Statistical Analysis of the Unemployment Rate of the 50 United States in 2016

Karsten Cook & Makayla Underwood

April 18, 2018

Table of Contents

Module 1: Introduction and Data Collection

Introduction

The Great Depression of the 1920s and subsequent economic recessions, like the one experienced in 2008, have been historically difficult times for the United States and its people. During these economic troughs, the negative effect can be felt far and wide. Rising unemployment results in a loss of income and employment for many individuals. The government faces increased financial stress as tax revenue plummets and social welfare spending skyrockets. Spending power is decreased dramatically, businesses become less profitable, and families struggle to make ends meet. It is not difficult to imagine the consequences of poor economic health. These consequences, just to name a few, highlight the importance of maintaining a healthy economy which helps to minimize the subtle but long-lasting repercussions of high unemployment.

Even at its best, an economy will experience a non-zero, natural level of unemployment as individuals move between jobs within a flexible labor market. A person is considered unemployed if he or she is does not currently have a job, has been actively searching for work within the last four weeks, is currently available for work, and above the age of 16. This strict definition of unemployment allows us to get an accurate picture of the current workforce and ultimately the state of the economy. The unemployment rate is ubiquitously used as an indicator of an economy's health, and is the measure we will choose to investigate throughout this project.

In this report, we will use data from 2016 that relates to the unemployment rate for each of the 50 United States. The purpose of examining this data is to answer a few of the following overarching questions through statistical analysis.

-What are the largest contributors to unemployment on a state-level?

-Is the United States' economy healthier than the rest of the world?

-Does having a high minimum wage equate to high unemployment rate in a state?

-Does political majority have an effect on a state's unemployment rate?

-Can we use the high school diploma rate to predict the unemployment rate of a state?

Data Acquisition and Cleaning

Our data was collected from government databases including the Census Bureau, Department of Labor, and the Bureau of Labor Statistics. We gathered our welfare spending reports from a U.S. debt tracking online database. The remainder of our data we gathered from Wikipedia and their cited sources. After collecting our data for the year 2016, we compiled it into a single data set in order to compare the relationships between our variables.

Additionally, we removed the District of Colombia and any U.S. territories from all data sets, as we only wanted to investigate the fifty states of the United States of America.

We defined our observational unit to be each of the fifty United States. We gathered data specific to an individual state in the year 2016 and created our spreadsheet accordingly. In total, we have fifty observations and eight variables. The variables and their information are in the table below:

Variable Table

```
#using ktable to create variable table

text_tbl <- data.frame(
  Variable = c("Unemployment Rate", "Minimum Wage", "Median Income", 'College
Diploma Holders', 'High School Diploma Holders', 'State Region', 'Political M
ajority', 'Welfare Spending'),
  Description = c('The percent of the labor force that is without gainful emp
loyment.', 'The state mandated lowest amount an employer may legally pay thei
r employees per hour', 'Median income of state population', '% of state popul
ation with college diploma', '% of state population with high school diploma'
, 'Classified as one of the Census Bureau designated regions.', 'The state ma
jority political party affiliation evaluated by looking at the last 5 preside
ntial election voting results for each state, and determining whether the maj
ority was Democratic or Republican', '% of total state GDP spent on welfare b
roken down into 8 equally indexed categories'),
  R_code = c('un_rate', 'min_wage', 'med_inc', 'dip_college', 'dip_hs', 'st_r
eg', 'pol_maj', 'welf_spend'),
  Type = c('Continuous','Continuous', 'Continuous', 'Continuous', 'Continuous
', 'Nominal', 'Binary', 'Ordinal'),
  Range = c('0%-100%', '$5.15-$10.00 $/hr', '$41,754.00- $78,945.00 $/yr', '0
%-100%', '0%-100%', 'Northeast, Midwest, South, West', 'Democratic, Republica
n', '0.4-0.6%, 0.6-0.8%, 0.8-1.0%, 1.0-1.2%, 1.2-1.4%, 1.4-1.6%, 1.6-1.8%, 1.
```

```
8-2.0%'),
  Measurement = c('% of Labor Force', 'U.S. Dollars', 'U.S. Dollars', '% of p
opulation', '% of population', 'N/A', 'N/A', 'N/A')
)

kable(x=text_tbl) %>%
  kable_styling(full_width = F, bootstrap_options = "striped") %>%
  column_spec(1, bold = T, border_right = T) %>%
  column_spec(2, border_right = T) %>%
  column_spec(3, border_right = T) %>%
  column_spec(4, border_right = T) %>%
  column_spec(5, border_right = T) %>%
  column_spec(6, border_right = T) %>%
  column_spec(2, width = "30em")
```

Variable	Description	R_code	Type	Range	Measurement
Unemployment Rate	The percent of the labor force that is without gainful employment.	un_rate	Continuous	0%-100%	% of Labor Force
Minimum Wage	The state mandated lowest amount an employer may legally pay their employees per hour	min_wage	Continuous	$5.15-$10.00 $/hr	U.S. Dollars
Median Income	Median income of state population	med_inc	Continuous	$41,754.00-$78,945.00 $/yr	U.S. Dollars
College Diploma Holders	% of state population with college diploma	dip_college	Continuous	0%-100%	% of population
High School Diploma Holders	% of state population with high school diploma	dip_hs	Continuous	0%-100%	% of population
State Region	Classified as one of the Census Bureau designated regions.	st_reg	Nominal	Northeast, Midwest, South, West	N/A
Political Majority	The state majority political party affiliation evaluated by looking at the last 5 presidential election voting results for each state, and determining whether the majority was Democratic or Republican	pol_maj	Binary	Democratic, Republican	N/A
Welfare Spending	% of total state GDP spent on welfare broken down into 8 equally indexed categories	welf_spend	Ordinal	0.4-0.6%, 0.6-0.8%, 0.8-1.0%, 1.0-1.2%, 1.2-1.4%, 1.4-1.6%, 1.6-1.8%, 1.8-2.0%	N/A

In the following modules, we will begin answering these questions using a number of statistical analysis tools, including one and two sample hypothesis tests, linear regression modeling, and more.

Module 2: Modeling Data with Probability distributions

In this module, we will examine four different types of variables within our data set: dip_college (continuous), welf_spent (manipulated to be ordinal), pol_maj (binary), and st_reg (categorical nominal). We will take each variable and summarize its characteristics numerically and graphically and then model it with a specific probability distribution, while estimating parameters as needed. We will perform various hypothesis tests on each variable to either reject or fail to reject our belief in the strength of the proposed probability model's representation of our sample data.

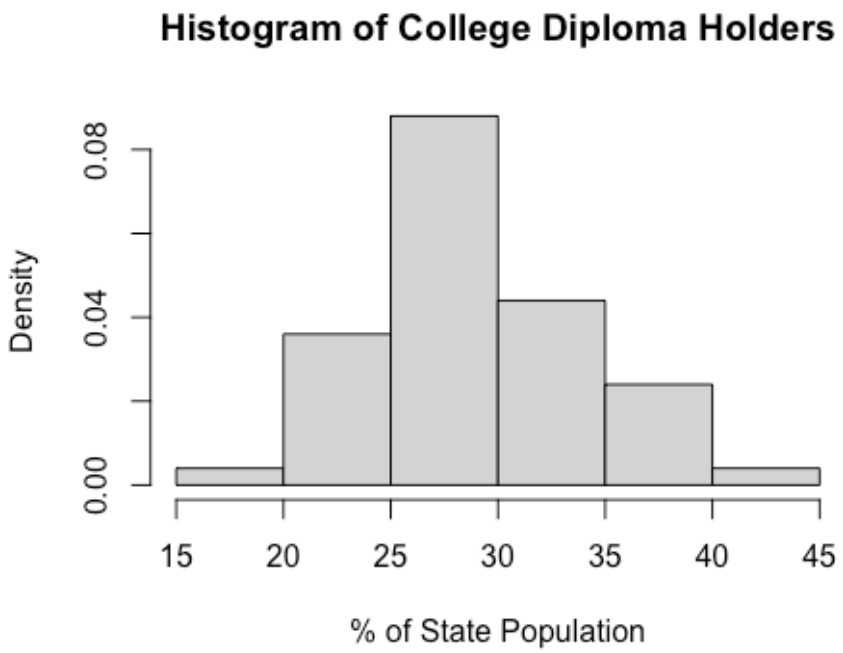
Analysis of a Continuous Variable - College Diplomas (dip_college)

The following section explores and summarizes the continuous variable of college diploma holders, which is measured as a percentage of the state population who received a college diploma.

To begin, a histogram and boxplot of the continuous variable will allow us to graphically assess the distribution of the data.

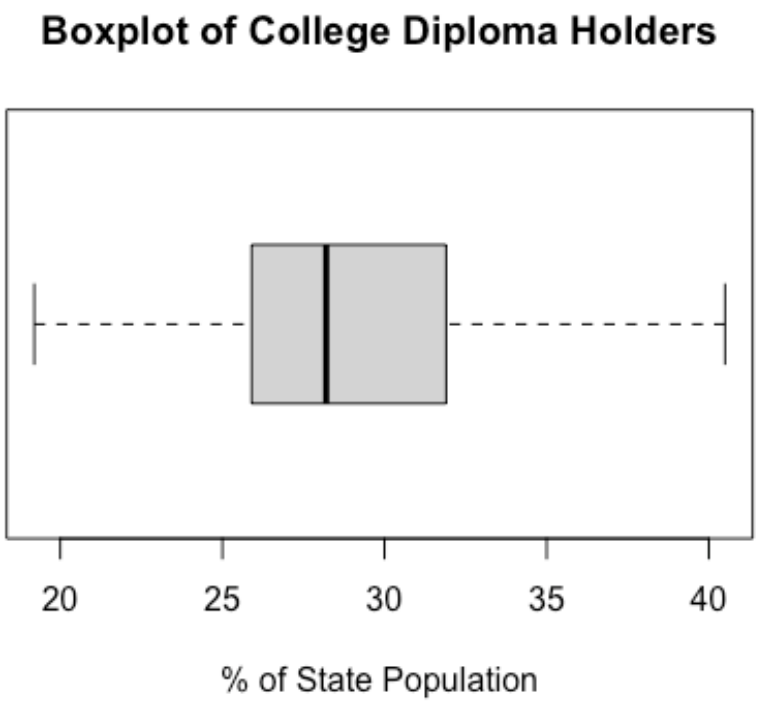
Creation of Histogram:

```
hist(dip_college, col = 'lightgrey', breaks = 7, xlab = '% of State Populatio
n', main = 'Histogram of College Diploma Holders', prob = TRUE)
```



The histogram of college diploma holders has a shape similar to the Normal distribution. The majority of the data are clustered in the middle of the graph and fall off to either side in a somewhat, symmetric manner. It is important to note that the aforementioned drop off is a bit steeper than we would expect a Normal distribution to be, which leaves a large bin in the center of the graph that is much higher than all the rest. The tails thin out to each side of the graph and are approximately equal, as we would expect.

```
boxplot(dip_college, horizontal = TRUE, main="Boxplot of College Diploma Hold
ers", xlab="% of State Population", col = "lightgrey")
```



The box plot of the continuous variable shows that the distance from Q_2 to Q_1 is smaller than the distance from Q_3 to Q_2, which is an indication of asymmetry. Typically, asymmetry would be a contradiction to the Normal distribution being a good fit for the data, but the asymmetry is minimal in this instance. Furthermore, the whiskers are close to the same length. From our investigation of outliers within the Normal distribution in our homework, we know that the probability that a point is an outlier is 0.0035. Therefore, we would expect, in a sample size of 50, that there would be 0 or 1 outliers, which is consistent with what we observe in our box plot. The box plot highlights some areas of concern that the Normal distribution might not model the variable perfectly; however, overall, it does seem to be a reasonable fit.

A numerical summary of the dip_college variable provides the exact quantile values that are graphically displayed in the box plot above.

```
summary(dip_college)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   19.20   25.95   28.20   29.01   31.77   40.50
```

```
#standard deviation of dip_college
sd(dip_college)
```

```
## [1] 4.93404
```

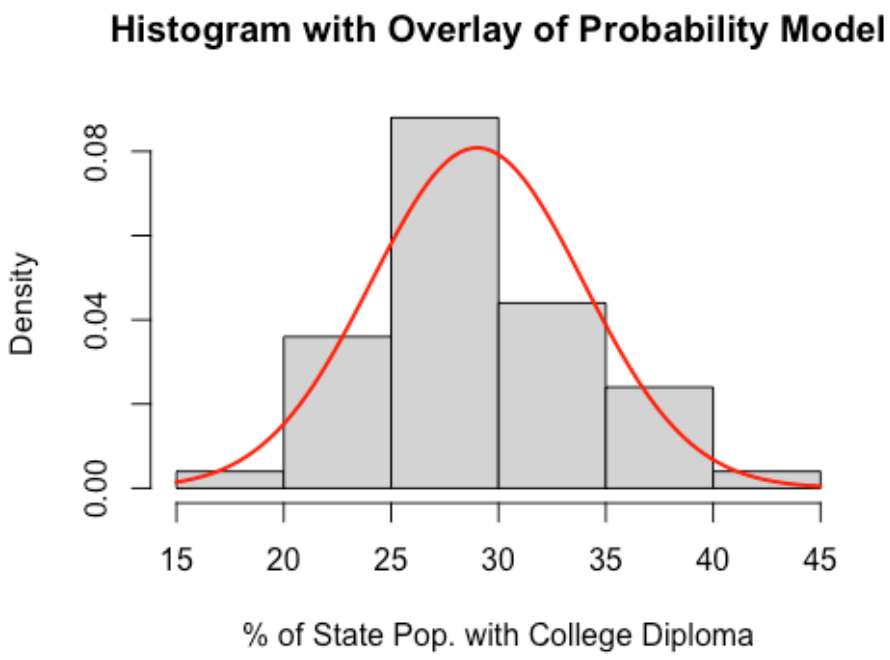
The summary shows that the first, second, and third quartiles are 25.95, 28.20, and 31.77 respectively. The mean, 29.01, is fairly close to the median value, 28.20, which supports the argument in favor of the Normal distribution being an adequate model for our variable. Since the data does not display a strong skew to the left or right, the mean is the best measure for central tendency and the standard deviation is the best indicator for variability or spread.

Choosing a Probability Model- College Diplomas (dip_college)

Because we argued that the Normal distribution would be an adequate probability model for dip_college, we will choose it to model our variable. We will use the sample mean as one of our parameter values as it is the Minimum Variable Unbiased (MVU) estimator for the population mean. We will choose sample variance as our second parameter value as it is an unbiased estimator for the population variance. Therefore, we will model our variable with a Normal(29.01, 4.932).

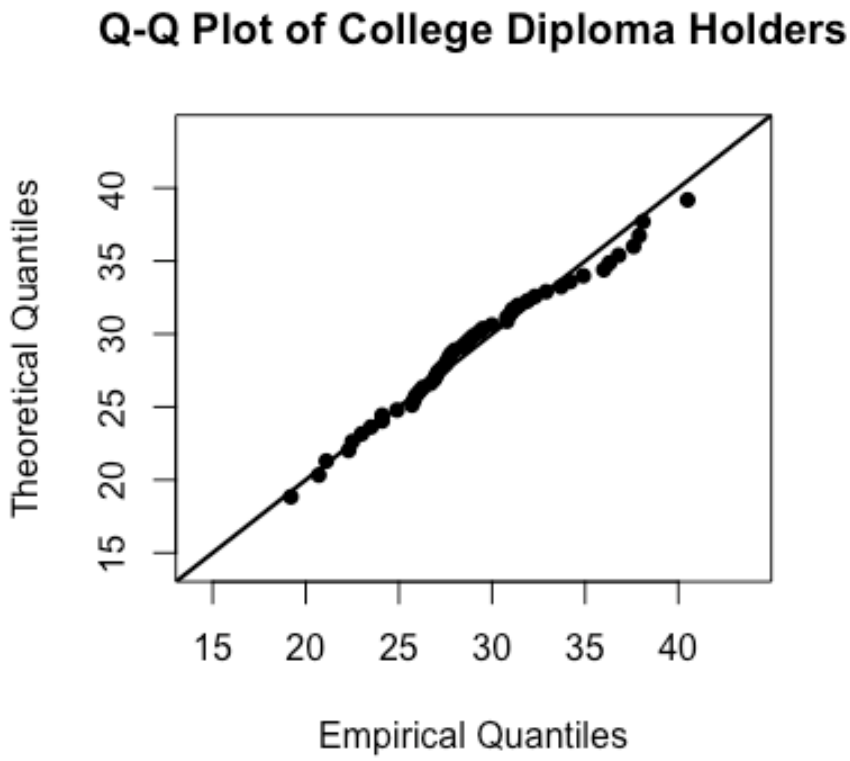
The probability model overlaid onto the histogram of the college diploma data scaled to density is shown below:

```
hist(dip_college, col = 'lightgrey', breaks = 7, xlab = '% of State Pop. with
College Diploma', main = 'Histogram with Overlay of Probability Model', prob
= TRUE)
curve(dnorm(x, mean=mean(dip_college), sd=sd(dip_college)), add=TRUE, lwd=2,
col = 'red')
```



While the overall shape of the data is similar to the normal curve, the curve does not fit the data perfectly. The curve underestimates the data in the center of the graph as well as the data just to the left of the center bin. For a more rigorous assessment of the fit of the model to the data, we will create a Q-Q Plot.

```
#creation of Q-Q Plot
x = dip_college
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 Q
uantiles", main="Q-Q Plot of College Diploma Holders",xlim=limits, ylim=limit
s, pch=16)
abline(0,1, lwd =2)
```



The Q-Q plot indicates that the Normal distribution is an adequate model for our data. The data points in the middle of the graph fit tightly to the identity line, which is good considering this is where the majority of our data points lie. As we move away from the center to the right, the data points decrease in frequency and stray a bit from the identity line but not enough to be significant. As we move away from the center to the left, the data points decrease in frequency and fit well to the identity line, which is what we would expect from Normally distributed data. Overall, the data fits the line in a way that shows that the Normal distribution is a reasonable model for our data.

Finally, we will conduct a Goodness of Fit test to further confirm the validity of this distribution as a model for dip_college. We will use the Shapiro - Wilk test of normality on our variable. In this test, H_0, or the null hypothesis, shows the Normal(29.01, 4.932) to be an adequate model for our data. In contrast, H_A, or the alternative, shows the Normal(29.01, 4.932) as an inadequate model for our data.

We hope to fail to reject the null hypothesis, which will support our argument in the previous section. We will choose alpha to be 0.05, consistent with the industry standard. Using R, we conducted the Shapiro test on dip_college:

```
shapiro.test(dip_college)

##
##  Shapiro-Wilk normality test
##
## data:  dip_college
## W = 0.9788, p-value = 0.5024
```

Our p-value is 0.5024 which is significantly greater than our alpha value of 0.05. Therefore, we conclude that we fail to reject the null hypothesis, thus further indicating that the Normal(29.01, 4.932) is a reasonable model for our variable describing the percentage of college diploma holders in the United States.

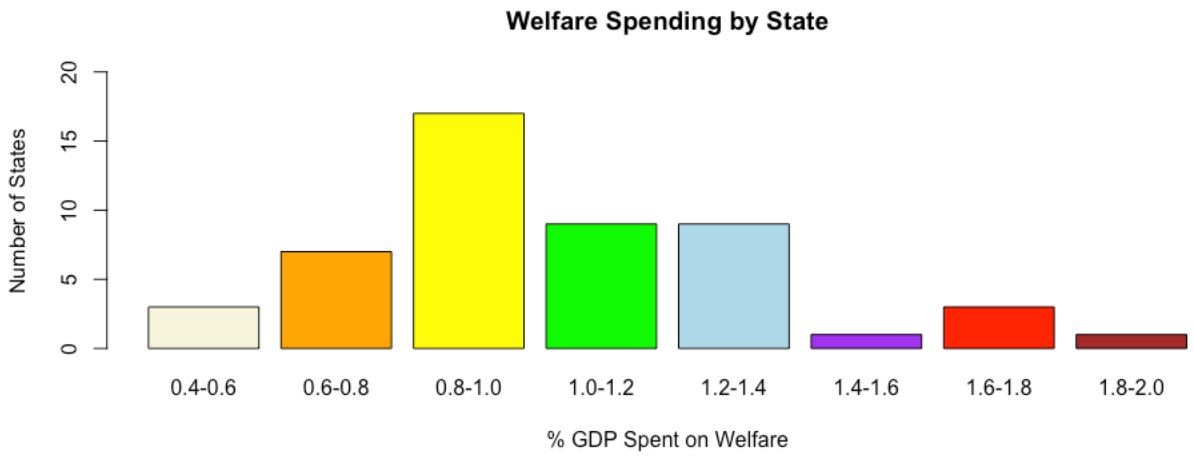
Analysis of an Ordinal Variable - Welfare Spending (welf_spent)

The following section explores and summarizes the categorical ordinal variable of state welfare spending. This variable is measured by the percentage of total state GDP spent on welfare. The range was from 0.4% to 2.0% of their total annual GDP. We ordered and then divided the variable into 8 equally indexed brackets of spending (Bin 0 - 7). The numerical summary for our observations can be found below:

```
table(welf_spent)

## welf_spent
## 0.4-0.6 0.6-0.8 0.8-1.0 1.0-1.2 1.2-1.4 1.4-1.6 1.6-1.8 1.8-2.0
##       3       7      17       9       9       1       3       1
```

Next, we will create a barplot of the data which will allow us to graphically assess the distribution of the data.

```
tb = table(Unemployment_Data$welf_spent)
barplot(tb, col=c("beige","orange", 'yellow', 'green', 'lightblue', 'purple',
'red', 'brown'),ylab="Number of States", xlab = '% GDP Spent on Welfare', mai
n = "Welfare Spending by State", ylim=c(0, 20))
```



According to the barplot, most of our observational units fall between bins 1-4. In fact, approximately 84.0% of our data falls between these bins. Bin 2 has the highest frequency with 17 states, and bins 5 and 7 have the lowest frequency with 1 state in each.

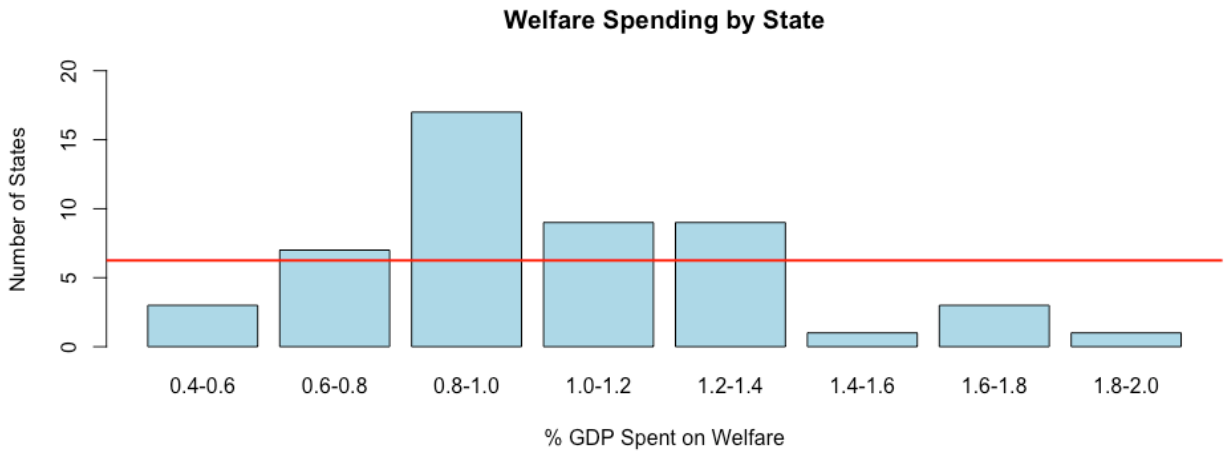
In conclusion, our data seems to be Uniformly distributed, therefore, we will propose a Discrete Uniform distribution with welf_spent ~ U[0,7] where the parameters 0 and 7 are chosen based off of the number of bins in the model.

Choosing a Probability Model- Welfare Spending (welf_spent)

According to our proposed theoretical model, the Discrete Uniform distribution U[0,7] with parameters 0 and 7. These parameters are not estimated values as they come from the number of bins in our model. We can calculate the probability that a state will fall in each bracket, and then calculate the expected number of states in each bracket. Because we have 50 observational units, the probability, p, that a given state will fall into a particular bin is p = 0.125. Therefore, the expected number of states in each bracket is 6.25.

In order to determine if our distribution resembles our proposed theoretical model of a Uniform distribution, we will overlay the theoretical pdf of Uniform[0,7] onto our barplot and compare the results.

```
tb = table(Unemployment_Data$welf_spent)
barplot(tb, col=c('lightblue'),ylab="Number of States", xlab = '% GDP Spent o
n Welfare', main = "Welfare Spending by State", ylim=c(0, 20))
abline(h = 6.25,lwd=2, col = 'red')
```



The overlay of the uniform pdf fits well for some of the bins, but we cannot confidently conclude that there is enough evidence to support the data falling under a Uniform model.

Therefore, we will perform a Goodness of Fit test on welf_spent to verify if the Uniform distribution is a sufficient theoretical model for our data. We will move forward with the following statements:

H_0: The Uniform distribution Model U[0,7] is an adequate model for the data.

H_A: The Uniform distribution Model U[0,7] is not an adequate model for the data.

First, we will choose a standard alpha value of 0.05. Next, we will calculate our Chi-Squared statistic, degrees of freedom, and p-value.

The Chi-Squared test statistic is $X_0^2 = \sum_{n=1}^{4} \frac{(O_j - E_j)^2}{E_j}$ and is distributed as a Chi-Squared distribution with 7 degrees of freedom, calculated by number of bins, 8, minus number of estimated parameters, 0, minus 1. The following table shows our work calculating the Chi-Square test statistic:

```
#using ktable to create chi squared test statistic table

text_tbl <- data.frame(
  Bin = c('0.4-0.6%', '0.6-0.8%', '0.8-1.0%', '1.0-1.2%', '1.2-1.4%', '1.4-1.
6%', '1.6-1.8%', '1.8-2.0%'),
  Bin.Index = c('0', '1', '2', '3', '4', '5', '6', '7'),
  O_j = c('3', '7', '17', '9', '9', '1', '3', '1'),
  E_j = c('6.26','6.26','6.26','6.26','6.26','6.26','6.26','6.26'),
  Oj.Ej.2.Ej = c('1.69','0.09', '18.49', '0.984064', '0.984064', '4.41', '1.6
9', '4.41')
)

kable(text_tbl) %>%
  kable_styling(bootstrap_options = "striped") %>%
  column_spec(1, bold = T, border_right = T, width = '7em') %>%
  column_spec(2, border_right = T) %>%
  column_spec(3, border_right = T, width = "5em") %>%
  column_spec(4, border_right = T, width = "5em") %>%
  column_spec(5, border_right = T, width = "5em") %>%
  column_spec(2, width = "5em")
```

Bin	Bin.Index	O_j	E_j	Oj.Ej.2.Ej
0.4-0.6%	0	3	6.26	1.69
0.6-0.8%	1	7	6.26	0.09
0.8-1.0%	2	17	6.26	18.49
1.0-1.2%	3	9	6.26	0.984064
1.2-1.4%	4	9	6.26	0.984064
1.4-1.6%	5	1	6.26	4.41
1.6-1.8%	6	3	6.26	1.69
1.8-2.0%	7	1	6.26	4.41

By summing the values in the right most column, we observe that our Chi-Squared test statistic is X_0^2 = 33.2. The corresponding p-value is calculated by P($X_0^2 \geq 33.2$) and is approximately equal to 0.0000243. The following R code verifies the Chi-Squared test above:

```
chisq.test(table(welf_spent))

##
##  Chi-squared test for given probabilities
##
## data:  table(welf_spent)
## X-squared = 33.2, df = 7, p-value = 2.43e-05
```

The p-value of 0.0000243 is significantly lower than our alpha value of 0.05. Therefore, we can confidently reject the null hypothesis in favor of our alternative hypothesis that U[0,7] is not an adequate model for the data. We can conclude that the model does not fit our data well because not all states spend the same percentage of their GDP on their respective welfare programs. Each state spends differing amounts, and they cannot be modeled well by a Uniform model. If our data were to be altered in order to fit the model U[0,7], more states would need to fall under bins 5-7, which means they would have to spend more money on welfare than they currently are spending.

Analysis of a Binary Variable - Political Majority (pol_maj)

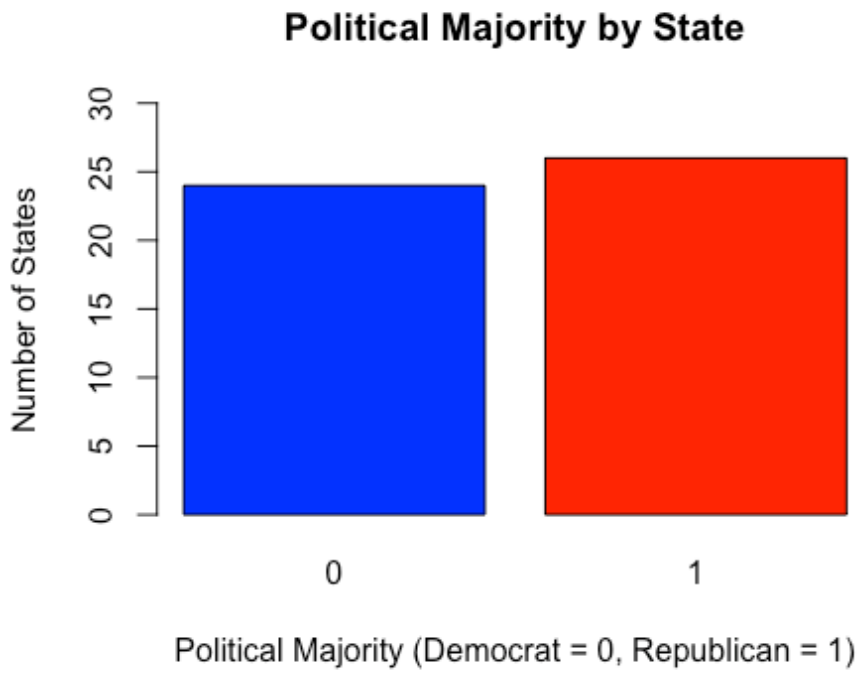
The following section explores and summarizes the categorical binary variable of state political majority. This variable is measured by looking at the last 5 presidential election voting results for each state, and determining whether the majority was Democratic or Republican. States with a Democratic majority were labeled as 0, and states with a Republican majority were labeled as 1. We will approximate the distribution of our binary variable with the Bernoulli distribution. The numerical summary for our observations can be found below:

```
table(pol_maj)
```

```
## pol_maj
##  0  1
## 24 26
```

Take note that Republican states totaled to 26 while Democratic states totaled to 24. Below is a graphical summary of the data using a barplot:

```
barplot(table(pol_maj), col=c('Blue','Red'),ylab="Number of States", xlab = '
Political Majority (Democrat = 0, Republican = 1)', main = "Political Majorit
y by State", ylim=c(0, 30))
```



This barplot shows us the visual distribution of states between Democratic and Republican majorities. The frequencies of each category are very close to being equal.

Choosing a Probability Model- Political Majority (pol_maj)

This data can be modeled by a Bernoulli distribution, because it is coded in binary values of 0 and 1. Let's say that Republican (1) is a "success." Therefore, the probability for a success in our sample is equal to $\frac{26}{50}$ = 0.52 or 52.0%. Let's set a random variable X to represent the political majority of a state. Then for any given state, X ~ Bernoulli(0.52). Conversely, we can find the probability that any given state is Democratic by taking 1- P{any given state is Republican} = 1 - 0.52 = 0.48. In conclusion, the probability of a state being Republican is 52% and the probability of a state being Democratic is 48%. If you consider a 1 (Republican) as a "success," this data can be modeled by the Bernoulli distribution of Bernoulli(0.52).

Analysis of a Categorical Nominal Variable - State Region (st_reg)

The following section explores and summarizes the categorical nominal variable regions, which describes the part of the United States in which a particular state lies. There are four U.S. Census Bureau designated regions: Northeast, South, Midwest, and West.

A numerical summary of the data is below:

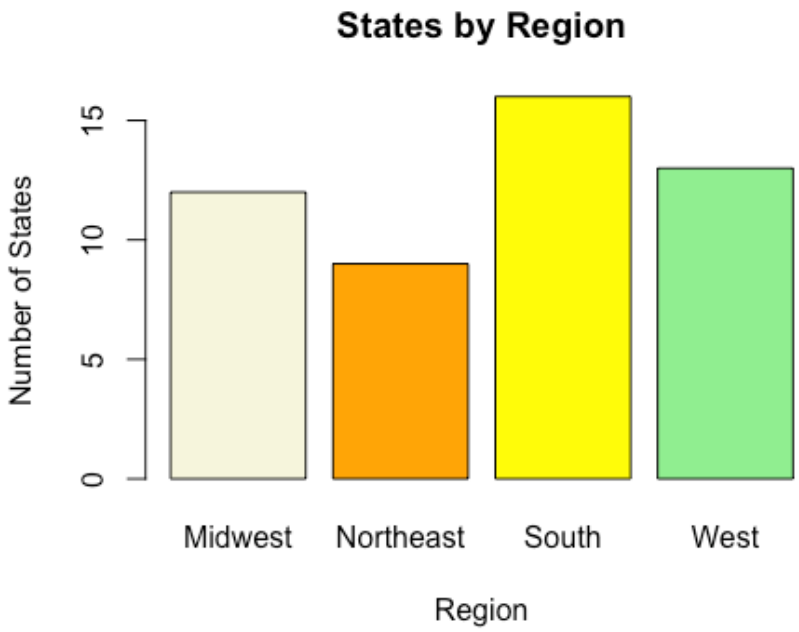
```
table(st_reg)

## st_reg
##   Midwest Northeast     South      West
##        12         9        16        13
```

A nominal variable separates the data into non-ordered, non-overlapping subsets. Therefore, the mean and standard deviation are not meaningful measurements. This means that numerical summaries of this sort are not helpful to tell us information about the data. Rather, numerical summaries showing the most frequent and least frequently occurring categories helps give a better picture of the data. In our data set, the most frequent region category is the South containing 16 states, and the least frequent region category is the Northeast containing 9 states.

A barplot of st_reg will allow us to graphically assess the distribution of the data.

Creation of barplot in R:

```
tab = table(Unemployment_Data$st_reg)
barplot(tab, col=c("beige","orange", 'yellow', 'lightgreen'),ylab="Number of
States", xlab = 'Region', main = "States by Region")
```



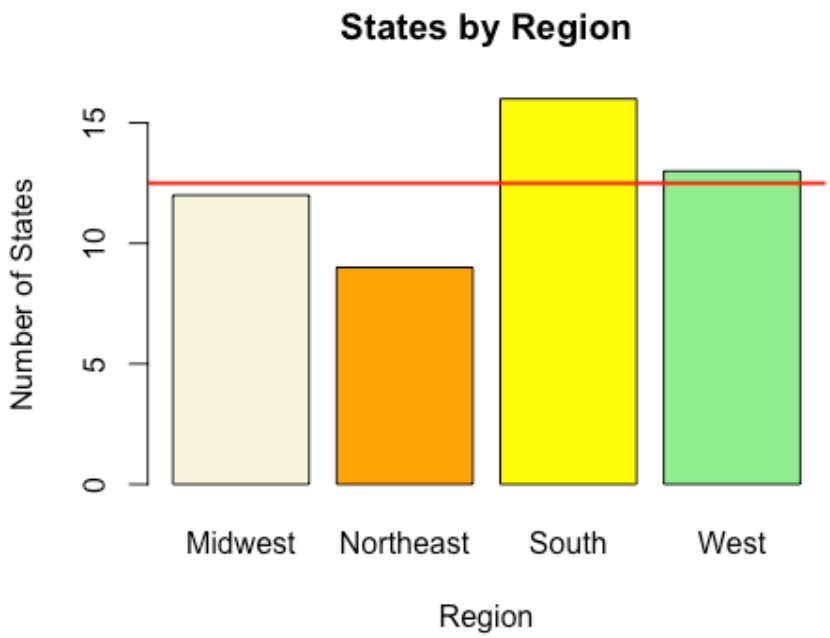
The barplot shows the four regions within the United States with the number of states within each region.

Choosing a Probability Model- State Region (st_reg)

Since our variable is categorical, we can expect the distribution to be a Discrete Uniform distribution. At first glance, the Discrete Uniform distribution looks to be a reasonable model for our data since all of the frequencies are numerically similar. Let X be the region of the United States, then we will propose $X \sim U[1,4]$ where the parameters 1 and 4 are chosen based off the number of bins in the model. These parameters are not estimated, as they come from the number of bins in the model. The probability, p, that a given state will fall into a particular bin is $p = 0.25$. Therefore, the expected number of states in each category is 12.5.

In order to determine if our distribution resembles our proposed theoretical model of a Uniform distribution, we will overlay the theoretical pdf of Uniform[1,4] onto our barplot and compare the results.

```
tab = table(Unemployment_Data$st_reg)
barplot(tab, col=c("beige","orange", 'yellow', 'lightgreen'),ylab="Number of
States", xlab = 'Region', main = "States by Region")
abline(h = 12.5,lwd=2, col = 'red')
```



The overlay of the uniform pdf fits well for the West and Midwest bins, but it overestimates a bit for the Northeast and underestimates for the South bin. Overall, it is a reasonable model, but it is certainly not a perfect fit. For a more definitive analysis, we will perform a Chi-Squared Goodness of Fit test on X to verify if the Uniform distribution is a sufficient theoretical model for our data. We will move forward with the following statements:

H_0: The Uniform distribution Model U[1,4] is an adequate model for the data.

H_A: The Uniform distribution Model U[1,4] is not an adequate model for the data.

We will choose a standard alpha value of 0.05. Next, we will calculate our Chi-squared test statistic, degrees of freedom and p-value. The Chi-Squared test statistic is distributed as a Chi-Squared distribution with 3 degrees of freedom, calculated by number of bins, 4, minus number of estimated parameters, 0, minus 1. The following table shows our work calculating the Chi-Squared test statistic:

```r
#using ktable to create chi squared test statistic table

text_tbl <- data.frame(
  Bin = c('Midwest', 'Northeast', 'South', 'West'),
  O_j = c('12', '9', '16', '13'),
  E_j = c('12.5','12.5','12.5','12.5'),
  Oj.Ej.2.Ej = c('0.02', '0.98', '0.98', '0.02')
)

kable(text_tbl) %>%
  kable_styling(bootstrap_options = "striped") %>%
  column_spec(1, bold = T, border_right = T, width = '7em') %>%
  column_spec(2, border_right = T) %>%
  column_spec(3, border_right = T, width = "5em") %>%
  column_spec(4, border_right = T, width = "5em") %>%
  column_spec(2, width = "5em")
```

Bin	O_j	E_j	Oj.Ej.2.Ej
Midwest	12	12.5	0.02
Northeast	9	12.5	0.98
South	16	12.5	0.98
West	13	12.5	0.02

By summing the values in the right most column, we observe that our Chi-Squared test statistic is $X_0^2 = 2$. The corresponding p-value calculated by $P(X_0^2 \geq 2)$ is approximately equal to 0.5724. The following R code verifies the Chi-Squared test above:

```r
chisq.test(table(st_reg))
```

```
##
##  Chi-squared test for given probabilities
##
## data:  table(st_reg)
## X-squared = 2, df = 3, p-value = 0.5724
```

As a result of our Chi-Squared test, our p-value of 0.5724 is greater than our alpha value of 0.05. Therefore, we fail to reject the null hypothesis indicating that the Uniform[1,4] is an adequate model for the data. We conclude that the model fits our data well, because the 50 states are equally represented in every region.

Module 2 Conclusion

In summary, we failed to reject our proposed probability models for the variables dip_college and st_reg, and rejected our probability model for welf_spent. For our variable pol_maj, we chose the Bernoulli distribution because the variable was binary.

Now that we have summarized our variables numerically and graphically and explored probability models for each, we can move forward with our investigation into the underlying relationships that all our variables have with the unemployment rate in the 50 states.

Module 3: Single Sample Analysis

In this module, we will take a closer look at our unemployment variable by performing a series of hypothesis tests. We will compare the states' unemployment rates to unemployment in the rest of the world, research studies on the economic predictions, and the political majority in any given state. Our first test will be on the true mean of the 50 states' unemployment rate in comparison with the world's unemployment rate in 2016. We would like to investigate whether or not the U.S. unemployment rate is in fact lower than the rest of the world's unemployment rate. Our second test will be a comparison of the U.S. 2016 unemployment rate to the U.S. 2015 unemployment rate to investigate whether or not the current economy is in fluctuation. Our third test will be a test of independence on our unemployment rate variable and our political majority variable. We will investigate if our two variables of interest are independent or dependent of one another.

Hypothesis Test 1: Does the United States have a lower unemployment rate than the rest of the world?

The United States has been touted as one of the greatest countries in the world to live, work, and play. The United States boasts a powerhouse economy with the highest GDP out of all the nations in the world. Consequentially, a plausible assumption could be that the U.S. would also have a lower unemployment rate than the world average unemployment rate, which was 5.739% in 2016 (The World Bank, 2017). The objective of this section is to answer the following question: Does the United States have a lower unemployment rate than the rest of the world?

Let U be the average unemployment rate for all the states in the U.S. in the year 2016. The true population mean of U will be represented by μ_u.

The following hypothesis will be used to answer the questions presented above:

H_0: $\mu_u < 5.739$

H_A: $\mu_u \geq 5.739$

Here, the null hypothesis is suggesting that the U.S. has a lower unemployment rate than the rest of the world, which we expect to be the case. Our alternative hypothesis suggests that the U.S. has an equal or higher unemployment rate than the rest of the world, which would be a surprising result because we anticipate the U.S. economy to be healthier than the rest of the world. To properly test these hypotheses, we will conduct a one-sample, one-sided hypothesis test on the population mean on a Normally distributed population with an unknown variance. The standard alpha value of 0.05 will be used for this test.

Assumptions

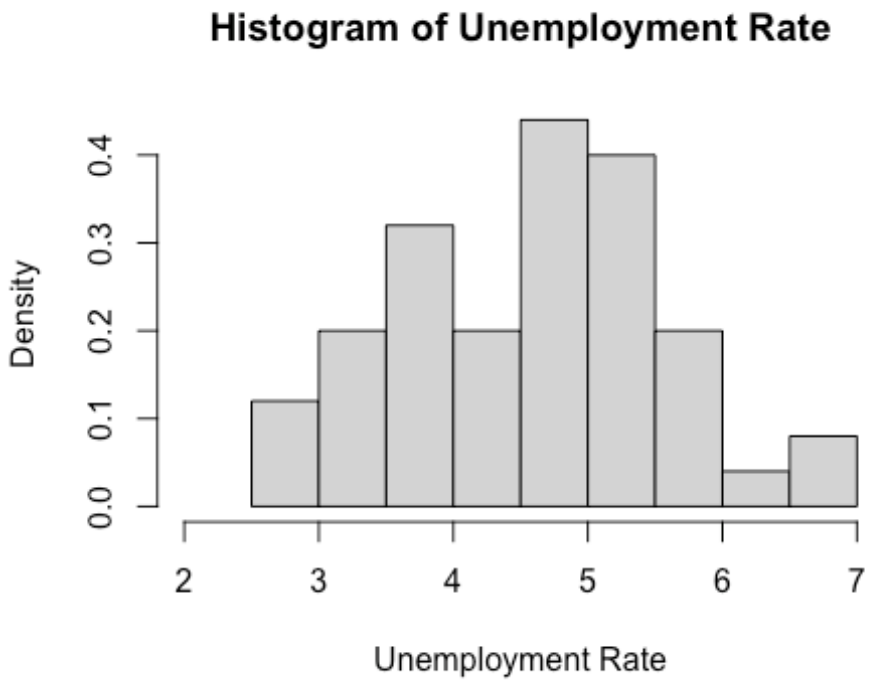
Before beginning our hypothesis test, we made three assumptions that are stated below:

1. The sample is a random sample.
    - The observational units of the population we are interested in examining is the 50 states of the United States of America in 2016. We will use all 50 data points representing each state's unemployment rate in 2016 and consider it a random sample in time of the probability model that represents the distribution of the average U.S. unemployment rate since the U.S. began recording the rate itself.
2. The observational units (state unemployment rates) are independent of each other.
    - This assumption may not be true, because each of the 50 states are a part of one sovereign nation, the United States. For example, if corporate tax laws are increased on a federal level, this might affect many states' unemployment rates. Additionally, if Coca-Cola (headquartered in Atlanta, Georgia) were to shut down, Georgia's unemployment rate would increase and many other states' rates would increase as well due to Coca-Cola employing many citizens from different states nation-wide. However, we will continue with our hypothesis test as if this assumption is true.
3. The data is normally distributed.
    - We will investigate this assumption below by looking at a histogram and Q-Q plot of the data set.

Histogram of un_rate variable

The histogram below has been created to help give a visual representation of how the data is distributed.

```
hist(un_rate, col = 'lightgrey', breaks =7, xlab = 'Unemployment Rate', main
= 'Histogram of Unemployment Rate', xlim = c(2, 7), prob = TRUE)
```



Because we are arguing that the Normal distribution would be an adequate probability model for the unemployment rate variable, we will choose to overlay the probability model over the histogram of the un_rate variable.

We will use the sample mean as one of our parameter values as it is the Minimum Variable Unbiased (MVU) estimator for the population mean. We will choose sample variance as our second parameter value as it is an unbiased estimator for the population variance. Therefore, we will model our variable with a Normal(4.64, 0.99).

The probability model overlaid onto the histogram of the unemployment rate data scaled to density below:

```
hist(un_rate, col = 'lightgrey', breaks = 7, xlab = 'Unemployment Rate', main
= 'Unemployment Rate', prob = TRUE)
curve(dnorm(x, mean=mean(un_rate), sd=sd(un_rate)), add=TRUE, lwd=2, col = 'r
ed')
```

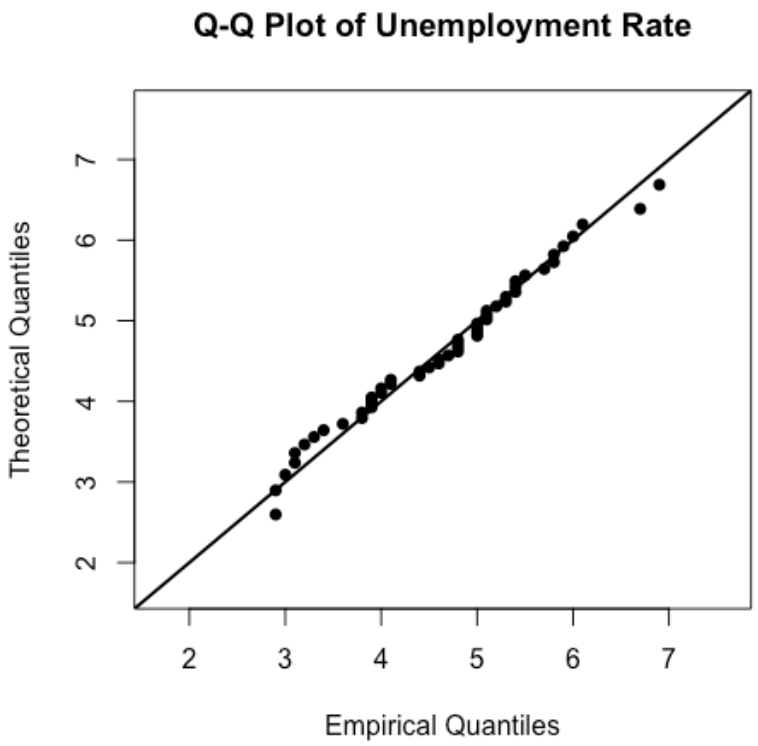
We conclude from the model above that the data has a small right tail and centers right around 4.60 (the true sample mean is 4.64). The data does appear to be Normal, and we know by the Central Limit Theorem that the limiting distribution of the sample mean will approach the standardized Normal distribution. Therefore, the histogram and model of un_rate does support the assumption that our data is normally distributed.

Q-Q Plot of un_rate

Additionally, we would like to construct a Q-Q plot to further test the assumption of the normality of the unemployment rate variable.

```
x = un_rate
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 Q
uantiles", main="Q-Q Plot of Unemployment Rate",xlim=limits, ylim=limits, pch
=16)
abline(0,1, lwd =2)
```



The Q-Q plot indicates that the Normal distribution is an adequate model for our data. The data points in the middle of the graph fit tightly to the identity line, which is good considering this is where the majority of our data points lie. As we move away from the center and towards the left side, the data points stray a bit from the identity line but not enough to be significant. As we move away from the center towards the right side, the data points fit fairly well to the line. Overall, the data fits the line in a way that shows that the Normal is a reasonable model for our data.

In conclusion, the visual and graphical representations we have constructed to test our data against the Normal distribution give us enough confidence in our assumption of the normality of the data. We will move forward with our hypothesis test.

Continuation of Hypothesis Test 1

Our test statistic for a one-sample, one-sided hypothesis test on the population mean of a Normally distributed population with an unknown variance is equal to:

$T_0 = \frac{\overline{x}-\mu}{\frac{S_S}{\sqrt{n}}} \sim t_{n-1}$

Where $\overline{x}$ is equal to the sample mean, S_s is equal to the sample standard deviation and n is equal to the sample size.

We can calculate $\overline{x}$ = 4.64, S_s = 0.99, and n = 50.

$T_0 = \frac{4.64-5.739}{\frac{0.99}{\sqrt{50}}}$ = -7.849

We know that T_0 is equal to 7.849 and has a t-distribution with (n-1) degrees of freedom:

T_0 = -7.849 ~ t_{49}

We can then calculate our critical value, which we will compare to our calculated test statistic. Because the hypothesis test we are performing is one-sided, we will have one critical value denoted as t*.

Before we calculate our critical value, we know that if our test statistic is less than our calculated critical value, we will fail to reject H_0 in favor of H_A. However, if the our test statistic is greater than our calculated critical value, we will reject H_0 in favor of H_A.

In order to calculate t*, we must use alpha to determine the rejection region that contains all of the critical values in our t distribution. Therefore, we know that the probability of our t_{49} distribution being less than our t* is equal to 0.95:

$P\{t_{49} < t^*\} = 0.95$

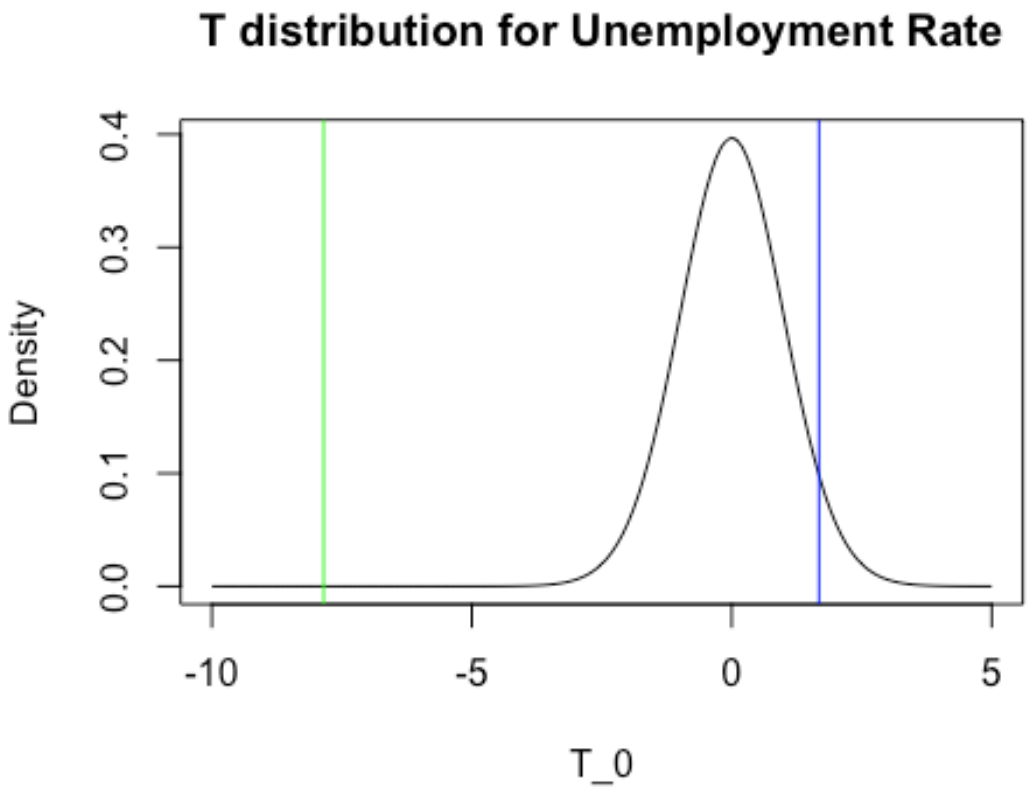
Next, we can use the invT function on a Ti-84 calculator, or equivalent, to calculate the value of t*:

invT(0.95,49) = 1.677

Thus, $T_0 = -7.849 < t^* = 1.677$

Below is a graph showing the t_{49} distribution. The plotted blue line is our critical value, t*, and the plotted green line is our test statistic, T_0.

```
x=seq(from=-10,to=5,by=0.1)
y=dt(x,df=49)
plot(x,y,type="l", main="T distribution for Unemployment Rate", xlab = "T_0",
ylab = "Density")
abline(v=1.677,col="blue")
abline(v=-7.849,col="green")
```



Next, we can calculate the p-value of our model. Our p-value is defined as the probability of getting the data we got or something more extreme. The area under the T distribution curve to the right of the green line is our p-value. We can calculate our p-value by using the Tcdf function on a Ti-84 calculator or equivalent.

Tcdf(7.849,99999,49) = 0.999

p-value = 0.999 > alpha = 0.05

The calculated p-value is much greater than 0.05, and our test statistic was less than our critical value, so we fail to reject our null hypothesis in favor of our alternative hypothesis.

T0 = -7.849 < t* = 1.677 and p-value = 0.999 > alpha = 0.05 ⇒ Fail to Reject H0

In conclusion, we fail to reject the hypothesis that the mean of the unemployment rate of the United States is lower than the mean value of the world's unemployment rate. This is not surprising, as the United States has one of the world's most powerful economies and is one of the top nations in terms economic growth and development.

95% Confidence Interval of Population Mean

In this section, we will create a two-sided 95% confidence interval on the population mean of the state unemployment rate in the United States. Our sample of the 50 states in 2016 results in following values:

$\overline{x}$ = 4.64

S_s = 0.99

n = 50

where $\overline{x}$ is equal to the sample mean, S_s is equal to the sample standard deviation and n is equal to the sample size.

As stated above, we will assume the distribution of the data is Normal.

Quantity of Interest: average rate of unemployment for the 50 states of the United States of America

Our population parameter can be defined as: μ

Because we have an unknown variance, we will use a t statistic in order to calculate our critical value.

Using an alpha of 0.05 for the 95% confidence interval and 49 degrees of freedom for our T-test, we will compute the value of t* (critical value) using the invT function on a Ti-84 calculator:

invT(0.025,49) = -2.00

Because of the two-sided confidence interval and the symmetric nature of the T distribution, we also know our upper critical value is equal to 2.00.

We will use the following formula to calculate the lower and upper bounds of our two-sided 95% confidence interval on the population mean:

$P(\overline{x}$ - ($t_{(\frac{\alpha}{2},n-1)}$ * standard error) $\leq \mu \leq \overline{x}$ + ($t_{(\frac{\alpha}{2},n-1)}$ * standard error)) = 1 - α

where standard error is defined as:

standard error = $\frac{S_s}{\sqrt{(n)}}$

P(4.64 - (2.00 * 0.14) $\leq \mu \leq$ 4.64 + (2.00 * 0.14)) = 0.95

P(4.36 $\leq \mu \leq$ 4.92) = 0.95

In conclusion, the 95% confidence interval on the population mean of the average state unemployment rate based on this data is [4.36, 4.92]. This means if we repeated this experiment of drawing a random sample, of size 50, thousands of times, and created a confidence interval for each data set, and knew the true population mean, approximately 95% of the confidence intervals would contain the true population mean.

Power Calculation

Additionally, we will conduct an analysis on the power of this hypothesis test. Power is defined to be $1 - \beta$ where β is the probability that the test rejects the null hypothesis when the null hypothesis is indeed false. In other words, there was a missed opportunity to shout to the world the surprising findings of the test. As a specific alternative, we will use the value μ = 6.5, which is an unemployment rate higher than one we would like to see in the United States as an indicator for the health of our economy.

By defining effect size, number of observations, and the significance level, we can calculate the power of our specific alternative.

Effect size = $\frac{delta}{sd} = \frac{(6.5-5.739)}{0.99}$

Number of Observations = 50

Significance level = $\alpha = 0.05$

Using R to calculate the power of the hypothesis test:

```
library(pwr)
num_obs = 50
delta = 6.3 - 5.739
alpha = 0.05
pwr.t.test(n = num_obs, d = delta, sig.level = alpha, power = NULL, type = c(
"one.sample"), alternative = c("greater"))

##
##      One-sample t test power calculation
##
##               n = 50
##               d = 0.561
##       sig.level = 0.05
##           power = 0.9882869
##     alternative = greater
```

The power of this hypothesis test is 0.9882869. Since power measures the sensitivity of a statistical test, if the true mean unemployment rate is really 6.5% this test will correctly reject H_0 and detect this difference 98.82% of the time. This is a high value of power, meaning the probability of getting a type II error is low.

Hypothesis Test 2: Has the U.S. economic expansion leveled off?

The unemployment rate is often used as an indicator of the health of an economy. The government, economists, and citizens alike have a vested interest in improving their economy's health, and at minimum, maintain it. Some economists forecast that the dramatic expansion the United States is currently experiencing is starting to level off to a more stable rate compared to the notable economic fluctuation we have experienced in the past couple of years. According to the U.S. Bureau of Labor Statistics, the unemployment

rate was 5.3% in 2015. We will look at the year over year unemployment rate of the United States to investigate the following question: Has the U.S. economic expansion leveled off?

Let U be the average unemployment rate for all the states in the U.S. in the year 2016. The true population mean of U will be represented by μ_S.

The following hypothesis will be used to answer the questions presented above:

H_0: $\mu_S = 5.27$

H_A: $\mu_S \neq 5.27$

Here our null hypothesis suggests that the U.S. unemployment rate from 2016 is equal to the unemployment rate from 2015, which indicates that the economists are predicting correctly that the economy is entering a more stable rate with less fluctuation. The alternative hypothesis suggests that the 2016 U.S. unemployment rate is different from the 2015 rate, which would suggest that, in fact, the U.S. economy is still in a state of fluctuation and has not leveled off. In order to properly test these hypotheses, our team will conduct a one sample, two sided hypothesis test on the population mean of Normally distributed data with an unknown variance. The standard alpha value of 0.05 will be used for this test.

Assumptions

As explained above in hypothesis test 1, we made three assumptions that are stated below:

1. The sample is a random sample.
2. The observational units (state unemployment rates) are independent of each other.
3. The data is normally distributed.

Continuation of Hypothesis Test 2

We can calculate $\overline{x}$ = 4.64, S_s = 0.99, and n = 50.

Our test statistic is $T_0 = \frac{(4.64-5.27)}{\frac{0.99}{\sqrt{(50)}}} = -4.48$.

Therefore, $T_0 = -4.48 \sim t_{49}$.

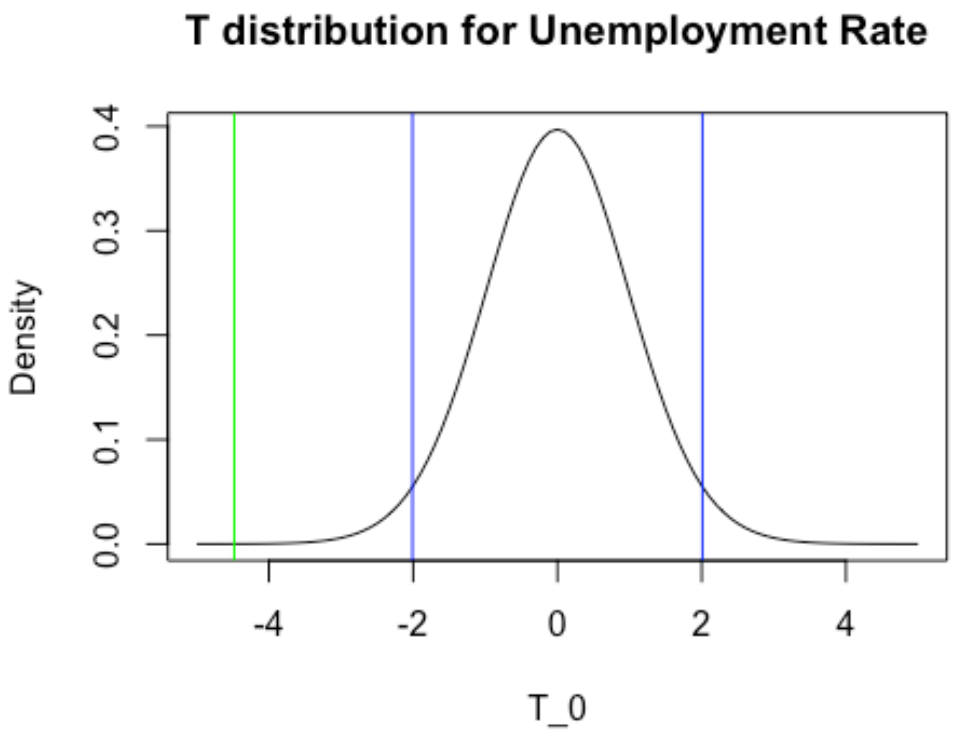
Since the test is two-sided, we will have two critical values, denoted -t* and t*, and we will fail to reject the null hypothesis in favor of the alternative when $T_0 > -t^*$ and $T_0 < t^*$.

Using the invT function on our calculator, we find that invT(0.025, 49) = t* = 2.01 and, by symmetry, -2.01.

$T_0 = -4.48 < -t^* = -2.01 < t^* = 2.01$

Below is a graphical representation of the t_{49} distribution. The blue lines indicate our critical values which are the outer bounds of our fail to reject region, and the plotted green line is our test statistic.

```
x=seq(from=-5,to=5,by=0.1)
y=dt(x,df=49)
plot(x,y,type="l", main="T distribution for Unemployment Rate", xlab = "T_0",
ylab = "Density")
abline(v=2.01,col="blue")
abline(v=-2.01, col = 'blue')
abline(v=-4.48,col="green")
```



Because our T_0 value is much smaller than our -t* value and falls outside of the fail to reject region, we comfortably fail to reject the null hypothesis in favor of the alternative hypothesis suggesting that the United States economy is in a state of fluctuation and has not yet leveled off year over year from 2015.

We can use R to confirm our calculations and conclusion:

```
t.test(un_rate, mu=5.27, alternative = c('two.sided'))
```

```
##
##  One Sample t-test
##
## data:  un_rate
## t = -4.4763, df = 49, p-value = 4.536e-05
## alternative hypothesis: true mean is not equal to 5.27
## 95 percent confidence interval:
##  4.360069 4.923931
## sample estimates:
## mean of x
##     4.642
```

The R analysis confirms that our calculations were correct and the true population mean is not equal to 5.27%. R further provides us with a p-value of 4.536e-05 which is smaller than our alpha value of 0.05; therefore, we can comfortably reject the null hypothesis in favor of the alternate hypothesis and say that the mean unemployment rate from 2015 is not equal to the mean unemployment rate from 2016, which indicates that the economic fluctuation period has not yet leveled off.

Hypothesis Test 3: Is a state's unemployment rate independent of its political majority?

In this section, we would like to use two variables from our data set, un_rate and pol_maj, to determine whether or not a state's unemployment rate is independent of its political majority. Politicians from both major political parties in the United States make many promises to their citizens and run their campaigns on the idea that they will improve the quality of life for their voters. Creating jobs and minimizing unemployment is often a major topic of discussion during their campaigns. Does a state having a particular political majority coincide with a higher or lower unemployment rate? We will further analyze this

relationship by performing a Chi-Squared test of independence on these two variables below.

Define Notation

Let un_rate denote the percent of the labor force that is without gainful employment by state.

Let pol_maj denote the state majority political party affiliation evaluated by looking at the last 5 presidential election voting results for each state, whether the majority was Democratic or Republican, labeled as 0 and 1 respectively.

State Hypothesis

H_0: un_rate and pol_maj are independent.

H_A: un_rate and pol_maj are not independent.

Choose Alpha

By default, our alpha value will be equal to 0.05, as that is the standard value for alpha.

Assumptions

Before beginning our hypothesis test, we made three assumptions that are stated below:

1. The sample size is sufficiently large
    - The observational units of the population we are interested in examining is the 50 states of the United States of America in 2016. We will use all 50 data points representing each state's unemployment rate in 2016 and consider it a random sample in time of the probability model that represents the distribution of the average U.S. unemployment rate since the U.S. began recording the rate itself. The standard number of data points in order to satisfy this assumption is 20, and we have 50, so this assumption is satisfied.
2. The observational units (states) are independent of each other

- This assumption may not be true for the un_rate variable, as explained in Hypothesis Test 1 above.
- This assumption may not be true for the pol_maj variable, because each of the 50 states are apart of one sovereign nation, the United States. Political preference is not always bounded by state borders, so we cannot say with confidence that each state is independent of another for the pol_maj variable. However, we will continue with our hypothesis test as if this assumption is true.

Convert Continuous Variable to Ordinal (Unemployment Rate)

In order to examine our state unemployment rate variable, we will need to break the data up into equally indexed bins. Because we have 50 data observations, we will use 7 bins because the wisdom of the ages says to take the square root of the number of your observations as your bin count, and 7 is close to 7.07.

```
newdata = cut(sort(un_rate), breaks = 7)
table(newdata)

## newdata
##  (2.9,3.47] (3.47,4.04] (4.04,4.61] (4.61,5.19] (5.19,5.76] (5.76,6.33]
##           8           8           7          12           8           5
##  (6.33,6.9]
##           2

bin1 = length(un_rate[un_rate > 1.99 & un_rate < 2.76])
bin2 = length(un_rate[un_rate >= 2.76 & un_rate < 3.51])
bin3 = length(un_rate[un_rate >= 3.51 & un_rate < 4.27])
bin4 = length(un_rate[un_rate >= 4.27 & un_rate < 5.03])
bin5 = length(un_rate[un_rate >= 5.03 & un_rate < 5.79])
bin6 = length(un_rate[un_rate >= 5.79 & un_rate < 6.54])
bin7 = length(un_rate[un_rate >= 6.54 & un_rate <= 7.31])
```

Create a Table of Counts

First, we need to find the counts of Democratic and Republican states within each bin.

```
Dbin1 = 2
Rbin1 = 2
Dbin2 = 7
Rbin2 = 8
Dbin3 = 6
Rbin3 = 4
```

```
Dbin4 = 11
Rbin4 = 7
Dbin5 = 0
Rbin5 = 1
Dbin6 = 1
Rbin6 = 0
Dbin7 = 0
Rbin7 = 1
```

Next, we can create a table which displays our observations based on unemployment bracket and political majority.

```
x = rbind(c(Dbin1, Dbin2, Dbin3, Dbin4, Dbin5, Dbin6, Dbin7),c(Rbin1,Rbin2, R
bin3, Rbin4, Rbin5, Rbin6, Rbin7))
rownames(x) <- c("Democratic (0)", "Republican (1)")
colnames(x) <- c("2.00-2.75","2.76-3.50","3.51-4.26","4.27-5.02","5.03-5.78",
"5.79-6.53","6.54-7.31")

x
```

```
##                2.00-2.75 2.76-3.50 3.51-4.26 4.27-5.02 5.03-5.78 5.79-6.53
## Democratic (0)         2         7         6        11         0         1
## Republican (1)         2         8         4         7         1         0
##                6.54-7.31
## Democratic (0)         0
## Republican (1)         1
```

Perform a Chi-squared Test

Finally, we will perform a Chi-squared test on our two variables to test for independence.

Let O_{jk} denote the number of observations in cell (j,k). Let E_{jk} denote the number of observations in cell (j,k).

Under the null hypothesis, if un_rate and pol_maj are independent, then the probability of each cell is the product of the marginal probabilities:

P{un_rate = j and pol_maj = k} = P{un_rate = j} * P{pol_maj = k} for all j,k

Observations in each cell and marginal probabilities:

	2.00-2.75	2.76-3.50	3.51-4.26	4.27-5.02	5.03-5.78	5.79-6.53	6.54-7.31	Total	Marg. Prob
Democratic	2	7	6	11	0	1	0	27	27/50
Republican	2	8	4	7	1	0	1	23	23/50
Total	4	15	10	18	1	1	1	50	
Marg. Prob	4/50	15/50	10/50	18/50	1/50	1/50	1/50		

The expected number of observations in each cell is the sample size times the cell probability.(E_{jk} = n * P{un_rate = j and pol_maj=k})

Degrees of Freedom = (r-1)*(c-1) Since r = 2 and c = 7, DoF = (2-1)(7-1) = 6

Create rejection region:

According to the Chi-Squared lookup table,

P{X_6^2 > 12.59} = 0.05; therefore, reject if test statistic > 12.59.

Calculate the value of the test statistic:

E_{jk} = P{un_rate = j}P{pol_maj=k}*50 for j = 1,2 and k = 1,2,3.

		D/R	Un. Rate	Expected	Observed	(E-O)^2/E
Democratic	2.00-2.75	24	4	2.16	2.00	0.01185185
Democratic	2.76-3.50	24	15	8.10	7.00	0.14938272
Democratic	3.51-4.26	24	10	5.40	6.00	0.06666667
Democratic	4.27-5.02	24	18	9.72	11.00	0.16855967
Democratic	5.03-5.78	24	1	0.54	0.00	0.54000000
Democratic	5.79-6.53	24	1	0.54	1.00	0.39185185
Democratic	6.54-7.31	24	1	0.54	0.00	0.54000000
Republican	2.00-2.75	26	4	1.84	2.00	0.13913044
Republican	2.76-3.50	26	15	6.90	8.00	0.17536231
Republican	3.51-4.26	26	10	4.70	4.00	0.10425532
Republican	4.27-5.02	26	18	8.28	7.00	0.19787439
Republican	5.03-5.78	26	1	0.46	1.00	0.63391304
Republican	5.79-6.53	26	1	0.46	0.00	0.46000000
Republican	6.54-7.31	26	4	0.46	1.00	0.63391304
			sum	50.00	50.00	4.06150000

Because $X_2^0 = 4.0615 < 12.59$, we fail to reject the null hypothesis, H_0. The data supports the claim that a state's unemployment rate is independent of the political majority of the state.

Verify test with R:

In order to ensure we performed our Chi-Squared test correctly, we will verify our calculations with R code.

```
chisq.test(x)
```

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

```
##
##  Pearson's Chi-squared test
##
## data:  x
## X-squared = 4.0615, df = 6, p-value = 0.6683
```

Perform a Chi-squared Test with simulated p-value (based on 2,000 replicates)

Because the above warning our approximation may be incorrect, we will further test our variables by simulating the p-value for 2,000 replicates.

```
chisq.test(x, simulate.p.value = TRUE)
```

```
##
##  Pearson's Chi-squared test with simulated p-value (based on 2000
##  replicates)
##
## data:  x
## X-squared = 4.0615, df = NA, p-value = 0.8341
```

Note that the p-values are similar enough to provide evidence of the same conclusion, so in this situation the initial Chi-squared approximation is most likely adequate. However, because we are using simulation to back this, there is still a small probability that the approximation is incorrect.

The relatively large p-values imply that we fail to reject the null hypothesis that the variables un_rate and pol_maj are independent. Therefore, according to our results, we can conclude that the unemployment rate in a state is independent of the political majority of that state.

Conclusion

In conclusion, we failed to reject our null hypothesis that the unemployment rate of a state is independent of that state's political majority. This is not surprising as it is very difficult to capture the true political majority of a state with just a single variable. When calculating political majority of states, we simply took the average party vote in the last 5 presidential elections. This does not take into account any state or local voting records or any midterm elections held in the state. While this data set does not provide evidence of any dependency between the two variables, perhaps a further study can be done to determine a more accurate and in-depth measurement of each state's political majority and how it is related to the state unemployment rate.

Module 3 Conclusion

In this module we investigated a few of our variables from the data and attempted to answer some questions by performing hypothesis tests on the population mean of our variables as well as the independence between two variables.

In our first test, we performed a single sample, one-sided hypothesis test on the population mean of state unemployment data with the population variance unknown. We ultimately failed to reject the null hypothesis that the United States' average unemployment rate is lower than the rest of the world's unemployment rate. As mentioned before, this did not surprise us as the United States is one of the economic powerhouses in the modern world. Additionally, we calculated a 95% confidence interval of the population mean and also calculated the power of the hypothesis test.

In our second test, performed a single sample, two-sided hypothesis test on the population mean of state unemployment data in 2016 with the population variance unknown. We comfortably rejected the null hypothesis in favor of the alternate hypothesis and concluded that the population mean unemployment rate from 2015 is not equal to the unemployment rate from 2016, which indicates that, contrary to the studies done by the economists, the economic fluctuation period has not yet leveled off for the United States.

In our third and final test, we performed a Chi-Squared hypothesis test of independence on two of our variables, un_rate and pol_maj, in order to determine whether or not these two variables were independent. We ultimately failed to reject the null hypothesis that these were independent, but suggested that a further study be done while using a more comprehensive calculation of political majority by state.

Overall, our results were close to what we expected in tests 1 and 3, but we were surprised by the results of our second test. We rejected the null hypothesis in this test while we failed to reject in the other two. Next, we will attempt to answer questions about state unemployment rate while performing hypothesis test on 2 samples within our data set.

Module 4: Two Sample Analysis

In this module, we will investigate our unemployment variable (un_rate) by performing a series of hypothesis tests on two samples from within the our data set. We will attempt to answer 3 questions about our data set in this module.

In our first test, we ask the question, "Does a higher minimum wage raise the unemployment in a state?" We will conduct a one-sided hypothesis test on the difference in the population mean of two samples, states with a minimum wage greater than $7.25 per hour, and states with a minimum wage that is less than or equal to $7.25 per hour. We expect that states with higher minimum wages suffer from higher unemployment rates due to the economic impact of a higher minimum wage on business within a state. The proposed explanation behind the theory of this idea is that the more you force an employer to pay their employees, the fewer employees and employer can employ in their workforce.

In our second test, we ask the question, "Does the unemployment rate differ between western states and eastern states?" We will conduct a two-sided hypothesis test on the difference in the population variances of unemployment rate of two samples, broken up into "western" and "eastern" states. Our team would like to investigate the differences in the population variance between these two samples to determine if location has an effect on the variance of the unemployment rates of states.

In our third test, we ask the question, "Is there a statistically significant difference between the proportions of states with higher unemployment rate that have higher welfare spending vs. states that have lower welfare spending?" We will conduct a two-sided hypothesis test on the equality of proportions of high unemployment rate between two samples from our data: states that spend high amounts of their annual GDP on welfare and states that spend low amounts of their annual GDP on welfare.

Hypothesis Test 1: Does a higher minimum wage raise the unemployment rate in a state?

How does the minimum wage in a particular state affect that state's unemployment rate? In this section, we will investigate whether or not having a higher minimum wage negatively effects the unemployment rate in any given state. In order to do this, we will perform a 2 sample, one-sided hypothesis test on the difference in population means of our two samples with the population variance unknown.

Creation of 2 Distinct Samples using min_wage

First, our team decided to take a look at the distribution of our current variable that accounts for the minimum wage of each state, min_wage.

```
hist(min_wage, col = 'lightgrey', breaks = 7, xlab = 'State Minimum Wage (U.S
. Dollars) ', main = 'Minimum Wage of States in the U.S.')
```

The value of the mode of our variable min_wage is equal to $7.25. As this value appears to be the standard value for the majority of the states, we decided to group our sample into

two distinct groups of states: those with a minimum wage at or below $7.25 (22 states total) and those with a minimum wage above $7.25 (28 states total). We shall define these samples as LowWage and HighWage respectively.

The summaries of our 2 samples is shown below where M represents the sample mean of un_rate, S represents the sample standard deviation of un_rate, and N represents the sample size, each being denoted with a subscript "l" or "h" in order distinguish between the other.

LowWage:

$M_l = 4.64$

$S_l = 0.93$

$N_l = 22$

HighWage:

$M_h = 4.47$

$S_h = 1.08$

$N_h = 28$

Assumptions

Before beginning our hypothesis test, we made three assumptions that are stated below:

1. Each sample is a random sample
    - The observational units of the population we are interested in examining is the 50 states of the United States of America in 2016. We will use all 50 data points representing each state's unemployment rate in 2016 and consider it a random sample in time of the probability model that represents the distribution of the average U.S. unemployment rate since the U.S. began recording the rate itself. We then split the random sample into two samples by the min_wage variable in order to perform our hypothesis test.

2. The observational units (state unemployment rates) and two samples (less than or equal to 7.25 and greater than 7.25) are independent of each other.
    - This assumption may not be true, as explained in the assumptions of Module 3, hypothesis test 1.
3. The data is distributed normally.
    - Refer to module 3 analysis on the normality of the data.

Claim and Hypothesis Test

For this test, our claim is that states with a higher minimum wage will have a higher unemployment rate than states with a lower minimum wage.

In order to test this, we will perform a one-sided, 2 sample hypothesis test on the difference of population means with the population variance unknown.

Claim: population mean of un_rate @ higher min_wage > population mean of un_rate @ lower min_wage

H_0: The population mean unemployment rate of states with higher minimum wages is less than or equal to the population mean unemployment rate of states with lower minimum wages.

H_A: The population mean unemployment rate of states with higher minimum wages is higher than the population mean unemployment rate of states with lower minimum wages.

H_0: $\mu_h - \mu_l \leq 0$

H_A: $\mu_h - \mu_l > 0$

We will use a standard alpha value of 0.05 as this is standard in the industry.

Before we continue any further with our hypothesis test, we must first take a closer look at our assumption that the population variance is unknown. This is a crucial step in determining our degrees of freedom for our T distribution.

$S_h = 1.08$

$S_l = 0.93$

Because the sample variances are not equal, we cannot assume equal variance and thus we cannot assume we know the population variance.

Therefore, our test statistic for a two sample, one-sided hypothesis test on the difference in population means with the population variance unknown is equal to:

$$T_0 = \frac{(x_h - x_l) - (\mu_h - \mu_l)}{\sqrt{\frac{s_h^2}{N_h} + \frac{s_l^2}{N_l}}}$$

with $T_0 \sim t_v$ where v is equal to the degrees of freedom calculated by the formula:

$$v = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(\frac{s_1^2}{n_1})^2}{n_1 - 1} + \frac{(\frac{s_2^2}{n_2})^2}{n_2 - 1}}$$

Using our formula for the test statistic, we can calculate our T_0 using $M_h = 4.47$ $M_l = 4.64$, $S_h = 1.08$, $S_l = 0.93$, $N_h = 28$, and $N_l = 22$.

$$T_0 = \frac{(4.47 - 4.64) - (0)}{\sqrt{\frac{1.08^2}{28} + \frac{0.93^2}{22}}} = -0.5974$$

We know that T_0 is equal to -0.5974 and has a t-distribution with v degrees of freedom.

$v = 4.0559 \Rightarrow 4.00$ (by taking the floor of v)

$T_0 = -0.5974 \sim t_4$

We can then calculate our critical value, which we can compare to our calculated test statistic. Because the hypothesis test we are performing is one-sided, we will have one critical value denoted as t*.

Before we calculate our critical value, we know that if our test statistic is less than our calculated critical value, we will fail to reject the null hypothesis in favor of our alternative hypothesis. However, if the inverse is true, we will reject H_0 in favor of H_A.

In order to calculate t*, we must use set value of alpha to determine the rejection region that contains all of the critical values in our t distribution. We have set our alpha equal to 0.05. Therefore, we know that the probability of our t_4 distribution being less than our t* is equal to 0.95.

$P\{t_4 < t^*\} = 0.95$

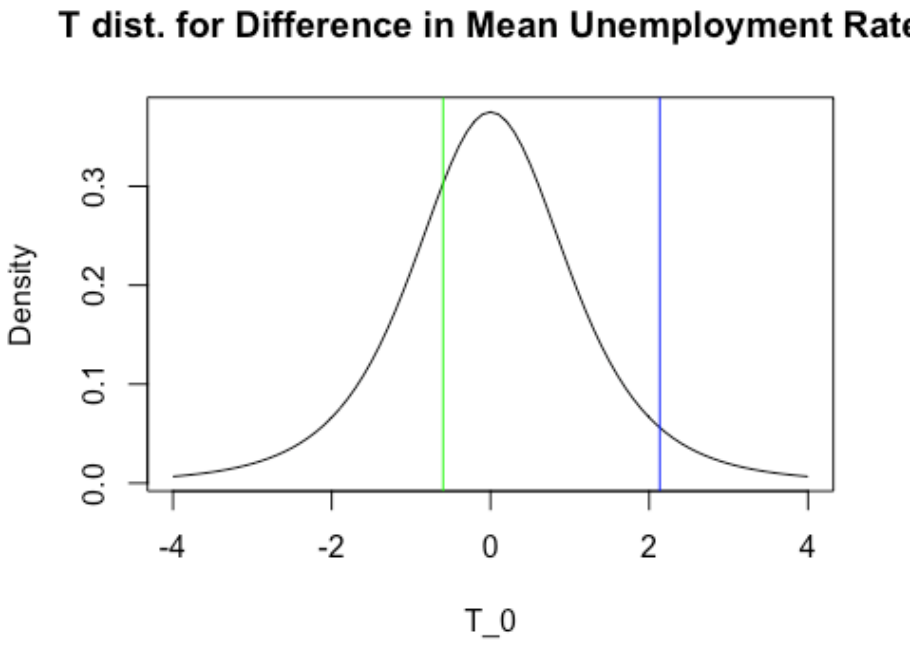
Next, we can use the invT function on a Ti-84 calculator or equivalent to calculate the value of t*.

invT(0.95,4) = 2.132

$T_0 = -0.5974 < t^* = 2.132$

Below is a graph showing the t_4 distribution. The plotted blue line is our critical value, t* and the plotted green line is our test statistic, T_0.

```
x=seq(from=-4,to=4,by=0.1)
y=dt(x,df=4)
plot(x,y,type="l", main="T dist. for Difference in Mean Unemployment Rate", y
lab = 'Density', xlab="T_0")
abline(v=2.132,col="blue")
abline(v=-0.5974,col="green")
```



Next, we can calculate the p-value of our model. Our p-value is defined as the probability of getting the same data we got or something more extreme. The area under the t distribution curve to the right of the green line is our p-value. We can calculate our p-value by using the Tcdf() function on a Ti-84 calculator or equivalent.

Tcdf(-0.5974,99999,4) = 0.7088

p-value = 0.7088 > alpha = 0.05

The calculated p-value is much greater than 0.05, and our test statistic was less than our critical value, so we fail to reject our null hypothesis in favor of our alternative hypothesis.

T_0 = -0.5974 < t* = 2.132 and p-value = 0.7088 > alpha = 0.05 ⇒ Fail to Reject H_0

In conclusion, we fail to reject the hypothesis that the population mean unemployment rate of states with higher minimum wages is less than or equal to the population mean unemployment rate of states with lower minimum wages. Ultimately, we were surprised at this conclusion as we expected the inverse to be true. However, the data does not support the claim that states with higher minimum wages suffer higher unemployment rates than states with lower minimum wages.

95% Confidence Interval on the Difference in Population Means

In this section, we will create a two-sided 95% confidence interval on the difference in population means of the state unemployment rate with the population variance unknown and assumed to not be equal. Our 2 sample of the 50 states in 2016 results in following values:

LowWage:

M_l = 4.64

S_l = 0.93

N_l = 22

HighWage:

$M_h = 4.47$

$S_h = 1.08$

$N_h = 28$

where m represents the sample mean of un_rate, s represents the sample standard deviation of un_rate, and n represents the sample size, each being denoted with a subscript "l" or "h" in order distinguish between the other.

As stated above, we will assume the distribution of the data is normal.

Quantity of Interest: average rate of unemployment for the 50 states of the United States of America

Our population parameter can be defined as: $\mu_h - \mu_l$

Because we have an unknown population variance and the two are assumed to be unequal, we will use a t statistic in order to calculate our critical value.

Using an alpha of 0.05 for the 95% confidence interval and 4 degrees of freedom for our T-test, we will compute the value of t* (t critical value) using the invT function on a Ti-84 calculator:

invT(0.025,4) = -2.776

Because of the two-sided confidence interval, we also know our upper critical value is equal to 2.776.

We will use the following formula to calculate the lower and upper bounds of our two-sided 95% confidence interval on the population mean:

$P(\mu_h - \mu_l - (t_{(\frac{\alpha}{2},\nu)} * \text{standard error}) \leq \mu_h - \mu_l \leq \mu_h - \mu_l + (t_{(\frac{\alpha}{2},\nu)} * \text{standard error})) = 1 - \alpha$

where standard error is defined as:

$$\text{standard error} = \sqrt{\frac{S_h^2}{N_h} + \frac{S_l^2}{N_l}}$$

$P(-0.17 - (2.776 * 0.2846) \leq \mu_h - \mu_l \leq -0.17 + (2.776 * 0.2846)) = 0.95$

$P(-0.9599 \leq \mu_h - \mu_l \leq 0.6199) = 0.95$

In conclusion, the 95% confidence interval on the difference in population means of the state unemployment rate with the population variance unknown and assumed to not be equal based on this data is [-0.9599, 0.6199]. This means if we repeated this experiment of drawing a random samples of size 28 and 22, thousands of times, and created a confidence interval for each data set, and knew the true difference in population means, approximately 95% of the confidence intervals would contain the true difference in population means.

Hypothesis Test 2: Does the unemployment rate differ in western states versus eastern states?

How does the unemployment rate change as you move from western states to eastern states? In this section, we will investigate whether or not the geographical location of a particular state affects the unemployment rate in that state. In order to do this, we perform a 2 sample, two-sided hypothesis test on the difference in population variance of the unemployment rate of our two samples with the population mean and variances of each sample being unknown.

Creation of 2 Distinct Samples using st_reg

Before beginning our analysis, we created 2 samples of states from our initial data set, Unemployment_Data. We chose to study the differences between states we considered to be "western" states and "eastern" states. For any states with the value of "Midwest" or "West" for their st_reg variable, we placed in the "western" sample. Any states with the values of "South" or"Northeast" we placed in the "eastern" sample. The summary of each sample is shown below, where V is equal to the sample variance and N is equal to the sample size, each denoted with either a "w" or "e" to denote classification:

Western:

$V_w = 1.30$

$N_w = 25$

Eastern:

$V_e = 0.67$

$N_e = 25$

Assumptions

Before beginning our hypothesis test, we made three assumptions that are stated below:

1. Each sample is a random sample
   - Refer to Module 3
2. The observational units (state unemployment rates) and two samples (eastern versus western are independent of each other.
   - This assumption may not be true, as discussed in Module 3, hypothesis test 1.
3. The data is distributed normally.
   - Refer to Module 3 analysis on the normality of the data.

Hypothesis Test Calculations

For this test, we would like to investigate the differences in population variance of unemployment rate between our two samples: eastern and western states of the United States. We expect that the variances will not be significantly different and that they will actually be very similar.

In order to test this, we will perform a two-sided, 2 sample hypothesis test on the difference of population variances of unemployment rate, with the population means and variances of both samples unknown.

H_0: The population variance of the unemployment rate of western states is equal to the population variance of the unemployment rate of eastern states. H_A: The population variance of the unemployment rate of western states is not equal to the population variance of the unemployment rate of eastern states.

H_0: $V_w = V_e$

H_A: $V_w \neq V_e$

We will use a standard alpha value of 0.05 as this is the industry standard.

Our test statistic for a two sample, two-sided hypothesis test on the difference in population variances with the population means and variances unknown is equal to:

$F_0 = \frac{V_w}{V_e}$

with $F_0 \sim f(u,v)$ where u,v are equal to the degrees of freedom calculated by the formula:

$u = N_w - 1$

$v = N_e - 1$

Using our formula for the test statistic, We can calculate our F_0 using $V_w = 1.30$, $V_e = 0.67$, and $N_w = N_e = 25$.

$F_0 = \frac{1.30}{0.67} = 1.940$

We know that F_0 is equal to 1.940 and has a f-distribution with u,v degrees of freedom.

$u = 24$

$v = 24$

$T_0 = 1.940 \sim f(24,24)$

We can then calculate our critical value which we can compare to our calculated test statistic. Because the hypothesis test we are performing is two sided, we will have two critical values denoted as f1* and f2*.

Before we calculate our critical values, we know that if our test statistic is within the bounds of our calculated critical values, we will fail to reject the null hypothesis in favor of our alternative hypothesis. However, if the inverse is true, we will reject H_0 in favor of H_A.

In order to calculate our critical values, we must use the set value of alpha to determine the rejection region that contains all of the critical values in our f distribution. We have set our alpha equal to 0.05. Because the test is two-sided, we will divide our alpha value by two giving us an alpha value of 0.0250 for each side. Therefore, we know that the probability of our f(24,24) distribution being less than our f1* is equal to 0.250, and the probability of our f(24,24) distribution being greater than our f2* is equal to 0.975.

P{f(24,24) > f1*} = 0.250

P{f(24,24) < f2*} = 0.250

Next, we can use the F distribution look up tables to find the values of our f1 based on our degrees of freedom, u and v.

P{f(24,24) > 2.27} = 0.250 (using look up table)

Then, we will use the following formula to calculate the value of f2, given f1:

$$f_{1-\alpha,u,v} = \frac{1}{f_{\alpha,v,u}}$$

$$f_{1-\alpha,u,v} = \frac{1}{2.27} = 0.4405$$

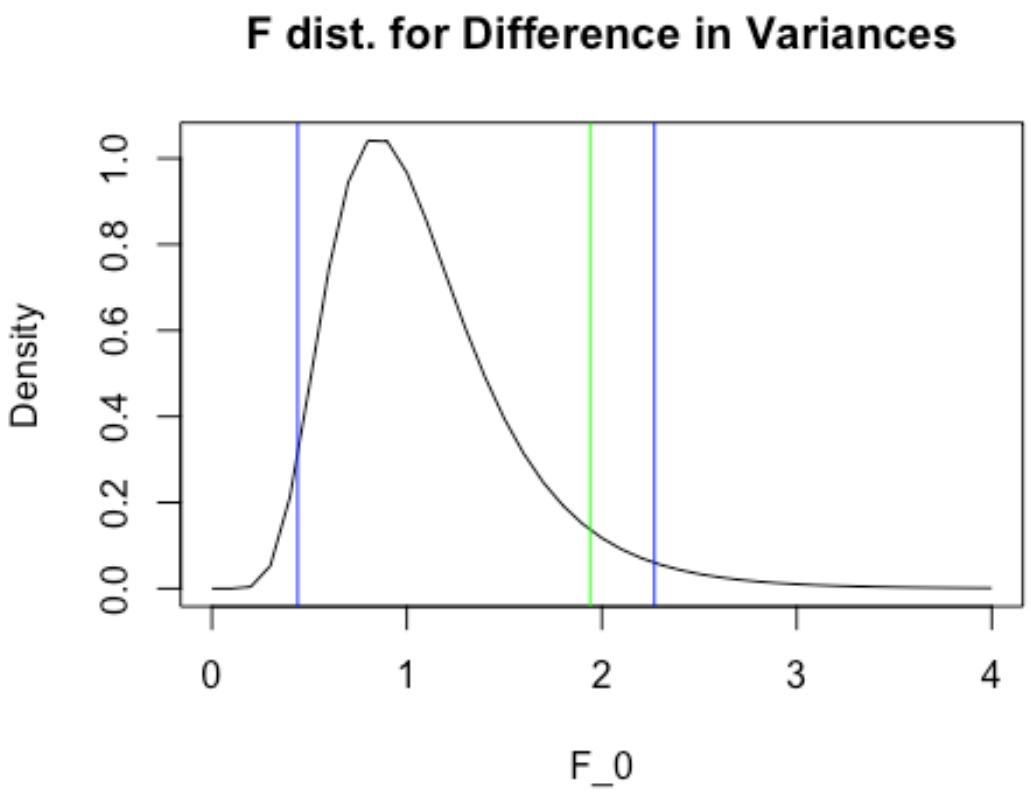
P{f(24,24) < 0.4405} = 0.250

$f2^* < F_0 < f1^*$

0.4405 < 1.940 < 2.27

Below is a graph showing the f(24,24) distribution. The plotted blue lines are the critical values, f1* and f2* and the plotted green line is our test statistic, F_0.

```
x=seq(from=0,to=4,by=0.1)
y=df(x, 24, 24)
plot(x,y,type="l", main="F dist. for Difference in Variances", xlab = "F_0",
ylab = 'Density')
abline(v=c(0.4405,1.940, 2.27), col=c("blue", "green", "blue"))
```



Next, we can calculate the p-value of our model. Our p-value is defined as the probability of getting the same data we got or something more extreme. The area under the f distribution curve to the right and left of the green line is our p-value. We can calculate our p-value by using a statistical software package or online calculator. We will use the p-value calculator for an F-test from www.danielsoper.com.

p-value = 0.0557

The calculated p-value is greater than our alpha, and our test statistic fell within the region of our critical values [0.4405, 2.270], so we fail to reject our null hypothesis in favor of our alternative hypothesis.

F_0 = 1.940 < f1*, 1.940 > f2*, p-value = 0.0557 > alpha = 0.05 ⇒ Fail to Reject H_0

In conclusion, we fail to reject the hypothesis that the population variance of the unemployment rate of western states is equal to the population variance of the

unemployment rate of eastern states. The data does not support the claim that the population variances of unemployment rate for western and eastern states are different. However, based on our results from the test, we have reason to question our final answer as our F_0 was very close to our f1* value and our p-value was very close to our alpha value of 0.05. A mathematician would conclude to move forward and fail to reject the null hypothesis, but as engineers, we are not satisfied with our conclusion because our alpha of 0.05 is simply a number that has become an industry standard for the value of a type I error and should not be considered as an absolute cut off point. We recommend that further analysis be done to challenge the null hypothesis and to continue to study the question whether or not location (west vs. east) does indeed affect a state's unemployment rate.

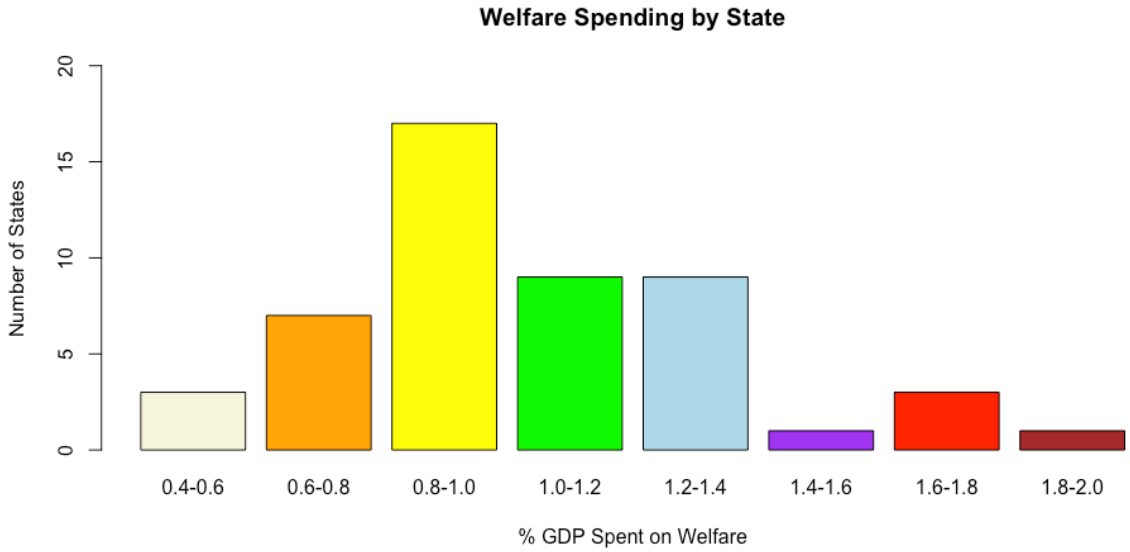
Hypothesis Test 3: Is there a statistically significant difference between the proportions of states with higher unemployment rate that have higher welfare spending vs. states that have lower welfare spending?

In this section, we will investigate whether there is a statistically significant difference in the proportions of states with higher unemployment rate based on proportions recorded from two samples: states with higher welfare spending and states with lower welfare spending. We will do this by performing a two-sided, two sample hypothesis test on the difference between population proportions.

Determining cutoff points and creating proportions

In order to determine what is considered "low" or "high" welfare spending, we must first determine what a sensible cutoff point would be for our data set. We will do this by looking at a barplot of our welf_spent variable.

```
tb = table(Unemployment_Data$welf_spent)
barplot(tb, col=c("beige","orange", 'yellow', 'green', 'lightblue', 'purple',
'red', 'brown'),ylab="Number of States", xlab = '% GDP Spent on Welfare', mai
n = "Welfare Spending by State", ylim=c(0, 20))
```



In order to split at a point where we achieve similar sample sizes, we will choose to split the states into two groups: states who spent less than or exactly 1% of their annual GDP in 2016 on welfare, and states that spent more than 1% of their annual GDP on welfare.

Now, our two sample sizes consist of the following:

Low Welfare Spending:

$N_l = 27$

High Welfare Spending:

$N_h = 23$

where N stands for sample size and is subscripted "l" or "h" to distinguish "low" or "high".

Next, we will determine the cutoff for what is to be considered "high" and "low" unemployment rate. According to an article written in 2011 by economists Justin Weidner and John C. Williams of the Federal Reserve Bank of San Fransico, the current "normal" unemployment rate is approximately equal to 5%. We will use 5% as the cutoff to determine whether or not a state has a "high" or "low" unemployment rate.

Upon examining the data, we determined the following counts of "high" unemployment states (states with an un_rate ≥ 5.00) for each of our samples:

Low Welfare Spending:

$Y_l = 13$

High Welfare Spending:

$Y_h = 9$

where Y stands for the number of "Yes" responses (states with an un_rate ≥ 5.00) and is subscripted "l" or "h" to distinguish "low" or "high".

Therefore, our two sample proportions can be defined as follows:

Low Welfare Spending: $P_l = \frac{Y_l}{N_l} = \frac{13}{27}$

High Welfare Spending: $P_h = \frac{Y_h}{N_h} = \frac{9}{23}$

where "P" stands for sample proportion and is subscripted "l" or "h" to distinguish "low" or "high".

Assumptions

Before beginning our hypothesis test, we made three assumptions that are stated below:

1. Each sample is a random sample
   - This assumption discussed in Module 3, hypothesis test 1.
2. The observational units (state unemployment rates) and two samples (eastern versus western are independent of each other.
   - This assumption may not be true, explained in Module 3, hypothesis test 1.
3. The data is distributed normally.
   - Refer to module 3 analysis on the normality of un_rate.
   - Refer to module 2 analysis on the distribution of welf_spent. We chose to model this variable with a Uniform distribution in Module 2 and ultimately decided that the Uniform is not an adequate model. However, we do know that a larger sample size would cause welf_spent approach a Normal distribution. Therefore

we will move forward as if this assumption about welf_spent being Normally distributed is true to show our ability to conduct this hypothesis test.

Claim and Hypothesis Test Calculations

Our claim for this scenario is that there is a statistically significant difference in proportions of high unemployment rate between states with high and low welfare spending.

Let our null and alternative hypotheses for this hypothesis test be the following:

H_0: $P_l = P_h$

H_A: $P_l \neq P_h$

We will compare the test to an alpha value of 0.05 as that value is the industry standard.

Additionally, we will not factor in the Yates continuity correction into our proportion test as our sample size is sufficiently large.

```
prop.test(c(13,9),c(27,23),alternative="two.sided",conf.level=0.95, correct=F
ALSE)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  c(13, 9) out of c(27, 23)
## X-squared = 0.4099, df = 1, p-value = 0.522
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.1842349  0.3645892
## sample estimates:
##    prop 1    prop 2
## 0.4814815 0.3913043
```

The proportion test calculated a p-value of 0.522 which is higher than our alpha value of 0.05. Therefore, we conclude to fail to reject the null hypothesis that the two proportions are equal. Ultimately, the data does not support the claim that there is a statistically significant difference in proportions of high unemployment rate between states with high and low welfare spending.

Module 4 Conclusion

In this module, we investigated our unemployment variable (un_rate) by performing a series of hypothesis tests on two samples from within the our data set. We attempted to answer the following three questions about our data set:

1. Does a higher minimum wage raise the unemployment in a state?
2. Does the unemployment rate differ between western states and eastern states?
3. Is there a statistically significant difference between the proportions of states with higher unemployment rate that have higher welfare spending vs. states that have lower welfare spending?

While it would be impossible to answer each of these questions with high certainty by performing a single hypothesis test on our sample of a much larger population, we still were able to draw conclusions that provided suggestions of possible answers to all of these questions.

In our next module, we will attempt to utilize our knowledge of linear regression to build a linear model that will predict the unemployment rate of a particular state based on a predictor variable from our data set.

Module 5: Linear Regression

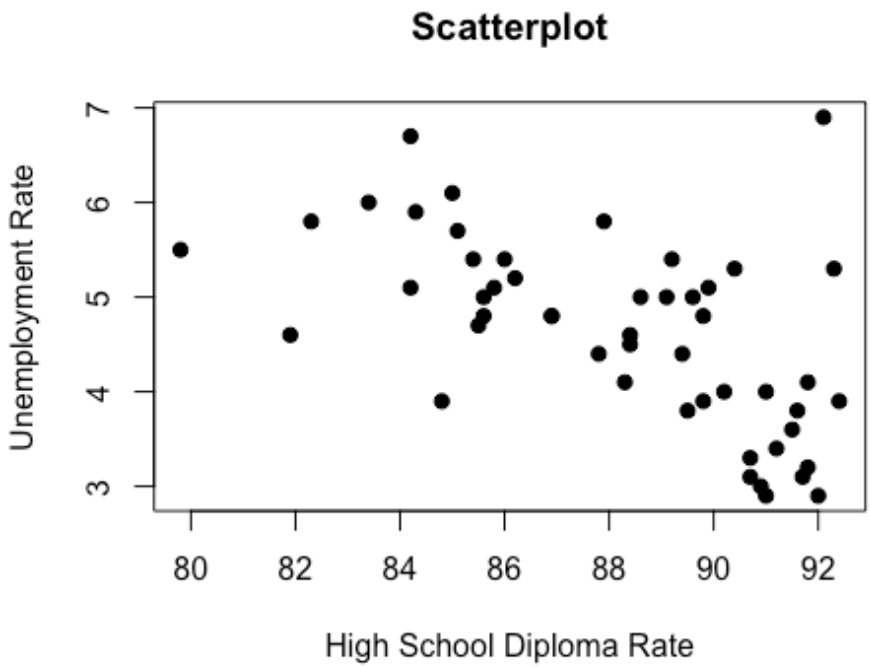
Linear regression helps us to learn about the relationship between two variables and create a numerical method to predict theoretical observations. Specifically, linear regression is an analysis technique to predict the value of a response variable as a linear function of predictor variables. In this module, we will perform a simple linear regression using unemployment rate as our response variable. To display our understanding of both simple and multiple linear regression, we will also include a multiple regression model at the end of the module to investigate the relationship of multiple predictor variables to our response variable. We would like to analyze the relationship between the unemployment rate and the percentage of the population who received a high school diploma. With linear regression, we hope to answer the following question: can we use the high school diploma rate to predict the unemployment rate? This is a valuable question to ask because it could further support the argument that an unhealthy economy (one with high unemployment) affects more than just the current workforce. In fact, it could have effects on our younger generations in a way we have not noticed before. This analysis could help shed light on the far reaching effects of increased unemployment rates.

Can we use the high school diploma rate to predict the unemployment rate of a state?

Our predictor variable will be the high school diploma rate, and our response variable will be the unemployment rate. First, let's plot the unemployment rate as a function of high school diploma rate. This step is important because we can visually check if there might be a correlation between the two variables.

A scatterplot of the variables along with the calculated correlation coefficient is shown below:

```
plot(dip_hs, un_rate, main = 'Scatterplot', xlab = 'High School Diploma Rate'
, ylab = 'Unemployment Rate', pch = 19)
```

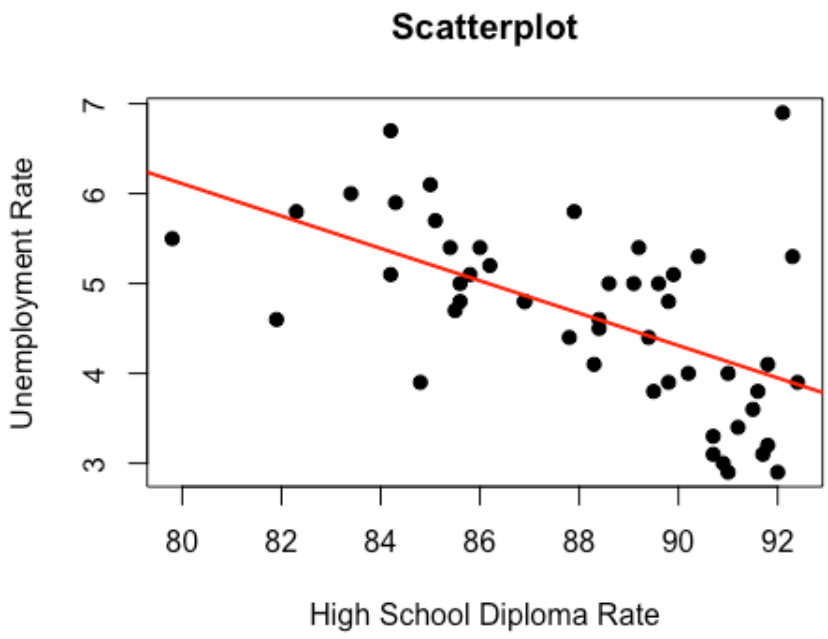



```
correlation_coeff = cor(dip_hs, un_rate)
correlation_coeff
```

```
## [1] -0.5725387
```

We conclude from the graph that the two variables appear to have a negative moderately, negative relationship to one another. The calculated correlation coefficient of -0.5725387 confirms the visual interpretation that the variable in fact have a moderately strong, negative relationship to one another. Below is the same scatterplot with the line of best fit added:

```
plot(dip_hs, un_rate, main = 'Scatterplot', xlab = 'High School Diploma Rate'
, ylab = 'Unemployment Rate', pch = 19)
model2 = lm(un_rate ~ dip_hs)
abline(model2, lwd = 2, col = "red")
```



Assumptions

When fitting a regression model, there are a few assumptions that we must make about the errors up front in order to perform the regression model. An analysis of the residuals will be helpful to check our assumptions and thus evaluate the adequacy of our linear regression model. We will perform that analysis towards the end of this module. We will make the following three assumptions about the residuals that are stated below:

1. The residuals are normally distributed.
   - We will investigate this assumption at the end by looking at a histogram and Q-Q plot of the data set.
2. The mean of the residuals is 0.
   - We know this always to be true, so verifying it at the end will not be necessary.
3. The residuals have a constant variance.
   - We will also investigate this assumption at the end by looking at a plot of the residuals to confirm no perceivable pattern arises, which would indicate a non-linear relationship between our predictor and response variables.

Linear Regression

To view the linear regression model in R:

```
model2 = lm(un_rate ~ dip_hs)
summary(model2)

##
## Call:
## lm(formula = un_rate ~ dip_hs)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.34717 -0.49883 -0.06871  0.48888  2.96841
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 20.52955    3.28588   6.248 1.05e-07 ***
## dip_hs      -0.18022    0.03725  -4.838 1.40e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8218 on 48 degrees of freedom
## Multiple R-squared:  0.3278, Adjusted R-squared:  0.3138
## F-statistic: 23.41 on 1 and 48 DF,  p-value: 1.396e-05

cor(dip_hs, un_rate)

## [1] -0.5725387
```

Using the information above, our least squares regression line is the following:

$\widehat{un\_rate} = 20.5296 + -0.18022 * dip\_hs$

This equation suggests that for one unit increase in high school diploma rate, we can expect the unemployment rate to decrease by -0.18022.

The model also does an analysis on β_1, or the slope, via a hypothesis test. The HT tests the following hypotheses:

H_0: $\beta_1 = \beta_2 = \ldots = \beta_k = 0$

H_A: $\exists_j$ s.t. $\beta_j \neq 0$

If we fail to reject the null hypothesis in favor of the alternative hypothesis, this would mean we do not have strong enough evidence to confirm that there is actually a correlation between the variables. The F test statistic is 23.41, and the p-value for this test statistic is approximately zero. Since the p-value is statistically significant (as indicated by the three asterisks) at 1.396e-05 and is less than our alpha value of 0.05, we reject Ho in favor of Ha. Our slope is significant enough to conclude that there is a relationship between high school diploma rate and unemployment rate.

Furthermore, the adjusted R-squared value is 0.3138. This value is interpreted to mean that 31.38% of the variation observed in unemployment rate is accounted for by our linear regression model. We know that it is very difficult to get an adjusted R-squared close to 1 using real-life data, so our value of 31.38% is good considering we are using real-life data.

We can also use the information from the linear model to create a two-sided 95% confidence interval on the slope estimate. The formula we will use to create a confidence interval is:

CI = estimate $\pm$ critical value * standard error

The estimate value is -0.18022 and the standard error value is 0.03725. We need to calculate the critical t value using our TI-84 calculator's invT function. We will use the standard alpha value of 0.05, but we will need to divide that value by two since the confidence interval is two-sided. Additionally, we know from above that we have 48 degrees of freedom. To calculate the critical value:

invT(0.025, 48) = -2.011

Thus, our CI is:

[-0.18022 $\pm$ 2.011*0.03725] = [-0.2551, -0.1053]

Therefore the two-sided 95% confidence interval for the slope estimate is [-0.2551, -0.1053]. We can interpret the confidence interval in the following way: if we repeated the experiment of drawing random samples of size n thousands of times, created a confidence interval for each data set, and knew the true value of the slope, when we checked those

thousands of confidence intervals, approximately 95% of those would contain the true slope value.

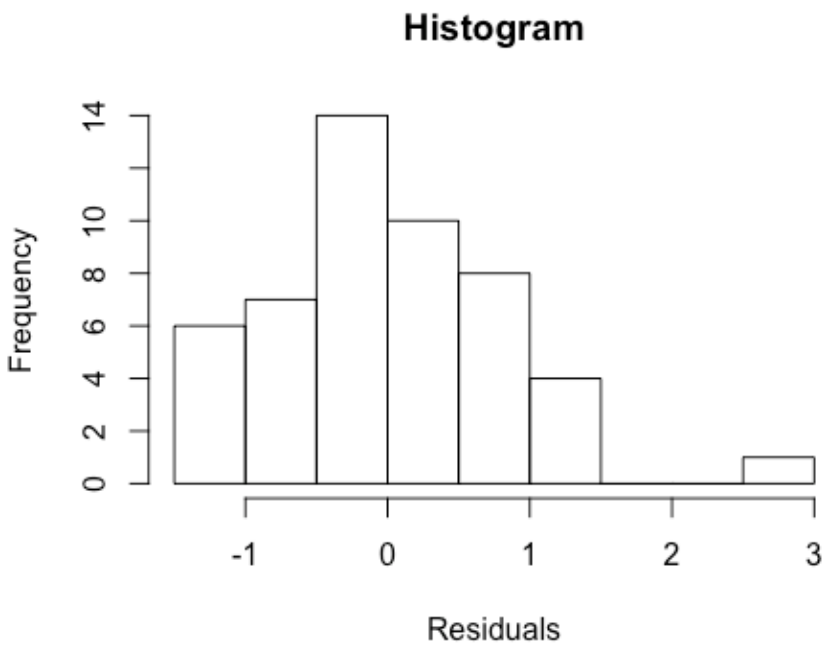
Checking the Assumptions

The above linear regression model was performed under the assumptions that our residual values are normally distributed, have a mean of 0, and have a constant variance. We will now investigate the residuals to ensure that our assumptions about them hold, which will evaluate the adequacy of our linear regression model. To verify that the residuals are normally distributed, we will analyze a histogram and Q-Q plot.

```
res = residuals(model2)
summary(res)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## -1.34717 -0.49883 -0.06871  0.00000  0.48888  2.96841
```

```
hist(res, main = "Histogram", xlab = "Residuals")
```



The histogram of the residuals has the general shape that we would expect normal data to have, with the exception of the bin to the far right of the graph. The majority of the points are located at the center but this data only has one tail to the right, rather than two tails

trailing off to each side. At first glance, the normal distribution appears to be an okay model for the residuals, but is certainly not a great fit.

To further confirm this assumption, we will look at a Q-Q plot of the residuals.

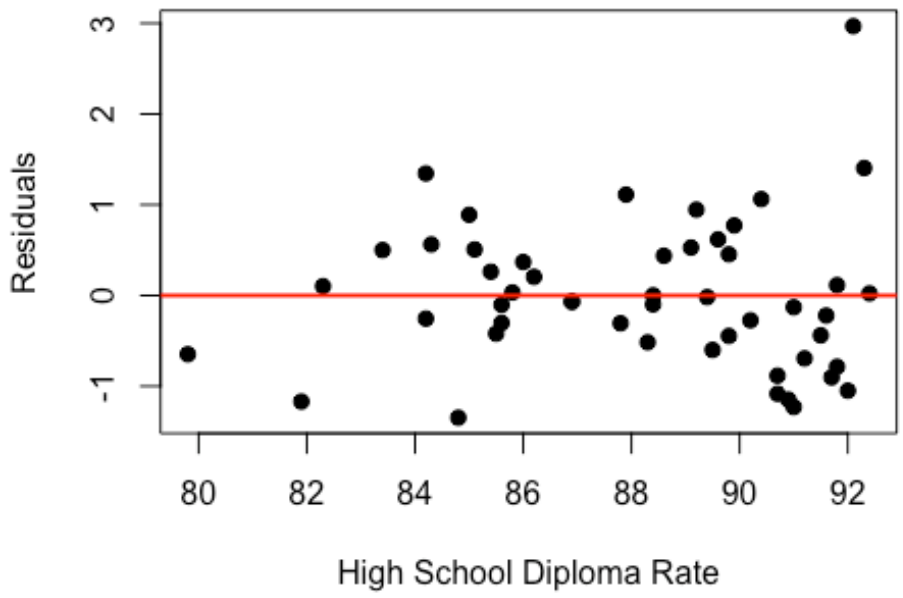
```
res = residuals(model2)
n = length(res)
mean = mean(res)
sd = sd(res)
limits = c(mean-3*sd,mean+3*sd)
probs = (1:n)/(n+1)
norm.quants = qnorm(probs, mean, sd)
plot(sort(res), sort(norm.quants), main = "QQ Plot", xlab = "Empirical Quanti
les", ylab = "Theoretical Quantiles", xlim=limits, ylim=limits, pch=16)
abline(0,1)
```

The Q-Q plot indicates a similar conclusion about the residuals. The data points in the middle of the graph fit tightly to the identity line, which is good considering this is where the majority of our data points lie and a requirement for the normal distribution. As we

move away from the center, the data points stray a small bit from the identity line but not enough to be significant. There are a few points at the left end that are a little farther from the line than we would want, but we make the judgement call that these points, alone, are not enough to say that the Normal distribution does not adequately enough model the residuals. The normal distribution appears to be an okay fit for the residuals but there are a few red flags that we would like to investigate further.

Now to determine if the variances are constant, we are going to plot the residuals as a function of the high school diploma rate. We are hoping to see data that looks randomly scattered about the graph with no distinguishable pattern, which would confirm that our predictor and response variables do not possess some kind of non-linear relationship. Patterns in the data indicate inconsistent variance values across the data, which would mean our assumption of constant variance does not hold among the residuals.

```
res = residuals(model2)
plot(dip_hs, res, pch = 19, xlab = "High School Diploma Rate", ylab = 'Residu
als')
abline(0,0, col = "red", lwd = 2)
```

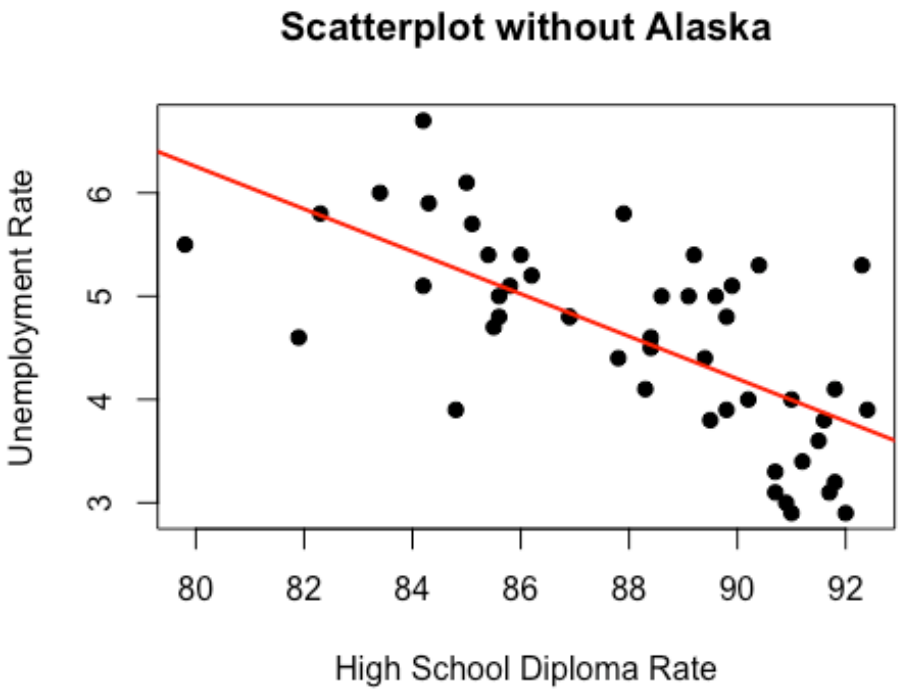


The residual plot shows a scatter of points that looks somewhat random. However, the points are clustered in the mid-to-bottom right hand corner of the graph with no points at all in the top left region of the graph. We do not feel totally comfortable using this graph to confirm the assumption that the residuals have a constant variance. As a result, failing to comfortably confirm the assumption of the residuals' variance calls into question the adequacy of the regression model. We will continue to investigate our model in hopes of increasing this adequacy.

Influential Points

It is important to check the model for influential points, which could help explain the residual's behavior in the above investigation. Influential points are ones that have a disproportionate effect on the slope of the least squares regression line. From our scatterplot above, we note one point, in particular, in the top right corner that could be influential, because it is much further from the line than other points. It lies roughly around (92, 7). After some investigation, we determined this point to be the state of Alaska. We will do a quick analysis to assess the impact of this point on our model, then decide whether or not it makes since to remove it moving forward. Below is a scatterplot without this point and the corresponding correlation coefficient:

```
new_data = Unemployment_Data[-c(2), ]
plot(new_data$dip_hs, new_data$un_rate, main = 'Scatterplot without Alaska',
xlab = 'High School Diploma Rate', ylab = 'Unemployment Rate', pch = 19)
model3 = lm(new_data$un_rate ~ new_data$dip_hs)
abline(model3, lwd = 2, col = "red")
```




```
new_correlation_coefficient = cor(new_data$dip_hs, new_data$un_rate)
new_correlation_coefficient
```

```
## [1] -0.680112
```

```
summary(model3)
```

```
##
## Call:
## lm(formula = new_data$un_rate ~ new_data$dip_hs)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.36970 -0.45019 -0.02963  0.49197  1.57212
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     22.70255    2.84869   7.969 2.83e-10 ***
## new_data$dip_hs -0.20558    0.03232  -6.360 7.66e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7014 on 47 degrees of freedom
## Multiple R-squared:  0.4626, Adjusted R-squared:  0.4511
## F-statistic: 40.45 on 1 and 47 DF,  p-value: 7.658e-08
```

Without including Alaska in the model, the correlation coefficient changes from -0.5725387 to -0.680112, a significant increase of magnitude of the correlation. Out least squares regression line has now changed from:

$\hat{un\_rate}$ = 20.5296 + -0.18022 * dip_hs

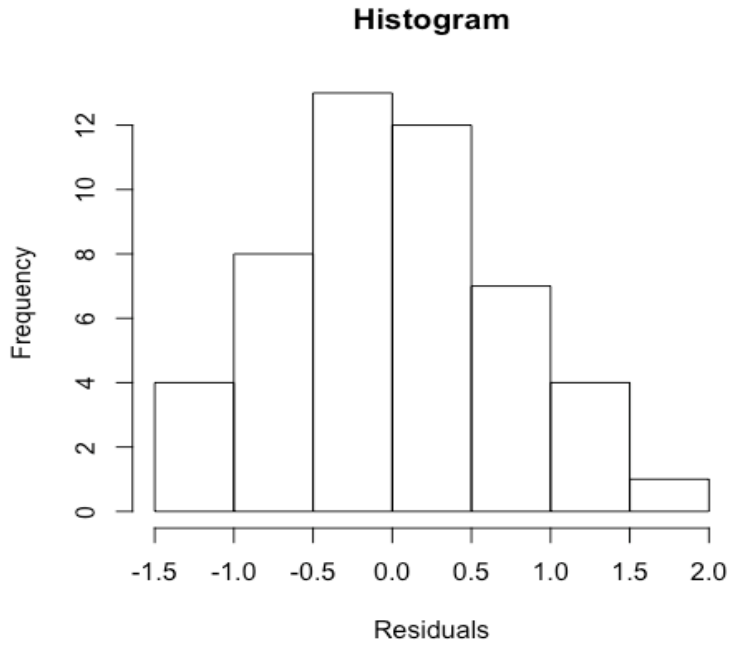
to

$\hat{un\_rate}$ = 22.70255 + -0.20558 * dip_hs

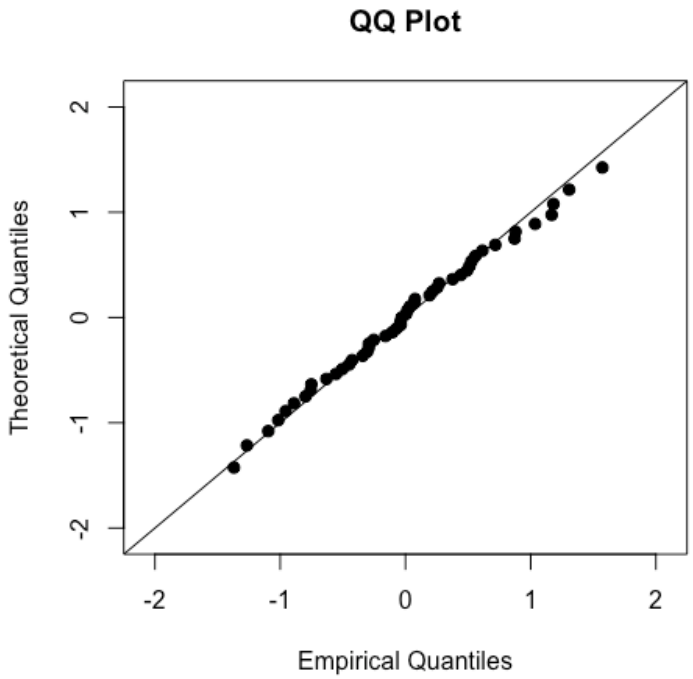
Clearly, although the data for the state of Alaska is not "bad data", it has a significant impact on the regression model. Alaska has some unique characteristics that make it very different from the rest of the continental states. It is frequently finds itself at the top of the U.S. unemployment rate list year after year. Alaska's main industries are dependent upon the land itself. The petroleum, fishing, mining, and timber industries rule the economy almost exclusively. These professions are largely seasonal, which contributes to a generally higher unemployment rate than other states. Additionally, a high number of employees in these industries are commuters from other states who flock to Alaska during the busy season for work and return to their home state in the off season. On top of this, Alaska is home to thousands of individuals who live primarily off the land with no true connection to the U.S. economic system. The bartering exchanges and small scale sales that occur between these people never make their way into the statistics. Alaska's location and unique industry characteristics explain why its unemployment rate is so different from the rest of the United States. Because it is so different, we feel comfortable removing this influential observation from our model moving forward.

To evaluate the residuals of the model with Alaska removed:

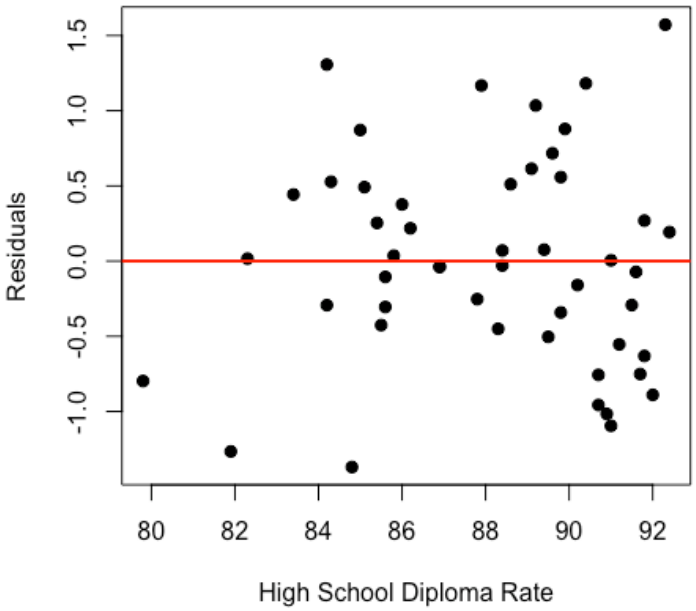
```
model2 = lm(new_data$un_rate ~ new_data$dip_hs)
res = residuals(model3)
#to check normality
hist(res, main = "Histogram", xlab = "Residuals")
```




```
n = length(res)
mean = mean(res)
sd = sd(res)
limits = c(mean-3*sd,mean+3*sd)
probs = (1:n)/(n+1)
norm.quants = qnorm(probs, mean, sd)
plot(sort(res), sort(norm.quants), main = "QQ Plot", xlab = "Empirical Quanti
les", ylab = "Theoretical Quantiles", xlim=limits, ylim=limits, pch=19)
abline(0,1)
```

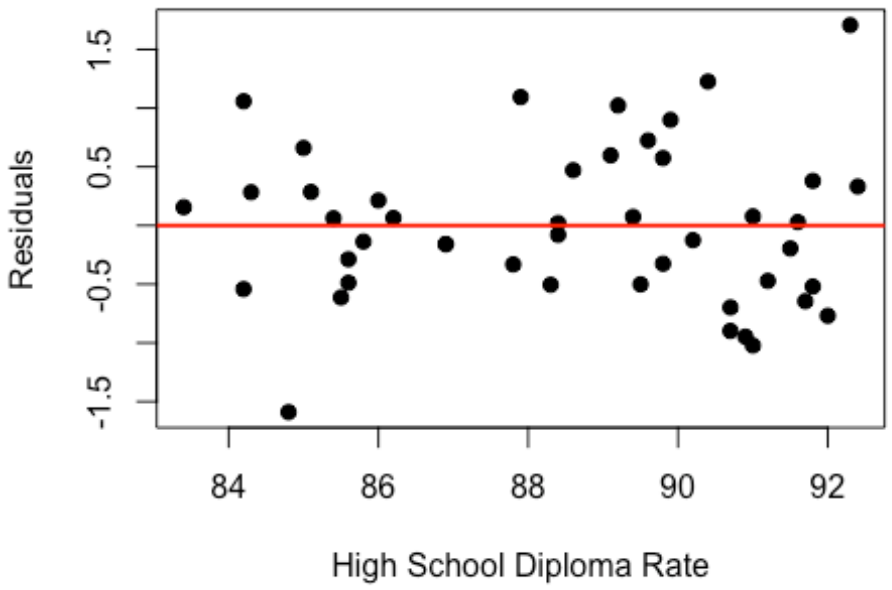



```r
#to check constant variance
plot(new_data$dip_hs, res, pch = 19, xlab = 'High School Diploma Rate', ylab
= 'Residuals')
abline(0,0, col = "red", lwd = 2)
```



The residuals are more normally distributed as evidenced by the normal histogram shape and the congregation of points in the middle and their fit to the identity line on the Q-Q plot. The removal of the influential point significantly improved the residual plot. The data points are now arranged in a pretty random pattern across the graph, which is what we are looking for. Although there is no clear pattern among the residuals, the graph does still show the points congregated to the right of the graph with very few points to the left of 82% high school diploma rate. This is happening because there is only one point occurring around 79%, 81%, and 82% high school diploma rate meaning that there are no other points at those values which we can use to compare the variance. Because of this, it is hard for our eyes to pick up on the true randomness that this graph is displaying. We will remove these lowest three high school diploma rates in order to better visualize the residual plot.

```
new_data_2 = Unemployment_Data[-c(2, 24, 43, 5), ]
model4 = lm(new_data_2$un_rate ~ new_data_2$dip_hs)
res = residuals(model4)
plot(new_data_2$dip_hs, res, pch = 19, xlab = 'High School Diploma Rate', yla
b = 'Residuals')
abline(0,0, col = "red", lwd = 2)
```



Clearly, there is no distinguishable pattern among the residual data points, which indicates that their variances are constant. We now feel comfortable confirming the original assumption that the variance of the residuals is constant. Since we now know that the residuals are normally distributed, have a mean of zero, and have a constant variance, we can say that high school diploma rate is an adequate predictor for unemployment rate.

Linear Regression Model in Action

To see our linear regression model in action, we will use it to predict the unemployment rate of the United States, as a whole. The high school diploma rate for the U.S. in 2016, according to the Bureau of Labor Statistics, was 86.7%. Therefore, we can predict the unemployment rate using our model:

$\hat{un\_rate}$ = 22.70255 + -0.20558 * 86.6 = 4.87%

Therefore, we predict the unemployment rate of the United States to be 4.87%.

Multiple Linear Regression

We will run a multiple linear regression model with our four continuous variables: high school diploma rate, minimum wage, median income, and college diploma rate. The model help us determine which variables are considered to be good predictors for our response variable, unemployment rate.

```
model1 = lm(un_rate ~ dip_hs + min_wage + med_inc + dip_college)
res = residuals(model1)
summary(model1)
```

```
##
## Call:
## lm(formula = un_rate ~ dip_hs + min_wage + med_inc + dip_college)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.69580 -0.40759 -0.00187  0.45912  2.07596
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.942e+01  3.568e+00   5.445 2.07e-06 ***
## dip_hs      -1.684e-01  4.239e-02  -3.972 0.000255 ***
## min_wage     8.787e-02  1.240e-01   0.708 0.482344
## med_inc      3.224e-05  2.214e-05   1.456 0.152228
## dip_college -8.622e-02  4.301e-02  -2.005 0.051050 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8122 on 45 degrees of freedom
## Multiple R-squared:  0.3844, Adjusted R-squared:  0.3297
## F-statistic: 7.026 on 4 and 45 DF,  p-value: 0.0001751
```

The above model indicates that high school diploma rate is the best predictor for unemployment rate. This variable is the only one with a significant p-value, one less than our alpha value of 0.05. Minimum wage and median income have rather large p-values. College diploma rate has a p-value that, as engineers, we would have to investigate as to whether or not this value is statistically significant within the context of this data. Although, a mathematician would consider the value to be non-significant (0.051 is numerically larger than 0.05). If we were to employ backwards elimination on this multiple regression model, we would remove one value at a time, starting with the one with the highest p-value (minimum wage), and work our way backwards to the variable that is most significant.

The least squares regression line for this model is:

$\hat{un\_rate}$ = 1.942e+01 + -1.684e-01 * dip_hs + 8.787e-02 * min_wage + 3.224e-05 * med_inc + -8.622e-02 * dip_college

This model supports our decision to investigate high school diploma rate as the predictor variable in our single linear regression model above, as it proves to be the best predictor for the unemployment rate response variable within our date set.

Module 5 Conclusion

In this module, we used a simple linear regression to show that the unemployment rate of a state can be adequately predicted by the high school diploma rate. The correlation between the two variables is -0.68, which indicates a negative, linear relationship. Our analysis is evidence that there is a clear, economic benefit to higher high school diploma rates. Even those individuals not directly, or currently, affected by high school graduation rates have a vested interest in seeing to it that our youth to complete high school. Additionally, we would argue that investing in our school systems and educators to promote high school success rates will have a direct and positive impact on the health of our economy.

Module 6: Summary and Conclusion

As we mentioned earlier, the unemployment rate is a strong indicator of the health of an economy. As individuals who would be directly, and negatively, impacted by a poor economy, we have a vested interest in maintaining and improving our respective state's unemployment rate. Throughout the last 5 modules, our goal has been to investigate the current U.S. economy by analyzing unemployment rate at a state level and its relationship to other, potentially related, variables. By utilizing the information we have learned throughout this semester, we were able to arrive at some interesting conclusions.

Unsurprisingly, the United States economy has a reputation of being one of the best and most powerful in the world. We investigated this assumption by conducting a one-sample hypothesis test on the mean unemployment rate of the U.S. compared to the mean unemployment rate for the rest of the world. We failed to reject the hypothesis that the mean value of the unemployment rate of the United States is lower than the mean value of the world's unemployment rate. According to our investigation, at least in 2016, the U.S. economy is healthier than the rest of the world. This was the conclusion that we anticipated, as it is consistent with our perception of the U.S. economy on the world stage.

We were hoping that our project would shed some light on some variables that are related to the unemployment rate in order to pinpoint potential areas to target that could improve the overall health of the economy. Therefore, we included variables in our data set that we thought might be large contributors to unemployment on a state-level. We were interested to know if high minimum wage, a specific political majority, a particular location, or a high amount of welfare spending in a state equates to high unemployment. We were surprised to find that none of these factors were significantly related to raising the unemployment rate. We failed to reject the hypothesis that the mean unemployment rate of states with higher minimum wages is less than or equal to the mean unemployment rate of states with lower minimum wages. We failed to reject the hypothesis that unemployment rate is independent of the political majority of a state. We failed to reject the hypothesis that the population variances of unemployment rate for western and eastern states are different, which would suggest that one region of the U.S. tends to have a healthier economy than the

other. Additionally, we failed to reject the hypothesis that there is a statistically significant difference in proportions of high unemployment rate between states with high and low welfare spending.

After four variables proved to be statistically unrelated to unemployment rate, our linear regression model, in module 6, was successful in locating a variable that was correlated to unemployment rate. We investigated whether or not high school diploma rate could be used to predict the unemployment rate of a state. There was a moderately strong correlation between the two variables. The correlation coefficient was -0.5725387. After running our linear regression model, we discovered that high school diploma rate does seem to be an adequate predictor for the unemployment rate in the United States. This was a bit surprising to us, as we anticipated the other variables to be more closely related than this one.

To continue the investigation of unemployment rate in Phase II of our project, we suggest repeating some of the tests with data from a wider range of years. Our analysis only considered one year, 2016, and the addition of a larger amount of time might reveal more insight, which could lead to stronger conclusions regarding the unemployment rate and various variables. Additionally, a more in-depth political party investigation might be interesting to fully explore a state's political party majority and the affect it has on unemployment rates in the United States. Because of time limitations, we were only able to look into the past 5 presidential elections and determine a party majority from that information. We are aware that this interpretation might be a too broad a view to for a relationship to unemployment rate to show itself. Perhaps gathering data from elections on a more local level or even exploring the influence high density cities have on the political majority and how that affects unemployment could add value to this investigation. Finally, we think a deeper dive into the high school diploma and graduation rate in a state would be worth while, as it has proven to be correlated with unemployment rate. The conclusions we make in our project suggest that increased attention, and possibly spending, to improve high school success rates has a direct, positive impact on the unemployment rate, and thus the health of the economy.

References

Bureau of Labor Statistics: Local Area Unemployment Statistics Map. (2017, June) Retrieved from https://data.bls.gov/map/MapToolServlet?survey=la

Census Regions and Divisions of the United States (2013, January) Retrieved from https://www2.census.gov/geo/pdfs/maps-data/maps/reference/us_regdiv.pdf

Chantrill, C. (2018, April 16) *State Welfare Spending Rank.* Retrieved from https://www.usgovernmentdebt.us/compare_state_spending_2016p40C

Guzman, G. (2017, September) *Household Income: 2016.* Retrieved from https://www.census.gov/content/dam/Census/library/publications/2017/acs/acsbr16-02.pdf

Political Party Strength in U.S. (2018, April 10) Retrieved from https://en.wikipedia.org/wiki/Political_party_strength_in_U.S._states

Purple Color Group. (2018, February 19) *Peer Review*. Personal Communication.

The World Bank: Unemployment ILOSTAT database. (2017, March) Retrieved from https://data.worldbank.org/indicator/SL.UEM.TOTL.ZS?end=2016&start=1991

United States Census Bureau: Educational Attainment, 2012-2016 American Community Survey 5-year Estimates. (2016, December) Retrieved from https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=ACS_15_5YR_S1501

United States Department of Labor: Wage and Hour Division (WHD). (2017, December) Retrieved from https://www.dol.gov/whd/state/stateMinWageHis.htm

Weidner, J. & Williams, J. (2011, February 14). *What Is the New Normal Unemployment Rate?* Retrieved from https://www.frbsf.org/economic-research/publications/economic-letter/2011/february/new-normal-unemployment-rate/

Appendix A: Data

state	un_rate	min_wage	med_inc	dip_hs	dip_college	st_reg	pol_maj	welf_spent
Alabama	5.9	7.25	46,257.00	84.30	23.50	South	1	0.6-0.8
Alaska	6.9	9.75	76,440.00	92.10	28.00	West	1	1.8-2.0
Arizona	5.4	8.05	53,558.00	86.00	27.50	West	1	0.8-1.0
Arkansas	3.9	8.00	44,334.00	84.80	21.10	South	1	0.8-1.0
California	5.5	10.00	67,739.00	79.80	31.40	West	0	1.6-1.8
Colorado	3.3	8.31	65,685.00	90.70	38.10	West	0	0.8-1.0
Connecticut	5.1	9.60	73,433.00	89.90	37.60	Northeast	0	1.2-1.4
Delaware	4.5	8.25	61,757.00	88.40	30.00	South	0	0.8-1.0
Florida	4.8	8.05	50,860.00	86.90	27.30	South	1	0.6-0.8
Georgia	5.4	5.15	53,559.00	85.40	28.80	South	1	0.4-0.6
Hawaii	2.9	8.50	74,511.00	91.00	30.80	West	0	1.0-1.2
Idaho	3.8	7.25	51,807.00	89.50	25.90	West	1	0.8-1.0
Illinois	5.8	8.25	60,960.00	87.90	32.30	Midwest	0	0.8-1.0
Indiana	4.4	7.25	52,314.00	87.80	24.10	Midwest	1	0.8-1.0
Iowa	3.6	7.25	56,247.00	91.50	26.70	Midwest	0	0.6-0.8
Kansas	4	7.25	54,935.00	90.20	31.00	Midwest	1	0.6-0.8
Kentucky	5.1	7.25	46,659.00	84.20	22.30	South	1	0.8-1.0
Louisiana	6	7.25	45,146.00	83.40	22.50	South	1	0.8-1.0
Maine	3.8	7.50	53,079.00	91.60	29.00	Northeast	0	1.6-1.8
Maryland	4.4	8.75	78,945.00	89.40	37.90	South	0	1.2-1.4
Massachusetts	3.9	10.00	75,297.00	89.80	40.50	Northeast	0	1.0-1.2
Michigan	5	8.50	52,492.00	89.60	26.90	Midwest	0	1.0-1.2
Minnesota	3.9	7.75	65,599.00	92.40	33.70	Midwest	0	1.2-1.4
Mississippi	5.8	7.25	41,754.00	82.30	20.70	South	1	0.8-1.0
Missouri	4.6	7.65	51,746.00	88.40	27.10	Midwest	1	0.4-0.6
Montana	4.1	8.05	50,027.00	91.80	29.50	West	1	1.2-1.4
Nebraska	3.1	9.00	56,927.00	90.70	29.30	Midwest	1	1.0-1.2
Nevada	5.7	7.25	55,180.00	85.10	23.00	West	0	0.8-1.0
New Hampshire	2.9	7.25	70,936.00	92.00	34.90	Northeast	0	1.0-1.2
New Jersey	5	8.38	76,126.00	88.60	36.80	Northeast	0	1.0-1.2
New Mexico	6.7	7.50	46,748.00	84.20	26.30	West	0	0.6-0.8
New York	4.8	9.00	62,909.00	85.60	34.20	Northeast	0	1.2-1.4
North Carolina	5.1	7.25	50,584.00	85.80	28.40	South	1	0.8-1.0
North Dakota	3.1	7.25	60,656.00	91.70	27.70	Midwest	1	1.0-1.2
Ohio	5	8.10	52,334.00	89.10	26.10	Midwest	1	1.0-1.2
Oklahoma	4.8	7.25	49,176.00	86.90	24.10	South	1	0.8-1.0
Oregon	4.8	9.75	57,532.00	89.80	30.80	West	0	1.2-1.4
Pennsylvania	5.4	7.25	56,907.00	89.20	28.60	Northeast	0	1.2-1.4
Rhode Island	5.2	9.60	60,596.00	86.20	31.90	Northeast	0	1.2-1.4
South Carolina	5	7.25	49,501.00	85.60	25.80	South	1	0.6-0.8
South Dakota	3	8.55	54,467.00	90.90	27.00	Midwest	1	1.0-1.2
Tennessee	4.7	7.25	48,547.00	85.50	24.90	South	1	0.8-1.0
Texas	4.6	7.25	56,565.00	81.90	27.60	South	1	0.4-0.6
Utah	3.4	7.25	65,977.00	91.20	31.10	West	1	0.8-1.0
Vermont	3.2	9.60	57,677.00	91.80	36.00	Northeast	0	1.6-1.8
Virginia	4.1	7.25	68,114.00	88.30	36.30	South	0	0.8-1.0
Washington	5.3	9.47	67,106.00	90.40	32.90	West	0	0.8-1.0
West Virginia	6.1	8.75	43,385.00	85.00	19.20	South	1	1.2-1.4
Wisconsin	4	7.25	56,811.00	91.00	27.80	Midwest	0	1.4-1.6
Wyoming	5.3	5.15	59,882.00	92.30	25.70	West	1	0.6-0.8

Appendix B: Billing Invoice

04/18/2018

Invoice No. 0001

To
Bureau of Labor Statistics
Postal Square Building
2 Massachusetts Ave., N.E.
Washington, DC 20212-0001

Quantity	Description	Price/Hour	Total
5	Pre-planning, Research	$30.00	$150.00
3	Data Acquisition and Cleaning	$30.00	$90.00
4	Module 1 Analysis	$30.00	$120.00
12	Module 2 Analysis	$50.00	$600.00
15	Module 3 Analysis	$50.00	$750.00
10	Module 4 Analysis	$50.00	$500.00
9	Module 5 Analysis	$50.00	$450.00
8	Module 6/Compile Modules	$20.00	$160.00
	Subtotal		$2820.00
	Sales Tax		$197.40
	Total Due		**$3.017.40**

Due upon receipt

Thank you for your business!

DatBuzz Inc.

Tel 770-123-4567
Fax 770-765-4321

350 Ferst Dr. NW,
Atlanta, GA 30318

www.datbuzz.com
purchasing@datbuzz.com

Senior Consultants: Karsten Cook & Makayla Underwood

Appendix C: Consulting Log

Date	Time (hrs)	Description
2/6/18	2	Planning project
2/7/18	1	Planning
2/5/18	2	Gathering Data
2/15/18	1	Finalized Spreadsheet
2/19/18	1	Planning Module 2
2/20/18	4	Module 2 - continuous variable & nominal variable
2/22/18	3	Module 2 - ordinal variable & bindary variable
2/22/18	7	Completed Module 2 - finalized HTs & transferred work to one Rmd file
3/21/18	4	Module 3 - first HT on mean
3/25/18	2	Module 3 - second HT on mean - two-sided
3/29/18	2	Module 3 - power analysis
3/30/18	2	Module 5 - linear regression model
3/29/18	2	Module 3 - tidy up
4/5/18	5	Completed Module 3
4/5/18	3	Module 5 - F HT test and confidence intervals
4/10/18	4	Completed Module 5
4/11/18	1	Completed Module 1 - introduction
4/9/18	2	Planning Module 4
4/9/18	3	Module 4 - first HT
4/10	3	Module 4 - second HT
4/12/18	2	Proofing Module 1 & 2
4/12/18	2	Planning Module 6
4/13/18	2	Proofing Module 3 & 4
4/15/18	4	Completed Module 6 & Appendices
4/17/18	2	Final Review & Prep for Turn In
Total Hours:	66	

Appendix D: Acknowledgements

We would like to thank our fellow 2028 peers, specifically the purple color group, for their helpful suggestions, contributions, and encouragement throughout the duration of this project, as well as their speedy responses in the GroupMe. We would like to thank Alicia Will for her help addressing our questions, concerns, or hiccups that arose throughout the semester. Additionally, we would like to give a big thanks to Dr. Alisha Waller for her availability, encouragement, suggestions, and dedication to teach us the material through this project. We wish her the best of luck as she trudges through hours of grading.