Hughey\_Jallow\_Raw\_Knit

Hughey\_Jallow

4/13/2018

knitr::opts\_chunk$set(echo = TRUE)

library(readxl)  
data <- read\_excel("finaldata.xlsx")  
attach(data)

# Module 1: Introduction and Data Collection

## General Area of Focus

Police shootings have become a serious point of discussion today, thus developing insight on the ongoing phenomenon of these fatal encounters was worthy of observation. We chose to focus on fatal police shootings from year 2015 until recently in February 6th, 2018. We focused on several factors that were present at the time of the police shooting, which ultimately resulted in victim fatally.

The use of deadly force by police officers is a very important subject in today’s society. Many consider the use of deadly force excessive in most cases. However, there are many aspects to look at when considering this topic, such as: Why was deadly force required? When did the officer feel it necessary to apply deadly force? What will be the implications for the officer after the fact? How does the use of deadly force affect society as a whole? In order to answer these questions, we must first define deadly force. According to the Legal Dictionary, “use of deadly force” is often granted to police forces when a person poses a significant threat to a law enforcement officer or believed to be an immediate danger to people around them. For example, an armed man in a shopping mall shooting at civilians without regard to the safety of anyone around him, and refusing or being unwilling to negotiate, would warrant usage of deadly force, as a means to prevent further danger to the community. In regards to the aforementioned questions, these inquiries will be answered through investigation by analyzing what recurring variables were present at the time of each fatal police shooting in each state of the United States. Those variables include intrinsic factors such as race, sex and age; as opposed to external factors such as presence of police body camera, average median income, homicide rate, and unemployment rate of the state.

The grey area lies where the excessive force by police officers are given a significant amount of discretion simply due to the nature of the job. Officers are faced with many threatening situations forcing them to react quickly, yet appropriately. They have the power to infringe upon any citizen’s rights to freedom and therefore they must use this power effectively. One major concern with the amount of discretion officers have is their power to decide when to use lethal force. Therefore, given the importance of the American walk of life, a review of what is unknown represents an essential place to begin.

In recent years, there have been hundreds of fatal encounters between civilians and police officers. Many untold and one-sided stories left up to imagination and ambiguity due to the inability to hear the late person’s testimony. Were they at fault or was lethal force warranted? Decide for yourself, after taking a read into a few high profiled fatal police shootings listed below.

July 27, 2015 – Jean Paul Falgout Law enforcement officers fatally shot a man who they said had pointed a realistic-looking pellet gun at them. The 45 year-old man, Jean Paul Falgout, who was wanted on nonviolent criminal charges, led officers on a car chase until he was cornered, and was shot after he got out of his car. Eight officers from three police agencies fired at him; it is not clear how many of the shots hit him.

February 27, 2015 – Ernesto Canepa Diaz Police officers fatally shot unarmed, Ernest Canepa Diaz, who they suspected of committing a robbery. Santa Ana police officers approached Ernesto J. Canepa Díaz, 27, a native of Mexico, because he was sitting in a vehicle that matched that of a man who had taken a gold chain from a woman, officials said. But they have not said what prompted the officers to open fire, and they have not publicly identified the officers. Family members say that Mr. Canepa was unarmed, which the authorities have not refuted. Officials have said that there was a replica pistol – possibly a BB gun – in the vehicle, but have not said whether he was holding it, or whether it played any part in the shooting or the earlier robbery.

July 5, 2016 – Alton Sterling Police were responding to a report that a man in a red shirt was selling CDs and that he had used a gun to threaten a man outside a convenience store. Officers were attempting to control Sterling’s arms while his back was on the ground, and he was shot after allegedly reaching for the officer’s gun in his pocket. The shooting was recorded by multiple bystanders and went viral across multiple news networks and social media platforms. The shooting led to protests in Baton Rouge and a request for a civil rights investigation by the U.S. Department of Justice.

July 6, 2016 – Philano Castile Castile was in a car with his girlfriend, Diamond Reynolds, and her four-year-old daughter when he was pulled over by Yanez and another officer. Philando Castile was shot and killed by Jeronimo Yanez, a St. Anthony, Minnesota, police officer, after being pulled over in a suburb of Saint Paul. The shooting achieved a high profile from a live stream video on Facebook made by Reynolds in the immediate aftermath of the shooting. It shows her interacting with the armed officer as a mortally injured Castile lies slumped over, moaning slightly and his left arm and side bloody. According to a police dashcam video/audio, after being asked for his license and registration, Castile told the officer he had a firearm, to which the officer replied “Don’t reach for it then”. After saying “Don’t pull it out” twice, the officer shot at Castile seven times. Diamond Reynold’s testimony was that Castile was shot while reaching for his ID after telling Officer Yanez he was armed.

July 7, 2016 – Dallas Sniper Attack In the aftermath of those consecutive fatal police shootings in the summer of 2016, there was immediate retaliation and social discontent expressed through riots, protests, and marches. Following the above occurrences, a lone-wolf sniper gunned down police officers in the streets of Dallas, Texas on the night of July 7, 2016. Micah Xavier Johnson was responsible for killing five officers and injuring nine others, while wounding two civilians.

For a country that prides itself on being a nation of laws, these are clearly troubling trends but they also beg the question as to whether these incidents represent a dramatic increase in the number of shootings involving law enforcement authorities or likelihood of being shot due to variables present at the time of the shooting.

## Big Questions

The comprehensive question that was under investigation throughout the project was: What circumstances are present where the probability of being fatally shot by a police officer is significantly increased?

1. Are there certain factors that lead to more deaths of a certain race than another?   
2. Are there more police shootings in areas of higher homicide rates or lower homicide rates?   
3. Should body cameras be worn by the police more often?   
4. Is there a relationship between the unemployment rate and the homicide rate?   
5. If a population has a higher unemployment rate, would there be more or less crime in that area?

These are all essential questions that were analyzed by dissecting the data to provide support and credence for the claim.

## Data Collection

When our data was collected, the observational unit was the 50 states with the District of Columbia included. The variables of interest include data on all the fatal police shootings within the area of study.

Within each state, underlying factors transpire before an observation of a police shooting. For example, there’s an occurrence of events that happen leading up to some police shooting that occurs in Florida and another shooting in Massachusetts perhaps at the same moment in time, yet neither observation influences the other shooting in the respective state; thus making the police shootings independent. The randomness of the sample comes from each point in time of the exact moment the shooting occurred, since there is no structured time for each observation to occur. From state to state the occurrence of the shootings are time dependent; however, what occurs in each state is independent of another. There are covariances that reside over all the states such as federal law, but the covariance effects are minimal in relation to every individual shooting in each state resulting in sufficient independence.

Furthermore, data was collected by visiting trusted websites and downloading spreadsheets. The main source of data was obtained from The Washington Post. Then we compiled the data from the spreadsheets into one excel file that fit our parameters, essentially transformed in cleaned data. Moreover, the desired variables represented the findings, however there were was an excess of categorical variables and a deficiency of numerical variables to examine all spectrums of the data. Thus, we created a new excel table with respective ordinal, continuous, nominal and binary variables that were pertinent to the general area of focus below.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable | Concept | Type | Range of Reasonable Values | Unit |
| num\_killed | The total number of killings in each state | Continuous | 0 - 600 | People |
| avg\_age | The average age of all the victims shot and killed by the police per sate | Continuous | 30 - 50 | Years |
| med\_income | The median household income of each state | Continuous | 40000 - 80000 | Dollars |
| homicide\_rate | Per 10,000 people, the rate at which homicide crime occurs in each state | Continuous | 0 - 20 | N/A |
| law\_officers | The number of sworn in and employed law enforcement officers in each state | Continuous | 0 - 50000 | People |
| maj\_race | The most common / majority race killed amongst all police induced fatalities in each state | Nominal | B (Black) A (Asian) W (White) H (Hispanic) N (Native American) O (Other) | N/A |
| maj\_armed | The majority type of arms (if any) the victim had on their person at the time of the incident per state | Nominal | Gun Vehicle Knife | N/A |
| unemp\_rate | The unemployment rate per state based upon the labor force | Ordinal | 2 ≤ x < 3 3 ≤ x< 4 4 ≤ x < 5 5 ≤ x < 6 6 ≤ x < 7 | N/A |
| geo\_region | The geographical region of the state defined by the US Census Bureau broken up into 9 divisions | Nominal | 1 – Northeast 2 – Midwest 3 – South 4 – West | N/A |
| body\_camera | Whether or not police officers are required to wear body cameras at all times | Binary | 0 - No 1 - Yes | N/A |
|  |  |  |  |  |

## Layout to Rest of the Project

Through the analysis of our data, we came to various conclusions and hypothesized recommendations in regards to fatal police shootings. Through each module, we investigated variables to gain a deeper understanding of what they truly represent about our sample. In Module 2, modeled our data with probability distributions. Whether a good or bad fit, we dove into the best models for our data, their perspective parameters and much more. In Modules 3 and 4, we looked into hypothesis tests for both single and two sample analyses to potentially answer a few of our questions outlined above. Lastly, in Module 5 we utilized a linear regression model to investigate the correlation that exists within two variables. In all, this statistical report outlined many different tactics used to analyze data in the statistics world.

# Module 2: Modeling Data with Probability Distributions

To analyze our data and the different types we choose to consider, this module investigated of summaries of data, probability distribution models, parameter estimations, and the fit of those models. It was first important to understand the various types of data we will be operating on in order to accurately utilize the correct methods of analysis for the respective type of each variable.

## Continuous Variable Analysis - num\_killed

summary(num\_killed)

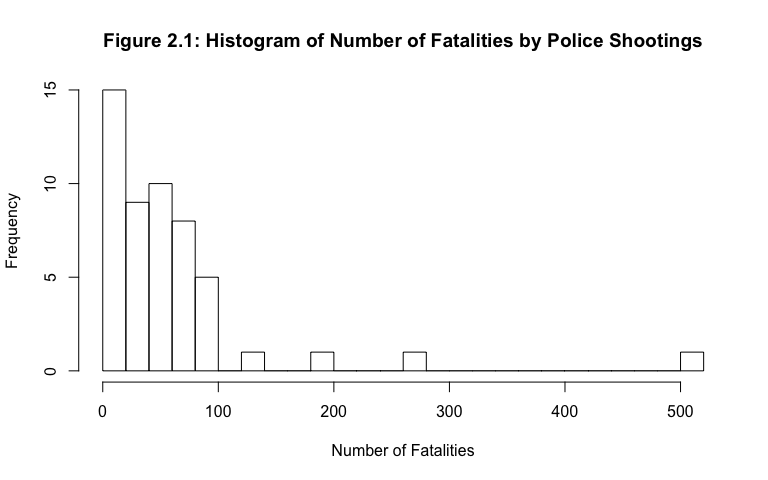
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 3.00 16.00 44.00 59.86 68.50 501.00

sd(num\_killed)

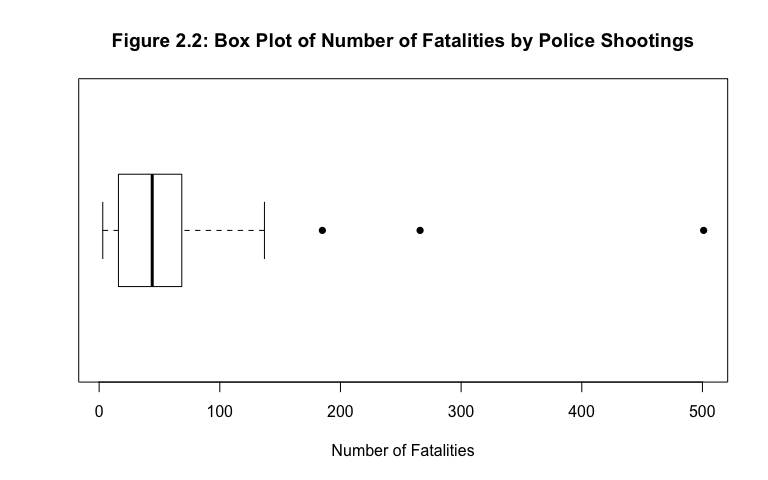
## [1] 78.65368

Continuing on, the continuous variable we analyzed was num\_killed. We believed this variable was an important factor in many of the conclusions that we drew as the victims killed vary, in many cases, extremely from state to state. First, the range of the data was quite large, from a minimum of 3 people to a maximum of 501. Between this range of values, the median was approximately 44 with a mean of 59.86. With just this information, we immediately disregarded a few types of distribution families (ie: normal, uniform) and began to focus in on the ones that made adequate sense for our data.

hist(num\_killed, breaks=25, xlab ="Number of Fatalities", main = "Figure 2.1: Histogram of Number of Fatalities by Police Shootings")

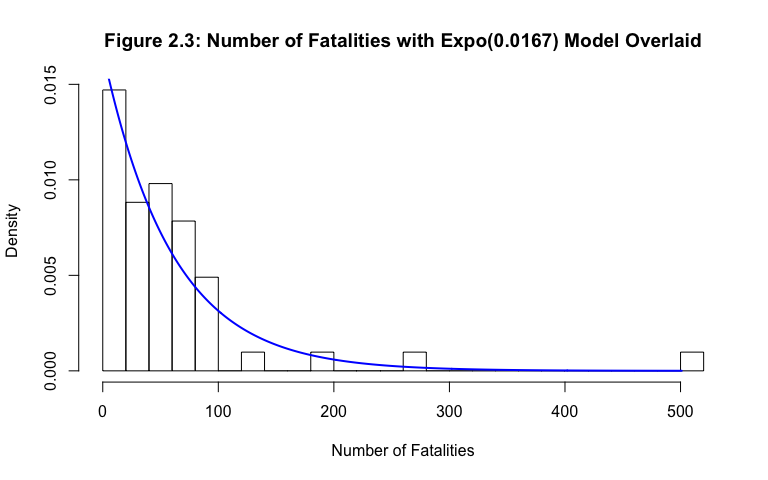


boxplot(num\_killed,horizontal = TRUE, xlab = "Number of Fatalities",pch = 16,main = "Figure 2.2: Box Plot of Number of Fatalities by Police Shootings")



The histogram and box plot of the variable made it quite obvious which distribution family should be chosen as a model (with hopes that it would be a good fit). The trends and characteristics of the visuals embodied those similar to the standard exponential distribution. More specifically, the values of the histogram in Figure 2.1 decreased with frequency as you moved to the left; the mean was greater than the median which implies a right skew (hence the long tail towards the right); the value of those shot and killed by the police started at 0 (since negative values made absolutely no sense); in the differences between the quartiles of the box plot in Figure 2.2; and versus and , more values fell towards 0 and the number of values towards the maximum decreased (which also explained the shorter left wisker and longer right wisker as well); and visually, outliers exist. Both the box plot and the histogram gave more justification as to why the exponential distribution family was ideal for this data, rather than the observation of one or the other.

hist(num\_killed, breaks=25, xlab ="Number of Fatalities", main = "Figure 2.3: Number of Fatalities with Expo(0.0167) Model Overlaid", freq = FALSE)  
x <- num\_killed  
xfit<-seq(min(x),max(x))  
yfit<-dexp(xfit,rate=1/mean(x))  
lines(xfit, yfit, col="blue", lwd=2)



The fit was not visually perfect for the model overlaid in Figure 2.3, as there was over represenation of the bins. Although the model included outliers, it appeared to be exponential, the shape of the histogram had the potential change with the absence of those outliers.

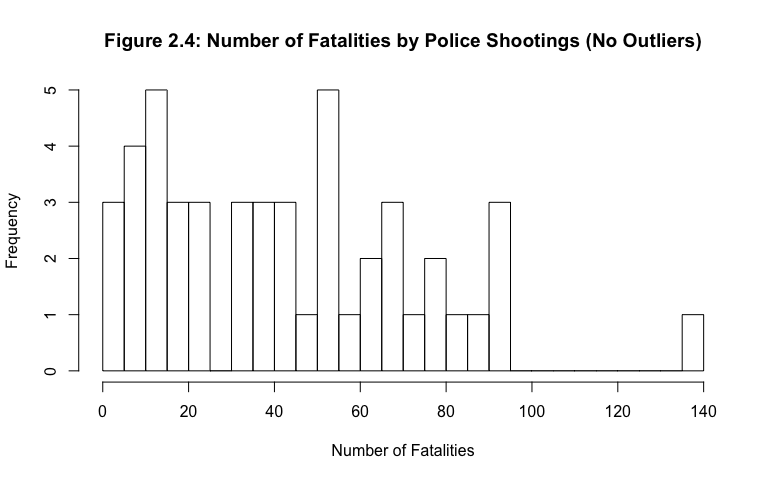
q1 = 16  
q2 = 44  
q3 = 68.50  
iqr = q3 - q1  
ub = q3 + 1.5\*iqr  
lb = q1 - 1.5\*iqr  
outlier = data$"num\_killed">ub  
noutliers = data$"num\_killed"[!outlier]  
summary(noutliers)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 3.00 15.75 40.50 43.77 65.25 137.00

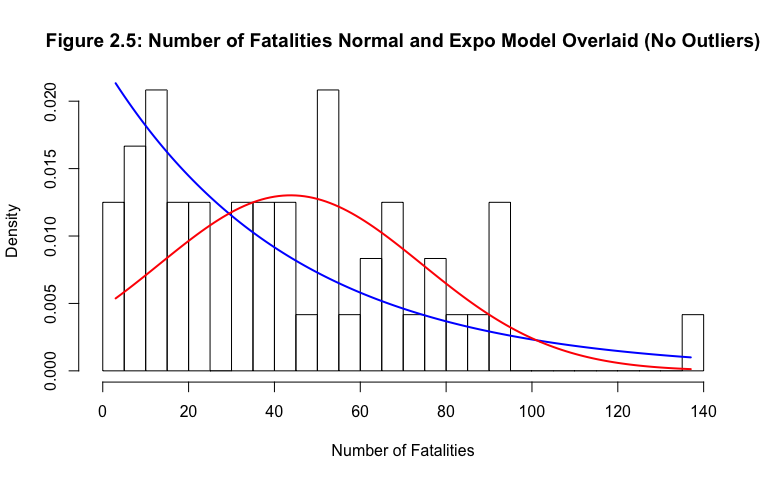
sd(noutliers)

## [1] 30.64431

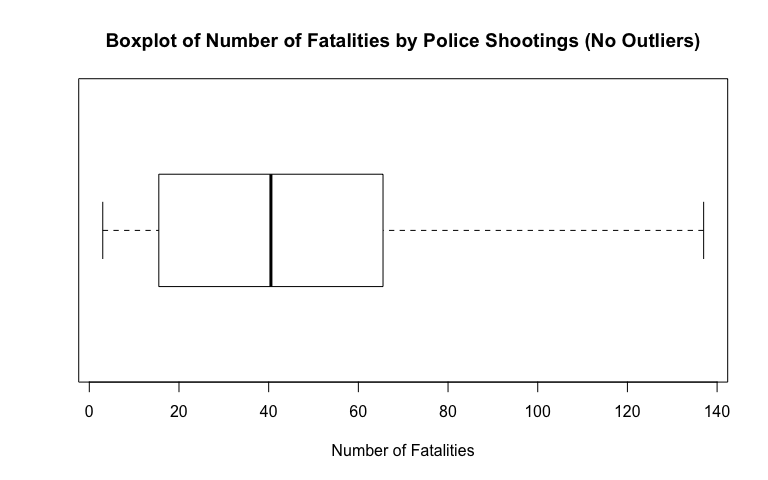
hist(noutliers, breaks=25, xlab ="Number of Fatalities", main = "Figure 2.4: Number of Fatalities by Police Shootings (No Outliers)")



hist(noutliers, breaks=25, xlab ="Number of Fatalities", main = "Figure 2.5: Number of Fatalities Normal and Expo Model Overlaid (No Outliers)", freq = FALSE)  
  
x <- noutliers  
xfit<-seq(min(x),max(x))  
yfit<-dexp(xfit,rate=1/mean(x))  
lines(xfit, yfit, col="blue", lwd=2)  
  
xfit2<-seq(min(x),max(x))  
yfit2<-dnorm(xfit,mean = mean(x), sd = sd(x))  
lines(xfit2, yfit2, col="red", lwd=2)

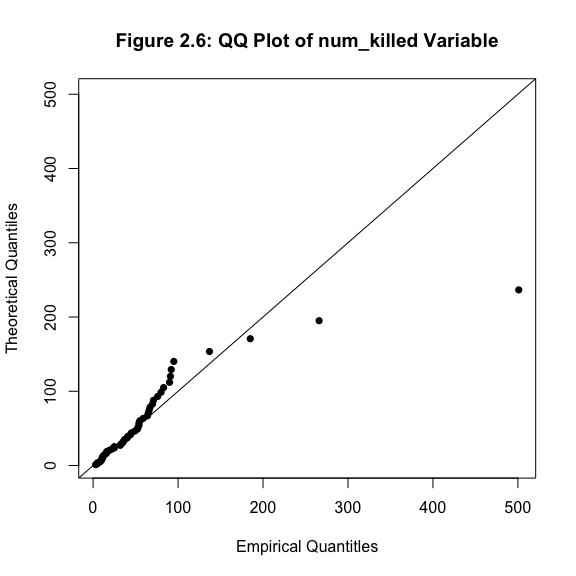


boxplot(noutliers,horizontal = TRUE, xlab = "Number of Fatalities",main="Boxplot of Number of Fatalities by Police Shootings (No Outliers)")



To ensure that the exponential model was indeed the best fit, it was not a bad idea to look at the histogram without those extreme values. To calculate the upper and outlier boundaries, the values of the summary above was used. The code used the formula and . Clearly, there were no lower outliers and based on the data vales there were 3 upper outliers: 185, 266, 501. Removing those values, the mean decreased drastically from 51.86 to 43.77 and the standard deviation also had a large decrease from 78.65 to 30.64, which implied that the outliers had a large impact on the skew of this data set. Therefore, one could visually argue either a normal or an exponential distribution family fit for the histogram. Clearly, the shape of the distributionchanged since Figure 2.4. When both models were overlaid, the blue being exponential and the red being normal, we argued that the exponential was still the better fit (even without the outliers). Visually, it appeared that the normal distribution did not capture as much data as the exponential. Furthermore, looking at the histogram, it shared similar characteristics to that of the boxplot originally analyzed in Figure 2.5. Considering this, our analysis of the num\_killed variable continued on with the exponential distribution family.

x = num\_killed  
n = length(data$"num\_killed")  
m = mean(x)  
r.exp = 1/m  
limits = c(min(x),max(x))  
probs = (1:n)/(n+1)  
exp.quants = qexp(probs,r.exp)  
plot(sort(x),sort(exp.quants),  
 ylab="Theoretical Quantiles",  
 xlab="Empirical Quantitles", main="Figure 2.6: QQ Plot of num\_killed Variable",xlim=limits, ylim=limits, pch=16)  
abline(0,1)



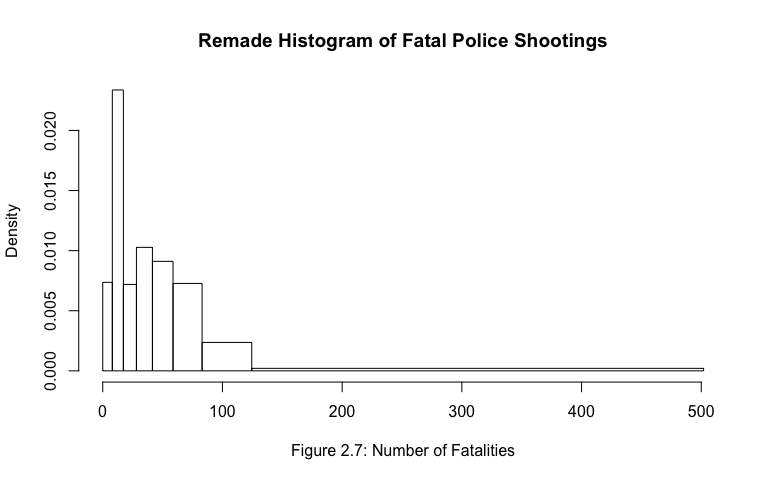
In order to estimate the parameter value of this exponential distribution, the method of moments was a justifiable way to determine the values. According to Dr. A’s table of probability distribution formulas, the mean (E[X]) of a exponential random family was . The mean of this data (including outliers) was 59.86 by the numerical summary of the data. Therefore, we employed the 1st moment of X, .

Then, E[expo()] = which corresponded to the right hand side of the MOM equation. The sample mean (sum of all the observations to the first power divided by the number of observations), 59.86 was the right hand side of the equation. Finally, = 59.86 so = 0.0167.

Based on this lambda value, a QQ-plot provided a more consolidated conclusion of our choice of distribution family in Figure 2.6. There was evidently a tight fit towards the lower end of the graph, which we expected as there should be more values that fell towards the left of the distribution. Moreover, as the quantities (both theoretical and empirical) got larger, the tightness decreased considering the probability of values appearing in such higher quantities decreased with more movement towards the right end of the graph. Theoretically, our measured outliers most definitely did not fit for this distribution; however, we were confident enough in our modeling to finalize a conclusion with a confidence interval analysis followed by a goodness of fit hypothesis test.

In order to gauge the confidence of both our lambda and mean values, a 95% confidence interval was applied. Utilizing *Applied Statistics and Probability for Engineers by Douglas C. Montgomery and George C. Runger*, in order to conduct a CI on an exponential distribution, percentage points against the Chi-Squared Distribution with degrees of freedom 2n (n being the same size) were used. = . The lower tail (LT) and the upper tail (UT) were found using the Chi-Squared distribution table with probabilities ; however, with the mentioned above, we want df = 2n = 102 and the areas of the LT and UT to be 0.975 and 0.025, respectively. Therefore, the equation was written as . Using the data of the num\_killed variable, ; and so, and . The 95% confidence interval of then became . Since lambda was equal to the 95% confidence interval of became . If we ran this experiment, gathering data for the fatal police shootings in the 50 states (including DC) thousands of times, and conduct a confidence interval on either the mean or the value of lambda each time; then, 95% of the intervals will contain the true value of lambda or the mean (depending on which interval is checked).

b = c(0,7.99,17.22,28.13,41.49,58.71,82.98,124.47,502)   
h = hist(num\_killed, breaks = b, xlab = "Figure 2.7: Number of Fatalities",main="Remade Histogram of Fatal Police Shootings")



bin = c(0,1,2,3,4,5,6,7)  
oj = h$counts  
ej = c(6.38,6.38,6.38,6.38,6.38,6.38,6.38,6.38)   
calculation = c()  
for(i in 1:8)  
{  
 calculation[i] = (oj[i] - ej[i])\*\*2/ej[i]  
}  
chisqtest = data.frame(bin,oj,ej,calculation)  
show(chisqtest)

## bin oj ej calculation  
## 1 0 3 6.38 1.79065831  
## 2 1 11 6.38 3.34551724  
## 3 2 4 6.38 0.88783699  
## 4 3 7 6.38 0.06025078  
## 5 4 8 6.38 0.41134796  
## 6 5 9 6.38 1.07592476  
## 7 6 5 6.38 0.29849530  
## 8 7 4 6.38 0.88783699

teststat = sum(calculation)

Utilizing a Chi-Squared Test on Goodness of Fit, we first binned our data to ensure that there was equal probability of the values that fell in each bin. To determine the bin boundaries for the eight bins, the CDF of an exponential distrubtion, Expo() = 1 - . With the value of 0.0167, and the increased proability by .125 per bin, the values of x (the bin boundaries) were determined. In the chart below, the procedure is laid out.

|  |  |  |
| --- | --- | --- |
| Bin | Formula | “x” Value |
| 1 | 1 - e^(0.0167)x = 0.125 | 7.99 |
| 2 | 1 - e^(0.0167)x = 0.25 | 17.22 |
| 3 | 1 - e^(0.0167)x = 0.375 | 28.13 |
| 4 | 1 - e^(0.0167)x = 0.5 | 41.49 |
| 5 | 1 - e^(0.0167)x = 0.625 | 58.71 |
| 6 | 1 - e^(0.0167)x = 0.75 | 82.98 |
| 7 | 1 - e^(0.0167)x = 0.875 | 124.47 |

Dependent on the Figure 2.7 above, there were two possibilities for this GOF test: = The exp(0.0167) model was an adequate model for our data and = The exp(0.0167) model was not an adequate model for our data. When performing the test, our value was 0.05, chosen against the most common value used. A Chi-Squared test stastic where was the observed values based on the frequencies of the bins, was the expected value of in this case. The degrees of freedom (df) = number of bins - number of parameters - 1; in this case df = 8 - 1 - 1 = 6. Therefore, we knew that the distribution of the test statistic ~ for the num\_killed variable. To calculate the critical value of the test statistic, the determination of the rejection region was P{} = 1 - . Using Table IV in M&R, the value of t = 12.59; and so, P{} = 0.95 therefore = 12.59. To calculate , the formula outlined above along with R code was used to determine the value to be 8.75. Finally, the comparison of , 8.75 < 12.59 therfore, we fail to reject . The exp(0.0167) model was an adequate model for our data. The data of the num\_killed variable fit well for the exponential distribution family. This fell in line with our expectations as explained above.

## Binary Variable Analysis: body\_camera

In recent times, the conversations regarding police usage/requirements of a body camera increased in many different areas. With that in mind, it was definitely interesting to see which observations out of our population require police to wear a body camera and which do not. Since there are only 2 choices for our variable, 0 or 1, the best distribution family for the variable was the Bernoulli distribution family. Our parameter, p, can be accurately identified as the probability of 1’s, calculated by the mean of the data. In the numerical summary below, the mean of the distribution was equal to 0.098 which was also the total number of 1’s divided by of the total number of observations (5/51 = 0.098). On the other hand, the probability of 0’s was 1 - p (the probability of 1’s) = 1 - 0.098 = 0.902 (46/51 = 0.092).

summary(body\_camera)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.00000 0.00000 0.00000 0.09804 0.00000 1.00000

counts <- table(data$body\_camera)  
barplot(counts, main = "Figure 2.8: Boxplot Body Camera Requirement", names.arg = c("Not Required (O) ","Required (1) "), xlab = "Requirement", ylab = "Frequency", ylim = c(0,50))

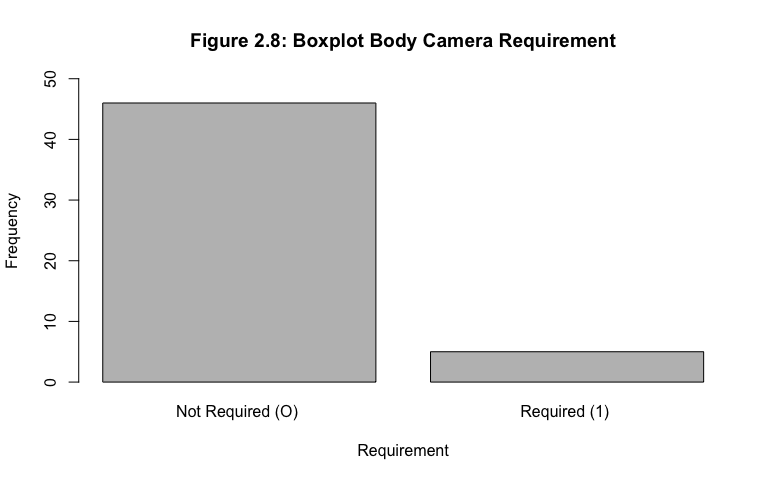
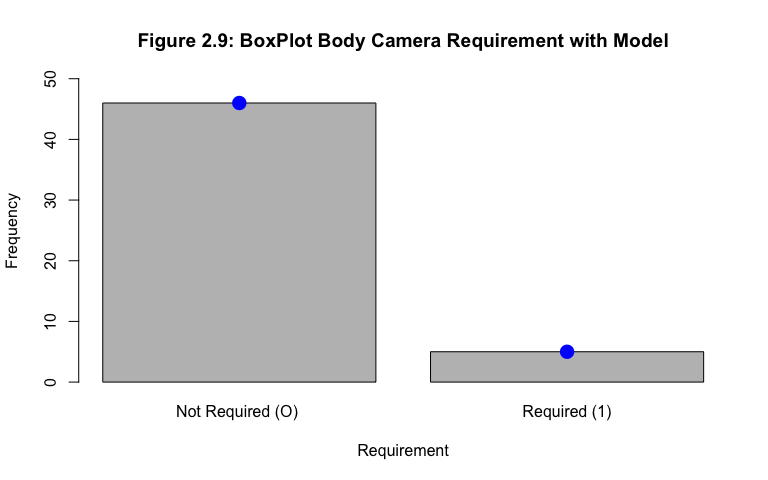


Figure 2.8 was created to clearly illustrate the shape of the distribution for categorical data. A QQ plot was not necessary for this anlaysis since the variable was not continuous.

bplot = barplot(counts, main = "Figure 2.9: BoxPlot Body Camera Requirement with Model", xlab = "Requirement", ylab = "Frequency", ylim = c(0,50), names.arg = c("Not Required (O) ","Required (1)"))  
points(x = bplot, y = c(46,5), pch=16, col ="Blue",cex =2)



For our data to fit a perfect model of the Bernoulli distribution, we would have expected there to be a perfect model overlaid with the parameters found with the data. In that case, the model overlaid as the 2 dots - the left (above 0) representing p and the right representing 1 - p (above 1) the model was in fact a perfect fit to our data in Figure 2.9. This did not go against our expectation as Bernoulli random variables only have 2 options which were almost always perfectly easy to calculate. For this reason, a goodness of fit test was not necessary as it was clear that the data of the fits the model well.

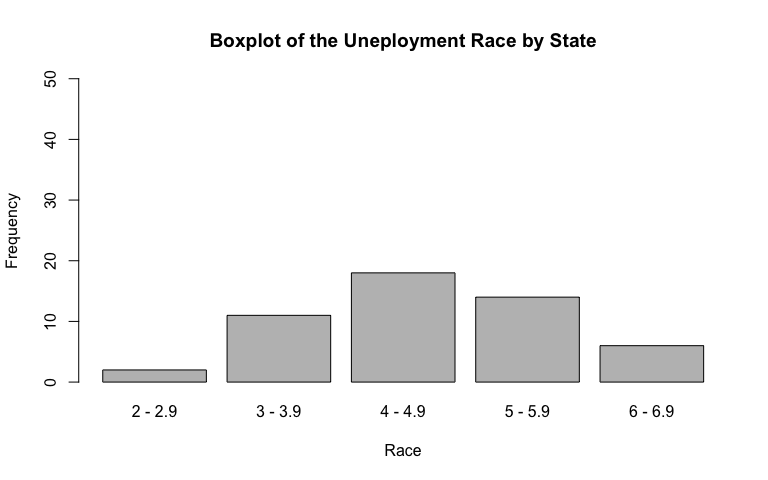
# Ordinal Variable Analysis: unemp\_rate

When choosing an ordinal variable, the unemp\_rate variable was broken up into five different, yet ordered categories each with a particular range of values. Within each range, there existed frequencies which were determined by the length of the respective range. Since ordinal variables were broken up into such categories, continuous distribution families would not be adequate. On the other hand, the discrete families made sense to use - in doing so, it was important to consider the shapes and trends of all of the families before heading towards one direction.

summary(data$unemp\_rate)

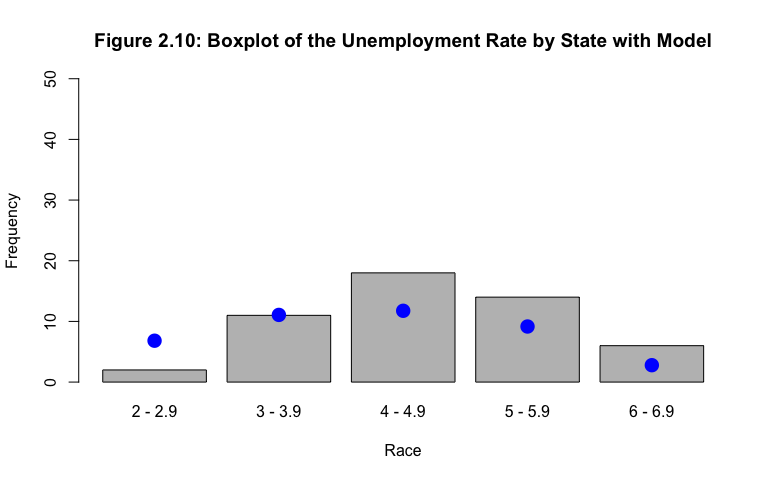
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 2.800 3.950 4.800 4.667 5.350 6.700

x1 = c(2.8,2.8)  
x2 = c(3.0,3.2,3.2,3.3,3.3,3.4,3.7,3.7,3.8,3.9,3.9)  
x3 = c(4.0,4.0,4.1,4.1,4.2,4.3,4.4,4.4,4.5,4.6,4.8,4.8,4.8,4.9,4.9,4.9,4.9,4.9)  
x4 = c(5.0,5.0,5.1,5.1,5.3,5.3,5.3,5.4,5.4,5.4,5.4,5.7,5.8,5.9)  
x5 = c(6.0,6.0,6.0,6.1,6.6,6.7)  
counts = c(length(x1),length(x2),length(x3),length(x4),length(x5))  
barplot(counts, main = "Boxplot of the Uneployment Race by State", xlab = "Race", ylab = "Frequency", ylim = c(0,50),names.arg = c("2 - 2.9","3 - 3.9","4 - 4.9","5 - 5.9","6 - 6.9"))



In the pool of the discrete family, we could have possibly overlaid the data with a discrete uniform, geometric, binomial, negative geometric, negative binomial, or poisson distribution family. Looking through the PDF’s of all of the families, the one that appeared to be most visually similar to the data in the barplot below was Binomial(n,p). There were a few reasons why this data fits that distribution well; the first being that it was a discrete random variable, as Binomial was as well. Secondly, our data was unimodal. As the value of n increases, the PDF emulated a Normal distribution more and more (although it can never be perfectly symmetric). Lastly, there was a “bound” on our value of n, it should not reach infinity but rather, 51 (the number of states in the US with the addition of DC). This lead to a hypothesis that the Binominal distribution was the best fit. Since the variable was non-continuous, it didn’t make sense to analyze the fit of the distribution with a QQ plot; however a Goodness of Fit test can be conducted.

x1 = c(2.8,2.8)  
x2 = c(3.0,3.2,3.2,3.3,3.3,3.4,3.7,3.7,3.8,3.9,3.9)  
x3 = c(4.0,4.0,4.1,4.1,4.2,4.3,4.4,4.4,4.5,4.6,4.8,4.8,4.8,4.9,4.9,4.9,4.9,4.9)  
x4 = c(5.0,5.0,5.1,5.1,5.3,5.3,5.3,5.4,5.4,5.4,5.4,5.7,5.8,5.9)  
x5 = c(6.0,6.0,6.0,6.1,6.6,6.7)  
counts = c(length(x1),length(x2),length(x3),length(x4),length(x5))  
bplot2 = barplot(counts, main = "Figure 2.10: Boxplot of the Unemployment Rate by State with Model", xlab = "Race", ylab = "Frequency", ylim = c(0,50),names.arg = c("2 - 2.9","3 - 3.9","4 - 4.9","5 - 5.9","6 - 6.9"))  
points(x = bplot2, y = c(6.8187,11.0721,11.751,9.1596,2.7846), pch=16, col ="Blue",cex =2)



With a Binomial(n,p) distribution, np = E[X]. In this case, we knew that n is equal to 51 and we must only estimate . In order to determine the expected values per the pdf of the distribution. The blue dots overlaid on the barplot correspond to the probability of each bin multiplied by n = 51.

oj = c(2,11,18,20)  
ej = c(6.8187,11.0721,11.751,11.9442)   
calculation\_ord = c()  
for(i in 1:4)  
{  
 calculation\_ord[i] = (oj[i] - ej[i])\*\*2/ej[i]  
}  
chisqtest\_ord = data.frame(oj,ej,calculation\_ord)  
show(chisqtest\_ord)

## oj ej calculation\_ord  
## 1 2 6.8187 3.4053220834  
## 2 11 11.0721 0.0004695053  
## 3 18 11.7510 3.3231215216  
## 4 20 11.9442 5.4332574505

teststat\_ord = sum(calculation\_ord)  
teststat

## [1] 8.757868

First, the value of . Using the Method of Moments, it was understood that E[X] = np. To calculate E[X] using the data, we multiplied the value of the bin by its proability. In this case, E[X] = = 3.118. Therefore, np = 51(p) and from calculations above, 3.118 = 51(p) = = 0.061. As any GOF HT, our : The Binomial(51, 0.061) distribution was an adequate model for the data and our : The Binomial(51, 0.061) distribution was not an adequate model for the data. With alpha being 0.05 and a Chi-Squared test statistic of where was the observed values based on the frequencies of the bins, was the expected value dependent on the probability density function. The degrees of freedom (df) = number of bins - number of parameters - 1; in this case df = 5 - 1 - 1 = 3. Therefore, we knew that the distribution of the test statistic ~ for the unemp\_rate variable. In the table below, the value of the test statistic from our data was calculated as follows.

|  |  |  |  |
| --- | --- | --- | --- |
| Bin | Observed Frequency | Probability | Expected Frequency |
| 1 | 2 | 0.1337 | 6.8187 |
| 2 | 11 | 0.2171 | 11.0721 |
| 3 | 18 | 0.2304 | 11.751 |
| 4 | 14 | 0.1796 | 9.1596 |
| 5 | 6 | 0.0546 | 2.7846 |

Since the expected frequency of bin 5 is less than 5, it was important that we combined its frequncy with that of bin 4 so we adequately conducted the test.

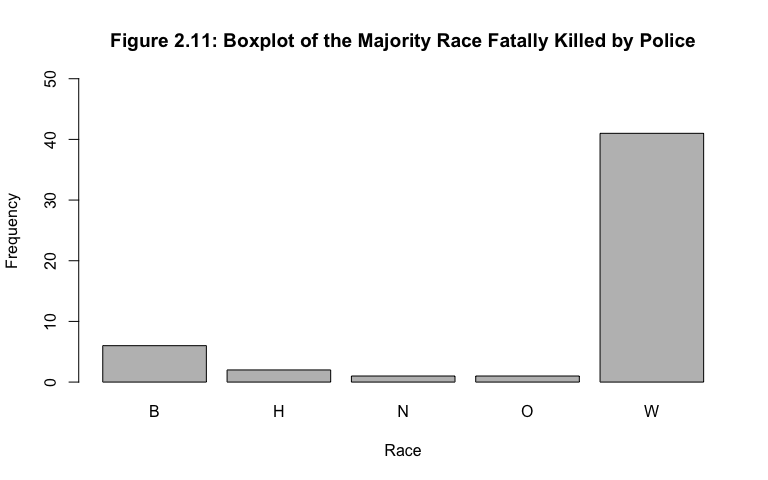
|  |  |  |  |
| --- | --- | --- | --- |
| Bin | Observed Frequency | Probability | Expected Frequency |
| 1 | 2 | 0.1337 | 6.8187 |
| 2 | 11 | 0.2171 | 11.0721 |
| 3 | 18 | 0.2304 | 11.751 |
| 4+ | 20 | 0.2342 | 11.9442 |

To calculate the critical value of the test statistic, the determination of the rejection region was P{} = 1 - . Using Table IV in M&R, the value of t = 7.815; and so, P{} = .95 therefore = 7.815. To calculate , the formula outlined above along with R code was used to determine the value to be 12.162. Finally, comparing , 12.162 > 7.815 therefore, we rejected in favor of . The Bin(051,0.061) model was not an adequate model for our data. The data of the unemp\_rate variable was unfortunately not a good fit of the binomial distribution family.

## Categorical Variable Analysis: race

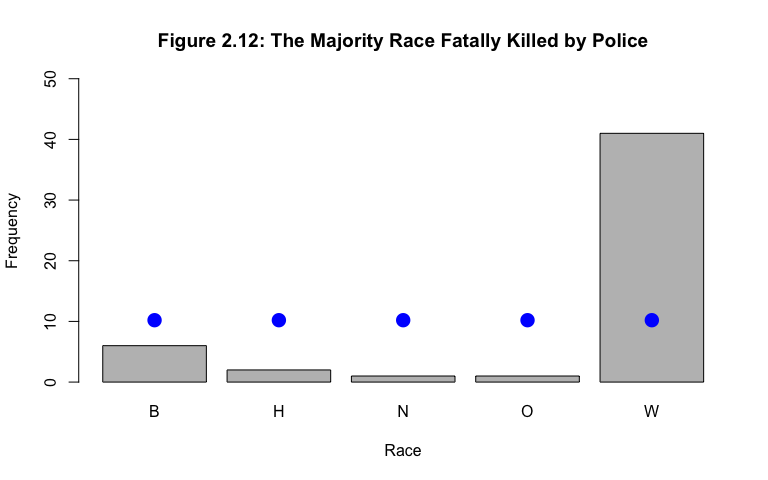
Race has been quite a controversial topic for many, many years. In the analysis of a categorical variable, the race of those fatally shot by the police - there lies an extreme difference between the minority races and those of the majority race.

countsrace <- table(as.factor(maj\_race))   
barplot(countsrace, main = "Figure 2.11: Boxplot of the Majority Race Fatally Killed by Police", xlab = "Race", ylab = "Frequency", ylim = c(0,50))



Since the data was categorical, the only distribution family that would make sense to model this data was discrete uniform. Since the categories do not have a specific order, we were able to re-arrange them in any order we chose and have the same probability for each bin.

bplot3 = barplot(countsrace, main = "Figure 2.12: The Majority Race Fatally Killed by Police", xlab = "Race", ylab = "Frequency", ylim = c(0,50))  
points(x = bplot3, y = c(10.2,10.2,10.2,10.2,10.2), pch=16, col ="Blue",cex =2)



Therefore, in order to overlay the Discrete Uniform(1,5) model, we wanted all 5 bins to have the same probability of falling at a certain spot on the graph. Thus the discrete uniform model had a frequency of 51/5 = 10.2 for each bin. Figure 2.12 portrayed that the minority races are underrepresented (especially Hispanic, Native American and Other) where as the White race was extremely overrepresented. Nonetheless, a GOF HT on the Discrete Uniform (1,5) model would still be conducted as this was the only distribution family that made sense to model the categorical data.

For this GOF test, the typical value of 0.05 was used. In this case, similar to the other GOF tests conducted, = The Discrete Uniform(1,5) model was an adequate model for our data and = The Discrete Uniform(1,5) model was not an adequate model for our data. The Chi-Squared test statistic where was the observed values based on the frequencies of the bins, was the expected value of in this case.

oj = c(6,2,1,1,41)  
ej = c(10.2,10.2,10.2,10.2,10.2)   
calculation\_cat = c()  
for(i in 1:5)  
{  
 calculation\_cat[i] = (oj[i] - ej[i])\*\*2/ej[i]  
}  
chisqtest\_cat = data.frame(oj,ej,calculation\_cat)  
show(chisqtest\_cat)

## oj ej calculation\_cat  
## 1 6 10.2 1.729412  
## 2 2 10.2 6.592157  
## 3 1 10.2 8.298039  
## 4 1 10.2 8.298039  
## 5 41 10.2 93.003922

teststat\_cat = sum(calculation\_cat)

The degrees of freedom = number of bins - number of parameters - 1; in this case df = 5 - 0 - 1 = 4. Therefore, we knew that the distribution of the test statistic ~ for the num\_killed variable. To calculate the critical value of the test statistic, the determination of the rejection region was P{} = 1 - . Using Table IV in M&R, the value of t = 9.488; and so, P{} = .95 therefore = 9.488. To calculate , the formula outlined above along with R code was applied to evaluate the value to be 117.9216. Comparing , 117.9216 > 9.488, we reject in favor of . The Discrete Uniform(1,5) model was not an representative model for our data. The data of the maj\_race variable was extremely unfitting of the discrete uniform distribution family. This was not a surprise as it can be expected that distribution family was usually a poor fit as the probabilities of certain categories were almost never exactly equal.

In conclusion, our variables of choice are 50/50 when it comes to their goodness of fit with a certain probability model. The continuous variable, num\_killed, it fits the exponential distribution well - supported by the shape of the histogram, a QQ plot and a goodness of fit test. The binary variable, body\_camera, was well fitting of the bernoulli distribution family as there are only two possible values for the outcomes and choosing those values carefully resulted in a good fit. On the other hand, the ordinal variable, unemp\_rate, poorly fit the binomial distribution family and likewise, the categorical variable, maj\_race was a poor fit of the discrete uniform distribution family. In all, since statistics is never certain, the judgement calls made, whether well fitting or poor fitting are never the truth of reality.

# Module 3: Single Sample Analysis

As the data was analyzed, it appeared that California was an extreme outlier when compared to all other states for the num\_killed variable, with 501 fatal police shootings. With that, there may be underlying causes/influences to such a high value when compared to other states. With this idea in mind, we divvied into the mean number killed in California vs the US (Is California truly an outlier?), the homicide rate’s connection to the number of fatal police shootings (Is the oddly large number of victims in California conducive to its homicide rate?) and the proportion of body camera requirements throughout the US (Do body cameras really matter?).

## Test 1: Is California truly an outlier?

In order to understand why certain states in the US have higher num\_killed values than others, it was important to look into those outliers themselves and see if the mean of that area specifically was comparable to the sample without the observation of the outlier. For this test, our variable of interest was num\_killed. In the case of this data, California appeared to be the outlier with a value of 501. We treated California as the population for California itself and removed the observation in the sample data when the sample mean and sample variance was computed. To determine the value of , the data for num\_killed variable was first transformed as explained. With that, our value of , = 51.04 and s = 47.5579.

newdata = data[c(5),]  
mean(newdata$num\_killed)

## [1] 501

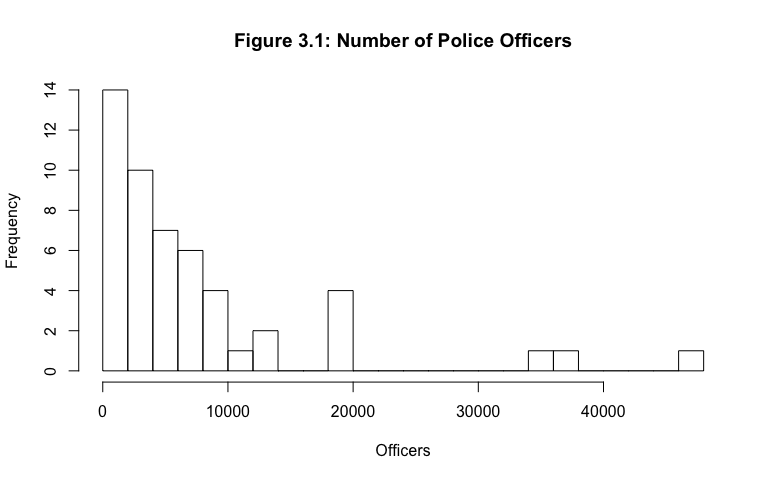
sd(newdata$num\_killed)

## [1] NA

For this hypothesis test, our = 0.05 as the standard alpha value. Before conducting the test, the assumptions were checked. Since this is a test on the population mean with variance unknown but estimated through the data as , a T-test was used. According to Chapter 9 of M&R, there were three assumptions that must hold:

1. The data is normal? It is important to note that the T-test is robust to the normality of the data.

hist(law\_officers, breaks=25, xlab ="Officers", main = "Figure 3.1: Number of Police Officers")

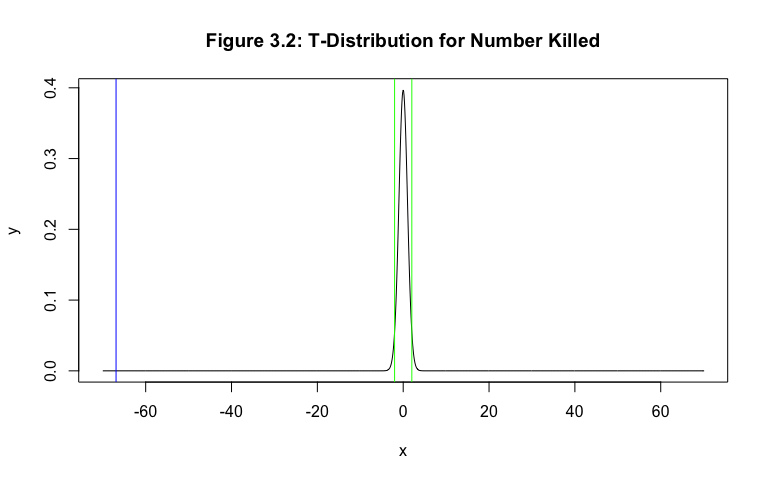


When Figure 3.1 was analyzed, it was clearly not Normal. Even though, this hypothesis test, the T-test, is robust to the normal assumption, we should still work with data that is Normal. However, in order to illustrate our understanding of the test, we continued the analysis and checked if the other 2 assumptions hold.

1. The data is of a random sample? As explained in Module 1, the nature of collecting data in regards to states is always random. Each instance is recorded at a point in time where they all independent of one another. That being said, the data observed about the state of California is and will always be random to the state of Maryland, for example.
2. The data is independent? Similarly to the data being a random sample, each observation does not depend on the other meaning the state of California, once again, is independent of the state of Maryland in terms of all the different variables we collected.

Since assumptions 2 and 3 held true, as mentioned above, we continued the test although we knew the data was not normally distributed as wanted. For the test, the : The true population mean of the number killed in the US is that of California’s () and the : The true population mean of the number killed in the US is not that of California’s ().

x=seq(from=-70,to=70,by=.1)  
y=dt(x,df=50)  
plot(x,y,type= "l", main="Figure 3.2: T-Distribution for Number Killed")  
abline(v=-66.9016,col="blue")  
abline(v=2.0086,col="green")  
abline(v=-2.0086,col="green")



The test statistic . The distribution of the test statistic was identified as: T ~ , more specifically, ~ . Therefore, = -66.9016 (blue line in Figure 3.2) . The value of was found using the invT fucntion of the TI-84 calculator with an area of 0.025, because it was a two-sided test with an alpha of 0.05. Therefore, = +/- 2.0086 (green lines in Figure 3.2) for the right and the left respectively. The graph below illustrated this idea, reinforcing that visually evident that the value of based on the data was in the reject in favor of region. To assure factualness, we determined the p-value of the distribution and compared it to our alpha value. The p-value based on the tcdf function on the TI-84 calculator was 0 which we expected because the area to the left of the data obtained was extremely small. Since 0 < 0.05 we reached the same conclusion.

According to M&R, the power of a hypothesis test is “the probability of rejecting the null hypothesis when the alterantive hypothesis is true.” To calculate its value, the formula 1 - was used where was the probability of a type II error, a missed opportunity of failing to reject in favor of when it should have been rejected. When multiple statistical tests around the same idea/question were completed, power was used to compare the significance of those tests relative to one other. According to the Pacific Standard, the average number of fatal police shootings per year is 487 cases. The difference in our hypothesis mean from that of the source was 14, which is a slight difference in value. So, we used the following test and calculated the power of the test against our alternative hypothesis that the true population mean of the number killed in the US was not that of California’s ().

power.t.test(n=50,delta=501-487,sd=47.5579,sig.level=.05,type="two.sample",alternative="two.sided")

##   
## Two-sample t test power calculation   
##   
## n = 50  
## delta = 14  
## sd = 47.5579  
## sig.level = 0.05  
## power = 0.3076627  
## alternative = two.sided  
##   
## NOTE: n is number in \*each\* group

With the output, the power of the test was 0.376637. With this value, we knew it touched the idea of the sensitivity of the test, with measured differences of a test with a mean of 501 from 487. Therefore, if the true mean of the average number killed in the US by police officers was really 487, then we knew that the test would reject a mean value of 501 and difference will be detected/identified 37.66% of the time. This value was not incredibly too high; however, M&R suggests that the alpha value or sample size can be increased to increase the power as well.

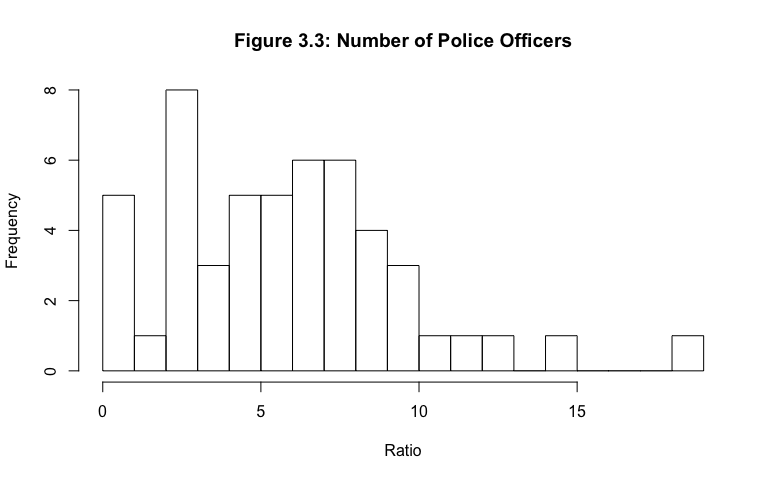
For the original test, in conclusion, we rejected in favor of . The data did not support the claim that the true population mean of the number killed in the US was that of California’s (501). We expected this to occur as there are an ample amount of reasons as to why California could possibly have a larger number killed such as larger population, more police officers, or even more crime related areas. To answer our question initially posed, California was an outlier by looking at the data itself; however, taking much caution in confidently making the statement based upon the hypothesis test. We knew that all of our assumptions did not hold to be true, since the num\_killed variable was not normally distributed. In all, we could not make any recommendations or further investigations about the question.

With that, we considered the possible relationship between number killed and the homicide rate. Would it be expected that more people were killed by police in an area with a higher homicide rate? Or is there another factor that should be considered?

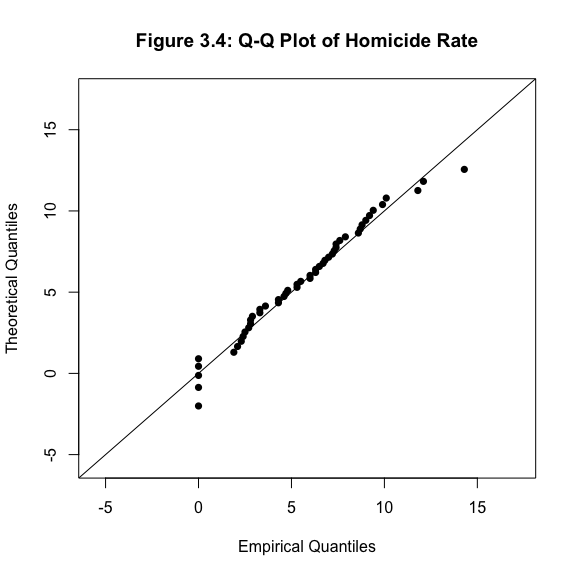
## Test 2: Is the oddly large number of victims in California conducive to its homicide rate?

To continue our investigation on California’s large number of victims shot in retrospect to the other states values, we looked further into the homicide rate both nationally and the population of California itself and compared the two. Homicides may relate to high crime areas, the population itself, socioeconomic deprivation and many other factors. It would be interesting to conclude that the homcide rate in California is greater than or equal to that of the national homicide rate as we could ask further questions relating to the outlier that California is when looking at the number killed - despite the fact that they have a comparable homicide rate to that of the national average.

hist(homicide\_rate, breaks=25, xlab ="Ratio", main = "Figure 3.3: Number of Police Officers")



n = length(homicide\_rate)  
m = mean(homicide\_rate)  
s = sd(homicide\_rate)  
x=homicide\_rate  
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="Figure 3.4: Q-Q Plot of Homicide Rate",  
 xlim=limits, ylim=limits, pch=16)  
abline(0,1)



To do so, we first checked our assumptions for the hypothesis test to make sure that our data was normal, a random sample and independent. As analyzed from the histogram (Figure 3.3), the homicide\_rate variable was moderately normal. To better gauge the normality of the data, a QQ plot (Figure 3.4) was also analyzed. The trend of the points in the plot follow the general trends that a normal distribution should follow. As expected, the large majority of the data was spread through the center of the line, there were less and less points towards the tails and lastly the data points were extremely close, if not laying directly on the line.

Secondly, as explained in the previous hypothesis test, the sample was both random and independent. We proved the random assumption to be true based upon the nature of how data collection was done in states while at the same time assuring that the data collected in Connecticut, for example, was independent of that collected in Georgia. Since all of the assumptions were checked and accurate for the variable of investigation, we can continue to conduct the test.

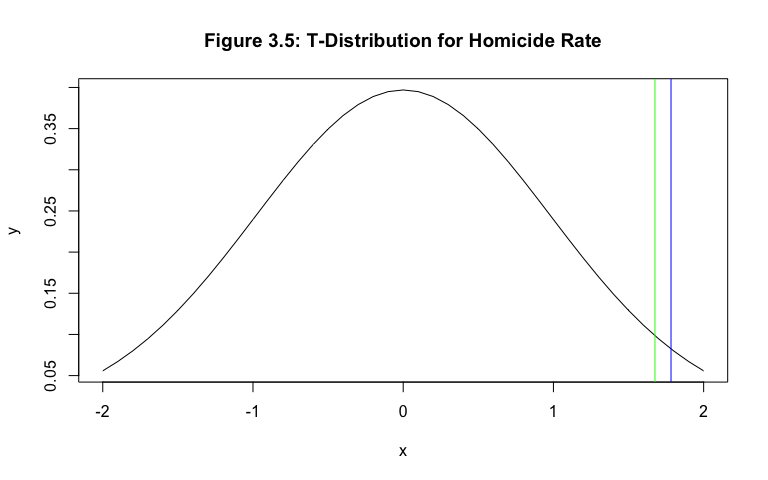
According to the FBI, the national homicide rate in the US during the year 2015 is 4.9, we treated this value as to see if it compared to the odd outlier of our data, California with a homicide rate of 5.3 as . The standard deviation of the test was determined by the standard deviation of the population that California was chosen from so that s = 3.79375. The sample size will still remain the same, s = 51.

For the test, the : The true value of the mean of California was less than or equal to the national average of 4.9 (). : The true value of the mean of California was greater than the national average of 4.9 (). Our alpha value will go across the standard, Our test statstic $T\_{o} = = = .752 $. T ~ ~ .

t.test(homicide\_rate,y=NULL,alternative="greater",4.9,conf.level=0.95)

##   
## One Sample t-test  
##   
## data: homicide\_rate  
## t = 1.7828, df = 50, p-value = 0.04035  
## alternative hypothesis: true mean is greater than 4.9  
## 95 percent confidence interval:  
## 4.956766 Inf  
## sample estimates:  
## mean of x   
## 5.847059

x=seq(from=-2,to=2,by=.1)  
y=dt(x,df=50)  
plot(x,y,type= "l", main="Figure 3.5: T-Distribution for Homicide Rate")  
abline(v=1.7828,col="blue")  
abline(v=1.6759,col="green")



To conduct the test itself, we utilized the one sample T-test from R alongside a 95% confidence interval. The t value, calculated from the test statistic with the data from homicide\_rate = 1.7828 (blue line in Figure 3.5) whereas the critical value, calculated through the invT function of the TI-84 calculator = 1.6759 (green line in Figure 3.5). Illustrated in Figure 3.5, the test statistic, t (the blue line), falls in the reject in favor of region. We concluded that the true value of the mean of California was greater than or equal to the national average of 4.9. That being said, speculation was raised to consider if California has such a high number of victims killed because of its higher rate of crime that occurs in comparison to that of national average. The answer was certainly not that simple as there are many factors that can relate to the number killed in California.

In recent times, the concern regarding the requirement of police wearing body cameras has been of much conversation. Substantial funding had been put into the program, aiming to equip law enforcement officers with cameras to record all of their interactions with the general population. There are both pros and cons that can attribute to requiring one to wear a body camera; however one can argue that they provide a better, unbiased, overall consensus about the situations one may have with a police officer.

## Test 3: Have the proportion of states require their law enforcement officers to have body cameras on them during times of service increased over the past few years?

About a year before the data for our sample was collected, the Bureau of Justice Statistics conducted an investigation to gather statistics about the law enforcement agencies across the United States. They found that about a third (30%) of local law enforcement/police officers used body-worn cameras in 2013. Whether that be the state requirement or not, this means that about 70% of police officers did not wear body cameras on the job. Since there has been much more conversation and advancement in technology allowing body-cameras to be a greater investment for the good of the population, one might expect the proportion of states whose police officers wear body cameras to increase over the span of a few years.

To begin to answer the question, we conducted a hypothesis test on population proportion. Before the analysis, the assumptions of must hold for the data. When performing a test on population proportion, 2 tests can be used: Z Approximation and Exact Binomial Test. According to M&R, for a Z Approximation, the value of p should not be extremely close to 1 or 0 and the sample size of the data must be relatively large. With the data set we analyzed, body\_camera, the sample size is 51 - what we considered to be relatively small. Similarly, the value of p is 0.30 which was not extremely close to 0; however, it was also not in the “middle range” of values between 0 and 1. By these observations, we will not conduct a Z-Approximation test but rather an Exact Binomial Test on the population proportion.

With that, our test concluded one of two statements: : The propotion of states that utilize body cameras is less than or equal to 0.3 () or : The proportion of states that utilize body cameras is greater than 0.3 (). The Exact Binomial Test does not have a test statistic therefore no distribution exists for the test statistic. In order to arise at our conclusion, we compared the alpha value to the p-value obtained by the data. Using the binom.test with x = 5 (the number of successes), n = 51 (the number of trials ), and p = 0.3 (the hypothesized probability of success) in R, we output a p-value of 0.9999.

Since = 0.05 < p-value = 0.9999, we failed to reject in favor of . The data supported the claim that the proportion of states that utilized body cameras was less than or equal to 0.3. With this conclusion at hand, we can certainly raise speculation towards the effects body cameras have. If our sample had more states required the use of body cameras would there be a less number of victims fatally shot and killed? Or more? It is definitely difficult to lean in one direction over the other; however, this hypothesis test does provide leeway into questions we can futher investigate.

binom.test(5,51,0.3,alternative = "greater",conf.level=0.95)

##   
## Exact binomial test  
##   
## data: 5 and 51  
## number of successes = 5, number of trials = 51, p-value = 0.9999  
## alternative hypothesis: true probability of success is greater than 0.3  
## 95 percent confidence interval:  
## 0.03943094 1.00000000  
## sample estimates:  
## probability of success   
## 0.09803922

This module took a dive into hypothesis tests for single samples, paying particular attention to the outlier in the data in regards to the num\_killed variable, California. We hesitantly concluded that California is indeed an outlier, the homicide rate of California was certainly larger than the national average, and our sample was not representative of the proportion of states we’d expect there to be a requirement of body cameras for their law enforcement officers. Considering the first test did not check off all the boxes for the three assumptions, we will not continue our investigation of variables based off of any conclusion made in the test. However, the other two tests certainly did raise questions that are worthy of thought. Perhaps a few of them can be answered when looking at two samples and conducting hypothesis tests on more than just a single sample.

# Module 4: Two Sample Analysis

In this module, we investigated two-samples at a time to answer a few overarching questions. Police shootings have been a topic at large in recent years and the concerns that one may have can certainly be looked at, of course with a smaller knowledge base, through the various hypothesis tests. We first looked into the race of the victims (maj\_race) and the location of the state (geo\_region) they were killed in to perhaps raise discussion of dependence that exists between the two variables. Then, we inspected into two samples of body camera requirements, to investigate if there was a difference in the means of the two. Finally, we compared the variances of two different samples to see how race can potentially effect the number killed in an area.

## Test 1: Does the location of the state and the majority race killed have any sort of relationship that should lead to further investigation? Is a victim of one race more likely to be killed by the police based upon where they reside?

In line with the standard value commonly used in the statistics field, = 0.05 was used for this hypothesis test. The test was designed to show the indepedence or dependence of two samples, X and Y. X assigned to the majority race of the victims per state = maj\_race with values B (Black), H (Hispanic), N (Native American), O (Other), W (White) and Y assigned to the geographical region of where the victim was shot = geo\_region with values 1 (Northeast), 2 (Midwest), 3 (South), 4 (West).

The assumptions of this test that we checked are that each sample in of itself has independent observations. With that, we made sure that the data collected for the maj\_race and geo\_region variable were all independent. As explained in a few of the hypothesis tests in Module 1 as well as Module 3, the nature of collecting data for a group of states is always independent of one another. The data for Georgia for example would have no influence over the data for Florida.

The is that X and Y are independent and the is that X and Y are not independent. To determine the test statistic, we used a table of counts below to correctly identify the counts and marginal probabilities to calculate the test statistic of the data.

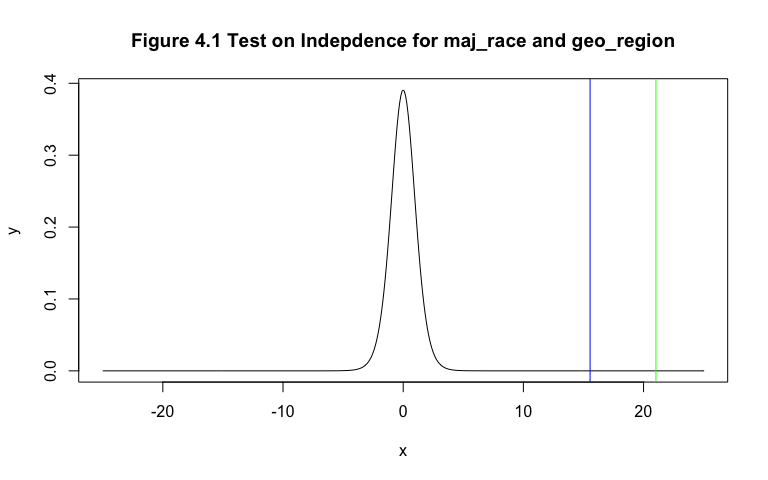
|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | B | H | N | O | W | total | marg prob |
| 1 | 2 | 0 | 0 | 0 | 7 | 9 | 9/51 |
| 2 | 1 | 0 | 1 | 0 | 10 | 12 | 12/51 |
| 3 | 3 | 0 | 0 | 0 | 14 | 17 | 17/51 |
| 4 | 0 | 2 | 0 | 1 | 10 | 13 | 13/51 |
| total | 6 | 2 | 1 | 1 | 41 | 51 |  |
| marg prob | 6/51 | 2/51 | 1/51 | 1/51 | 41/51 |  |  |

The Chi-Squared test statistic . . In particular to this HT, the df = (m - 1)(n - 1) = (5 - 1)(4 - 1) = 4 \* 3 = 12. Therefore, ~ . To determine the region that the data will fall in, the P{} = 1 - = 1. 0.05 = .95 was calculated using the Chi-Squared table in order to determine the critical value of = 21.03.

With the two formulas for the expected value and test statistic identified above, the number to the left in each cell was the expected value and the number to the right was Chi-Squared value for each possible combination in the table below.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | B | H | N | O | W |
| 1 | 1.508 .839 | .352 .352 | .176 .176 | .176 .176 | 7.24 .008 |
| 2 | 1.411 .1197 | .471 .471 | .235 .249 | .235 .235 | 9.63 .013 |
| 3 | 2 .5 | .667 .667 | .333 .333 | .333 .333 | 13.67 13.67 |
| 4 | 1.53 1.53 | .509 4.37 | .225 .225 | .225 2.67 | 10.45 .019 |

x=seq(from=-25,to=25,by=.1)  
y=dt(x,df=12)  
plot(x,y,type= "l", main="Figure 4.1 Test on Indepdence for maj\_race and geo\_region")  
abline(v=15.55,col="blue")  
abline(v=21.03,col="green")



After all the Chi-Squared values in each cell were added together, = 15.5597. Since < = 15.5597 < 21.03 we fail to reject . This conclusion was visually displayed in the graph with the critical value being the green line and the chi-squared value of the data as the blue line.

In all, this data does not support the claim that the majority race of the victim in one state was dependent on the geographical region of that state. There may be other variables that can be tested against majority race to see if there was any sort of independence/dependence relationship between the two. In our second test, we further continued the analysis on num\_killed to see if the difference in means of samples that require body cameras and that do not require body cameras id significant.

## Test 2: How significant are body camera laws to the number killed in the two different populations (states that do require body cameras versus states that don’t require body cameras)?

As proposed in Module 3, one may consider the requirement for law officers to wear and operate a body camera have substantial impact on the number of fatal shootings that occur. On the other side, one may believe that the requirement of body cameras causes more fatal police shootings to occur. Nonetheless, we analyzed the difference in means of num\_killed and broke the sample into two parts: states that require a body camera and the states that do no require a body camera. Although the two sample sizes are quite different, we believe this test can reach conclusion(s) and/or raise important questions of ongoing investigation. In addition, if more states in the US do adopt the requirement of body camera, this test should certainly be conducted once more to see if there is a change or consistency in the conclusion reached.

nocam = data[c(5,7,10,34,41),]  
cam = data[-c(5,7,10,34,41),]  
n\_nocam = 46  
n\_cam = 5  
mean\_noc = mean(nocam$num\_killed)  
mean\_c = mean(cam$num\_killed)  
s\_noc = sd(nocam$num\_killed)  
s\_c = sd(cam$num\_killed)

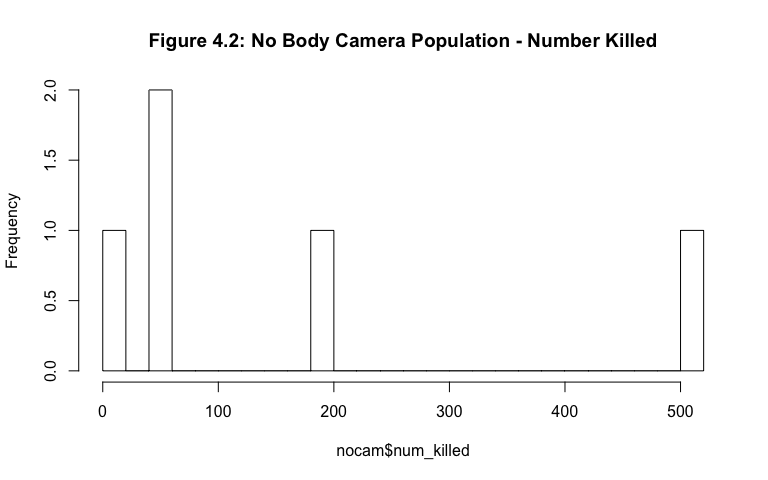
With that, we first split our sample into two components as explained above to gather the information needed for the test sample mean () sample size (n) and sample standard deviation (s). The measurements of variability and scatter within the data of the populations are as below:

|  |  |  |
| --- | --- | --- |
|  | Sample 1 No Body Camera | Sample 2 Body Camera |
| sample size | 46 | 5 |
| sample mean | 159.8 | 49 |
| sample standard deviation | 201.739 | 45.008 |

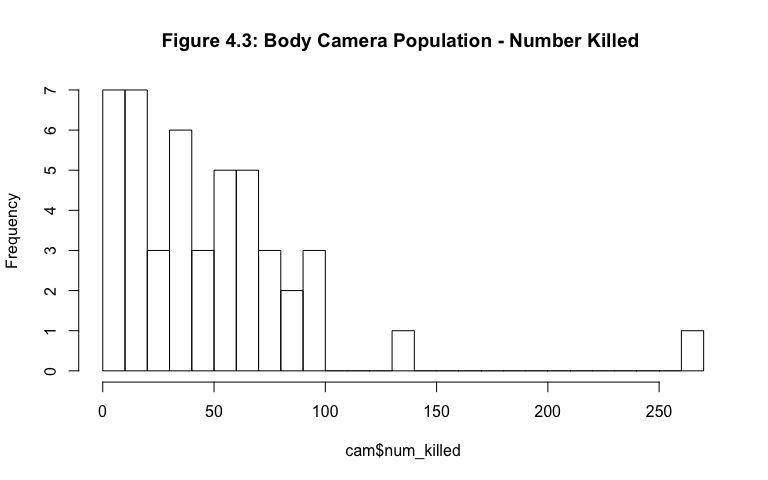
The claim throughout the test will be that body cameras reduce the mean of the number killed by the police. Meaning, the mean with the body camera requirement () will be less than the mean without the body camera requirement (). Using a standard alpha value of = 0.05 and the populations of all states that don’t require cameras versus all states that do require cameras; the for the test will be that the difference in the mean of the states that do require body cameras from the states that don’t require body cameras is greater than or equal to 0 () and the will be that the difference in the mean of the states that do require body cameras from the states that don’t require body cameras is less than 0 ().

Before conducting the test, we checked the assumptions. According to M&R, “the normality assumption is required to develop the test procedure, but moderate departures from normality do not adversely affect the procedure.” To determine how normal the data was, we analyzed the two histograms for both samples: requiring body cameras and not requiring body cameras. It was illustrated clearly in Figure 4.3 that the data of the num\_killed variable for states that do not require body cameras was not normal while the data for the num\_killed variable for states that do require body cameras appeared to be exponential in Figure 4.3. Despite the lack of normal distribution representation, the analysis continued to illustrate our understanding of the particular hypothesis test. The other two assumptions that we checked as in the other t-test’s are random sample and independent. As explained in the previous module based upon the nature that the data was collected within each state, we knew that these two assumptions also hold for this test.

hist(nocam$num\_killed,breaks = 25,main ="Figure 4.2: No Body Camera Population - Number Killed")



hist(cam$num\_killed, breaks = 25,main = "Figure 4.3: Body Camera Population - Number Killed ")

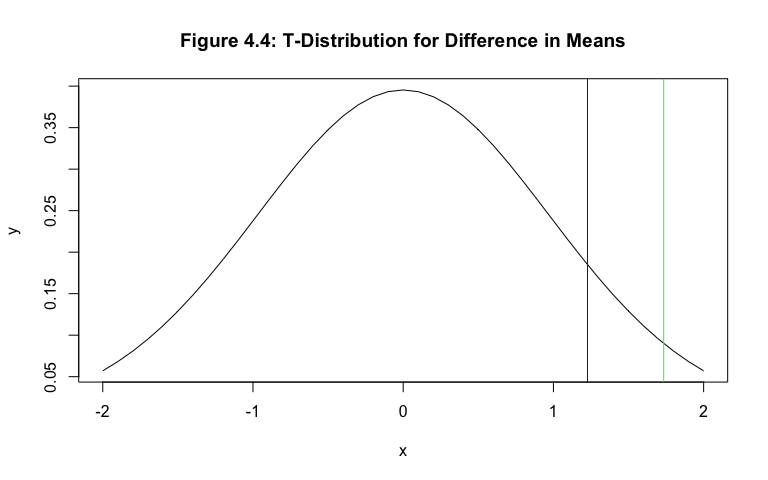


Although aware that two out of the three of the assumptions hold, we continued on to compute the test statistic and find its corresponding distribution. Since we knew that the variances are not equal since the two samples have different sample standard deviations and population variances unknown, the test statistic of will be use with a distribution of ~ where v is calculated by rounded down to the nearest integer. For this test, we calculated a value of v (degrees of freedom) to be 28.47533 which rounds down to 28. Therefore, the test statistic ~ .

t.test(nocam$num\_killed,cam$num\_killed,alternative = "less",var.equal = FALSE,conf.level = 0.95)

##   
## Welch Two Sample t-test  
##   
## data: nocam$num\_killed and cam$num\_killed  
## t = 1.2248, df = 4.0434, p-value = 0.8564  
## alternative hypothesis: true difference in means is less than 0  
## 95 percent confidence interval:  
## -Inf 303.0628  
## sample estimates:  
## mean of x mean of y   
## 159.8 49.0

x=seq(from=-2,to=2,by=.1)  
y=dt(x,df=28)  
plot(x,y,type= "l", main="Figure 4.4: T-Distribution for Difference in Means")  
abline(v=1.2268,col="blue")  
abline(v=1.734,col="green")



As outlined in the t.test function in R for the difference in means, sample unknown; assumed unequal unpaired t-test, the test statistic based on the data was equal to 1.2248. The critical value, found by the invT function of the TI-84 calculator with an area of 0.95 and degrees of freedom of 28 was 1.734. Since t = 1.2248 (the blue line in the Figure 4.4) < = 1.734 (the green line in Figure 4.4) we failed to reject in favor of . The data did not support the claim that the difference in the mean of the states that do require body cameras from the states that don’t require body cameras is greater than or equal to 0. We are not lead to believe that the use of body cameras reduces the number of fatal police shootings that occur. If the data was normally distributed and there were a large sample size of those states that do require body cameras, we would potentially reach a different conclusion.

## Test 3: How different is the number killed by police officers between states where minority races are killed versus states where the majority race is killed?

Race has been a topic of discussion for decades and although some areas of US had subsided the racial issues that have occurred in the past, there are certainly some areas where the concern has either increased or remained the same. For this test, we desired to see the difference in the num\_killed of states where the majority race killed in the state was of the minority races versus where it was the majority races. The majority race was considered White and the minority races were Black, Hispanic, Native American, Asian and Other. The decision was made to split the data into two samples this way according to the US Census, the majority race of the US population is White at 79% with other races at lower percentages.

To better answer the question, we conducted a test on variance to see how much variability exists within the states of the two samples. The test statistic used for this test was the F distribution. The distribution is comprised of the division of two Chi-Squared distributions. With that, we know there are 3 assumptions that must hold for the Chi-Squared test. The data is normally distributed. As indicated in the previous module we know that the data for the num\_killed variable was not normally distributed; however, it was both independent and of a random sample. For the purpose of the test, we will continue on but not make any strong conclusions based on the test.

majority <- subset(data, maj\_race == "W")  
min1 <- subset(data, maj\_race =="B")  
min2 <- subset(data, maj\_race =="H")  
min3 <- subset(data, maj\_race =="N")  
min4 <- subset(data, maj\_race =="O")  
min5 <- subset(data, maj\_race =="A")  
minority <- rbind(min1,min2,min3,min4,min5)  
num = sd(majority$num\_killed)  
denom = sd(minority$num\_killed)

Since we tested how different the two variances were, the test will be two sided. The sample size and sample standard deviation for each sample must be known.

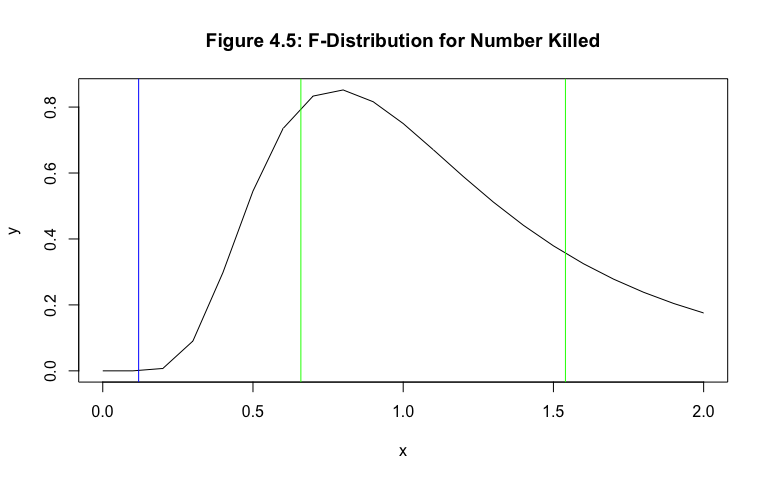
|  |  |  |
| --- | --- | --- |
|  | Sample 1 Maj (Majority) | Sample 2 Min (Minority) |
| n | 41 | 10 |
| s | 51.06618 | 147.7411 |

The for the test will be that meaning that the variability between states of the number fatally killed by police was the same despite the racial differences. The will be that meaning that the variability between states of the number fatally killed by police was not the same despite the racial differences.

var.test(majority$num\_killed,minority$num\_killed,alternative = "two.sided",conf.level = 0.95)

##   
## F test to compare two variances  
##   
## data: majority$num\_killed and minority$num\_killed  
## F = 0.11947, num df = 40, denom df = 9, p-value = 1.391e-06  
## alternative hypothesis: true ratio of variances is not equal to 1  
## 95 percent confidence interval:  
## 0.03408139 0.29293671  
## sample estimates:  
## ratio of variances   
## 0.1194714

x=seq(from=0,to=2,by=.1)  
y=df(x,df1=40,df2=9)  
plot(x,y,type= "l", main="Figure 4.5: F-Distribution for Number Killed")  
abline(v=0.11947,col="blue")  
abline(v=0.6594,col="green")  
abline(v=1.54,col="green")



With an value of 0.05 as standardly used, the test statistic for the test, was distributed f with degrees of freedom n - 1 for both samples. ~ . The test statistic, = the squared standard deviations divided by one another, if the null hypothesis were true, our statistic would be equal to 1. We compare F (from the data) equal to = = 0.11947 (the blue line in Figure 4.5) to the critical values from the F chart in M&R. Since our alpha value was 0.05, we looked for the value that corresponds to an area of 0.025 on each side. With that, the critical value to the right was 1.54 (the green line in Figure 4.5) to the right with degrees of freedom 40 for the numerator and 9 for the denominator. The critical value to the left was found by with the same degrees of freedom of 40 and 9: 0.6594 (the green line in Figure 4.5). Since the test statistic value of 0.11947 was less than 0.6594 and not in the range of values between the two critical values, we rejected in favor of . The data did no support the claim that the varability between states of the number fatally killed by police was the same despite the racial differences. As one may expect, the variation of the number killed by police between states was not the same if the majority race killed was White or if it were Black, Hispanic, or any other minority race.

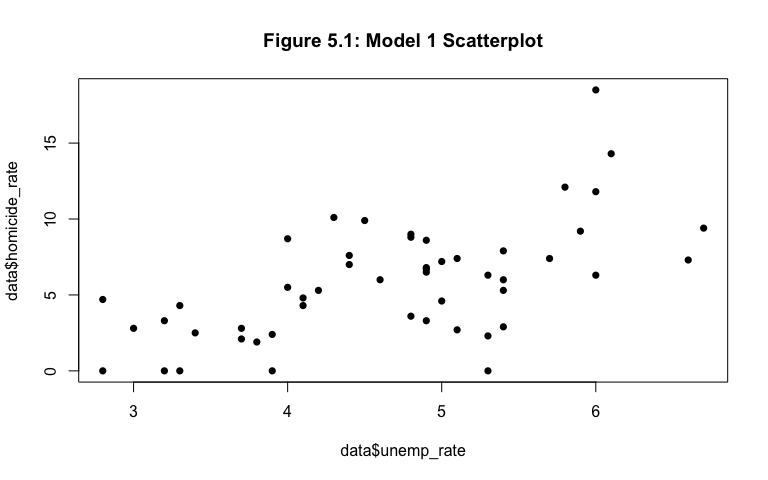
Throughout this module we investigated various tests to attempt to answer or even raise a few questions. We took a deeper look into the num\_killed variable, analyzing different speculations when splitting the observations into two different samples. In test 1, we concluded that the majority race of the victims killed was independent of the geographical location of the state. Test 2 raised the observation (not a conclusion because all the assumptions did not hold) that the use of body cameras may not reduce the number of fatal police shootings. Finally in test 3, although the data was not normally distributed once again, we arrived at a potential conclusion that there exists a difference in the variation of states depending on which race was the majority race killed in that state.

After conducting the six different hypothesis tests between Module 3 and 4, there are a few questions left unanswered that our next module - linear regression in Module 5 can provide a deeper investigation on the relationship of two variables. In line with the second hypothesis test in Module 3, we will use linear regression to determine the existence or nonexistence of a linear relationship between homicide rate and number killed per each state.

# Module 5: Prediction Using Linear Regression

One of our larger questions that was difficult to answer based upon assumptions, but touched was upon in our hypothesis testing, regarded the relationship that potentially existed between the homicide rate and the unemployment rate of a particular state. According to The Park Place Economist, there was “one basic notion that in order for an individual to maintain a certain standard of living during a period of unemployment he/she will become more likely to commit a criminal act.” The increase in criminal acts could potentially cause more citizen to police officer interaction which again can then lead to more fatal police shootings depending on the situation at hand. With that, we explored the linear relationship between the unemployment rate and homicide rate for the data collected.

plot(data$unemp\_rate,data$homicide\_rate,pch=16,main="Figure 5.1: Model 1 Scatterplot")



cor(data$unemp\_rate,data$homicide\_rate)

## [1] 0.6049735

Since a linear regression model investigated relationships between variables, we first analyzed the scatterplot of the unemployment rates on the x-axis the predictor and the homicide rate on the y-axis as the response in Figure 5.1. The scatterplot exhibits a linear, positive trend between the two variables. With that, we continued the model as a singular linear regression model. The determined correlation coefficient between the two variables was 0.6049735, a moderately strong correlation value. Considering the points are not completely fit to a perfectly linear trend, we expected such a value. The model used was homicide\_rate = + \* unemp\_rate where and both have minimized squared residuals.

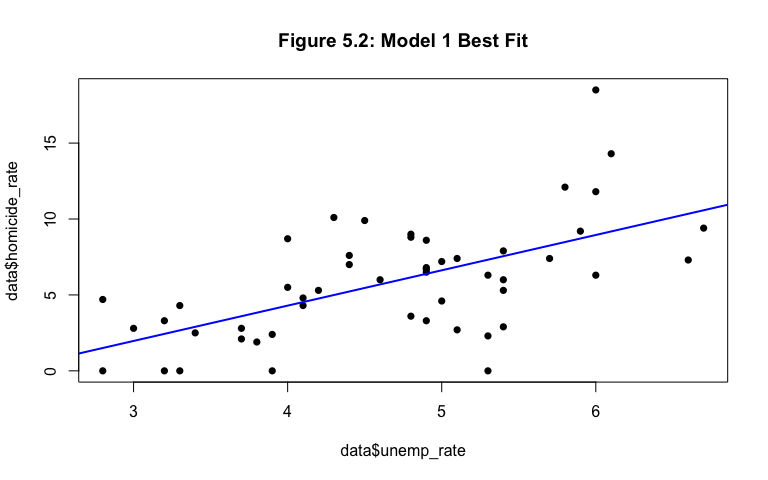
The linear model, called through the linear model function in R as lm(formula = hr~ur) with hr being the homicide rates and ur being the unemployment rates was summarized with the following output:

hr = data$homicide\_rate   
ur = data$unemp\_rate  
model1 = lm(formula = hr~ur)  
summary(model1)

##   
## Call:  
## lm(formula = hr ~ ur)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -7.3206 -1.9766 0.2714 1.7031 9.5507   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -5.0107 2.0858 -2.402 0.0201 \*   
## ur 2.3267 0.4375 5.318 2.57e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.051 on 49 degrees of freedom  
## Multiple R-squared: 0.366, Adjusted R-squared: 0.3531   
## F-statistic: 28.29 on 1 and 49 DF, p-value: 2.573e-06

There are certain parts of the output that must be analyzed to ensure that the model was in fact a good model or if changes need to be made. The values of the beta’s correspond to the coefficient estimates for both the intercept and the ur (unemployment rate). The model, with estimates: homicide\_rate = -5.0107 + 2.3267 \* unemp\_rate. The homicide rate for a particular state increased by 2.3267 percent for each additional percentage its respective unemployment rate increased. The positive slope of the formula holds true to our expectations based upon the scatterplot. When a state has an unemployment rate of 0, the homicide rate would be -5.0107, which ultimately means the homicide rate would practically be 0(negligible) as well since rates are usually never negative values.

plot(data$unemp\_rate,data$homicide\_rate,pch=16,main = "Figure 5.2: Model 1 Best Fit")  
abline(model1, col="blue", lwd=2)



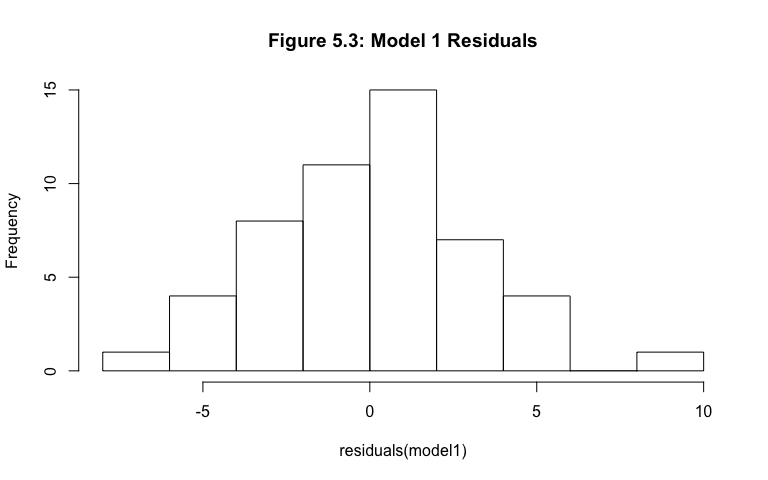
When we overlaid the scatterplot with its linear relationship model in Figure 5.2, the trends of a positive relationship still hold. There were points that perfectly fell on the line, others that were a bit far off, and one that raised a bit of concern. The point far above the blue line on the graph could potentially be an influential point. The point would make the model a better predictor if removed. In this case, the point towards the top right of the graph happened to be the District of Columbia with a homicide rate of 18.5 and an unemployment rate of 16. A new model can be compared to the first model when the point was removed - if there was legitimate reason as to why the data point didn’t fit will. When DC was compared to the rest of the observations, we knew that the other 50 observational units were actually states of the US whereas DC was the capital of US. With that, we removed its point from the data set and ran a new model, after we first analyzed model 1 with the point included.

With our knowledge of hypothesis testing, we tested the slope, of the line, to see if the linear model was better at predicting than the constant model. First, we created a 95% confidence interval for our estimate value of the slope, . The interval, found by, CI + critical value \* standard error. The critical value found through the invT function of the TI-84 calculator with parameters 0.025 (the alpha value divided by 2 for a two sided test) and n - 2 = 51 - 2 = 49 degrees of freedom (indicated on the last line of the model summary), was 2.0096. The 95% confidence interval 2.3267 2.0096 \*(0.4375) = [1.4475,3.2059]. If we were to conduct this experiment 1000’s of times and create a confidence interval under the same experiment each time, 95% of those intervals would contain the true value of the slope . Under our claim for the hypothesis test, the would be that = 0 and the would be that . With a standard alpha value of 0.05, we compared the p-value of the model to the alpha value to make the conclusion. Since 0.05 > 2.57e-06. We rejected in favor of . The data does not support the hypothesis that the slope of the model was equal to 0. The slope of this model was significant.

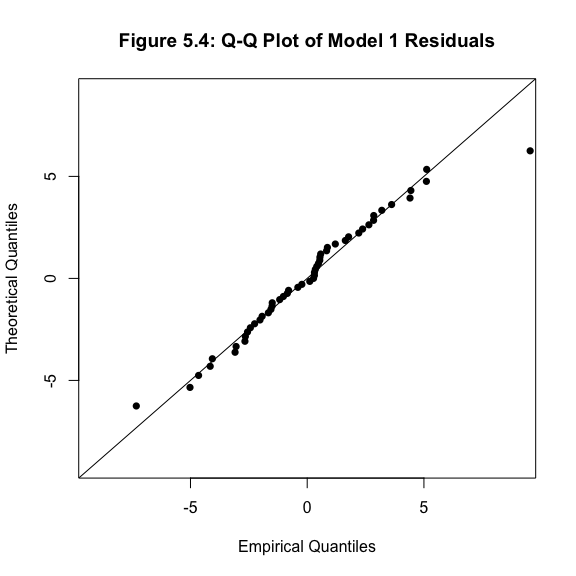
The residual standard error for the model was 3.051. A good model would hold a small residual standard error value. Relative to the unit of a homicide rate, ratio of homicides in the area relative to the population we expected to have a low residual standard error. The adjusted R-squared value reported that 35.21% of the variation in homicide rate was accounted for by the linear model. Relative to what a “good model” would have, our percentage of variation was within the range of the percentages that did not raise any flags.

Since the calculations for the first linear model have been checked and they were both reasonable and ideal values for a good model, we checked the assumptions of the linear model to make sure that they too held. Linear regression differs from hypothesis testing in many ways, but in terms of procedure, linear regression does the math first then checks the assumptions where as hypothesis testing is the other way around. There are three assumptions for a linear model that must hold to ensure that the model is adequate - if they do not hold, there are steps one can take to run a new model. In our case, we can decided to remove the influential point of DC, run a multiple linear regression model, or pair our predictor variable of unemployment rate with another response variable.

hist(residuals(model1),main = "Figure 5.3: Model 1 Residuals")

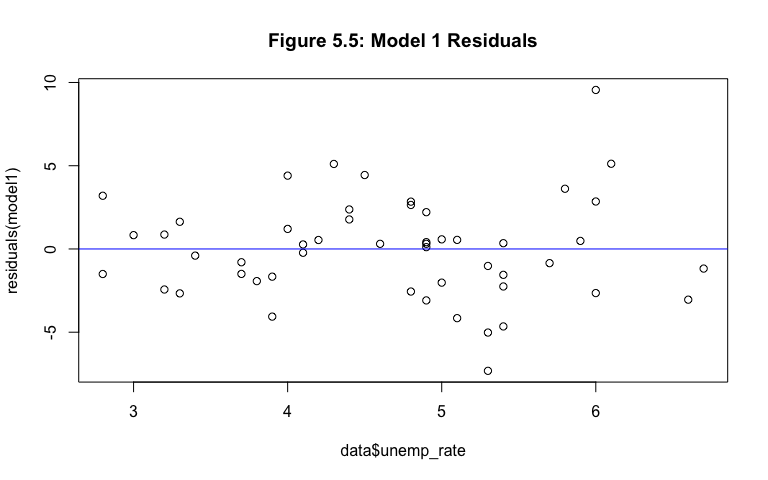


n = length(residuals(model1))  
m = mean(residuals(model1))  
s = sd(residuals(model1))  
x=residuals(model1)  
limits = c(m-3\*s,m+3\*s)  
  
probs = (1:n)/(n+1)  
norm.quants = qnorm(probs,m,s)  
plot(sort(x),sort(norm.quants),ylab="Theoretical Quantiles",   
 xlab="Empirical Quantiles", main="Figure 5.4: Q-Q Plot of Model 1 Residuals",  
 xlim=limits, ylim=limits, pch=16)  
abline(0,1)



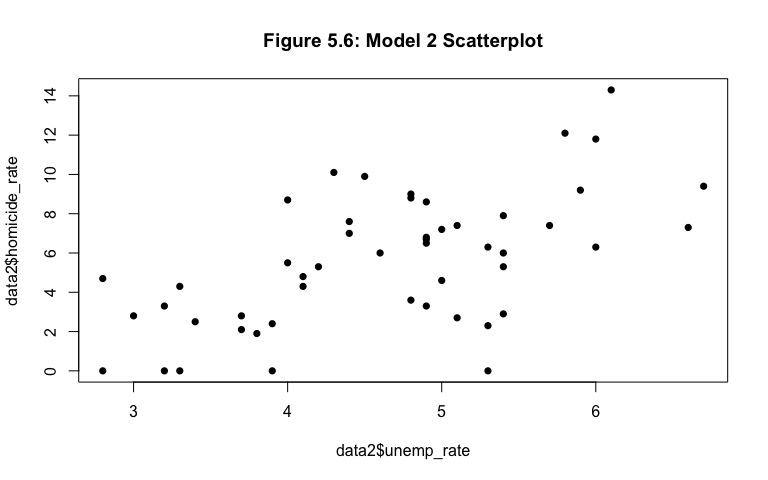
The first assumption was that the residuals are normally distributed. In Figure 5.3 the shape was unimodal, more points are towards the center, fewer towards the tails and it was roughly symmetric. To further ensure the normality of the residuals we analyzed a QQ plot in Figure 5.4 as well which carried the same trends, there are more points clustered towards the center, fewer towards the tails and the vast majority of points are located very close (or exactly on) the line.

plot(data$unemp\_rate,residuals(model1),main = "Figure 5.5: Model 1 Residuals")  
abline(0,0,col="blue")



The second assumption was that the distribution of the residuals was normal with a mean of 0. This assumption always holds true for any model because of the design of the situation. The final assumption was that the second parameter of the distribution, the variance, was constant across all residuals. To check this assumption, we created a plot of the residuals to identify if there was a pattern in the scatter of the points. Since our model did not have any pattern in Figure 5.5, we were confident in saying that a constant variance existed across all residuals. With that, all 3 assumptions hold for this first model. Nonetheless, we still removed DC’s observation from the data and ran a new model again to see if it was better fitting due to its influence.

data2 = newdata = data[-c(8),]  
plot(data2$unemp\_rate,data2$homicide\_rate,pch=16,main="Figure 5.6: Model 2 Scatterplot")



cor(data2$unemp\_rate,data2$homicide\_rate)

## [1] 0.5946495

hr2 = data2$homicide\_rate   
ur2 = data2$unemp\_rate  
model2 = lm(formula = hr2~ur2)  
summary(model2)

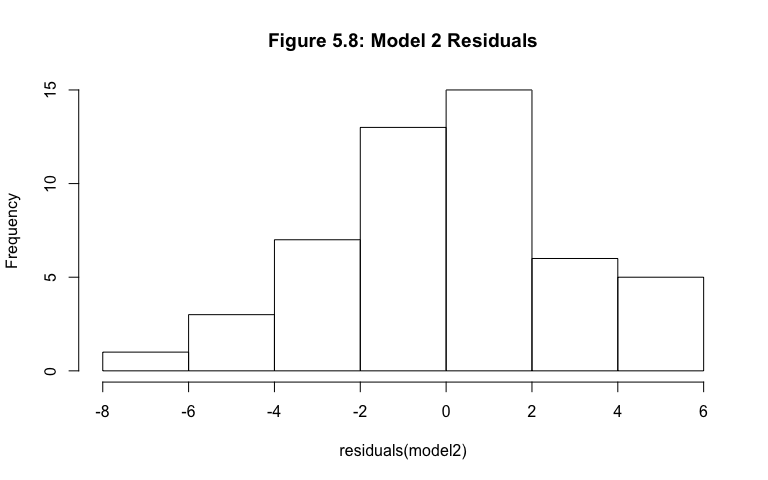
##   
## Call:  
## lm(formula = hr2 ~ ur2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -6.9466 -1.8444 0.3429 1.3935 5.7139   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -3.9150 1.8956 -2.065 0.0443 \*   
## ur2 2.0494 0.3999 5.124 5.27e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.737 on 48 degrees of freedom  
## Multiple R-squared: 0.3536, Adjusted R-squared: 0.3401   
## F-statistic: 26.26 on 1 and 48 DF, p-value: 5.273e-06

With the new data, after DC was removed, the scatter plot in Figure 5.6 still held the same trends in comparison to the first model. There was a positive, linear relationship between the unemployment rate and homicide rate. However, the correlation between the two variables decreased from 0.6049735 to 0.5946495. Although the difference wasn’t tremendously large, the correlation between the two variables weakened a bit. We still continued the model, to see the differences in its summary. The same model, homicide\_rate = + \* unemp\_rate was used, with new estimated values for the beta’s. The updated formula for model 2 was homicide\_rate = -3.9150 + 2.0494 \* unemp\_rate. The intercept increased a bit and the slope became a bit steeper as it increased as well.

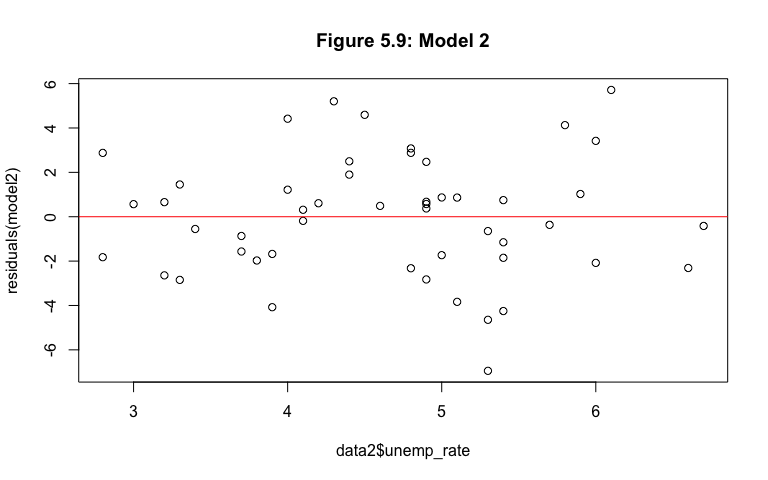
plot(data2$unemp\_rate,data2$homicide\_rate,pch=16,main = "Figure 5.7: Model 2 Best Fit")  
abline(model2, col="red", lwd=2)



hist(residuals(model2),main = "Figure 5.8: Model 2 Residuals")



plot(data2$unemp\_rate,residuals(model2),main = "Figure 5.9: Model 2")  
abline(0,0,col="red")



When the linear model was overlaid on the scatterplot in Figure 5.7, the similar trends hold: positive relationship, points close to the line, other points further away but not far enough to raise that much concern. The residual standard error for this second model was 2.737. The value decreased from the first model, one reason to advocate as to why this model was better than the first. The adjusted R-squared value decreased to 0.3401 meaning 34.01% of the variation in homicide rate was accounted for by the second linear model. This value was still within the range of the ideal values for what makes a model good. To check the three assumptions for this model: normality and constant variance (since the mean of 0 always holds), by analyzing the histogram and scatterplot of the residuals of the second model below, we can say that for the same reasons explained in the first model, all 3 assumptions hold. The difference in this histogram in Figure 5.8 compared to the first one would be that the center was shifted a bit to the right; however, it was still unimodal with most of the data spread in the center. Lastly, the residuals in Figure 5.9 did not have a clear pattern to indicate that the variance was not constant.

In conclusion, we believed that although both models were adequate in their linear relationship model between unemployment rate and homicide rate, we leaned toward the second model being the better of the two. Especially since its residual standard error value was smaller than the first, it can be considered the better model. As explored in various different research projects, a relationship exists between the two variables explored in this module. To analyze a more complex model, we conducted a multiple linear regression model to see what other variables respond to the predictor variable of unemployment rate. Out of the 11 variables in this project, it was not directly visible which can have a linear relationship with the unemployment rate, so the best method of multiple linear regression would be backwards elimination. We could have multiple different variables that we include in the model such as num\_killed, med\_income, law\_officers and avg\_age and run multiple models, removing the variable with the largest p-value until everything leftover in the model was useful. It was particularly interesting to see how two variables that related to such different ideals can be related in such a significant way that should be included in more discussion of police shootings. The concern of police shootings has run far and wide, and we believe that topic should remain one of many that statisticians continue to analyze.

# Module 6: Summary and Conclusion

For many years, police shootings have been a topic of discussion in many areas across the US. For some states, the rate of fatalities of its citizens due to police officer interaction is incredbily high, where as in others, they have as low as three shootings over the span of nearly three years. Across different races but majority the sex of men, our 51 observations outlined variables from number killed to unemployment rate to homicide rate to the arms that the victim was murdered with. Throughout the project, we took the time to dive into our major question: what circumstances are present where the probability of being fatally shot by a police officer is significantly increased? Although the answer was not directly found through the different statistical tests and analyses conducted, we were able to draw conclusions to shed let onto a new path of ideas to further investigate.

In Module 1, we identified a few underlying questions that we investigated throughout the other four modules. In order to set the layout for the project ahead, we clarified a few details about our data such as the way it was collected, how it was cleaned and a case by case look at the variables we would be working with. In modules 3 and 4 we worked with analyses of both single sample and two sample, respectively. We weren’t able to make as many firm conclusions for our first test regarding the number killed in California and its potential to be an outlier because the data was not normally distributed as the test calls for. However, in our other two tests, regarding homicide rate the proportion of states that require their law officers to carry and make use of body cameras, we concluded one of two spanning and mutually exclusive statements to further the investigation. In module 4, we worked with two samples of our data, broken up in various ways. This module also had two tests where we were unable to make firm and strict conclusions, because of the lack of normality of the data. The first test looked into the idea of the independence or dependence relationship between the location of the state and the majority killed in the respective state. The second and third tests were unable to make any firm conclusions because of the missing assumption as mentioned above. Overall, we were able to use our knowledge of hypothesis testing across a large scale to analyze and interpret different questions regarding fatal police shootings in the United States.

Finally, in module 5, we analyzed a linear regression model between what we believed to be two interesting variables that may have an indirect relationship to number killed (a variable we looked quite in depth throughout the project), homicide rate and unemployment rate. Although the two may not be directly related to the number killed in each state in a linear fashion, we hypothesized that a higher unemployment rate would lead to a higher homicide rate because there are acts that one is more likely to commit when they need to survive as unemployed. The rise in unemployment rate will lead to more police and citizen interaction which could potentially increase the number of fatal shootings that police find themselves acting upon. We were pleasantly able to confirm our hypothesis in the linear regression model found, finding a correlation to exist between the homicide rate and unemployment rate while at the same time running what we would consider a good model overall (without the area of DC included in the observations tested due to its influential nature).

Although police shootings are a topic of discussion on its own, homicides (the act of killing one person) has many different layers to it, one of them being police shootings. To continue on the investigation of this report, we can potentially look into individual situations of the people killed and determine more specific conclusions based upon in case instead of the state as a whole. When beginning this project, the data we first found, before transforming it, looked into many categorical variables in regards to fatal police shootings. There are certainly different tests and analyses that can be done on this type of data that differ from the tests we conducted on our data (primarily comprised of continuous variables). We believe that the combination of analyses on our data as well as the data of individual cases can lead to stronger conclusions made as well as more direction and depth into the realm of fatal police shootings.

detach(data)