Statistical Analysis of Obesity in America

Henry Hart & Colin McCormick

Wednesday, April 18th, 2018

ISyE 2028 – Alisha Waller

TABLE OF CONTENTS

Module 1

Obesity is beginning to spread through the United States like an epidemic. The American culture of being in a hurry has caused the country to eat out more often leading to the consumption of unhealthier food. Additionally, the rushed lifestyle leads to sedentary activities for free time instead of exercising. This has led to sharp rise in obesity among Americans, and this is not good. Obesity leads to an extremely high risk of developing heart disease, diabetes, breathing disorders, cancer, and having a heart attack or stroke. Furthermore, being obese leads to deterioration in quality of life because of the inability to enjoy certain activities and may even lead to depression and anxiety. A project researching the trends among obesity is of the utmost importance because understanding these trends will hopefully help campaigns to reduce obesity around the United States by providing them with tendencies and patterns among the United States population.

The big questions we will be investigating through statistical analysis are:

•Have policies aimed at lowering obesity rates been successful in the United States?

•Are obesity rates in the lower median income states higher than those with higher median incomes?

•Is there a relationship between a state's population and its obesity rate?

•Does a state's poverty rate help predict its obesity rate?

Because one of the major goals of this statistical analysis is to observe the connections between obesity and wealth, we will analyze data for this project from two main sources. First, we collected all our data on obesity rates and policies to reduce obesity from stateofobeisty.org. This website has an extensive database on various topics related to obesity, including obesity rates of states, race, and gender, along with these trends over the past few decades. This website's goal is much like our goals as well, analyzing trends in obesity data and spreading awareness of the obesity epidemic spreading across America. The other part of our data based on wealth and general state information is gathered from

a table found on the United States Census Bureau's website, census.gov, that contains the poverty rate, median income, and population of each state and the District of Columbia.

Based upon the spread of values for each variable and our attempt to analyze obesity trends across the entire country, we did not feel the need to clean our data to eliminate any of the states or the District of Columbia. Therefore, our data is not a random sample but is technically the population data of the United States split into different observational units. This technicality will cause some assumptions throughout the project to fail as they require the data set to be independent random samples, however, for the sake of the project and our curiosity, we will generally continue with the analysis performed in those sections of the project.

The observational unit is a state in the United States. Each state contains information that can be grouped into 10 variables outlined on the next page.

Table of variables:

Name	Description	Variable Type	Range of Reasonable Values
Obesity Rate	Percentage of population labelled obese	Continuous	20-40
poverty Rate	Percentage of population below poverty threshold	Continuous	6-22
Region	Region the State falls within	Categorical	Northeast, Midwest, South, West
White Obesity Rate	Percentage of white population labelled obese	Continuous	8-40
African American Obesity Rate	Percentage of African American population labelled obese	Continuous	15-45
Hispanic Obesity Rate	Percentage of Hispanic population labelled obese	Continuous	20-40
Population	Population of the state	Continuous	500,000-40,000,000
Median Income Range	Median income range of each state	Ordinal	1-4
State Policy	Whether or not the state has a policy in place to reduce obesity rates	Binary	Yes or No
2016 Vote	Whether the state is labelled a red or blue state based on 2016 election	Binary	red or blue

Median Income range

1 = $40,000-$50,000

2 = $50,000-$60,000

3 = $60,000-$70,000

4 = $70,000-$80,000

We will conduct a wide variety of statistical analyses to uncover many relationships and trends among obesity rates in the United States. In addition to the big questions we hope to answer in this project, some smaller questions we will attempt to answer are:

•Is the proportion of obese African-Americans equal to the national average?

•Do regional dieting habits affect obesity rates?

Using the tools we have learned throughout this entire course, we will go on a journey to discover the many hidden trends found with obesity and delve deep into the cultural, social, and economical issues that are associated with obesity rates. We will attempt to model each variable type with a theoretical distribution. We will perform one sample and two sample hypothesis tests on the mean and proportions, as well as a hypothesis test of independence between two categorical variables. We will also create a linear regression model on two of our continuous variables to predict obesity rates of states. Each module will focus exclusively on each type of statistical analysis as we answer many of our questions related to obesity to identify many important relationships and facts that should be communicated to the American public.

Module 2

Module Introduction:

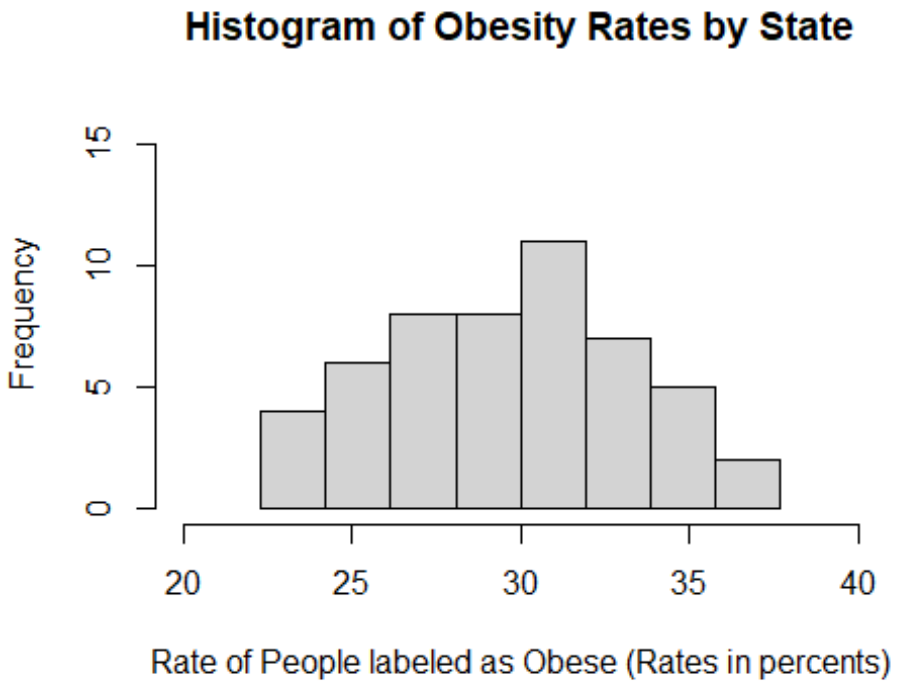
In Module 2, we will begin to analyze the distributions of some of our variables from our data set, specifically obesity rates, region, state policy regarding nutritional standards, and median income range. To better understand the nature of these variables, we will visually represent them and model each of them to a specific distribution family, with specific theoretical parameters. We use a variety of methods to evaluate the effectiveness of our distribution models on the variables, such as QQ Plots, Goodness of Fit Tests, and Normality Tests.

Analysis of Continuous Variable

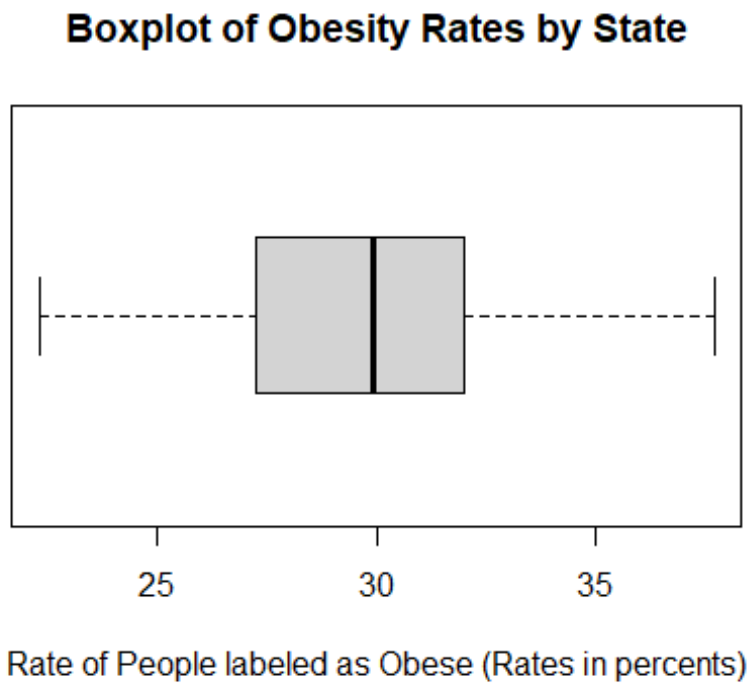
In this section we will explore the continuous random variable obs, which is the obesity rates of all the states in the US.

First, we begin with graphical summaries using a histogram and boxplot.

```
bins = seq(min(obs), max(obs), length=9)
hist(obs, breaks=bins, xlim=c(20,40), ylim=c(0,16), main="Histogram of
Obesity Rates by State", xlab="Rate of People labeled as Obese (Rates in
percents)", col="lightgrey")
```



```
boxplot(obs, horizontal = TRUE, main="Boxplot of Obesity Rates by State",
xlab="Rate of People labeled as Obese (Rates in percents)", col =
"lightgrey")
```



To further understand the distribution of our data we observe the numerical summary.

```
summary(obs)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   22.30   27.25   29.90   29.79   32.00   37.70

sd(obs)

## [1] 3.73736
```

Choosing a Probability Model

Our histogram shows that obesity rates distribution is unimodal and has tails that slope off. While the right tail drops more rapidly and appears more skewed than the left tail, our histogram appears roughly symmetric. While the histogram is not perfectly symmetric, it

shows a shape that is like a Normal distribution. This is evident because most of the data is located around the center and the tails slope off. While the shapes of the tails are not exactly equal, the difference in shape does not greatly affect the symmetry of the distribution.

The boxplot furthers our claim that our obesity rates distribution follows a Normal Distribution in that the distances between Q1 to Q2 and Q2 to Q3 are roughly equal and that both tails are roughly the same length. Furthermore, from homework #1 we know that the probability of a Normal distribution having an outlier is 0.0035, and with n = 51 we would expect less than 1 outliers on both tails, which is the case with our boxplot.

From our numerical summary we have a mean and standard deviation of 29.79 and 3.73736 respectively, and our quartiles are 27.25, 29.90, and 32. The mean and median are extremely close (29.79 vs 29.90) which furthers our statement of this distribution being roughly symmetric and a Normal distribution as our model.

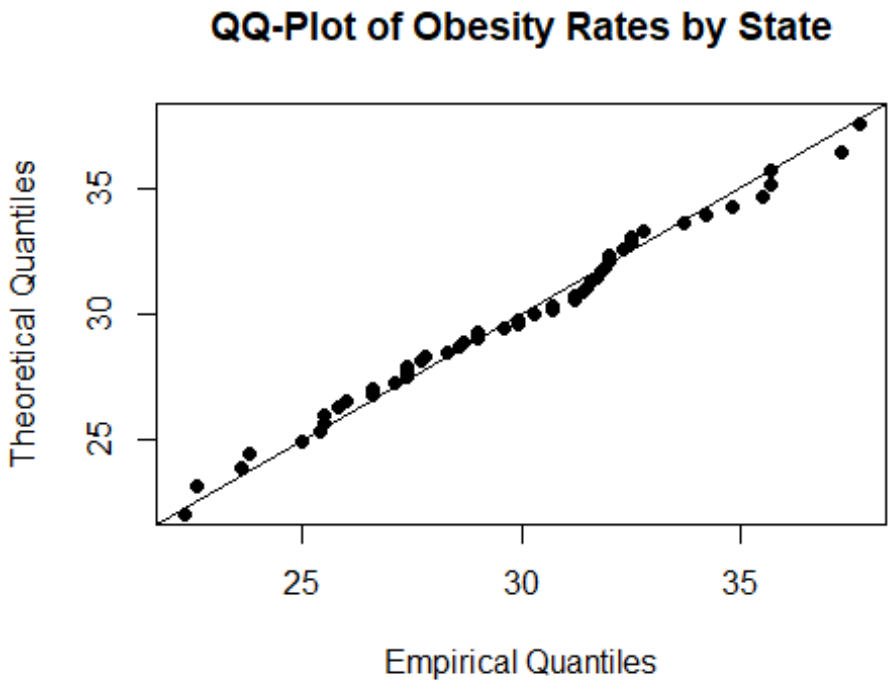
Based on our supporting graphical and numerical summaries, we argue that a Normal distribution would be a good model for our data. We will use the wisdom of the ages to estimate our parameters of u(mew-hat) and s^2(sigma2-hat) by using our sample mean and sample variance, so our probability distribution would be N(29.79, 13.9679). To visualize this model with our data, we recreate the histogram of our data and overlay the model.

```
bins = seq(min(obs), max(obs), length=9)
hist(obs, breaks=bins, freq = FALSE, xlim=c(20,40), main="Histogram of
Obesity Rates by State", xlab="Rate of People labeled as Obese (Rates in
decimals)", col="lightgrey")
curve(dnorm(x, mean=mean(obs), sd=sd(obs)), add=TRUE, lwd = 2)
```

The density curve seems to moderately fit our histogram, especially on the tails, but the largest bin causes the center to appear slightly off. To truly test if our model is a good fit we will use a QQ-plot and use a Shapiro-Wilk Normality test to see the goodness of fit of our data.

To create a QQ-plot, we must first create theoretical data from our model to compare to our actual data. We will use many factors from our data as parameters to generate our QQ-plot.

```r
n = length(obs) #sets number of values for theoretical data
mean = mean(obs) #sets mean of theoretical data
sd = sd(obs) #sets standard deviation of theoretical data
limits = c(min(obs), max(obs)) #sets the limits of theoretical data
#generating QQ-plot:
probs = (1:n)/(n+1)
norm.quants = qnorm(probs, mean, sd)
plot(sort(obs), sort(norm.quants), ylab="Theoretical Quantiles",
xlab="Empirical Quantiles", main="QQ-Plot of Obesity Rates by State",
xlim=limits, ylim=limits, pch=16)
abline(0,1)
```



This QQ-plot shows that our data roughly fits the x=y line, however we do not see most of the data clumped in the center of the graph. However, this shouldn't be too alarming because our standard deviation is rather large (~4) compared to the range of values (~15), which causes basically all our data to fall within two standard deviations of the mean, so

this large standard deviation makes our data appear to be more spread out. Therefore, we can still strongly argue that a Normal Distribution is a good fit for our data.

Next, we will use a Shapiro-Wilk Normality test to see how well our normal distribution model fits our data set. We will use a standard alpha value of 0.05 to compare the p-value generated by this test. If we generate a value less than 0.05 then our model is not an adequate model for our data set.

```
shapiro.test(obs)

##
##  Shapiro-Wilk normality test
##
## data:  obs
## W = 0.98456, p-value = 0.742
```

Based upon the very large p-value of 0.742, we can strongly claim that a Normal distribution, N(29.79, 13.9679), is an adequate model for our data set of obesity rates of states in the United States.

Analysis of Categorical Variable

In this section we will explore the categorical variable reg, which is the Region a state falls within in the United States. We divide the states into four categories based upon the United States Census Bureau region classifications of Northeast, Midwest, South, and West.

Northeast: Connecticut, Maine, Massachusetts, New Hampshire, Rhode Island, Vermont, New Jersey, New York, Pennsylvania

Midwest: Indiana, Illinois, Michigan, Ohio, Wisconsin, Iowa, Nebraska, Kansas, North Dakota, Minnesota, South Dakota, Missouri

South: Delaware, District of Columbia, Florida, Georgia, Maryland, North Carolina, South Carolina, Virginia, West Virginia, Alabama, Kentucky, Mississippi, Tennessee, Arkansas, Louisiana, Oklahoma, Texas

West: Arizona, Colorado, Idaho, New Mexico, Montana, Utah, Nevada, Wyoming, Alaska, California, Hawaii, Oregon, Washington

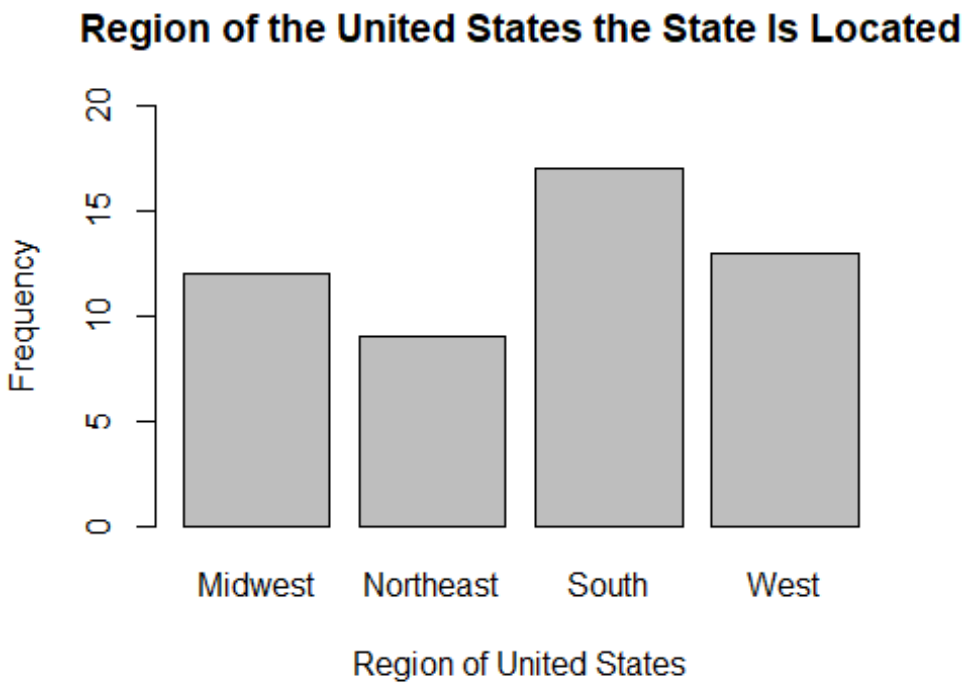
Our table of observed values is:

```
summary(as.factor(reg))

##   Midwest Northeast     South      West
##        12         9        17        13
```

Now we create a graphical summary of our data using a bar plot.

```
region = table((reg))
barplot(region, main = "Region of the United States the State Is Located",
xlab = "Region of United States", ylab= "Frequency", ylim = c(0,20))
```



Because the region data is categorical, the mean and standard deviation are not good measures of center and variability, and neither is the median or interquartile range. The only valid measures of center are the regions that have the most and least states, and those are the South with 17 states and the Northeast with 9 states respectively. We can observe

these measures in our bar plot, where the “South” is obviously the highest and the “Northeast” is the lowest. Additionally, the only viable option for measure of variability is range, which is 8.

Choosing a Probability Model

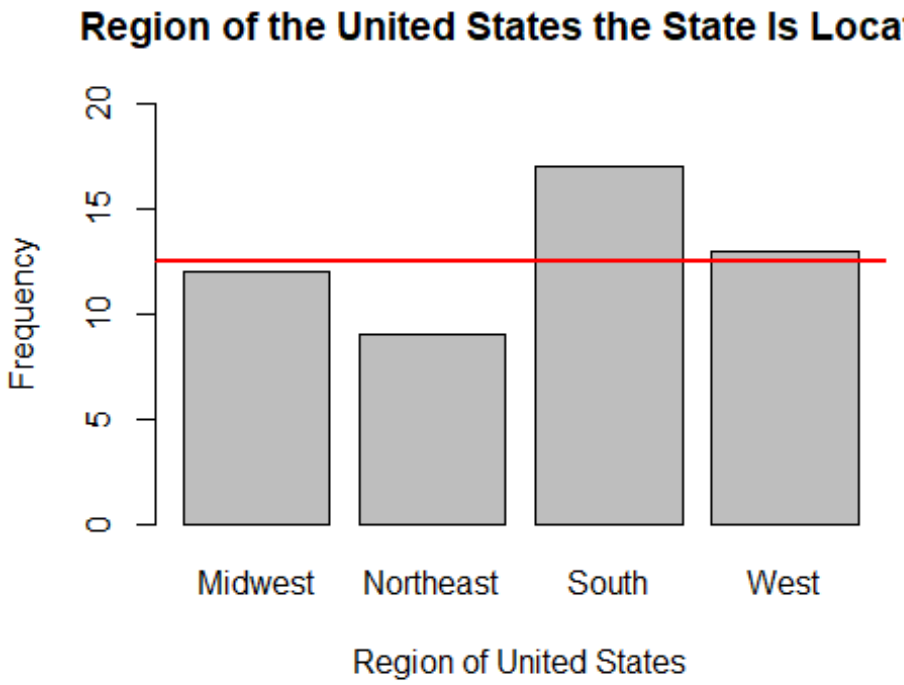
Our data set is a discrete finite categorical variable and beacuse of this the only possible theoretical distribution is a discrete uniform distribution. This fact is due to there being no correct order to our data.

We will also use the wisdom of the ages to estimate our parameters for our model because we know from ISYE 2027 that the probability of being placed in a bin is 1/#Bins. Furthermore, we know that the expected value of every bin is n*P(U(a,b)), so for our data we would expect there to be 12.5 states per region.

Our distribution model: $X \sim U(1,4) \sim \{1, 2, 3, 4\}$ Since $n = 50$, $P(X) = 1/4 = 0.25$ $E[X] = np = 50*1/4 = 12.5$

Now we must test to see if our model fits our data set. For starters we will overlay our model’s curve with our bar plot.

```
barplot(region, main = "Region of the United States the State Is Located",
xlab = "Region of United States", ylab= "Frequency", ylim = c(0,20))
abline(h=12.5, col="red", lwd = 2)
```



Based on the overlay we can obviously see that all the bars are not the same height, however, the Midwest and West are very close to our expected value of 12.5.

Now we will perform a chi-squared goodness of fit hypothesis test with our data and theoretical model. We will use a standard alpha value of 0.05.

degrees of freedom = #bins - 1 = 4-1 = 3

H_0: Our uniform distribution, U(1,4), is an adequate model for our data.

H_A: Our uniform distribution, U(1,4), is NOT an adequate model for our data.

Our test statistic is:

$$\chi_0^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

where E_j is the expected value of the bin and O_j is the observed value of the bin.

Our rejection region is calculated by X^2 3, and using a chi-squared table with a rejection region of 0.05 and df of 3 we get a critical value of 7.815.

Our table of observed values, expected values, and X^2 values is:

```
region = table((reg))
observed = as.vector(region)
expected = c(12.5, 12.5, 12.5, 12.5)
xsquared = (observed-expected)^2/expected

my.matrix = cbind(observed, expected, xsquared)
colnames(my.matrix)=c("Observed (O~j~)", "Expected (E~j~)", "Chi-Squared
Value")
rownames(my.matrix)=c("Midwest", "Northeast", "South", "West")

kable(my.matrix, col.names=c("Observed (O~j~)", "Expected (E~j~)", "X^2
Value"), row.names=TRUE)
```

	Observed (O_j)	Expected (E_j)	X^2 Value
Midwest	12	12.5	0.02
Northeast	9	12.5	0.98
South	17	12.5	1.62
West	13	12.5	0.02

Now that we have our table of Chi-squared values, we can calculate our test statistic.

```r
sum(xsquared)
```

```
## [1] 2.64
```

```r
chisq.test(x = observed, p = c(0.25, 0.25, 0.25, 0.25))
```

```
##
##  Chi-squared test for given probabilities
##
## data:  observed
## X-squared = 2.5686, df = 3, p-value = 0.463
```

After calculating the sum of our chi-squared values, 2.64, we can see that it is less than our critical value of 7.815, and we would therefore fail to reject our null hypothesis. Additionally, the p-value from our chi-square test gave us a p-value much larger than our alpha of 0.05, 0.463, which further cements our claim of failing to reject the null hypothesis. Based upon these tests and comparisons, we can concluded that a uniform distribution, U(1,4), is an adequate model for our data set of the region a state falls within.

Analysis of Binary Variable

In this section, the binary variable state_pol, which represents whether or not a state has an existing policy and guidelines in place on nutritional standards, will be analyzed and modeled. A "yes" value for the variable implies the presence of such guidelines, and a "no" implies the opposite.

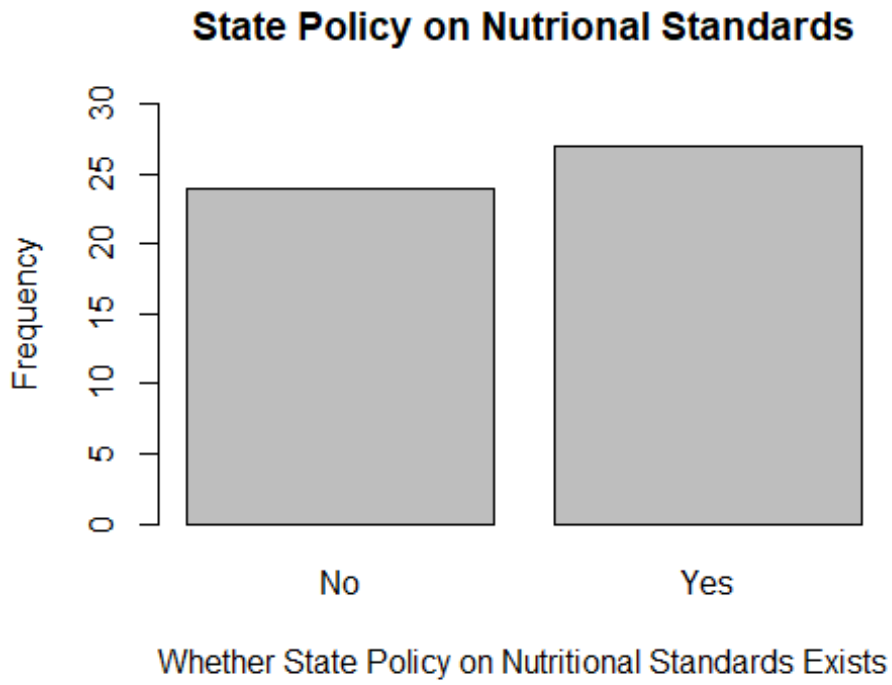
The table of observed values is:

```r
summary(as.factor(state_pol))
```

```
##  No Yes
##  24  27
```

Since our data is binary, which is essentially a categorical variable with only two possible values, it is best to graph the data using a bar plot, as shown in the following code:

```
state_policy = table((state_pol))
barplot(state_policy, main = "State Policy on Nutrional Standards", xlab =
"Whether State Policy on Nutritional Standards Exists", ylab= "Frequency",
ylim = c(0,30))
```



Since the state policy data is binary, the mean and standard deviation are not adequate measures of central tendency and variability. A viable statistic to measure is the probability that a state does have a state policy regarding nutritional standards, as well as the probability that a state does not have a state policy regarding nutritional standards. The probability of a state having such a policy would be equal to the number of states with a policy divided by the number of states total. This would be 27/51, which is approximately 0.53. For the probability of a state not having such a policy, the probability would be equal to the number of states without such a policy divided by the number of states total. This would be 24/51, which is approximately 0.47.

There is no logical measure of variability to calculate since the variable measured is binary; the range of values is only Yes or No.

Choosing a Probability Model

The Binomial Distribution Family is the only distribution family that can logically be applied to binary data. The parameters of the Binomial distribution are n and p. n is equal to the amount of times the experiment or variable is observed, and p is equal to the probability of success. In the case of this data, n would be equal to 51, since we are using data from the 50 states plus Washington, D.C. The p-value can be estimated to be 0.53, since 53% of the states have a state policy regarding nutritional standards.

To better understand the accuracy of the theoretical model to the actual data set, we will conduct a goodness of fit hypothesis test.

Goodness of Fit Hypothesis Test

For the goodness of fit hypothesis test, we will use the standard value for alpha, 0.05.

degrees of freedom = #bins - 1 = 2-1 = 1

H0: Our Binomial distribution, Bin(51, 0.53), is an adequate model for our data.

HA: Our Binomial distribution, Bin(51, 0.53), is NOT an adequate model for our data.

Our test statistic is:

$$\chi_0^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

Our rejection region is calculated by X^2 3, and using a chi-squared table with a rejection region of 0.05 and df of 1 we get a critical value of 3.841.

Our table of observed values, expected values, and X^2 values is:

```r
statepolicy = table((state_pol))
observed = as.vector(statepolicy)
expected = c(23.97, 27.03)
xsquared = (observed-expected)^2/expected

my.matrix = cbind(observed, expected, xsquared)
colnames(my.matrix)=c("Observed (O_j)", "Expected (E_j)", "Chi-Squared
Value")
rownames(my.matrix)=c("No", "Yes")
my.matrix
```

```
##     Observed (O_j) Expected (E_j) Chi-Squared Value
## No              24          23.97      3.754693e-05
## Yes             27          27.03      3.329634e-05
```

Now that we have our table of Chi-squared values, we can calculate our test statistic.

```
test_statistic = sum(xsquared)
test_statistic

## [1] 7.084327e-05

observed

## [1] 24 27

my.matrix

##     Observed (O_j) Expected (E_j) Chi-Squared Value
## No              24          23.97      3.754693e-05
## Yes             27          27.03      3.329634e-05

chisq.test(x = observed, p = c(0.47, 0.53))

##
##  Chi-squared test for given probabilities
##
## data:  observed
## X-squared = 7.0843e-05, df = 1, p-value = 0.9933
```

After calculating the sum of our chi-squared values, 7.084e-05, we can see that it is much less than our critical value of 3.841, and we would therefore fail to reject our null hypothesis. Also, the p-value that the test returned was 0.9933, which is much higher than the α value of 0.05. Based upon these tests and comparisons, we can concluded that a binomial distribution, Bin(51,0.53), is an adequate model for our data set of whether a state has a policy regarding nutritional standards exist.

Analysis of Ordinal Variable

In this section, the ordinal variable med_inc, which represents the median income range of a state, will be analyzed. The four median income ranges defined are:

40,000 - 50,000 (referenced as 1)

50,000 - 60,000 (referenced as 2)

60,000 - 70,000 (referenced as 3)

70,000 - 80,000 (referenced as 4)

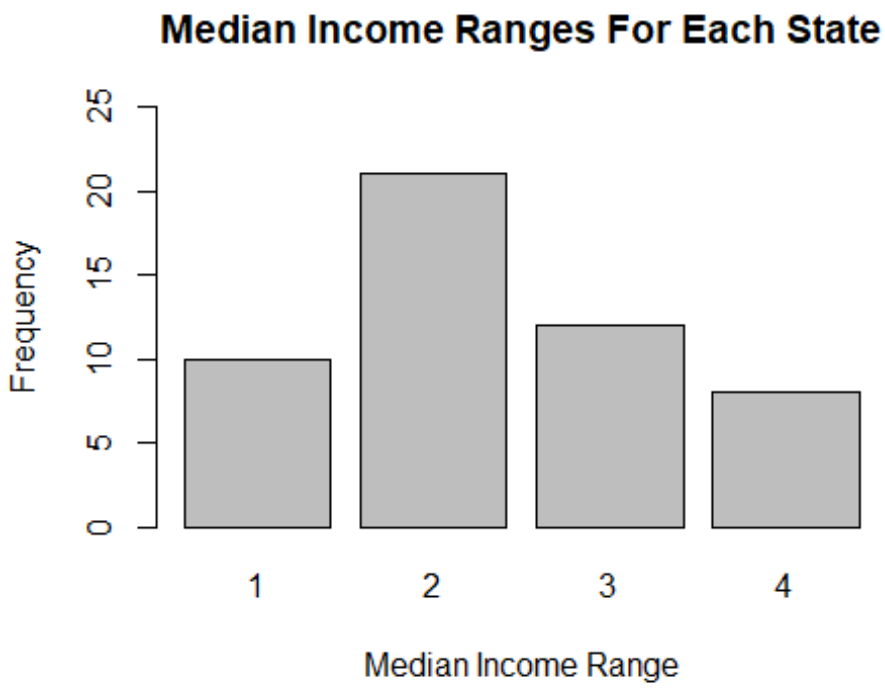
The table of observed values is:

```
summary(as.factor(med_inc))

##  1  2  3  4
## 10 21 12  8
```

Since our data is ordinal, which is a type of categorical data, it is best to graphically represent it with a bar plot. The following code creates a bar plot to display the median income range data:

```
median_income = table((med_inc))
barplot(median_income, main = "Median Income Ranges For Each State", xlab =
"Median Income Range", ylab= "Frequency", ylim = c(0,25))
```



Since the median income range data is ordinal, the mean and standard deviation are not good measures of central tendency and variability, and neither is the median or interquartile range. For ordinal data, a good measure of central tendency is the mode, or most occurring range. In this case, it is 2, or the income range from 50,000 - 60,000. It is worth noting that the median income range with the lowest frequency of 8 is 70,000 - 80,000. This means the range of the frequencies of the ordinal data would be 21 - 8 = 13. The range is the only viable option for a measure of variability for this type of data.

Choosing a Probability Model

The bar plot for the median income range data is roughly in the shape of the Poisson distribution. The parameter for the Poisson distribution is lambda, the expected value of the data being modeled. In this case, we estimate the expected value by taking the percentage of states in a category times the value (1,2,3,4) corresponding to the category. This calculation is as follows:

expected value = (10/51)*1 + (21/51)*2 + (12/51)*3 + (8/51)*4

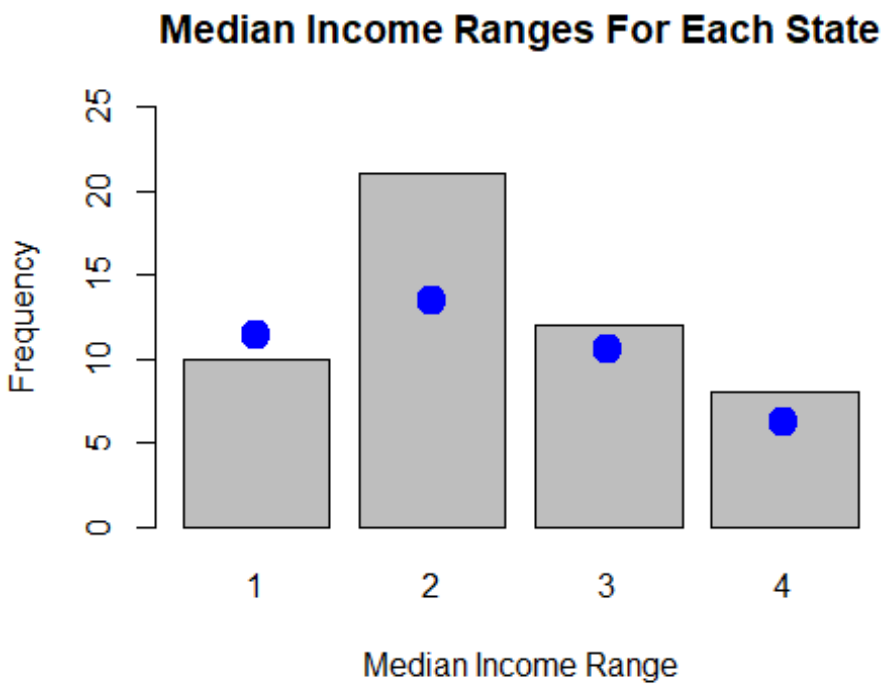
expected value = 2.35

This expected value of 2.35 corresponds to a median income in the range 50,000 - 60,0000, or approximately $53,500.

The following code will show the bar plot created to model the data, as well as a blue point in each bin that represents the estimated Poisson distribution with lambda equal to 2.35.

```
b = barplot(median_income, main = "Median Income Ranges For Each State", xlab
= "Median Income Range", ylab= "Frequency", ylim = c(0,25))
points(b,51*dpois(1:4, 2.35), col = "blue", lwd = 2, add = TRUE, pch = 19,
cex = 2)

## Warning in plot.xy(xy.coords(x, y), type = type, ...): "add" is not a
## graphical parameter
```



The blue dots represent the values of the Poisson(2.35) distribution for each income range, and the bars represent the actual data values collected. The distribution appears to be a rough match for the data set; we will conduct a Goodness of Fit Hypothesis Test to find out whether it is an adequate model for the data set.

Goodness of Fit Hypothesis Test

For the goodness of fit hypothesis test, we will use the standard value for alpha, 0.05.

degrees of freedom = #bins - #estimated parameters - 1 = 4-1-1 = 2

H_0: Our Poisson distribution, Poisson(2.35), is an adequate model for our data. H_A: Our Poisson distribution, Poisson(2.35), is NOT an adequate model for our data.

Our test statistic is:

$$\chi_0^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

Our rejection region is calculated by X^2 3, and using a chi-squared table with a rejection region of 0.05 and df of 2 we get a critical value of 5.991.

The following code calculates expected values for each bin, or median income range.

```
#for the first bin:
firstbin = 51*dpois(1, 2.35)

#for the second bin:
secondbin = 51*dpois(2, 2.35)

#for the third bin:
thirdbin = 51*dpois(3, 2.35)

#for the fourth bin:
fourthbin = 51*dpois(4, 2.35)
```

The following code creates variables for observed and expected values for median income, as well as the expected probabilities for each median income range. It returns the expected values and the expected p values in its output.

```
medianincome = table((med_inc))
observed = as.vector(medianincome)
expected = c(firstbin, secondbin,thirdbin, fourthbin)
expected
```

```
## [1] 11.42999 13.43024 10.52036  6.18071
```

```
expected_p = expected / 51
expected_p
```

```
## [1] 0.2241175 0.2633381 0.2062815 0.1211904
```

Now that we have the expected p values, we can run a chi-squared test using R.

```
chisq.test(x = observed, p = c(0.224, 0.263, 0.206, 0.121), rescale.p = TRUE)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  observed
## X-squared = 2.4876, df = 3, p-value = 0.4775
```

The X-squared value given by the hypothesis test is 2.4876, which is less than our critical value of 5.991, and we therefore fail to reject our null hypothesis. Also, the p-value from the test was 0.4775, which is much greater than our alpha value of 0.05. Based upon these tests and comparisons, we can conclude that a Poisson distribution, Poisson(2.35), is an adequate model for our data set of the median income ranges for each state in the United States.

Module Conclusion:

In Module 2, we modeled the distribution of some of our variables, and evaluated the accuracy of our theoretical distributions to the empirical data set. In Module 3, we will begin to answer some of the big questions we have for these variables by using single sample hypothesis testing to draw conclusions about our variables.

Module 3

Module Introduction:

In Module 3, we will use one-sample hypothesis tests to answer three questions regarding a comparison of obesity rates now versus in 2008, the effect of median income range on obesity rate, and a comparison of African American obesity rates to the national average. The two types of hypothesis tests that we will employ are a one-sided T-test on a population parameter and a Z approximation on a population parameter.

Hypothesis Test 1: Have obesity rates fallen since Michelle Obama's health initiatives in the late-2000s?

In the mid-2000s America began to seriously discuss its problem with obesity rates. During her time as first lady, Michelle Obama helped enact and push for many healthy reforms in nutritional standards and activity levels, especially among children and teenagers. We want to explore the question: Have obesity rates fallen since Michelle Obama's health initiatives in the late-2000s? For this project we collected most of our data on obesity rates from the year 2016, so we will perform a one-sided single sample hypothesis test on the mean obesity rate among states in the US.

If Michelle Obama actually helped create meaningful nutrition programs across the United States that help create healthy habits among America's youth, then one should expect the obesity rates to have decreased across the United States.

Our parameter of interest is: μ_{obs} = the mean obesity rate among all the states in the United States. We also have our value we are using to compare our data to: $\mu_o = 26$ = the mean obesity rate among all the states in the United States in 2008.

The average obesity rate for states in the United States in 2008 was 26%. Based on this, we will set up our hypotheses as follows:

$H_O: \mu_{obs} \geq 26$

$H_A: \mu_{obs} < 26$

We will use a standard alpha value of α = 0.05 to perform this hypothesis test. In addition, we must make the following assumptions before beginning our hypothesis test:

First, we must assume that our sample mean obesity rate is normally distributed around mean μ and an unknown variance. We explored this assumption in Module 2, where we used histograms, QQ-plots, and a Shapiro-Wilk Normality test to come to the conclusion that the obesity rates among states are modeled well by a Normal distribution N(29.79, 13.9679). Therefore, we can assume our data set is normally distributed.

Additionally, we must assume our data set is an independent random sample from the population. We cannot realistically make this assumption however because our population is all 51 states, and in this hypothesis test our sample is the entire population. Additionally, it is very difficult for each state's obesity data to be independent from another state's due to regional similarities between states and how easy it is to move between states, and one person's weight class to be taken with them to the new state. However, for the sake of this hypothesis test we will continue anyways, but we will make sure to take this failed assumption into account when we make our conclusions.

Our test statistic is:

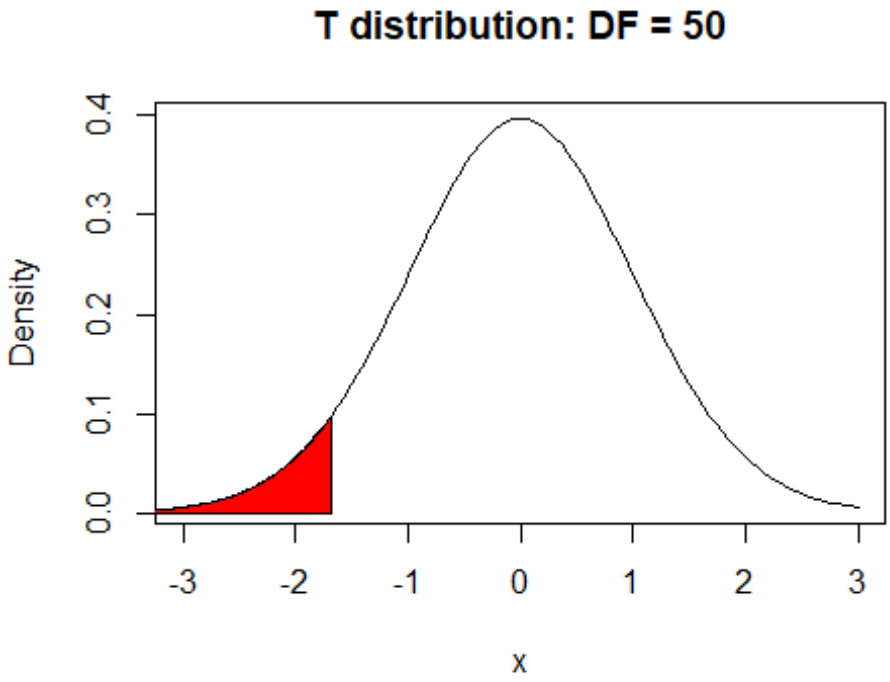
$$T_o = \frac{\overline{x} - \mu_o}{s/\sqrt{n}}$$

We have degrees of freedom equal to n-1 for t-distribution, so for our data we have 51 states and DC so df = 50. Our rejection region would be:

$T_o < T_{0.05,50} = -1.676$

We show this rejection region graphically in red for our t_{50} distribution below:

```
cord.x = c(-1.676,seq(-4,-1.676,0.01),-1.68)
cord.y = c(0,dt(seq(-4,-1.676,0.01),50),0)
curve(dt(x,50), xlim=c(-3,3), main='T distribution: DF = 50', ylab =
"Density")
polygon(cord.x,cord.y,col='red')
```



Now we must calculate our test statistic:

```
obs_mean = mean(obs)
obs_sd = sd(obs)
```

$\overline{x}$ = 29.7882353

s = 3.7373599

n = 51

$$T_o = \frac{\overline{x} - \mu_o}{s/\sqrt{n}} = \frac{29.79 - 26}{3.737/\sqrt{51}} = 7.239$$

Based on our test statistic value, we fail to reject H_o because 7.239 does not fall in our rejection region of T_o < -1.676. This means that obesity rates have not fallen across states in the United States since Michelle Obama helped enact nutrition reforms among children and teenagers. This extremely large positive t-value may indicate that the opposite may actually be true, and the obesity rates in the United States may have actually risen since 2008. However, that kind of statement can only be supported with further hypothesis tests.

We can use the t.test function in R to help support our claims we created based upon calculations done by hand.

```
t.test(obs, alternative = "less", mu = 26)

##
##  One Sample t-test
##
## data:  obs
## t = 7.2386, df = 50, p-value = 1
## alternative hypothesis: true mean is less than 26
## 95 percent confidence interval:
##     -Inf 30.6653
## sample estimates:
## mean of x
##  29.78824
```

This test matches our calculations of our t-value and supports our claim that we fail to reject H_o. Furthermore, this R function calculates our p-value, which is equal to the $P(\mu_{obs} \geq 26)$ and provides us with p-value = 1. This p-value provides an even stronger claim that obesity rates have not fallen since 2008, and may have in fact risen since then. However, again, we cannot make the claim they have in fact risen without performing more hypothesis tests.

Hypothesis Test 2: Are the obesity rates of states in the two lowest median income ranges higher than the national average?

For this hypothesis test, we will investigate if the states in the two lowest median income ranges have obesity rates that are higher than the national average. This is an intriguing question worth answering because the results will show whether lower income people tend to be more obese or less obese. On one hand, it can be argued that lower income people cannot afford to pay for food, so as a result, they will be less obese than the national average. However, on the other hand, it can be argued that lower income people can only afford lower quality, less healthy food, so they will be more obese than the national average. The average of the obesity rates among all states is computed below:

```
national_average = mean(obs)
```

This national average is calculated to be approximately 29.79%.

The μ referenced in the null and alternate hypotheses refers to the population obesity rate for the 31 states that have the two lowest income ranges, which are 40,000 - 50,000 and 50,000 - 60,000, referenced as "1" and "2" in the data.

H_0: μ <= 29.79 H_A: μ > 29.79

The following code will compile all of the states whose median income ranges are one of the bottom two ranges, as well as calculate their mean obesity rate and the standard deviation of their obesity rates:

```
national_average = mean(obs)
levelone = obs[which(med_inc=="1")]
leveltwo = obs[which(med_inc== "2")]
lowertwo = append(levelone, leveltwo)
u_hat = mean(lowertwo)
u_hat
```

```
## [1] 31.48387
```

```
s = sd(lowertwo)
s
```

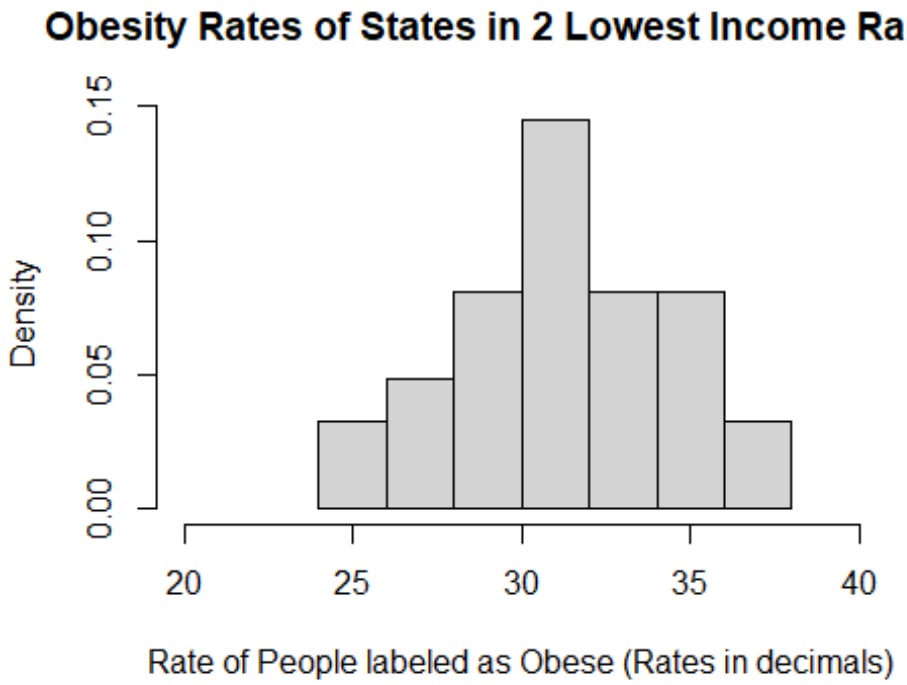
```
## [1] 3.191039
```

```
s2 = s^2
length(lowertwo)
```

```
## [1] 31
```

The assumption for this hypothesis test is that the data being used is part of a random sample with an approximate normal distribution. Since the data is from actual obesity rates from states in the US, we can not safely assume that the data is truly random in nature. However, for the purposes of this project, we will continue with the hypothesis test. To verify that the distribution is approximately normal, we will create a histogram of the obesity rates of the states in the two lowest median income range.

```
hist(lowertwo, breaks=5, freq = FALSE, xlim=c(20,40), main="Obesity Rates of States in 2 Lowest Income Ranges", xlab="Rate of People labeled as Obese (Rates in decimals)", col="lightgrey")
```



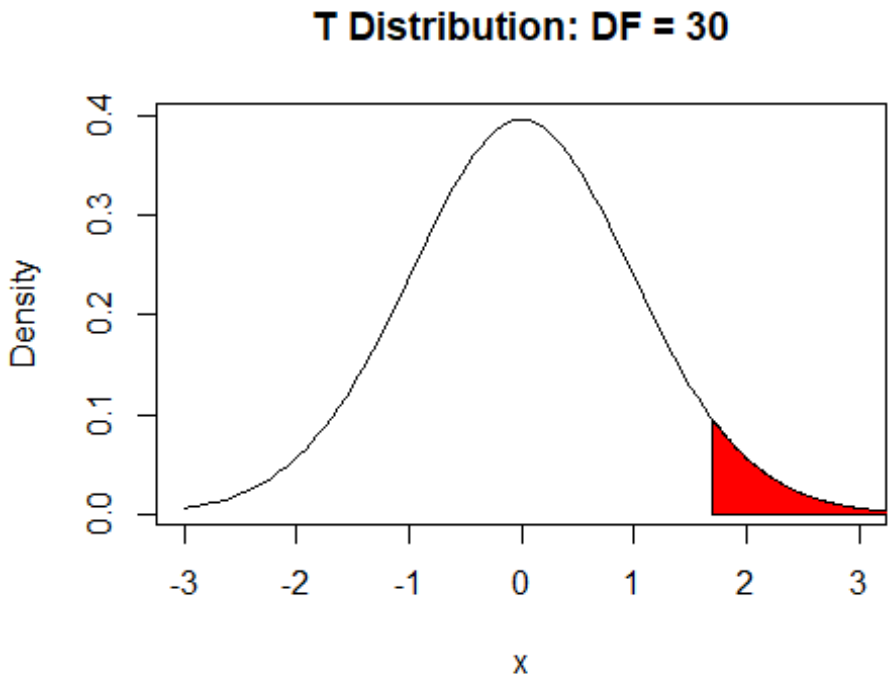
This histogram shows that the data generally follows the shape of a normal distribution, which verifies the other assumption. We can now proceed with the hypothesis test.

We must use a t-test in this circumstance because the population variance for all of the states with the two lowest median income ranges is considered unknown. We estimate it in this case by taking the variance of all the states that have these two median income ranges.

We then calculate the critical value for this t-test. We already established that the α value will be 0.05. The degrees of freedom for the T distribution is n - 1, where n is the sample size of the data. In this case, since 31 of the states fall in the lower two median income ranges, the degrees of freedom is 31 - 1 = 30. With α = 0.05 and 30 degrees of freedom, the critical value for the hypothesis test is 1.697.

The t distribution with 30 degrees of freedom, with the area to the right of the critical value shaded red, is shown below:

```
cord.x = c(1.697, seq(1.697, 4, 0.01), 1.697)
cord.y = c(0, dt(seq(1.697, 4, 0.01), 30), 0)
curve(dt(x, 30), xlim=c(-3, 3), main = "T Distribution: DF = 30", ylab =
"Density")
polygon(cord.x,cord.y,col = "red")
```



The following code conducts a right-tailed T test on the obesity rates from the states in the two lowest income ranges, with μ = 29.79, and a confidence level of 0.95, which reflects the α value of 1 - 0.95 = 0.05.

```r
t.test(lowertwo, alternative = "greater", mu = 29.79,conf.level = 0.95)

##
##  One Sample t-test
##
## data:  lowertwo
## t = 2.9555, df = 30, p-value = 0.003013
## alternative hypothesis: true mean is greater than 29.79
## 95 percent confidence interval:
##  30.51112      Inf
## sample estimates:
## mean of x
##  31.48387
```

The T test gave a t value of 2.956, which is greater than the critical value, 1.697. Also, the p value given by the T test is 0.00301, which is much less than the α value of 0.05. This is enough evidence to reject our null hypothesis in favor of the alternative hypothesis: $\mu >$ 29.79. In other words, the population obesity rate of states in the lower two median income ranges is greater than the national average obesity rate. This implies that if one's income is lower, they are more likely to obese. This helps answer the question proposed in the introduction to this hypothesis test; it appears that lower-income people cannot afford high quality or healthy food, and therefore tend to be more obese.

For this particular hypothesis test, we will calculate the power of our hypothesis test. The power is equal to 1 - β, where beta is the probability of making a Type II error. A type II error is an error where we failed to reject the null hypothesis, but the null hypothesis was false.

To calculate the power of our t-test, we will use the R function "power.t.test." For this function, n is the sample size, in our case 31. Delta is the difference between the $\u$ estimated in our hypotheses, which is 29.79, and the theoretical "true mean"" that we chose in order to calculate the power. In this case, we chose a "true mean" of 31; therefore, the delta value is 31 - 29.79 = 1.21. "sd" is the standard deviation of the sample, which was

calculated earlier to be 3.191. The significance level is equal to the α value we chose, which is 0.05.

```
power.t.test(n=31,delta=1.21,sd=3.191,sig.level=0.05,
        type="one.sample",alternative="one.sided",strict = TRUE)

##
##      One-sample t test power calculation
##
##               n = 31
##           delta = 1.21
##              sd = 3.191
##       sig.level = 0.05
##           power = 0.6621751
##     alternative = one.sided
```

The power.t.test function tells us the power of our hypothesis test is 0.662, or about 66.2%. This can be interpreted as a 66.2% chance that we falsely rejected a null hypothesis, if the true population mean is equal to 31. This is a very high power value, which implies that we should not be very confident about the results of this hypothesis test.

Since the power of a hypothesis test is equal to 1 - β, we can calculate the β value for a true population mean value of 31 to be 1-0.662 = 0.338. Therefore, we can say that in our hypothesis test, if the true population mean was 31, there was a 33.8% chance of making a Type II error. However, since we ended up rejecting our null hypothesis in favor of the alternative hypothesis, this is not necessarily an error we need to worry about.

Hypothesis Test 3: Are the majority of states' African American obesity rates greater than the national average obesity rate?

When we first compiled our data, at first glance, we observed an apparent pattern that the obesity rates for African American people seemed to be generally higher than the obesity rates for other races. This hypothesis test will test this observation by determining whether the proportion of African American obesity rates greater than the national average is greater than 0.5; in other words, if a majority of states' African American obesity rates are greater than the national average obesity rate.

First, we will determine the amount of states whose African American obesity rate is greater than the national average obesity rate. The following code calculates the national average obesity rate.

```
national_average = mean(obs)
national_average
```

```
## [1] 29.78824
```

The following code extracts of all the African American obesity rates that are greater than 29.788, returns these African American obesity rates, and returns the length of this list.

```
black_above_avg = black[which(black > "29.788")]
black_above_avg
```

```
##  [1] 44.6 44.6 44.1 44.2 42.9 45.1 42.4 42.4 36.6 37.4 41.7 41.7 37.2 32.1
## [15] 40.2 38.6 41.3 37.6 37.7 43.6 43.1 39.9 36.5 36.4 34.3 38.1 33.1 39.4
## [29] 30.2 34.2 34.4 30.4 35.2 37.2 30.3 31.9 37.7 30.8 32.1 30.1 31.0 31.3
## [43] 36.6 35.5
```

```
length(black_above_avg)
```

```
## [1] 44
```

There are 44 African American obesity rates above the national average.

For the hypothesis test on the population proportion, we will use the standard α value of 0.05. For the null and alternative hypotheses, p will refer to the population proportion of $\frac{StateswithAfricanAmericanObesityRatesGreaterThantheNationalAverage}{All51States(includesWashington,D.C.)}$, which will be considered unknown for the purposes of conducting this hypothesis test.

H_O: p <= 0.5

H_A: p > 0.5

The assumptions for this test are: - p is not extremely close to 0 or 1 - the sample size is relatively large - np and n(1 - p) are greater than or equal to 5

The first assumption is true for our particular test, since the p value of 0.5 is not extremely close to 0 or 1.

The second assumption is true as well, since we are working with a data set with 51 observations, which is relatively large. Since our sample size is relatively large, we have a higher value for the degrees of freedom, so we are able to approximate the binomial distribution with a Z, or normal, distribution.

For the third assumption, our n value is 51 and the p value is 0.5. np = 51 * 0.5 = 25.5 and n(1-p) = 25.5. Since 25.5 > 5, the third assumption is fulfilled as well.

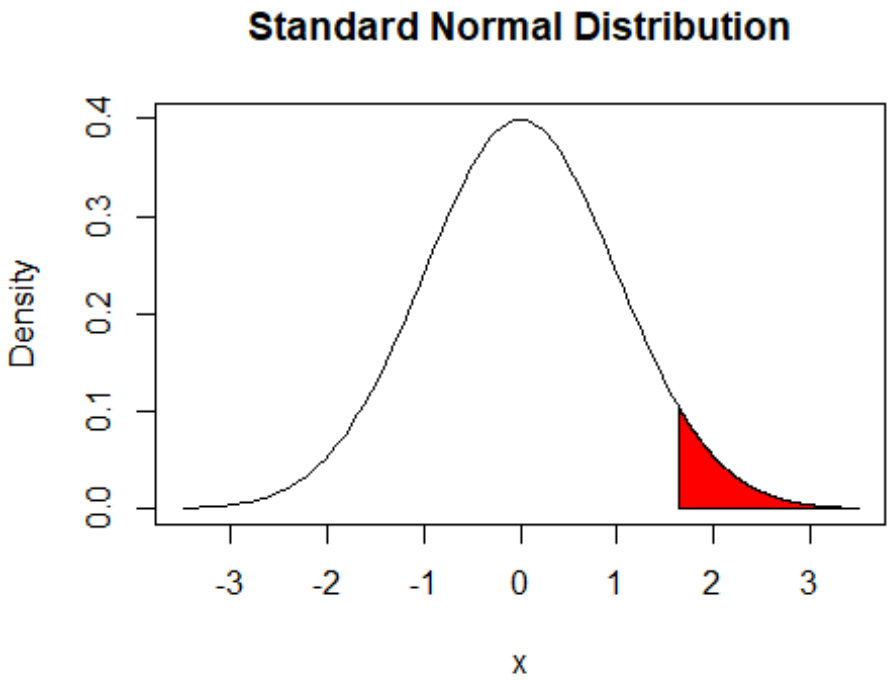
The test statistic used for this hypothesis test is:

$$Z = \frac{X - np_0}{\sqrt{np_0(1 - p_0)}}$$

The Z statistic being used has a standard normal distribution. For this Z statistic, X is the amount of observations in the sample that belong in the class associated with the numerator of p. In this case, X is equal to 44, the amount of states whose African American obesity rate is above the national average. “n” is the sample size, which in our case is 51. Although technically we are using population data, for the purposes of this project, we will treat it as sample data. p_0 is the p value from the null and alternate hypotheses; in our case, p_0 is equal to 0.5.

The critical value for a Z distribution, N(0,1), with an α value of 0.05 is 1.645. The standard normal distribution with the area to the right of this critical value shaded red is shown by the following code:

```
cord.x = c(1.645, seq(1.645, 3.5, 0.0001), 1.645)
cord.y = c(0, dnorm(seq(1.645, 3.5, 0.0001), 0), 0)
curve(dnorm(x), xlim=c(-3.5, 3.5), main = "Standard Normal Distribution",
ylab = "Density")
polygon(cord.x,cord.y,col = "red")
```



Next, we calculate the Z_0 value, which is the Z value for our specific values for X, n, and p_0, which are 44, 51, and 0.5, respectively. When plugged in the formula for Z, the resultant Z_0 value is:

$$Z_0 = \frac{44 - 51 * 0.5}{\sqrt{51 * 0.5(1 - 0.5)}} = 5.181$$

We will also run the following hypothesis test in R to determine our p value:

```r
prop.test(44, 51, 0.5, alternative = "greater", conf.level= 0.95, correct=
TRUE)

##
##  1-sample proportions test with continuity correction
##
## data:  44 out of 51, null probability 0.5
## X-squared = 25.412, df = 1, p-value = 2.315e-07
## alternative hypothesis: true p is greater than 0.5
## 95 percent confidence interval:
##  0.7537048 1.0000000
## sample estimates:
##         p
## 0.8627451
```

The Z_0 value of 5.181 is much greater than the critical value of 1.645. Also, the output from this code shows that the p-value from the hypothesis test is 2.315e-07, which is much lower than the α value of 0.05. Therefore, we can reject the null hypothesis in favor of the alternative hypothesis: p > 0.5. In the context of our data set, this means the the majority of states' African American obesity rates are above the national average. This confirms the pattern we initially observed in our data set to be true.

Module Conclusion

In Module 3, we drew conclusions from our data using one variable at a time in one sample hypothesis tests to provide meaningful answers to our questions. In Module 4, to better investigate relationships between multiple variables, we will use two sample hypothesis tests. Module 4 will prepare us to better discuss specific relationships between different variables.

Module 4

Module Introduction:

In Module 4, we will use two-sample hypothesis tests to investigate relationships between multiple variables from our data. We will make use of a one sided, two sample hypothesis test on the population mean and a hypothesis test of independence. By the end of this module, we will have a nuanced understanding on the relationships between state policies on nutritional standards and obesity rates, region and obesity rates, and median income level and region.

Hypothesis Test 1: Do states with policies in place to reduce obesity rates actually have lower obesity rates?

When we collected our data from the website stateofobesity.org, we noticed a section that provided statistics about which states had policies in place that should help prevent obesity rates from rising. We decided to explore some of these and found an interesting policy that required early childhood education programs to prepare meals and snacks that meet healthy dietary guidelines. These kind of policies are instituted in about half of the states in United States, 27 states in total. This makes it very easy to compare these two groups, so we want to answer the question: Do states with policies in place to reduce obesity rates actually have lower obesity rates? Similar to the initiatives inspired by Michelle Obama, one would expect these kind of nutritional guidelines for early childhood education programs to have a positive impact on obesity rates and cause the obesity rates to be lower in these states than those without them.

To answer our question posed, we will perform a one-sided, two sample hypothesis test on the population mean, with population variance unknown. Our parameters of interest are: μ_p = the mean obesity rate of the states with the policy in place μ_n = the mean obesity rate of the states without the policy in place

We set up our hypotheses as follows:

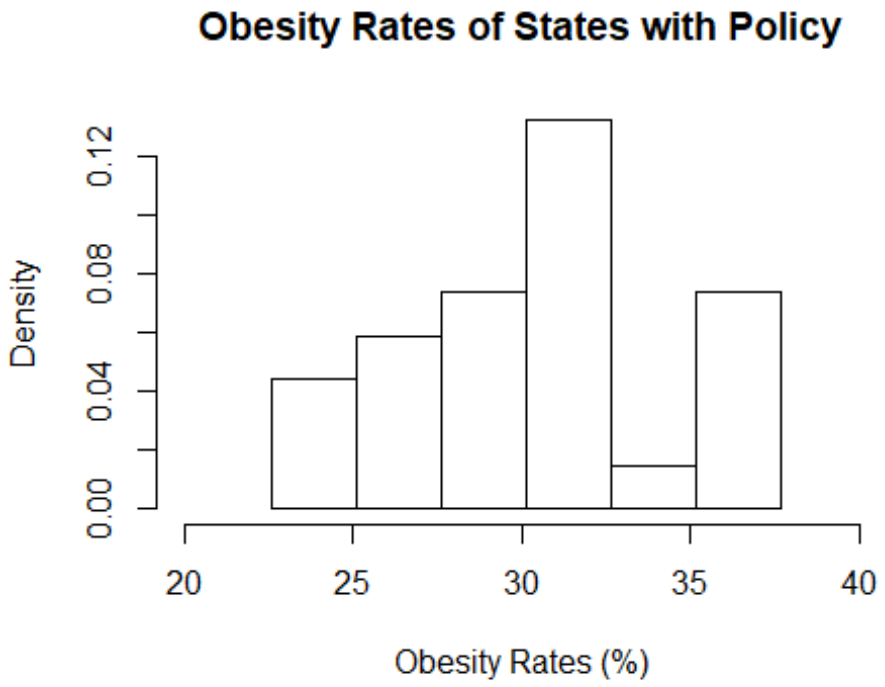
H_o: $\mu_p - \mu_n \geq 0$, The true difference in mean obesity rates between states with the policy and states without the policy is greater than or equal to 0.

H_A: $\mu_p - \mu_n < 0$, The true difference in mean obesity rates between states with the policy and states without the policy is less than 0.

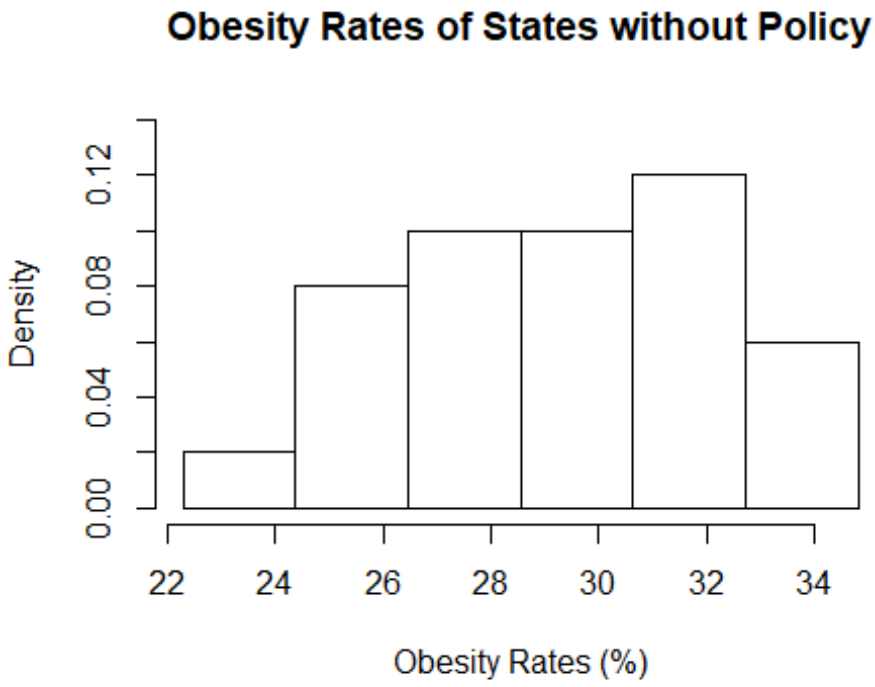
We will use a standard alpha value of α = 0.05 to perform this hypothesis test. In addition, we must make the following assumptions before beginning our hypothesis test:

First, we must make the assumption that our two data sets are normally distributed. To analyze this we will create histograms and QQ-plots of each data set versus a normal distribution.

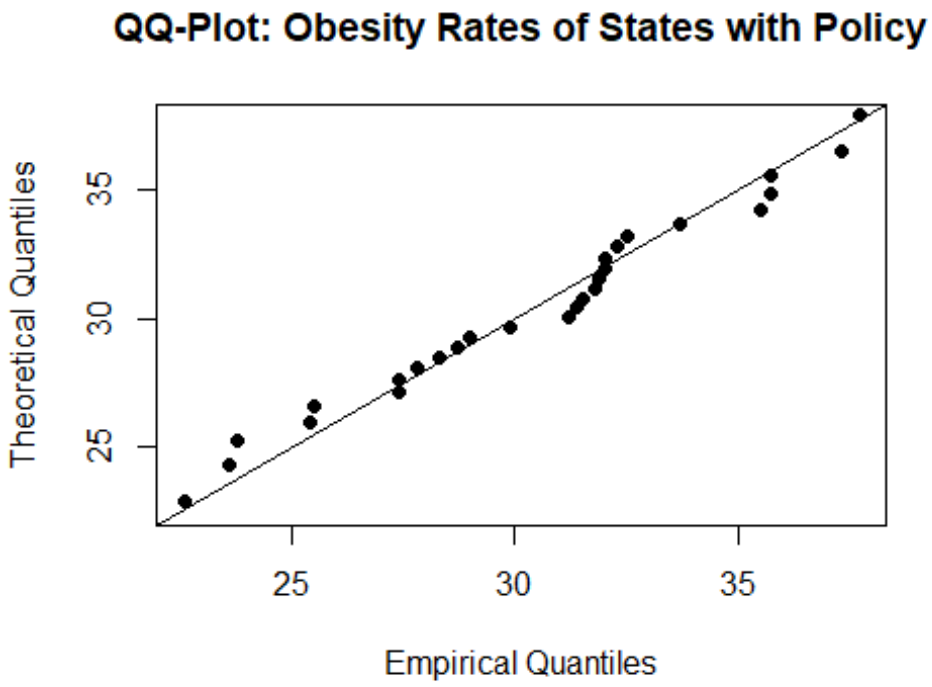
```
pol = obs[which(state_pol=="Yes")]
bins = seq(min(pol), max(pol), length=7)
hist(pol, breaks = bins, main = "Obesity Rates of States with Policy", xlab =
"Obesity Rates (%)", xlim = c(20, 40), freq = FALSE)
```



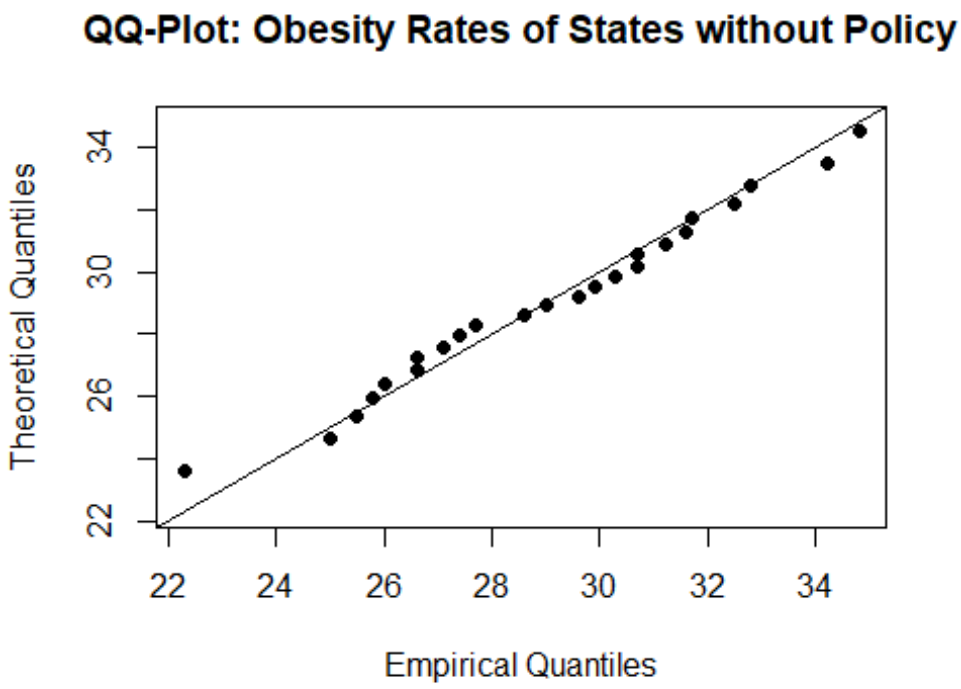
```r
no_pol = obs[which(state_pol=="No")]
bins = seq(min(no_pol), max(no_pol), length=7)
hist(no_pol, breaks = bins, main = "Obesity Rates of States without Policy",
xlab = "Obesity Rates (%)", freq = FALSE, ylim= c(0,.14))
```



```r
p = mean(pol)
psd = sd(pol)
n_p = length(pol)
limits = c(min(pol), max(pol))
probs = (1:n_p)/(n_p+1)
norm.quants = qnorm(probs, p, psd)
plot(sort(pol), sort(norm.quants), ylab="Theoretical Quantiles",
xlab="Empirical Quantiles", main="QQ-Plot: Obesity Rates of States with
Policy", xlim=limits, ylim=limits, pch=16)
abline(0,1)
```



```r
n = mean(no_pol)
nsd = sd(no_pol)
n_n = length(no_pol)
limits = c(min(no_pol), max(no_pol))
probs = (1:n_n)/(n_n+1)
norm.quants = qnorm(probs, n, nsd)
plot(sort(no_pol), sort(norm.quants), ylab="Theoretical Quantiles",
xlab="Empirical Quantiles", main="QQ-Plot: Obesity Rates of States without
Policy", xlim=limits, ylim=limits, pch=16)
abline(0,1)
```



Observing the two histograms, they each appear roughly symmetric on a varying spectrum of strength. Our histogram on states with the policy appears to be only slightly symmetric, and its shape is thrown off by the large central bin that is followed by a very small bin. I would not call this distribution a perfect Normal distribution, but the QQ-plot does show it is fairly close to Normal since most of the data is clumped to the center and the points do not vary too far from our slope line. Additionally, the histogram on states without the

policy appears to show a much stronger case for normality. It has a somewhat more obvious center with smaller tails, however, these tails are not that much smaller than the center. While the histogram may leave us skeptical on if it is a Normal distribution, the QQ-plot for states without the policy appears to strongly show that it follows a Normal distribution with most of our data falling around the center of the line and most of the points are tightly fit to the slope line. Based on our analysis on the graphical summaries of our two datasets we can conclude our assumption for the two datasets to be Normally distributed is true.

Furthermore, we must assume that our two data sets are independent random samples taken from a population. We cannot comfortably make this assumption for several reasons. First, both of these data sets are not randomly sampled, as we intentionally selected certain subsets of the population data set, states with the policy and states without the policy. Additionally, the two sets are not independent from each other, as explained in module 3, due to the interstate characteristic between every state in the United States that allows free movement between them. While we fail to make this assumption, for the sake of this project and our curiosity we will continue with this hypothesis test.

We must also decide whether or not the population variances of our two distributions are equal. Let σ_p^2 be the population variance of the obesity rates among states with the policy, and let σ_n^2 be the population variance of the obesity rates among states without the policy. Then we will have s_p^2 to represent the sample variance of the obesity rates among states with the policy, and we will also have s_n^2 to represent the sample variance of the obesity rates among states without the policy.

```
pvar = var(pol)
nvar = var(no_pol)
```

These sample variances are 17.233 and 9.857, respectively, which have 42.8% difference. With a percent difference that large we cannot comfortably make the assumption that the population variances of these two data sets are equal and will make the assumption that they are not equal.

We chose to do a one-sided hypothesis test because we want to see if the state policy has the desired effect of lowering obesity rates in a state. Additionally, because we do not know the population variances, our test statistic is

$$T_o = \frac{(\hat{\mu}_p - \hat{\mu}_n) - (\mu_p - \mu_n)}{\sqrt{\frac{s_p^2}{n_p} - \frac{s_n^2}{n_n}}}$$

Our sample statistics are given below:

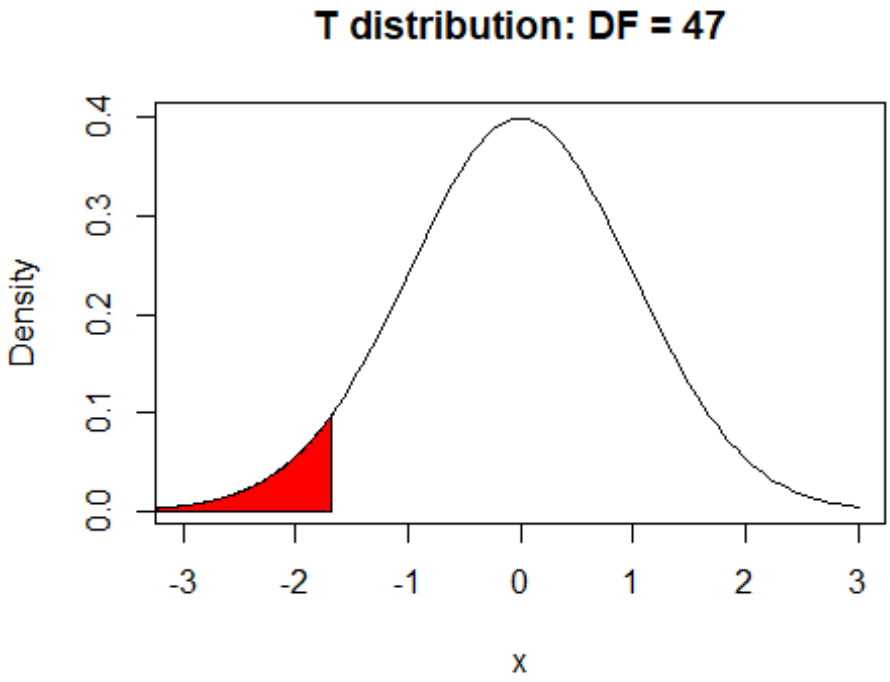
$\hat{\mu}_p$ = 30.4296296 $\hat{\mu}_n$ = 29.0666667 s_p^2 = 17.2337037 s_n^2 = 9.8571014 n_p = 27 n_n = 24 $\mu_p - \mu_n = 0$

Since we do not know the population variances, our degrees of freedom are

$$v = \frac{\sqrt{\frac{s_p^2}{n_p} - \frac{s_n^2}{n_n}}}{\frac{(s_p^2/n_p)^2}{n_p - 1} + \frac{(s_n^2/n_n)^2}{n_n - 1}} = 47$$

We will reject H_o if $T_o < T_{0.05,47} = -1.68$. To illustrate this, we create a graph of our T distribution and highlight the rejection region in red.

```
cord.x <- c(-1.68,seq(-4,-1.68,0.01),-1.68)
cord.y <- c(0,dt(seq(-4,-1.68,0.01),50),0)
curve(dnorm(x), xlim=c(-3,3), main='T distribution: DF = 47', ylab =
"Density")
polygon(cord.x,cord.y,col='red')
```



Now we must calculate our test statistic:

$$T_o = \frac{(\hat{\mu}_p - \hat{\mu}_n) - (\mu_p - \mu_n)}{\sqrt{\frac{s_p^2}{n_p} + \frac{s_n^2}{n_n}}} = \frac{30.43 - 29.07}{\sqrt{\frac{17.23}{27} + \frac{9.86}{24}}} = 1.33$$

Based upon our t-value of 1.33 given by our calculations we fail to reject null hypothesis that $\mu_p - \mu_n \geq 0$ because 1.33 does not fall within our rejection region of T_o < -1.68. This means that the mean obesity rate of states with a policy to enforce nutritional standards in early childhood education programs is greater than or equal to states without this policy. This result can possibly be explained by states creating this policy only after they have

become more obese than other states in the United States. We can confirm these conclusions by perform an analysis using the t.test function in R.

```
t.test(pol, no_pol, alternative = "less", mu = 0, paired = FALSE, var.equal = FALSE, conf.level = 0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  pol and no_pol
## t = 1.3308, df = 47.836, p-value = 0.9052
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##      -Inf 3.080902
## sample estimates:
## mean of x mean of y
##  30.42963  29.06667
```

This test confirms our calculations made by hand, with a t-value of 1.3308, and confirms our claim that we fail to reject H_o. Additionally, this function calculates our p-value, 0.9052, which is equal to $P(\mu_p - \mu_n \geq 0)$. This p-value provides even stronger support that obesity rates in states with a policy to reduce obesity are not lower than obesity rates in states without the policy, and they may actually be higher than states without the policy. However, we would have to perform more hypothesis tests to confirm that claim.

Hypothesis Test 2: Do regional dieting habits affect obesity rates?

While the United States may be viewed internationally as a very obese country with poor eating habits, many parts of the country don't necessarily follow the same dietary patterns. We will explore this by attempting to answer the following question: do regional dieting habits affect obesity rates?

We initially split the states into four regions based upon the United States Census Bureau's classification but, now we will split them into two groups based upon general dietary habits. Stereotypically, the South and Midwest are viewed as having poor dietary habits

and eating a lot of "soul" food that generally involves fried and salty foods and large portion sizes, and the Northeast and West are viewed as having better dietary habits and eating more vegetables and less salty foods. Based upon these stereotypes one would most likely expect their differences to not be equal to 0.

To answer this question, we will perform a two-sided, two sample hypothesis test on the population mean, with population variances unknown. Our parameters of interest are: μ_{smw} = the mean obesity rate of the states in the South or Midwest μ_{new} = the mean obesity rate of the states in the Northeast or West

Based upon the stereotypes mentioned in the earlier paragraph, we set up our hypotheses as follows:

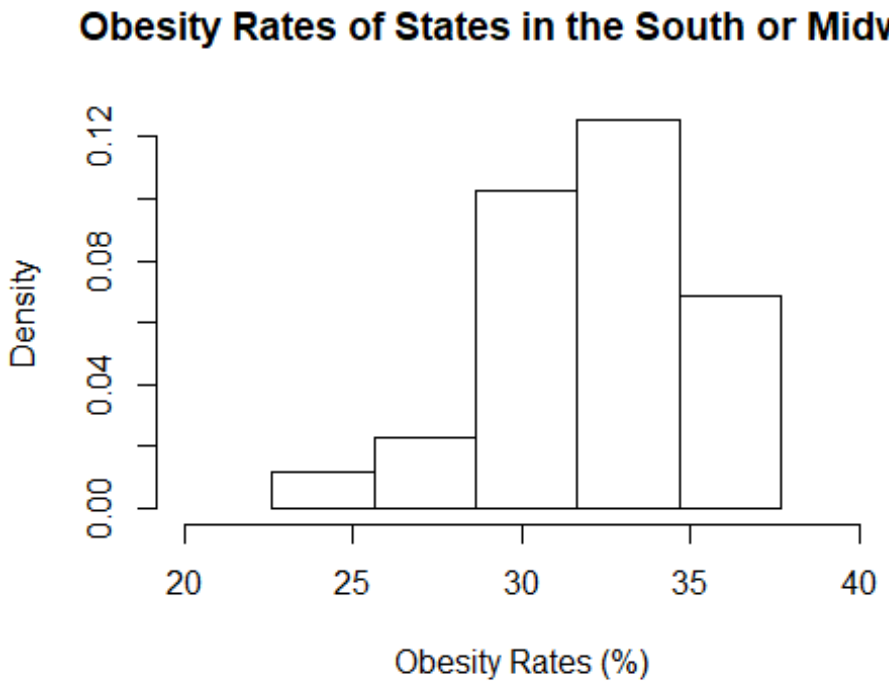
$H_o: \mu_{smw} - \mu_{new} = 0$, The true difference in mean obesity rates between states in the South and Midwest and states in the Northeast and West is equal to 0.

$H_a: \mu_{smw} - \mu_{new} \neq 0$, The true difference in mean obesity rates between states in the South and Midwest and states in the Northeast and West is not equal to 0.

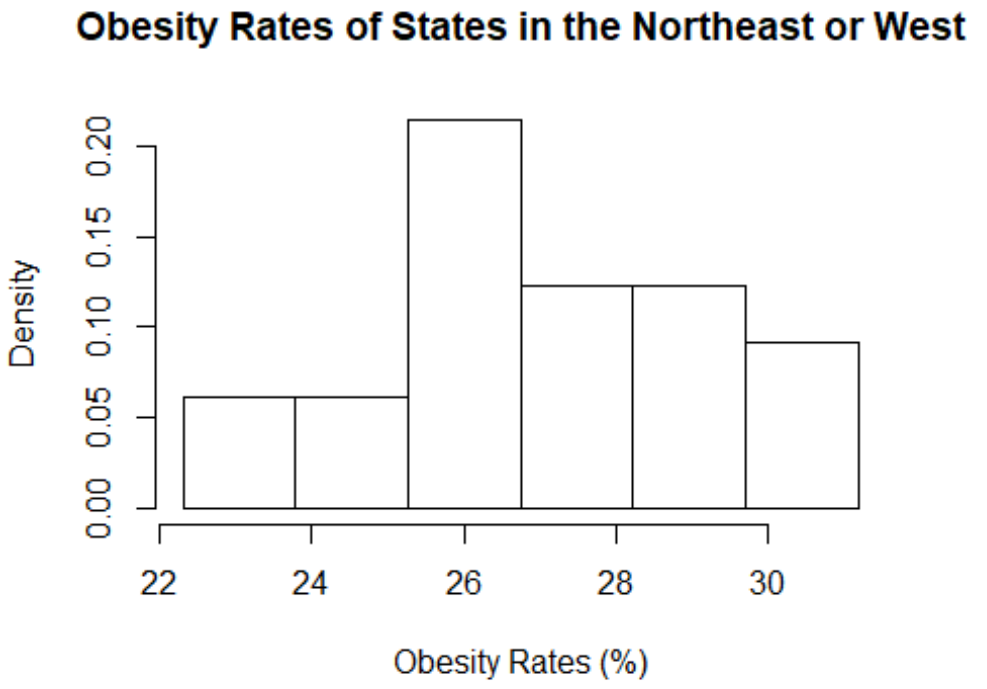
We will use a standard alpha value of α = 0.05 to perform this hypothesis test. In addition, we must make the following assumptions before beginning our hypothesis test:

First, we must make the assumption that our two data sets are normally distributed. To analyze this we will create histograms and QQ-plots of each data set versus a normal distribution.

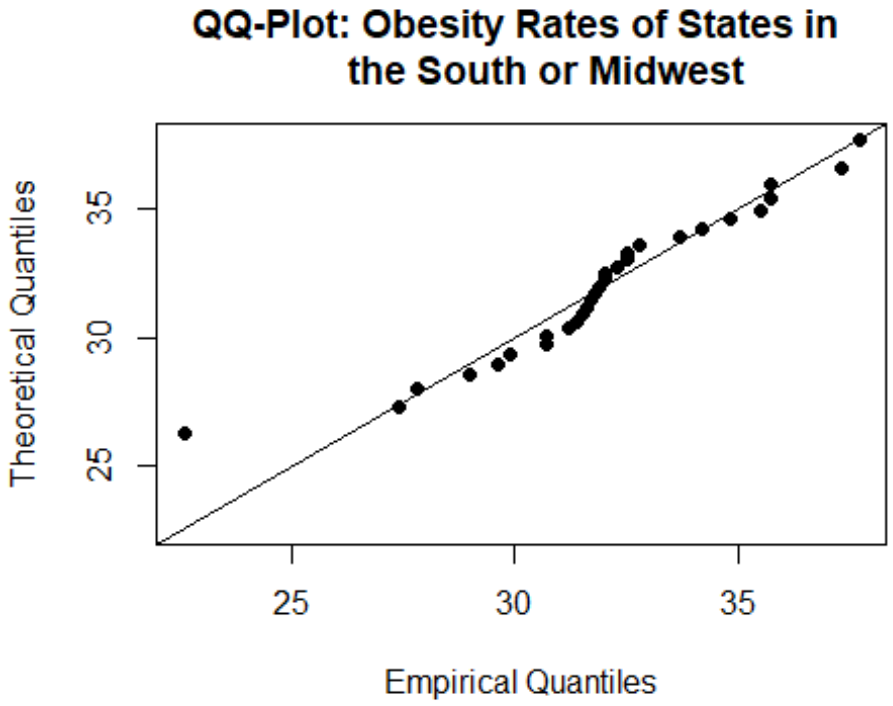
```
s = obs[which(reg=="South")]
mw = obs[which(reg=="Midwest")]
smw = append(s, mw)
bins = seq(min(smw), max(smw), length=6)
hist(smw, breaks = bins, main = "Obesity Rates of States in the South or
Midwest", xlab = "Obesity Rates (%)", xlim = c(20, 40), freq = FALSE)
```



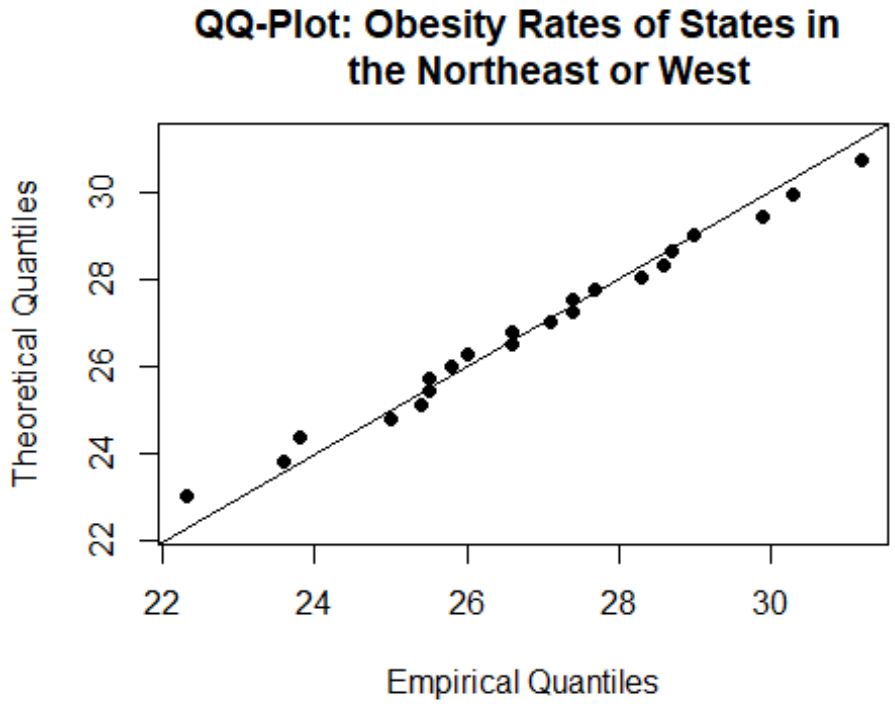
```
ne = obs[which(reg=="Northeast")]
w = obs[which(reg=="West")]
new = append(ne, w)
bins = seq(min(new), max(new), length=7)
hist(new, breaks = bins, main = "Obesity Rates of States in the Northeast or
West", xlab = "Obesity Rates (%)", freq = FALSE)
```



```r
sm = mean(smw)
smwsd = sd(smw)
n_smw = length(smw)
limits = c(min(smw), max(smw))
probs = (1:n_smw)/(n_smw+1)
norm.quants = qnorm(probs, sm, smwsd)
plot(sort(smw), sort(norm.quants), ylab="Theoretical Quantiles",
xlab="Empirical Quantiles", main="QQ-Plot: Obesity Rates of States in
     the South or Midwest", xlim=limits, ylim=limits, pch=16)
abline(0,1)
```



```
nwm = mean(new)
newsd = sd(new)
n_new = length(new)
limits = c(min(new), max(new))
probs = (1:n_new)/(n_new+1)
norm.quants = qnorm(probs, nwm, newsd)
plot(sort(new), sort(norm.quants), ylab="Theoretical Quantiles",
xlab="Empirical Quantiles", main="QQ-Plot: Obesity Rates of States in
     the Northeast or West", xlim=limits, ylim=limits, pch=16)
abline(0,1)
```



When we observe the first histogram, it is obvious that there is a fairly strong skew to the left since most of the data falls in the furthest 3 bins on the right, while there are much smaller frequencies in the bins on the left. This claim is supported by our QQ-plot that shows a pattern that does not fit the slope line, suggesting a different model is a better fit. The other histogram however does show a more symmetric pattern; however, the data is spread out fairly evenly. The QQ-plot does show a tight fit along the slope line, but there is

not an obvious trend that most of the data is clumped in the center of the graph. Based on our analysis of the graphical summaries, we cannot confidently claim that our two datasets follow a Normal distribution and therefore we do not meet the assumption for the hypothesis test. However, for the sake of continuing our project we will proceed anyways and keep this failed assumption in mind when analyzing our results.

Furthermore, we must assume that our two data sets are independent random samples taken from a population. We cannot comfortably make this assumption for several reasons. First, both of these data sets are not randomly sampled, as we intentionally selected certain subsets of the population data set, states in the South or Midwest and states in the Northeast or West. Additionally, the two sets are not independent from each other, as explained in module 3, due to the interstate characteristic between every state in the United States that allows free movement between them. While we fail to make this assumption, for the sake of this project and our curiosity we will continue with this hypothesis test.

We must also decide whether or not the population variances of our two distributions are equal. Let σ_p^2 be the population variance of the obesity rates among states in South or Midwest of the United States, and let σ_n^2 be the population variance of the obesity rates among states in Northeast or West of the United States. Then we will have s_p^2 to represent the sample variance of the obesity rates among states in South or Midwest of the United States, and we will also have s_n^2 to represent the sample variance of the obesity rates among states in Northeast or West of the United States.

```
smvar = var(smw)
nwvar = var(new)
```

These sample variances are 9.5807635 and 5.0652165, respectively, which have 48.21% difference. With a percent difference that large we cannot comfortably make the assumption that the population variances of these two data sets are equal and will make the assumption that they are not equal.

Our sample statistics are:

$\hat{\mu}_{smw}$ = 31.9827586 $\hat{\mu}_{new}$ = 26.8954545 s^2_{smw} = 9.5807635 s^2_{new} = 5.0652165

Because we do not know the population variances and we assume they are unequal our test statistic is

$$T_o = \frac{(\hat{\mu}_{smw} - \hat{\mu}_{new}) - (\mu_{smw} - \mu_{new})}{\sqrt{\frac{s^2_{smw}}{n_{smw}} + \frac{s^2_{new}}{n_{new}}}}$$

To perform the test we will use the help of R and use its t.test function to give us a p-value.

```
t.test(smw, new, alternative = "two.sided", mu = 0, paired = FALSE, var.equal
= FALSE, conf.level = 0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  smw and new
## t = 6.7945, df = 48.936, p-value = 1.389e-08
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  3.582609 6.591999
## sample estimates:
## mean of x mean of y
##  31.98276  26.89545
```

Based upon our p-value of 1.389e-08 being significantly lower than our alpha = 0.05, we can safely reject the null hypothesis in favor of the alternative. In context, we can conclude that the difference in the mean obesity rate of states in the South or Midwest and states in the Northeast or West are not equal to zero. Because our 95% confidence interval is extremely positive, [3.583,6.591], we can say that if we repeated this test thousands of times, the true difference in population means would fall within [3.583,6.591] 95% of the time. A confidence interval this far from zero may support the claim that states in the South or Midwest may actually have higher obesity rates than states in the Northeast or West.

However, to completely support this claim we would have to perform more hypothesis tests.

As we discovered while checking assumptions, we did not pass either of assumptions needed to perform this hypothesis test. This means that the results of this section should be taken with a grain of salt and not be interpreted as a serious conclusion.

Hypothesis Test 3: Is there relationship between the population in a state and its obesity rate?

So far, we have concluded through hypothesis testing that the region of a state and the median income of a state influence the obesity rate of the state. In this hypothesis test, we will examine the relationship between the region of a state and the state's median income range. Since these two variables have been established to be related to obesity rate, it is worth discovering whether they influence each other. To determine if the two variables are dependent or independent, we will conduct a hypothesis test of independence.

We will use the standard α value of 0.05 for this hypothesis test, and the null and alternative hypotheses will be:

H_0: Median income range and region are independent. H_A: Median income range and region are not independent.

The test statistic for a hypothesis test of independence is χ^2, which is defined below:

$$\chi^2 = \sum \frac{(O_{r,c} - E_{r,c})^2}{E_{r,c}}$$

$O_{r,c}$ and $E_{r,c}$ denote the observed and expected frequencies for each row r and column c of the chi-squared table.

The assumptions for the hypothesis test of independence are: - the data being tested is categorical - the groups for each categorical variable are mutually exclusive - at least 5 expected frequencies in each group of your categorical variable.

The first and second assumptions are met for our particular data, as both variables are categorical measurements, and each group in the categorical variables are mutually exclusive. However, the third assumption is not met in our particular case, since there are only four possible values for median income range and only four possible values for region. However, for the purposes of this project, we will continue to conduct the hypothesis test.

The degrees of freedom for this hypothesis is equal to (rows - 1)(columns -1) = (3)(3) = 9.

The following code conducts the goodness of fit ^{2} test with median income range and region with 9 degrees of freedom. The critical value for a chi-squared distribution with 9 degrees of freedom is 16.919.

```
tbl = table(reg, med_inc)
tbl
```

```
##             med_inc
## reg          1 2 3 4
##   Midwest    0 9 3 0
##   Northeast 0 3 2 4
##   South      9 4 2 2
##   West       1 5 5 2
```

```
chisq.test(tbl)
```

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

```
##
##  Pearson's Chi-squared test
##
## data:  tbl
## X-squared = 28.514, df = 9, p-value = 0.0007823
```

The chi-squared test gave a chi-squared value of 28.514, which is well above the critical value of 16.919. Also, the p-value is 0.000782, which is much less than our alpha value of 0.05. This is enough evidence to reject our null hypothesis in favor of the alternative

hypothesis; median income range and region are not independent. These results imply that there is a relationship between the two variables such that they depend on each other. Further study would have to be done to discover more detail on the nature of this relationship between the two variables, such as determining which regions of the United States are most likely to have the lowest or highest median income ranges.

Module Conclusion:

In Module 4, we used two-sample hypothesis tests to better understand relationships between different variables in our data set. In Module 5, we will continue to investigate relationships between multiple variables, specifically poverty rate and obesity rate. However, instead of using two-sample hypothesis tests, we will use a linear regression model.

Module 5

Module Introduction:

In Module 5, we will create a linear regression model with poverty rate as the predictor variable and obesity rate as the response variable. In Module 3, we determined that states in the two lowest median income ranges were more likely to have higher obesity rates. Module 5 will use another statistic that measures economic status, poverty rate, to further investigate this relationship between economic status and obesity. We hope to further cement our hypothesis that states in poorer economic states tend to have higher obesity rates.

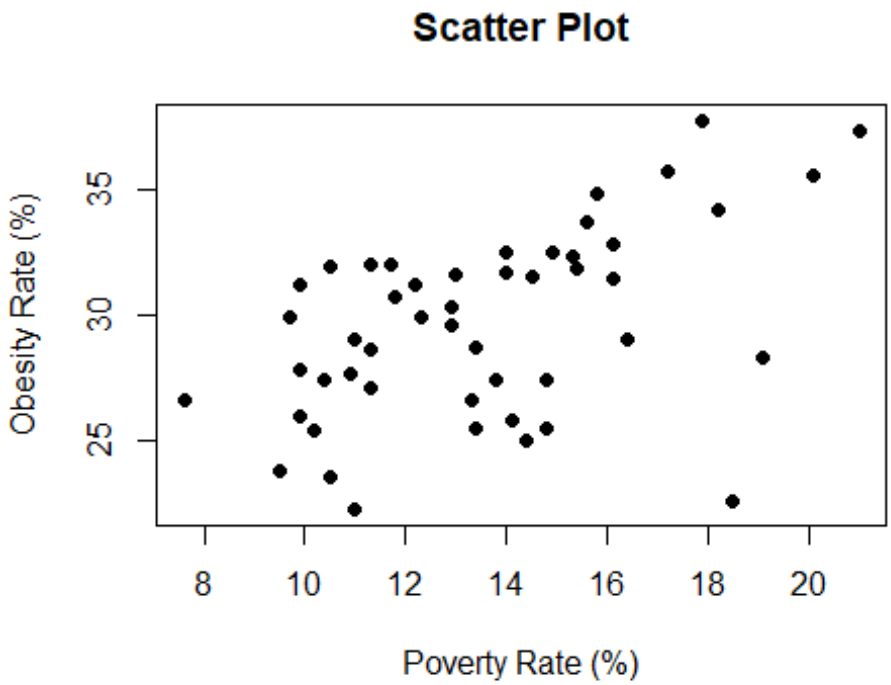
Does a state's poverty rate predict its obesity rate?

A major factor in determining someone's diet is how much money they have available to spend on food. In Module 2, we explored the obesity rates of the states in the two lowest median income ranges, and concluded that the obesity rates of these states are higher than the national average.

In Module 5, we want to further investigate this relationship between financial status and obesity. Another finance-related variable we have available in our data set is the poverty rate of each state. Therefore, we will study the correlation between the poverty rate of a state and the obesity rate of the state. Using a linear regression model, we will determine if the poverty rate of a state can help determine its obesity rate, as well as how much of a state's obesity rate is determined by its poverty rate. We hope to further prove our general hypothesis that states with citizens in worse financial situations (lower median income ranges, higher poverty rates) tend to have more obese citizens.

To investigate the relationship between poverty rates and obesity rates, we must first visually observe the association and possible correlation between them. Below we plot the scatterplot with poverty rates as the predictor variable and obesity rates as the response variable:

```
plot(pov, obs, xlab="Poverty Rate (%)", ylab = "Obesity Rate (%)",
main="Scatter Plot", pch=16)
```



Upon examination of the scatterplot, it appears that there is a moderately positive linear association between the two variables. However, there appears to be one influential point that strays from the linear pattern of the other variables. This influential point is the District of Columbia (18.5, 22.6), who has a very high poverty rate but one of the lowest obesity rates. This value will likely have a significant effect on the creation of our linear regression model and skew our regression line.

Linear Regression Model

```
lmmodel = lm(obs~pov)
```

To perform linear regression model, we must make the following assumptions:

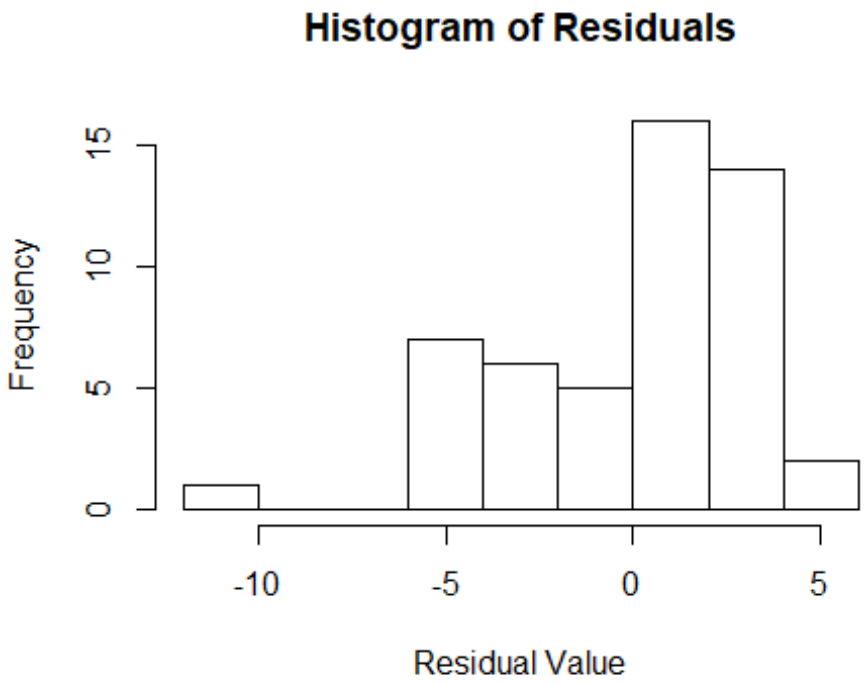
1.) The variables must be independent. This is true because our dataset is made up of observations made at a single point in time, and therefore contain independent observations. 2.) The residuals must be normally distributed with $\mu = 0$ and σ^2 = a

constant. We can confirm this by analyzing various graphical summaries consisting of a histogram, boxplot, and QQ-plot of our residuals.

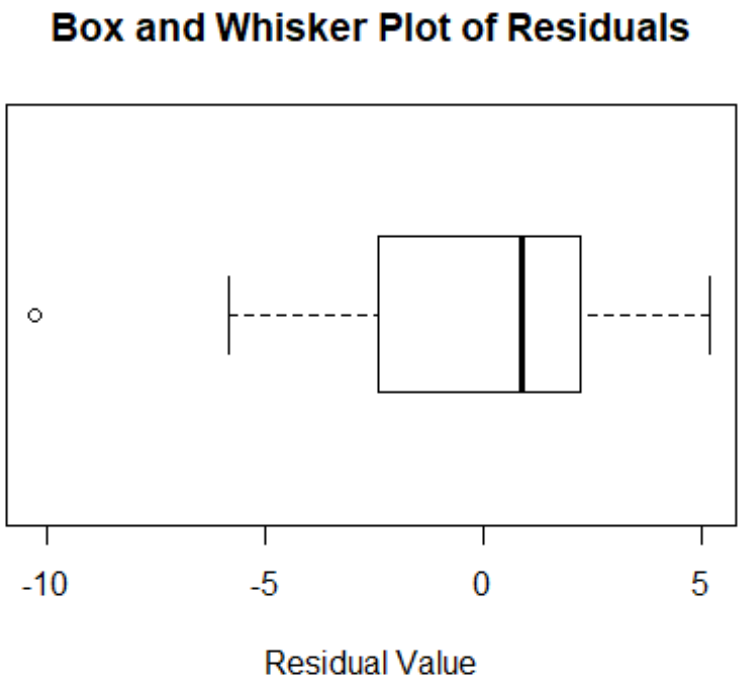
```
res = residuals(lmmodel)
summary(res)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -10.2986  -2.3858   0.8629   0.0000   2.2346   5.1811
```

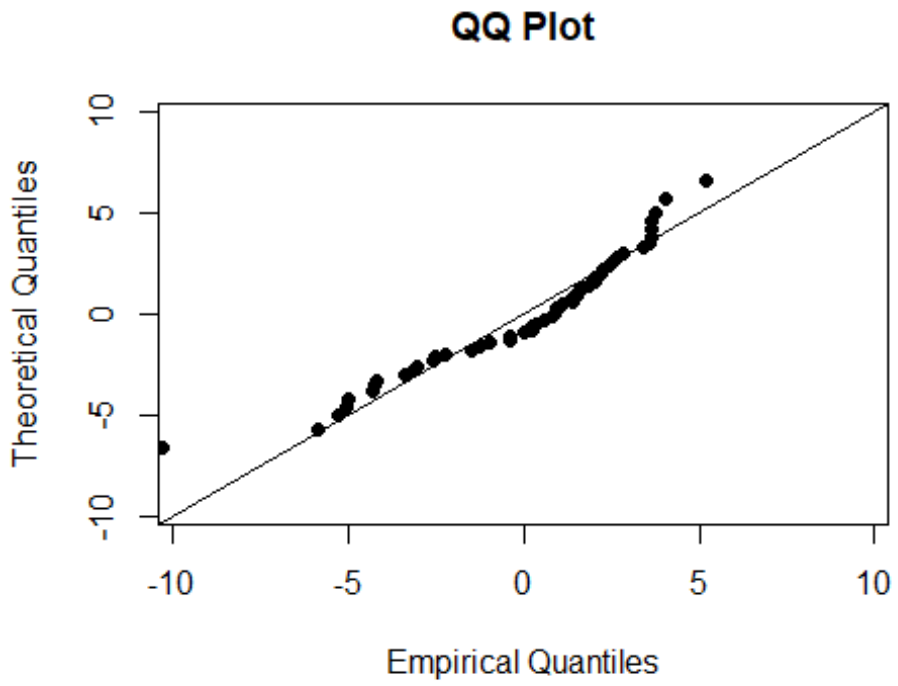
```
hist(res, main="Histogram of Residuals", xlab="Residual Value")
```



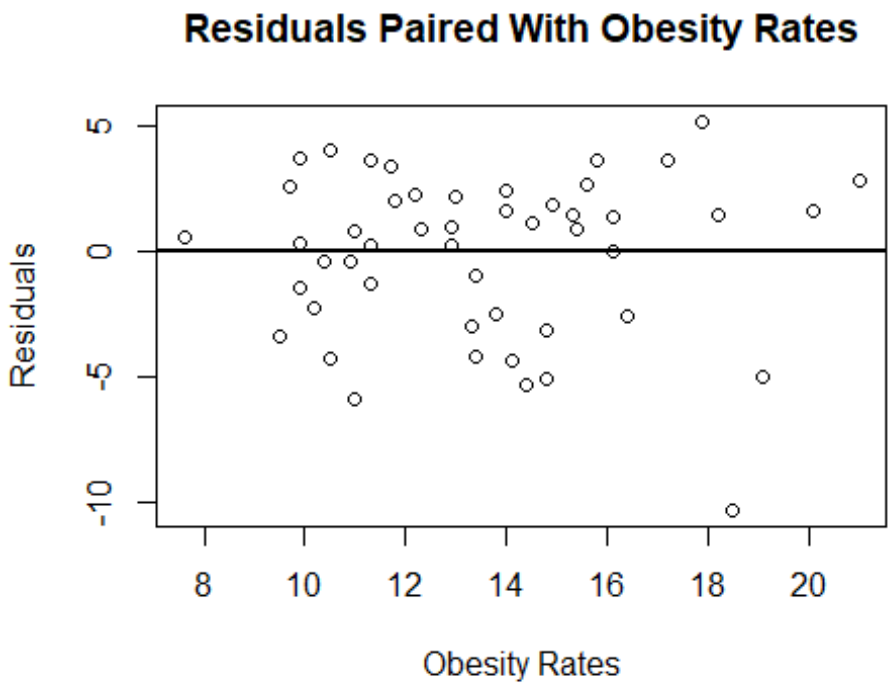
```
boxplot(res, main="Box and Whisker Plot of Residuals", xlab="Residual Value",
horizontal = TRUE)
```



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



```
plot(pov, res, main = "Residuals Paired With Obesity Rates", xlab =
"Obesity Rates", ylab = "Residuals")
abline(0,0, lwd = 2)
```



We begin by looking at the numerical summary of our residuals. We observe that the residuals are obviously centered at a mean of 0, but with a median value of 0.8629, we can see that there is some slight skew causing our median and mean to be different by that magnitude. Looking at the histogram we observe a very strong skew to left, however, this extremely large skew is mostly caused by the influential point of the District of Columbia, but there would still be a fairly obvious skew to the left, with most of our residuals falling in the last two bins. The box and whisker plot also shows the strong skew to the left with the distance between Q1 and Q2 being much larger than the distance between Q2 and Q3. The QQ-plot somewhat shows that our residuals have a Normal Distribution, but since it doesn't have its points tightly fit to the slope line and there is obvious skew in the other graphical summaries we can confidently say that our residuals are not Normally distributed. However, for the sake of this project and our curiosity we will continue with the linear regression model we have selected and analyze its summary.

Additionally, looking at our residual plot of residuals versus our predictor variable of poverty rates, we observe randomness across the plot, with a generally even spread between positive and negative residual values. There is one point extremely far from the rest of our residuals, which is the District of Columbia so this is not surprising as we label it as a influential point. This means if we were to say that our residuals are normally distributed we could confirm they have a constant standard deviation.

Now that we have verified the assumptions are true, we can begin to analyze our linear model.

```
summary(lmmodel)

##
## Call:
## lm(formula = obs ~ pov)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.2986  -2.3858   0.8629   2.2346   5.1811
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  21.1929     2.0989  10.097 1.46e-13 ***
## pov           0.6327     0.1509   4.194 0.000114 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.238 on 49 degrees of freedom
## Multiple R-squared:  0.2642, Adjusted R-squared:  0.2491
## F-statistic: 17.59 on 1 and 49 DF,  p-value: 0.0001145

cor(pov, obs)

## [1] 0.5139656
```

The correlation coefficient between the two variables is 0.514. As we predicted from the scatterplot, this implies a moderate positive correlation between obesity rates and poverty rates.

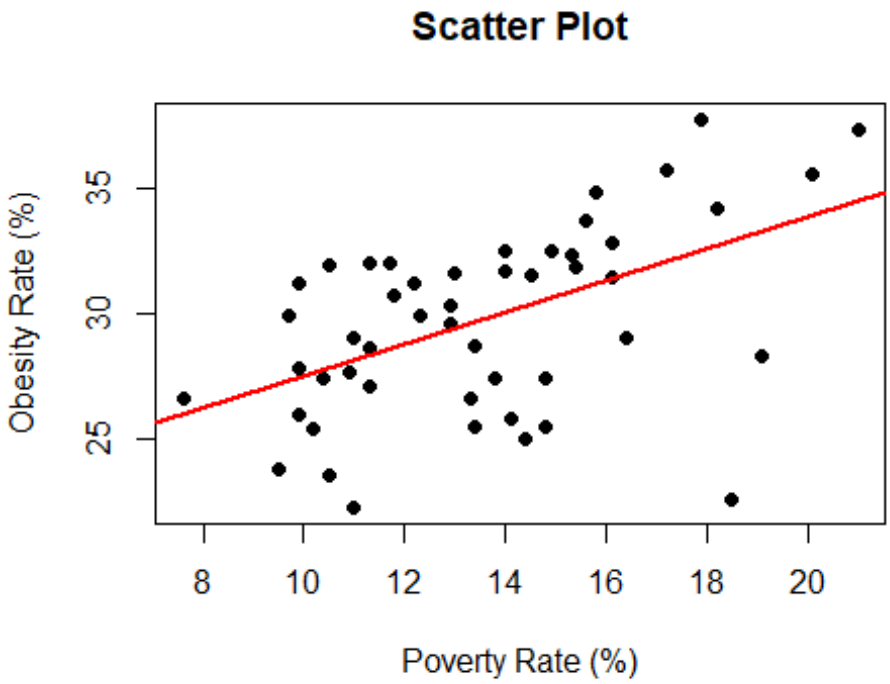
The p value for this coefficient of the predictor variable (poverty rate) in this linear regression model is 0.0001145. This p-value is the resultant p-value from the hypothesis test with the following null and alternative hypotheses, where β_0 represents the coefficient of poverty rate in the linear model, which is the slope of our model. H_0: $\beta_0 = 0$ H_A: $\beta_0 \neq 0$

In other words, this p value states that there is a 0.01145% chance that the predictor variable, poverty rate, is not a relevant variable for determining the response variable, obesity rate. This is a very low percentage, which implies that the poverty rate is an important variable to use in a linear model for obesity rate.

The R-Squared Value is 0.2491. This implies that 24.91% of the variation in a state's obesity rate can be attributed to the state's poverty rate. The other ~75% of the variation is due to other factors.

From our linear model function, we get an intercept coefficient of 21.193 and a slope coefficient of 0.633. We can interpret our intercept coefficient as meaning if a state had a poverty rate of 0%, then that state's obesity rate would be 21.193%, however, this interpretation isn't realistic as a state will never have a poverty rate at 0%. Similarly, the slope coefficient can be interpreted as meaning for every increase 1% of a state's poverty rate, the obesity rate will increase by 0.633%. The following is our equation we can employ to predict a state's obesity rate from its poverty rate: $\hat{obesity\ rate} = 21.193 + 0.633(poverty\ rate)$. Below is our scatterplot with the linear regression equation overlaid.

```
plot(pov, obs, xlab="Poverty Rate (%)", ylab = "Obesity Rate (%)",
main="Scatter Plot", pch=16)
abline(lmmodel, col="red", lwd=2)
```



Using our linear regression equation, we are able to predict a state's obesity rate from its poverty rate. Suppose we have a state with a poverty rate of 15%, using this model we predict this state's obesity rate to be: $21.193 + 0.633(15) = 30.688$ Therefore, the predicted obesity rate for a state with a poverty rate of 15% is 30.688%. The standard deviation of our predicted obesity rate models from our linear regression model is determined from our residual standard error value, which is 3.238. This means that this state's obesity rate could be slightly different from the value of 30.688% by an average of about 3.238%.

Our F-statistic value is 17.59, which is the mean squares of our predictor variable divided by the mean squares of our residuals. As this value increases in size, it becomes more unlikely for our β values to not have an effect in our model. The p-value of 0.0001145 that accompanies our F-Statistic is a good indicator that our linear regression model is an overall effective model, since it is less than a standard $\alpha = 0.05$, in predicting a state's obesity rates from a state's poverty rate in the United States.

Developing a More Complex Model:

As we observed in our first model, the District of Columbia is a very influential point in our linear regression analysis since it skewed our regression line and residual data greatly. We decided that this point should be thrown out of the dataset for a few reasons. First, because the District of Columbia is not a state and only consists of a major metropolitan area, we do not think this population composition is equivalent to other states. Additionally, the District of Columbia has one of the lowest populations in our dataset, so working with a state with this low of a population versus the other states may also affect how its data is represented. We believe that if we adjust our dataset we will have a more accurate linear regression model with higher correlation, higher adjusted r^2 values, and a generally better fitting linear model emulated by a lower p-value with our F-statistic. Below we clean our data and create a new linear model without the District of Columbia.

```
new_pov = pov[which(states != "District of Columbia")]
new_obs = obs[which(states != "District of Columbia")]
lmmodel2 = lm(new_obs~new_pov)
summary(lmmodel2)
```

```
##
## Call:
## lm(formula = new_obs ~ new_pov)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.8488 -1.8575  0.7668  2.0344  4.4525
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  19.8023     1.9074  10.382 7.32e-14 ***
## new_pov       0.7511     0.1382   5.437 1.80e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

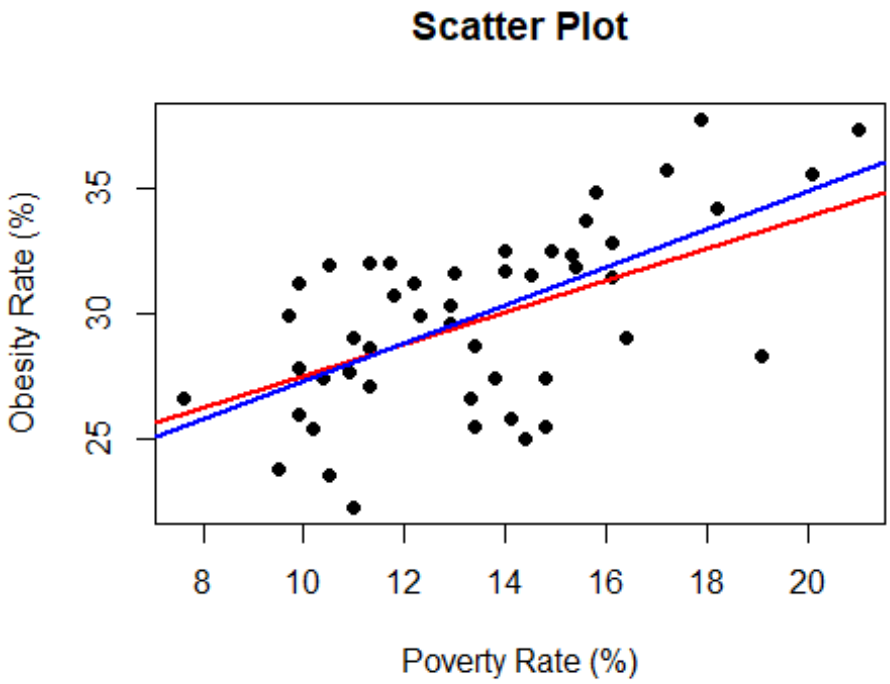
```
## Residual standard error: 2.885 on 48 degrees of freedom
## Multiple R-squared:  0.3811, Adjusted R-squared:  0.3682
## F-statistic: 29.56 on 1 and 48 DF,  p-value: 1.795e-06
```

```
cor(new_pov, new_obs)
```

```
## [1] 0.6173418
```

Based upon the summary of our new model, we can see that many of our predictions are actually true. We now have a stronger correlation of 0.617 versus 0.514. Additionally, our adjusted R-squared value increased from 0.2491 to 0.3682, this means that 12% more of the variation of our obesity rates can be attributed to our linear model for poverty rates when compared to the previous model. Furthermore, our p-value for both of our β values decreased as well meaning that they are more effective in our linear model to explain changes in obesity rates. Finally, the F-statistic p-value has also decreased, which illustrates this new model is better at predicting obesity rate of a state from its poverty rate than the previous model. Below we illustrate the change in our best fit line without the District of Columbia, with the new linear model colored blue.

```
plot(new_pov, new_obs, xlab="Poverty Rate (%)", ylab = "Obesity Rate (%)",
main="Scatter Plot", pch=16)
abline(lmmodel, col="red", lwd=2)
abline(lmmodel2, col="blue", lwd=2)
```



Module Conclusion:

In Module 5, we created a linear regression model with poverty rate as the predictor variable and obesity rate as a response variable. We discovered a moderate positive correlation between the two that became stronger with the removal of the influential point corresponding to District of Columbia. This further cements our claim that states in poorer economic states (lower median income range, higher poverty rate) tend to have higher obesity. In Module 6, we will summarize and reflect on all of our findings thus far in our project.

Module 6

In the last five modules we analyzed our variables in our data set to tell a story that we hope will help raise awareness for the obesity epidemic spreading across America. Our data set consisted of the 50 states and the District of Columbia with statistics on a variety of topics on obesity and economic trends. The goal of this project was to explore the connections between obesity and the many economical, social, and cultural differences in America. We hope that these connections can be communicated to important officials so that changes can be made for a healthier America.

Module 2

In Module 2, we created visual models and distributions for one of each type of variable we have. Specifically, the variables we investigated were obesity rates, region, state policy on nutritional standards, and median income range. For obesity rates, a continuous variable, we found that states' obesity rates in America can be modeled with the Normal distribution family, with estimated parameters 29.79 for population mean and 13.9679 for population variance. This N(29.79, 13.9679) proved to be a good fit for our data with the use of a QQ Plot. For region, a categorical variable, we found that the distribution of the regions of all the states can be modeled with the discrete uniform distribution family, with estimated parameters 1 and 4. The U(1,4) model proved to be a good fit for our data through the use of a goodness of fit hypothesis test. For state policy on nutritional standards, a binary variable, we found that the distribution of the existence or nonexistence of such a policy among all the states could be modeled with the binomial distribution family, with estimated parameters n equal to 51, and p equal to 0.53. This model was an excellent fit for our data, which was verified through another goodness of fit hypothesis test. For median income range, an ordinal variable, we found that the median income ranges of all the states in the US can be modeled with a Poisson distribution, with λ equal to 2.35. We verified that this distribution was a good fit for our data with another goodness of fit hypothesis test.

Module 3

In Module 3, we employed the use of one-sample hypothesis tests to answer three questions:

The first question we answered was: have obesity rates fallen since Michelle Obama's health initiatives in the late-2000s? We used a left-tailed T test to conclude that obesity rates have not fallen since Michelle Obama implemented nutrition reforms and health initiatives. In fact, we had such a high t-value that we predict the opposite may be true; obesity rates have probably increased since 2008. Further hypothesis testing would have to be done to verify this prediction.

The second question we answered was: Are the obesity rates of states in the two lowest median income ranges higher than the national average? We used a right-tailed T test to conclude that the population obesity rate of states in the lower two median income ranges is greater than the national average obesity rate. We also calculated the power value for this hypothesis test to be 66.2%, which is relatively high. This made us question the accuracy of our results. A similar hypothesis test with just the lowest median income range, instead of the lowest two income ranges, might help further prove our claim.

The third question we answered was: are the majority of states' African American obesity rates greater than the national average obesity rate? To investigate this question, we used a proportion test; specifically, a Z approximation on the population proportion. We concluded that the majority of states' African American obesity rates are above the national average, which we expected to be true based on initial patterns we observed in our data set.

Module 4

In Module 4, we moved on from one sample hypothesis tests and began to use two sample hypothesis tests to compare multiple variables from our data set at the same time.

The first question we investigated in Module 4 was: Do states with policies in place to reduce obesity rates actually have lower obesity rates? We performed a one sided, two

sample hypothesis test to conclude that obesity rates in states with a policy to reduce obesity are not lower than obesity rates in states without the policy, and they may actually be higher than states without the policy. More hypothesis testing would have to be conducted in order to verify this claim.

The second question we investigated in Module 4 was: Do regional dieting habits affect obesity rates? We performed a hypothesis test on the difference of means and concluded that the difference in the mean obesity rate of states in the South or Midwest and states in the Northeast or West are not equal to zero. We further predicted that states in the South or Midwest may actually have higher obesity rates than states in the Northeast or West by looking at the confidence interval, but we would have conduct more hypothesis tests to verify this prediction.

The third question we investigated in Module 4 was: Is there a relationship between the region of a state and the state's median income range? We performed a hypothesis test on independence between the two variables region and median income range, and concluded that median income range and region are not independent. Further study would have to be done to discover the exact nature of the relationship between the two variables.

Module 5

In Module 5, we investigated one of the most intriguing relationships in our project: the relationship between economic status and obesity rate. We created a linear regression model with obesity rate as the response variable and poverty rate as the predictor variable. The correlation coefficient was 0.514, which shows a moderate positive correlation. The linear model that we produced for predicting a state's obesity rate from it's poverty rate was $\hat{obesity\ rate} = 21.193 + 0.633(poverty\ rate)$. We used this equation to predict a state's obesity rate with a theoretical poverty level of 15%. We analyzed our residuals and made a QQ Plot to verify the accuracy of our linear model, and while we did not meet all of the assumptions to create a linear regression model, we proceeded anyways to demonstrate our knowledge and answer our questions. We created a more complex model

by removing the influential point corresponding to Washington, D.C. to create an alternative linear model with an even higher correlation coefficient of 0.617.

Conclusions

After performing a variety of statistical analyses to answer our questions we have posed, we discovered that, despite efforts by the government to decrease or slow down the obesity epidemic, obesity rates in America continue to increase. Also, we have realized there are a variety of factors that influence the rise of obesity rates, such as differences in region, economic status, and race. These results show that it is not necessarily up to the government to find ways to decrease obesity rates; it is up to the people to recognize the causes of this epidemic and make changes to our daily habits found in American culture. Additionally, if there was a universal solution to the high poverty rates across the country, this solution could inadvertently help stop the obesity epidemic, as evidence from our project shows that poverty rates account for almost 50% of the variation of obesity rates among states.

There are several great ideas for a potential phase II after the completion of this project. First, our initial idea was to study the same variables we analyzed for states but using county data instead. This idea was thrown out, as the data collection and sampling process would have been difficult and time consuming, however, if we were allotted more time and manpower to perform the same analyses in this project on county data, we believe we could find more concrete and accurate claims for regions and populations of the United States. We are extremely interested in the potential findings from analyzing the thousands of counties that make up the United States, instead of just data on 51 states.

Furthermore, our linear regression line did not pass all of the assumptions required to perform a legitimate linear regression analysis between obesity rates and poverty rates among states in America. We believe that using the county data instead would potentially create a far superior model in estimating obesity rates from the poverty rate of a population set. Additionally, a multiple linear regression model may also create a stronger linear model that would allow us to predict obesity rates accurately from confounding variables, such as poverty rates, that all influence the variability of our predictions. We

believe that an accurate and dependable linear regression analysis focused on economic and regional differences between communities would be the most beneficial way of attempting to solve the problem of rising obesity rates.

References:

The State of Obesity (2018). State Briefs Ranked by Highest Obesity Rate among Adults (2016). Retrieved from https://stateofobesity.org/states/

The State of Obesity (2018). Early Childhood Nutrition Standards. Retrieved from https://stateofobesity.org/state-policy/policies/usdastandards

United States Census Bureau (2013, April 8). Census Regions and Divisions of the United States. Retrieved from https://www2.census.gov/geo/pdfs/maps-data/maps/reference/us_regdiv.pdf

United States Census Bureau (2018, July 12). Small Area Income and Poverty Estimates. Retrieved from https://www.census.gov/data-tools/demo/saipe/saipe.html?s_appName=saipe&map_yearSelector=2016&map_geoSelector=aa_c&menu=grid_proxy&s_inclStTot=y&s_USStOnly=y

270 to Win (2018). 2016 Presidential Election Live Results. Retrieved from https://www.270towin.com/live-2016-presidential-election-results/state-by-state/

Appendices

A) One Page of Data Spreadsheet

states	obs	pov	reg	white	black	hispanic	pop	med_inc	state_pol	RvsD
West Virginia	37.7	17.9	South	36.0	44.6	37.7	1,779,953	1	Yes	R
Mississippi	37.3	21.0	South	31.9	44.6	22.3	2,892,926	1	Yes	R
Alabama	35.7	17.2	South	32.4	44.1	28.1	4,741,355	1	Yes	R
Arkansas	35.7	17.2	South	34.0	44.2	32.4	2,898,653	1	Yes	R
Louisiana	35.5	20.1	South	32.6	42.9	32.2	4,546,072	1	Yes	R
Tennessee	34.8	15.8	South	31.3	45.1	33.0	6,489,044	1	No	R
Kentucky	34.2	18.2	South	33.4	42.4	25.0	4,298,981	1	No	R
Texas	33.7	15.6	South	29.2	42.4	37.4	27,236,431	2	Yes	R
Oklahoma	32.8	16.1	South	32.3	36.6	36.7	3,809,426	1	No	R
Michigan	32.5	14.9	Midwest	30.7	37.4	38.4	9,702,326	2	Yes	R
Indiana	32.5	14.0	Midwest	31.8	41.7	28.7	6,432,851	2	No	R
South Carolina	32.3	15.3	South	28.9	41.7	28.2	4,820,435	1	Yes	R
Nebraska	32.0	11.3	Midwest	30.8	37.2	31.8	1,851,082	2	Yes	R
Iowa	32.0	11.7	Midwest	31.9	32.1	29.9	3,031,272	2	Yes	R
North Dakota	31.9	10.5	Midwest	31.5	15.9	37.2	731,713	3	Yes	R
North Carolina	31.8	15.4	South	28.1	40.2	31.2	9,885,985	2	Yes	R
Missouri	31.7	14.0	Midwest	30.7	38.6	32.7	5,911,099	2	No	R
Illinois	31.6	13.0	Midwest	29.2	41.3	36.3	12,502,043	3	No	D
Ohio	31.5	14.5	Midwest	30.8	37.6	27.8	11,287,358	2	Yes	R
Georgia	31.4	16.1	South	28.9	37.7	28.4	10,040,187	2	Yes	R
Alaska	31.2	9.9	West	28.7	43.6	27.9	647,652	4	Yes	R
Kansas	31.2	12.2	Midwest	31.5	43.1	35.2	2,825,887	2	No	R
Wisconsin	30.7	11.8	Midwest	30.5	39.9	31.2	5,628,526	2	No	R
Delaware	30.7	11.8	South	29.4	36.5	32.1	926,860	3	No	D
Pennsylvania	30.3	12.9	Northeast	29.5	36.4	39.5	12,368,248	2	No	R

B) Billing Invoice

INVOICE

Henry Hart & Colin McCormick
182 6th Street NW
Atlanta, GA 30313
[678-123-4567]
email@gmail.com

Invoice No : 100432
Date : 4/16/2018
Customer ID : ABC12345

Obesity Research of America
123 Main St
Atlanta, GA 30313
678-987-6543

Date	Hours	Description	Unit Price	Total
02/04/18 - 02/10/18	2	Research	$25.00	$ 50.00
02/11/18 - 02/17/18	4	Data Collection and Variables	$25.00	$ 100.00
02/18/18 - 02/24/18	3	Variable Theoretical Model Research	$25.00	$ 75.00
02/25/18 - 03/03/18	1	Drafting of Project	$25.00	$ 25.00
03/04/18 - 03/10/18	2	Drafting the Introduction	$25.00	$ 50.00
03/11/18 - 03/24/18	10	Creating Models for Variables	$25.00	$ 250.00
03/25/18 - 03/31/18	3	Drafting Hypothesis Tests	$25.00	$ 75.00
04/01/18 - 04/07/18	20	Conducting Hypothesis Tests	$25.00	$ 500.00
04/08/18 - 04/14/18	30	Conducting Linear Regression Analysis	$25.00	$ 750.00
04/15/18 - 04/18/18	20	Drawing Conclusions and Finalizing Analysis Report	$25.00	$ 500.00
			Subtotal	$ 2,375.00
			Sales Tax @ 8.00%	$ 190.00
			TOTAL	$ 2,565.00

Make all checks payable to Henry Hart & Colin McCormick.

THANK YOU FOR YOUR BUSINESS!

C) Consulting Log

Date	Hours spent working	What was accomplished
02/04/18 - 02/10/18	2	brainstormed project ideas and researched potential datasets
02/11/18 - 02/17/18	4	found final datasets, changed idea to obesity rates, created project page with idea, sources and variables
02/18/18 - 02/24/18	3	began work on module 2, chose variables to model and created summaries
02/25/18 - 03/03/18	1	drafted layout of Module 1
03/04/18 - 03/10/18	2	rough draft of module 1
03/11/18 - 03/17/18	5	worked independently on module 2, two variable types per person
03/18/18 - 03/24/18	5	Finalized module 2
03/25/18 - 03/31/18	3	drafted ideas for HTs in module 3 and 4, split it up 2 single sample and 1 two sample and vice versa per person
04/01/18 - 04/07/18	20	Finalized module 3 and began working on last HTs of module 4
04/08/18 - 04/14/18	30	Finalized module 4 began working on regression analysis of obesity and poverty, began working on reference list and appendices
04/15/18 - 04/18/18	20	Finalized module 5 and 6, finished references and appendices, created title page and table of contents, printed project to turn in

D) Acknowledgements

We would like to send our gratitude to everyone who helped us accomplish this statistical analysis project. Most importantly, we would like to thank Dr. Waller who has provided us with an excellent foundation knowledge in statistics to apply on our own and for answering any of our questions and providing example code snippets and projects. Additionally, we would like to thank our ISyE 2027 professor Dr. Andradottir who provided us with our background knowledge of probability. We would like to express our gratitude for Douglas Montgomery and George Runger, who wrote our statistitc text book Applied Statistics and Probability for Engineers, and Beth Chance and Allan Rossman, who wrote our textbook Investigating Statistical Concepts, Applications, and Methods. Without either of these resources, we would not have been able to complete this project, as they provided many answers to our small questions and cemented our understanding of statistics by providing examples, formulas, and definitions of relevant statistical topics to this project. Finally, we would like to thank our peers in ISyE 2028, who helped with questions involving code and statistical concepts.