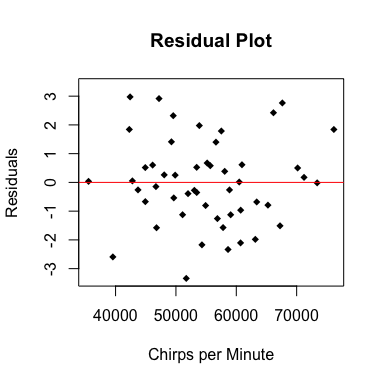
### 

### Examination of Residual Plots to Assess Whether the Technical Conditions for Inference Appear to be Satisfied

res1 = residuals(model1)  
  
x = res1  
n = length(x)  
m = mean(x)  
s = sd(x)  
limits = c(m - 3 \* s, m + 3 \* s)  
  
probs = (1:n)/(n+1)  
norm.quants = qnorm(probs, m, s)  
plot(sort(x), sort(norm.quants), ylab = "Theoretical Quantiles",   
 xlab = "Empirical Quantiles", main = "Q-Q Plot",  
 xlim = limits, ylim = limits, pch = 16)  
abline(0,1)



max\_abs\_res = max(abs(res1))  
plot(mi, res1, pch = 18, ylab = "Residuals", main = "Residual Plot", ylim = c(-1\*max\_abs\_res, max\_abs\_res),  
 xlab = "Chirps per Minute")  
abline(0, 0, col = 'red')



#Note these assumptions about residuals, Normally distributed (check this) with a mean 0, var 1, and lastly if there is a constant variance. The residual plot checks for constant variance, we do not want patterns, because this shows variance that is not constant. Watch out for "fan"/"sneeze pattern", this means that there is not constant variance. The qqplot checks for normality, we want most points on that close to the lin. Note that we check these assumptions after. If there is no linear pattern in scatter plot say that there is no linear pattern and linear regression cannot be done. If there is a pattern in your residual plot, keep working on the data .

Now that a Linear Regression Model has been implemented, we checked the assumptions for our model. There are four principal assumptions which justify the use of linear regression models for purposes of inference or prediction:

**Linearity:** This was shown by our scatterplot. The expected value of dependent variable was a moderately straight-lined function of each independent variable, holding the others fixed.

**Statistical Independence of the Errors**: The nature of the independence of these variables was expounded upon in Modules 1, 4, and 5.

**Constant Variance of the Errors**: The residual plot showed that there was no recognizable pattern formed, the points on it are random. Therefore, the constant variance assumption was met as well.

**Normality of the error distribution**: The mean of 0 was always present. However, we must establish whether or not the residuals are normally distributed. The QQ Plot checked for normality. We wanted most points on the QQ Plot close to the line and mostly concentrated in the center with few points on the ends. This QQ Plot shows this, as well as the fact that there were no outliers. Therefore the normality assumption passed the test.

### Hypothesis Test

Let represent the true slope in the relationship between incarceration rate and median income.

There is no relationship between incarceration rate and median income.

There is a relationship between incarceration rate and median income, we have a significant slope.

α = 0.05

The P-Value for 2.80e-06 was approximately zero. This value was small enough for us to reject our null hypothesis given our alpha value of 0.05. Therefore, we rejected our null hypothesis. If there was no real relationship between the median income and incarceration rate, then we would almost never observe a sample relationship such as this by chance. Therefore, our slope was significant enough to conclude that there was a relationship between median income and incarceration rate.

### Constructing a 95% Confidence Interval for the Population Slope Coefficient and Interpretation of this Interval.

We estimated two values so we must subtract 2 from our sample size to get our degrees of freedom. dof = 50 - 2 = 48 because we calculated the 95% confidence interval.

$t^{*} = $ 2.011 ± t* = -1.251e-04 ± 2.011(2.357e-05) = (-1.72e-04,-7.76e-05)

The 95% confidence interval for the population slope coefficient was (-1.72e-04,-7.76e-05). This means that if we repeated the process (select a random sample, gather the data, and construct the associated CI) many times (thousands of times), then approximately 95% of the constructed CIs will contain the population slope coefficient. However, we can never know which CIs actually contain the true value and which CIs do not.

### Steps for Developing a More Complex Model

A regression model that contains more than one regressor variable is called a multiple regression model. In Multiple Linear Regression, each x-variable can be a predictor variable or a transformation of predictor variables. Thus, to form a more complex model, one could analyze the impact of median income, unemployment rate, and the rate of poverty stricken familes to determine how they all impact the incarceration rate in each state.

### Conclusion

In all, this module was able to answer a significant question that was developed and created through the preceeding ones. We found out that there is indeed correlation between the median income of a state and its incarceration rate. This was an example of one of the bigger questions that we aimed to answer in Module 1. It was important to note that correlation does not imply causation. Correlation establishes the foundation for experiments and observational studies. Therefore, the results of this module would not tell one why states who have higher median incomes have lower incarceration rates, only that these two findings typically occur in tandem.

## Module 6: Summary and Conclusion

In Module 1, we established the most important question that we continued to revisit throughout the duration of the project; “What trends in states are commonly associated with high incarceration rates?” We sought to answer this question because in theory, there are socioeconomic and political variables in different states that seem to appear common amongst certain levels of incarceration rates. We wanted to further study this to determine correlations between socioeconomic/political statistics and crime-related statistics. While looking at this question, we also explored the change of incarceration rate from 2010-2014. 2010 was a pivotal year for census data, but, this year was also important because the United States was still recovering from a recession in the previous years. We expected 2010’s numbers to differ from 2014’s numbers because of the fact that many socioecomic/political trends were different in 2010 compared to 2014. Proceeding from the previous question, we analyzed at the difference in incarceration rates among Republican states versus Democratic states. From news outlets and claims of individuals living in the United States we have theorized that there may be a stronger bias towards the punishment of crimes coming from Republican states. Thus, we expected Republican states to have a higher incarceration rate on average. We asked a series of other questions to support our findings to ensure that the research conducted in this project was accurate.

Among all of these questions, we discovered that there indeed is a correlation between median income and incarceration rate within a given state. While analyzing this data, we must also remember, as stated above, that correlation does not determine causation, but could still be a good tool in determining further exploration for any portential causation. We concluded that the unemployment rate in the United States has not remained the same since 2010 as was expected over this 5-year period trailing from a nation-wide recession. We determined that the incarceration rate of Republican states was higher than the incaraceration rate of Democratic states as was predicted. As previously stated in module 5, a multiple regression would be sufficient for finding other factors that correlate with the number of prisoners in a state.

In observing and doing analysis on this data, we must also note that the investigation we have done does not fully explain the total relationship of these crime statistics amongst each other. It is simply a look into some of the more interesting questions that we have to ask about crime and possible correlations. Doing a multiple linear regression on all of these variables and selecting an elimination method to work by, we could learn much more about how each of the crime statistics relate to each other and how much each of them could be used to predict a rate of incarceration within a given state. This linear model would be able to tell us which variables relate to incarceration rate the most if we used incarceration rate as our response variable and the rest of the variables as predictor variables.

If a hypothetical phase 2 of this project were to be initiated, then it would involve more analysis with the use of data science. It would be benefical to actually run experiments, conduct surveys, and interview policymakers to understand the true cause of crime in the United States. Methods would include sentiment analysis of crime reports, analysis of of both structured and unstructured data, and metrics.

**Appendix A: References**

American FactFinder [Advertisement]. (2015). Retrieved February 19, 2018, from https://factfinder.census.gov/faces/nav/jsf/pages/index.xhtml

Crime in the United States by State, 2014. (2015, August 17). Retrieved February 18, 2018, from https://ucr.fbi.gov/crime-in-the-u.s/2014/crime-in-the-u.s.-2014/tables/table-5

G&A Staff. (2014, October 17). Best States for Gun Owners 2014. Retrieved February 19, 2018, from http://www.gunsandammo.com/network-topics/culture-politics-network/best-states-for-gun-owners-2014/

Kaeble, D., Glaze, L. E., Tsoutis, A., & Minton, T. D. (2015, December 29). Correctional Populations in the United States, 2014. Retrieved February 18, 2018, from https://www.bjs.gov/index.cfm?ty=pbdetail&iid=5519

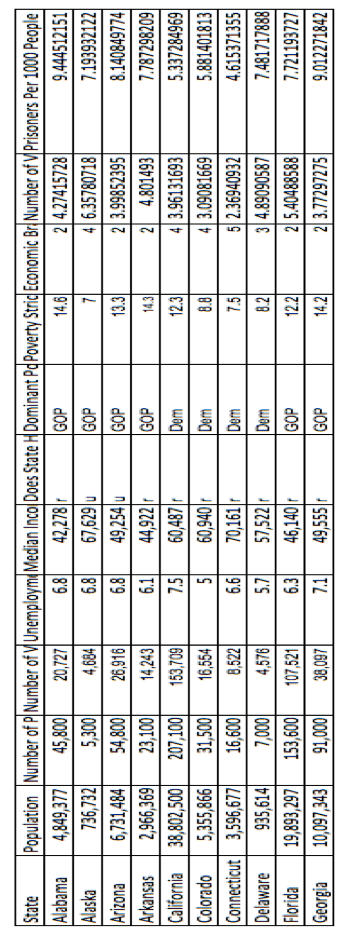
Kermitt. (2014, October 16). O Gerry, Stop Splitting Districts. Retrieved February 19, 2018, from https://demosrhino.wordpress.com/2014/10/16/o-gerry-stop-splitting-districts/

Montgomery, D.C., & Runger, G.C. (2018). Applied Statistics and Probability for Engineers[6th Edition]

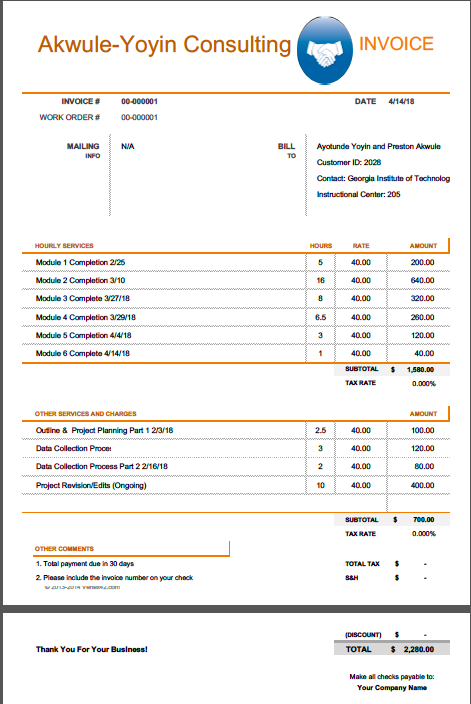
State-by-state and national crime estimates by year(s). (2015). Retrieved February 19, 2018, from https://www.ucrdatatool.gov/Search/Crime/State/StatebyState.cfm?NoVariables=Y&CFID=233562476&CFTOKEN=f5aedd6ad26b0374-BC914950-0200-0DC0-7C13EB8D5EF3EBCA

S. (2016, April 28). How do I interpret the Shapiro-Wilk test for normality in JMP? Retrieved April 14, 2018, from http://www.jmp.com/support/notes/35/406.html

**Appendix B: Data Spreadsheet**

****

**Appendix C: Billing Invoice**

****

**Appendix D: Consulting Hours Logged**

|  |  |  |
| --- | --- | --- |
| 2/3/18 | Outline & Project Planning Part 1: Do research on topic and have relevant idea ready to implement (Joint Delegation) | 2.5 |
| 2/10/18 | Data Compilation: Collected data from different sources, decided the final continuous and random variables to use (Delegated to Preston) | 3 |
| 2/16/18 | Reevaluated some of the data used. Compiled data for both 2014 and 2010 to allow for ease of comparison in future modules (Delegated to Ayo) | 1.5 |
| 2/17/18 | Established "big questions aimed to answer" for Module 1 (Delegated to Preston) | 1 |
| 2/19/18 | Compiled references to post on ISYE 2028 Canvas for Module 1 (Delegated to Ayo) | 2 |
| 2/25/18 | Wrote up Module 1 in RMD File | 2 |
| 3/5/18 | Attended office hours to plan module 2 | 2 |
| 3/7/18 | Wrote up most of the content for module 2 (joint delegation) | 6 |
| 3/9/18 | Coded tests in RMD file for Module 2 (Joint delegation) | 4 |
| 3/10/18 | Corrections made to module 2 | 4 |
| 3/19/18 | Questions about Module 3 asked Outline done at home. (Joint delegation) | 4 |
| 3/27/18 | Module coded and completed (Joint delegation) | 4 |
| 3/28/18 | Ayo attended office hours to ask about module 4 | 2 |
| 3/29/18 | Module 4 completed | 4.5 |
| 4/3/18 | Module edit and review session based off Dr. A’s recommendations (Joint delegation | 3 |
| 4/4/18 | Module 5 completed | 3 |
| 4/9/18 | Module 3 and 4 revisited for edits and power calculation | 3 |
| 4/14/18 | Module 6 completed and project final edits | 5 |

**Total Hours 56.5**

**Appendix E: Thanks**

There were many individuals who made this project successful. We would like to thank:

* Dr. Alisha Waller
* Alicia Will
* Ndeyanta Jallow
* Christian Hughey
* DeLauryn Brown
* Ashwin Haritsa

The assistance, guidance and wisdom was very helpful and the project would not be a success without the individuals.