STAT 345 - Chapter 6: Descriptive Statistics

Kellin Rumsey

September 30, 2017

An Illustrative Dataset

The CDI dataset contains Demographic information for the 440 most populated counties in the United States. Throughout this handout, we will be focusing on two variables.

1. `percentPoverty` - The percent of the population which is below poverty level.
2. `percentDiploma` - The percent of the (adult) population with a High School diploma.

```
CDI_data <- read.csv('http://math.unm.edu/~knrumsey/classes/spring17/MiniProjects/data.csv')
percentPoverty <- CDI_data$PercentBelowPoverty
```

Measures of Center

Mean

The **sample (arithmetic) mean**, is what most people mean when they say "average". In Statistics, we denote the mean of observations $x_1, x_2, \cdots x_n$ by $\bar{x}$, which we obtain by adding up the observations and dividing by the sample size.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

We can compute the sample mean easily using R.

```
mean(percentPoverty)
```

```
## [1] 8.720682
```

Median

The **sample median** of a set of numbers (also referred to as the *second quartile* or the 50^{th} *percentile*), is any number M that separates the data in half. That is, the number of of observations which are less than or equal to M should be the same as the number of observations which are greater than or equal to M. For instance, if $x_1 = 1$, $x_2 = 2$ and $x_3 = 3$, then the Median must be $M = 2$. However, if we add $x_4 = 4$, then any number between 2 and 3 fits the definition of a Median. In this case, we usually define M to be the midpoint between 2 and 3, hence $M = 2.5$.

In general, if we have n observations $x_1, x_2, \cdots x_n$, we can arrange them from smallest to biggest. We call the sorted sample, $x_{(1)}, x_{(2)}, \cdots x_{(n)}$. Then the Median is defined as follows.

$$M = \begin{cases} x_{(\frac{n+1}{2})}, & n \text{ is odd} \\ \frac{x_{(n/2)} + x_{(n/2+1)}}{2}, & n \text{ is even} \end{cases}$$

We can compute the sample median easily using R.

```r
median(percentPoverty)
```

```
## [1] 7.9
```

Mode

The **sample mode** of a set of numbers, is the value which occurs most frequently. Intuitively, the Mode of a distribution refers to the location of the highest peak of the distribution. Distributions sometimes have multiple modes, so it is always a good idea to plot the data using a histogram. As an example, we consider the following sample $[1, 2, 4, 4, 4, 5, 6, 6, 7]$. Since 4 occurs more often than any other value, it is the sample Mode. We can find this in R by creating a table of the values, and extracting the value which occurs the most often.

```r
names(which.max(table(percentPoverty)))
```

```
## [1] "9.8"
```

When the observations can only take a handful of values (think discrete), the sample mode works well. But in the `percentPoverty` example the data is nearly continuous. In these cases, the sample Mode can be misleading. One simple approach is to round the data before computing the sample mode.

```r
names(which.max(table(round(percentPoverty))))
```

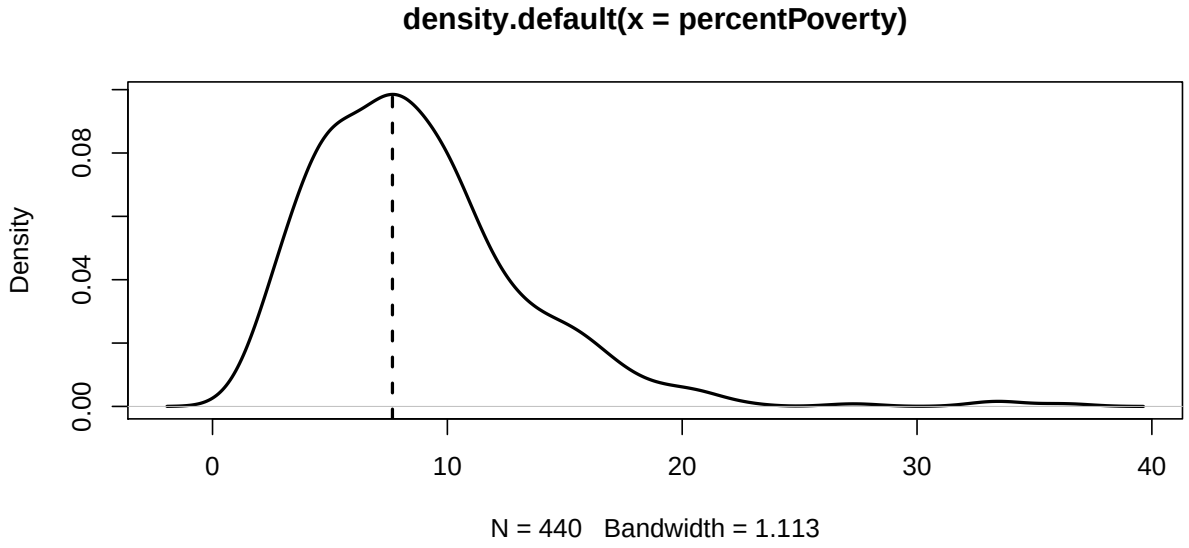
```
## [1] "8"
```

Without going into too much detail, a non-parametric approach may provide better results. A *kernel density estimate (KDE)* is a way of approximating the density curve of a data set. Since the population mode is the location of the peak of a density curve, we can estimate the sample mode by finding the value which maximizes the KDE of the dataset.

```r
kde <- density(percentPoverty)
mode_np <- kde$x[which.max(kde$y)]
mode_np
```

```
## [1] 7.661602
```

```r
#Plot density estimate and mode
plot(kde, lwd=2)
abline(v=mode_np, lwd=2, lty=2)
```



Resistance

Note that the sample mean, sample median and sample mode are NOT the same as the mean and median we discussed in chapters 3 and 4. We can refer to these as the "population" (or "true") mean, median and mode. Although we will discuss this more in chapter 7, it should be intuitive that the sample mean will be a reasonable "estimator" for the population mean (and similarly for median and mode).

We say that an estimator is *resistant*, if it doesn't change too much with an outlier. For instance, consider this simple example.

- In the mid-1980's, the average (sample mean) starting salary of geography students at UNC was about $150,000$.
- Seems suspiciously high... In fact, the Median was probably somewhere around $40,000$.
- We have Michael Jordan to thank for this.

Moral of the story, the sample mean is highly dependent on outliers, but the sample median is not. Another quick example, consider three data points $x_1 = 1, x_2 = 2, x_3 = 3$. Clearly the sample mean and median are both 2. Now add a data point $x_4 = 1000$, now $\bar{x} = 251.5$, but $M = 2.5$.

Measures of Spread

Sample Variance

The **sample variance**, denoted s^2, is a commonly used measure of spread defined by

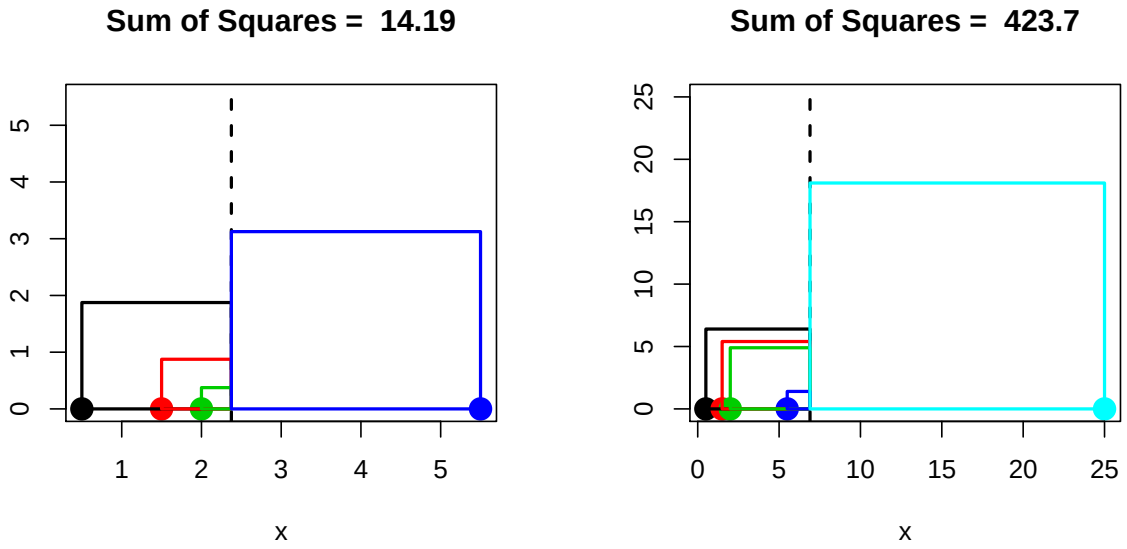
$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

The **sample standard deviation**, denoted s, is simply the square root of the sample variance. The term $\sum_{i=1}^{n}(x_i - \bar{x})^2$ is referred to as the **sum of squares**. The fact that we divide by $n-1$ rather than n is often unintuitive to people.

- Since the formula for s^2 involves $\bar{x}$, we have lost a *degree of freedom*. Thus we are dividing by the degrees of freedom $n-1$.

- The variance of a single observation is now undefined instead of 0.
- We will see in chapter 7, that dividing by n leads to an "biased" estimator of σ^2, but dividing by $n-1$ yeilds an "unbiased" estimator.

The following figure illustrates both *how* the sample variance measures spread, and also demonstrates that s^2 is NOT resistant to outliers. The sum of squares is just the sum of the areas of the squares in the figures below.



There is also a convenient identity for calculating s^2 which is sometimes called the **shortcut method**. This identity will be very useful in the next chapter.

$$s^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)$$

Order Statistics

Given a sample $x_1, x_2, \cdots x_n$, the order statistics are just the sorted observations $x_{(1)}, x_{(2)}, \cdots x_{(n)}$. That is $x_{(i)} \leq x_{(i+1)}$ for every $i = 1, 2, \cdots n-1$. This means that $x_{(1)}$ can be interpreted as the minimum observation, and $x_{(n)}$ is the maximum observation.

The **Five Number Summary** refers to 5 numbers (shocker) which describe the spread of the observations.

- Minimum: $x_{(1)}$
- Maximum: $x_{(n)}$
- Median (M): Defined above
- First Quartile (Q_1): The median of all observations to the left of M.
- Third Quartile (Q_3): The median of all observations to the right of M.

Intuitively, the quartiles divide the observations into 4 (approximately) equally sized groups. For example, 25% of the observations should be less than Q_1 and 25% of observations should be greater than Q_3.

As our first example, consider the observations $(3, 2, 5, 1, 4)$. The five number summary is found as follows:

- Minimum: $x_{(1)} = 1$
- Maximum: $x_{(5)} = 5$
- Median: $M = 3$

- First Quartile: There are 3 observations less than or equal to $M = 3$. Q_1 is the median of these observations, hence $Q_1 = 2$.
- Third Quartile: There are 3 observations greater than or equal to $M = 3$. Q_3 is the median of these observations, hence $Q_3 = 4$.

We can calculate the 5 number summary in R as follows.

```
quantile(c(1,2,3,4,5))
```

```
##   0%  25%  50%  75% 100%
##    1    2    3    4    5
```

```
quantile(percentPoverty)
```

```
##   0%  25%  50%  75% 100%
##  1.4  5.3  7.9 10.9 36.3
```

Finally, we define the **sample range** $R = x_{(n)} - x_{(1)}$ and the **Inter-quartile range (IQR)** $IQR = Q_3 - Q_1$.

Identifying Outliers

An outlier is an observation point which is far from the other observations. An outlier can be an extreme case, such as the Michael Jordan example from earlier, or it can indicate some kind of error (experimental, measurement etc.). For instance, if a Bio-Statistics study records the weight of a subject to be 1,500 pounds, we can assume that this outlier is due to mis-recorded data. In general, it is not okay to eliminate outliers from the dataset without good reason.

Outliers are difficult to describe mathematically. Most of the time, it is sufficient to plot a histogram and see if any of the points look weird. With that said, we will discuss two easy methods for flagging certain observations that may be outliers.

First Method - Assuming Normality

If we believe the data is normal, then we know that a z-score which is larger than 2 in absolute value should (theoretically) occur less than 5% of the time. Therefore, we can standardize each observation

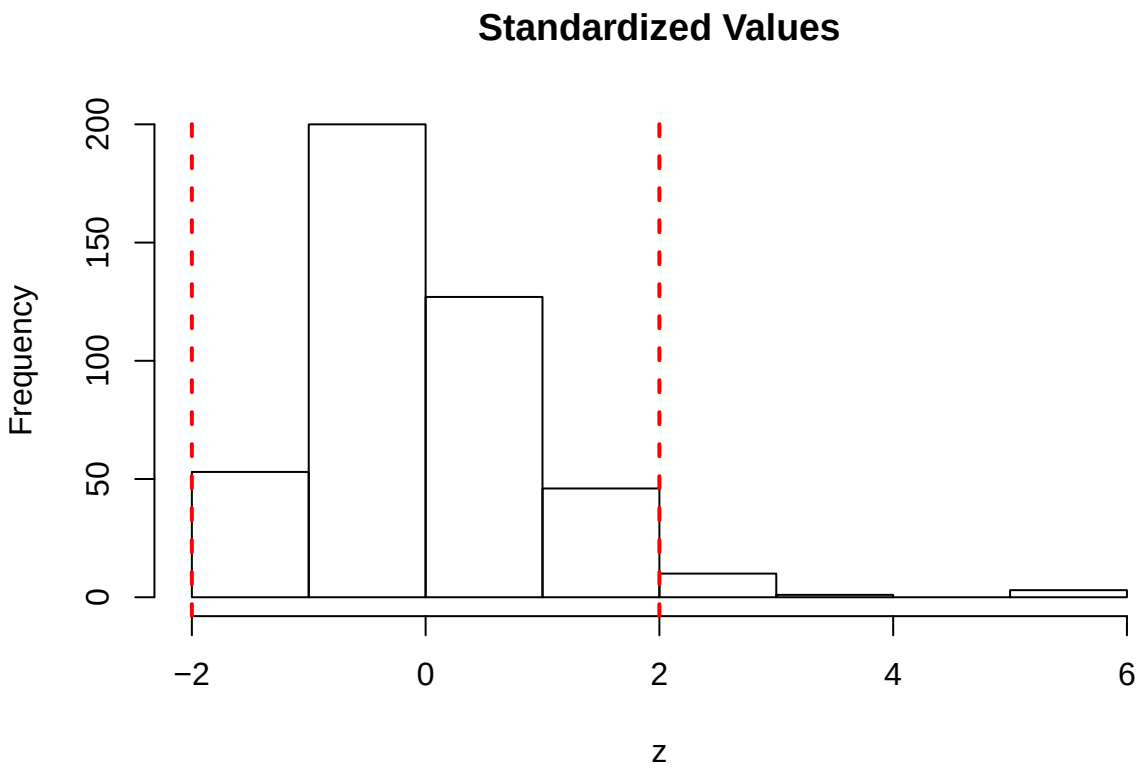
$$z_i = \frac{x_i - \bar{x}}{s}$$

and flag any observations such that $|z_i| > 2$.

```
z <- (percentPoverty - mean(percentPoverty))/sd(percentPoverty)
outliers <- which(abs(z) > 2)
percentPoverty[outliers]
```

```
##  [1] 19.5 22.4 27.3 20.6 36.3 33.7 19.1 19.6 20.7 33.1 18.6 18.7 20.8 20.7
```

```
hist(z, main='Standardized Values', xlab='z')
abline(v=c(-2,2), lwd=2, lty=2, col='red')
```



Therefore, according to this rule, we have flagged several observations as "upper" outliers. We can see that our data clearly isn't symmetric, so this should be taken with a grain of salt.

Bonferonni Improvement

In the previous approach, we flag an observation as an outlier if it is so "extreme" that it would have occured less than 5% of the time. This means that in a sample of 100 observations, we *expect* 5 of these points to be classified as outliers on average. With this logic, are these points really outliers?

A better approach is to let the critical value grow with the sample size. Without going into details, we can use Bonferonni's rule. This replaces the criteria $|z_i| > 2$ with $|z_i| > z^*$ where $z*$ is the $1 - \frac{\alpha}{2n}$ quantile of the $N(0, 1)$ distribution. Hence setting $\alpha = 0.05$ and using $n = 440$, we have $z^* = 3.86$. Roughly speaking, this ensures that Normally distributed data with n observations will flag no outliers with probability $1 - \alpha$.

```
z_crit <- qnorm(1-.05/(2*length(percentPoverty)))
z_crit
```

```
## [1] 3.859463
```

```
z <- (percentPoverty - mean(percentPoverty))/sd(percentPoverty)
outliers <- which(abs(z) > z_crit)
percentPoverty[outliers]
```

```
## [1] 27.3 36.3 33.7 33.1
```

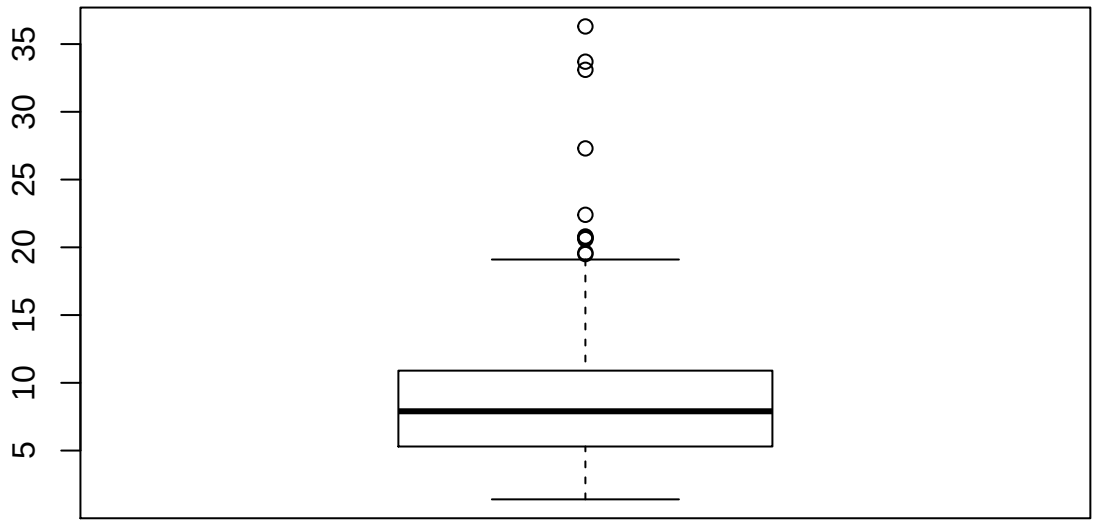
Second Method - Using Quantiles

There is another commonly use method of flagging potential outliers, which is subtly related to the previous method. As before, it works best if the data is Normal, or at least symmetric. First, we discuss Boxplots.

Boxplots

Boxplots are a convenient way to readily visualize the 5 number summary. R has a built in function for plotting these easily.

```
boxplot(percentPoverty)
```



The point symbols in the box-plot are potential outliers, according the $1.5 \times IQR$ rule.

$1.5 \times IQR$ Rule

We define the upper and lower fence by

$$UF = Q_3 + 1.5(IQR)$$

$$LF = Q_1 - 1.5(IQR)$$

any observations below the lower fence, or above the upper fence are flagged as potential outliers. We can accomplish this in R with the following code.

```r
UF <- quantile(percentPoverty, probs=0.75) + 1.5*IQR(percentPoverty)
LF <- quantile(percentPoverty, probs=0.25) - 1.5*IQR(percentPoverty)
outliers <- union(which(percentPoverty > UF), which(percentPoverty < LF))
percentPoverty[outliers]
```

```
##  [1] 19.5 22.4 27.3 20.6 36.3 33.7 19.6 20.7 33.1 20.8 20.7
```

Assessing Distributional Fit

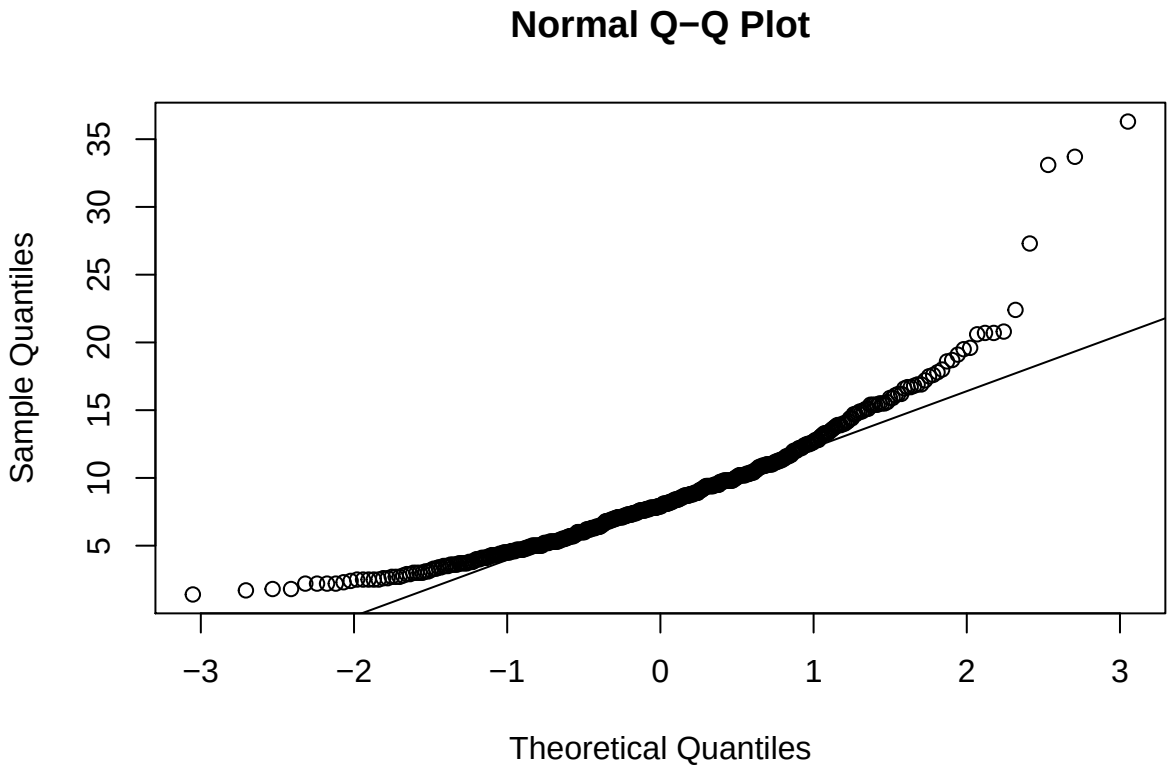
As we move along with the Statistics portion of this class. We will often have to assume a distribution for the data. If the distributional assumption is wrong, this can lead us to false conclusions.

Probability Plots

Often, we assume that our data is Normally distributed. Plotting a histogram of the data is a good way to determine if this assumption is reasonable. Probability plots are another good way to assess this fit, and they can be extended to other distributions as well. Simply put, a probability plot, creates a scatterplot of $x_{(i)}$ vs $E(X_{(i)})$ under some distributional assumption. Clearly, these expected values can be hard to find, and is beyond the scope of this class. Still, we can benefit from other peoples hard work.

If the distributional assumption is appropriate, then we expect these points to fall approximately on a line.

```r
qqnorm(percentPoverty)
qqline(percentPoverty)
```



Since these points don't fall on the line, they tell us (what we already knew) that these data are most likely not Normally distributed. In general,

- If the points are concave up, the distribution is skewed right.
- If the points are concave down, the distribution is skewed left.
- If the points are S shaped, the distribution has "heavy tails".

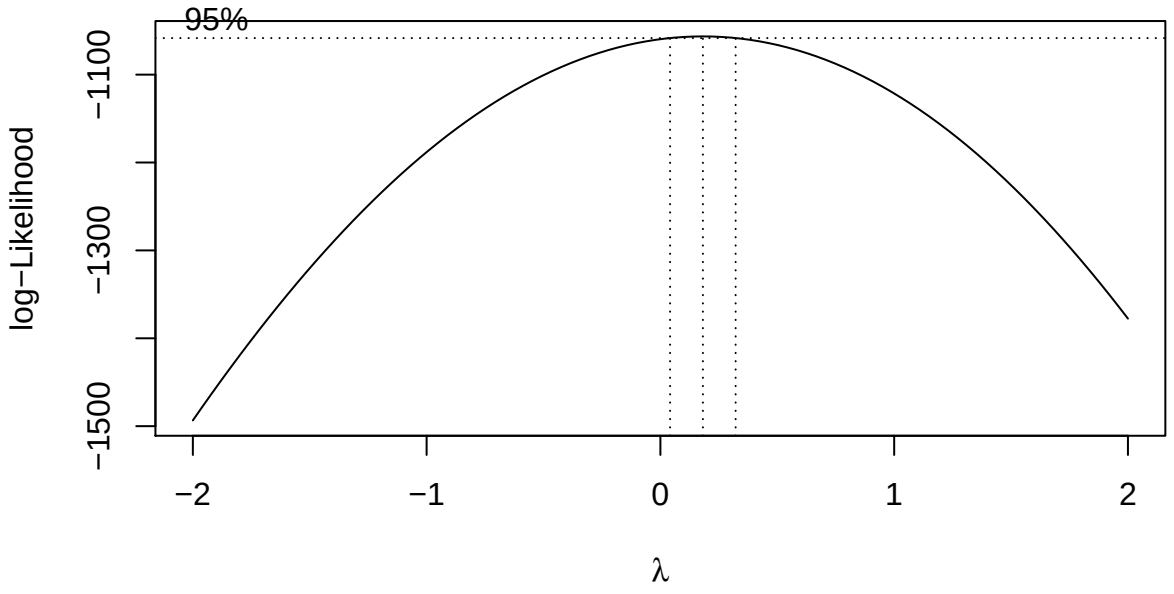
Transformations

Sometimes, we can transform our data $y_i = g(x_i)$ and acheive (approximate) normality. One useful family of transformations known as the Box-Cox Power Transformation is

$$y_i = \begin{cases} x_i^\lambda, & \lambda \neq 0 \\ \ln(x_i), & \lambda = 0 \end{cases}$$

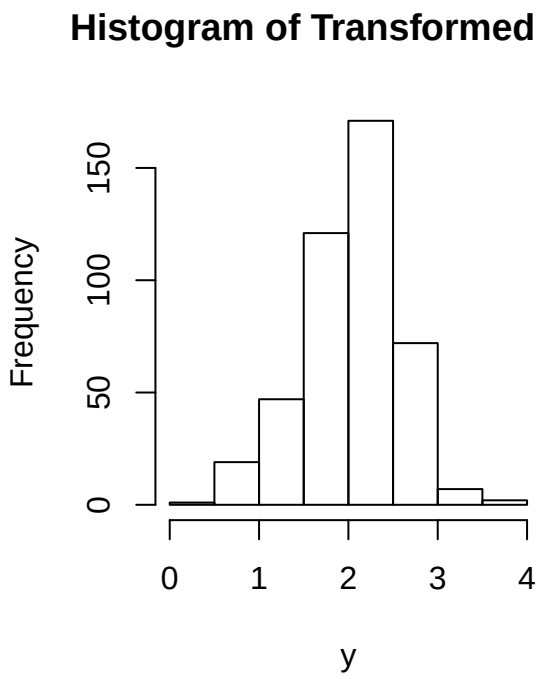
A Box-Cox plot tries out many of these transformations very quickly, and gives a curve indicating which values of λ are the most likely. To do this in R, we need to install the "MASS" package `install.packages('MASS')`.

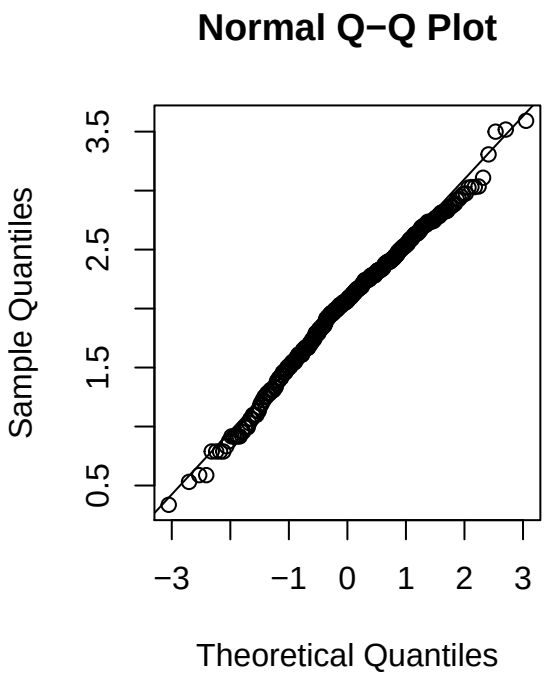
```
library(MASS)
boxcox(lm(percentPoverty~1))
```



We typically prefer to stick to transformations which make sense (such as $\lambda = 1/2$ or $\lambda \in \mathbb{Z}$). Since the dotted band is pretty close to 0, this indicates that a *log* transformation might help. We check this by creating a histogram and Normal probability plot of $y_i = \ln(x_i)$.

```
y <- log(percentPoverty)
par(mfrow=c(1,2))
hist(y, main='Histogram of Transformed Data')
qqnorm(y)
qqline(y)
```





The data appear much more normal than before. We may now be able to continue with our analysis assuming normality, but we can't forget to transform our results back to the original scale at the end.

Question: If a log-transformation makes our data normal, what does that say about the distribution of percent below poverty level?

Relationships Between Variables

Assume we have paired observations of two variables (x_i, y_i). For instance x_i and y_i could represent the height and weight of a particular person. Let $\bar{x}$, $\bar{y}$, s_x and s_y denote the means and standard deviations of the two variables. We define the **sample correlation** as

$$r = \frac{1}{n-1} \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{s_x s_y}$$

As with the population correlation ρ, we have that $-1 \leq r \leq 1$. The sample correlation is a quantitative measure of the strength of direction of the linear relationship between two variables.

For our example with the CDI dataset, we consider a second variable, "percent of population with HS diploma". The correlation can be calculated in R as follows.

```
percentGraduate <- CDI_data$PercentHSGraduates
cor(percentGraduate, percentPoverty)
```

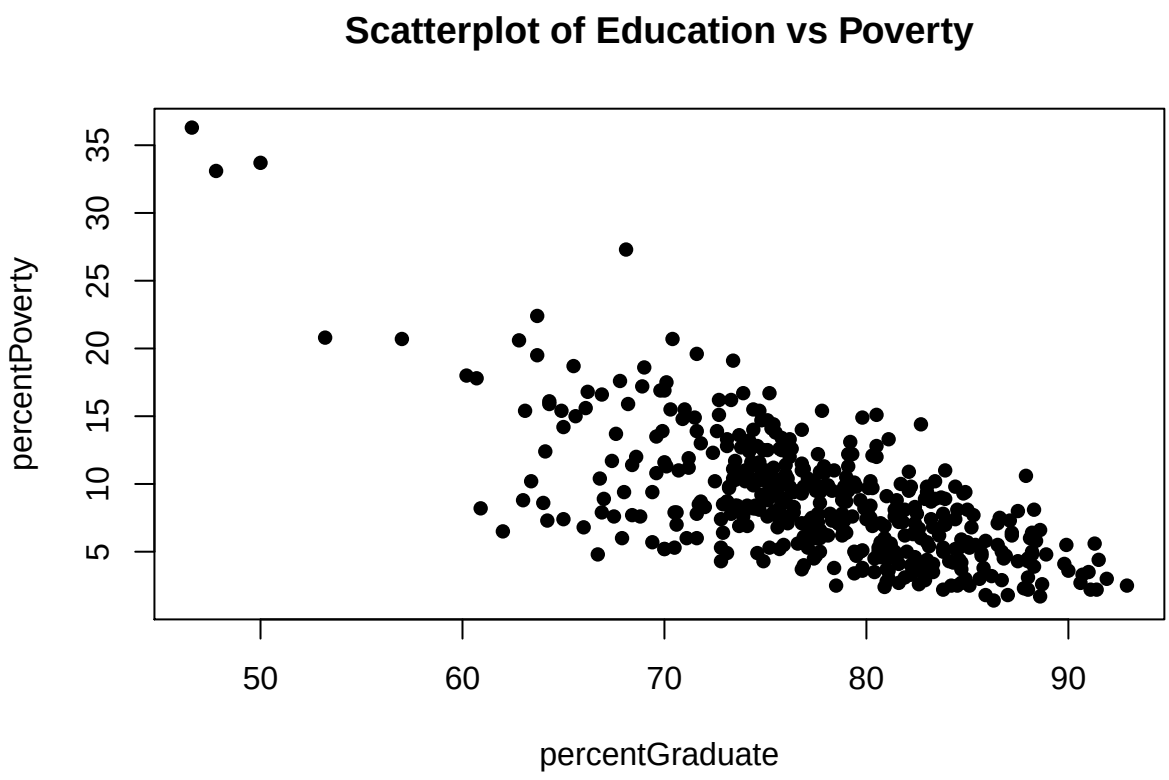
```
## [1] -0.6917505
```

Thus we can determine that the relationship between these variables is moderately strong, and they are inversely related (as expected).

Scatter Plots

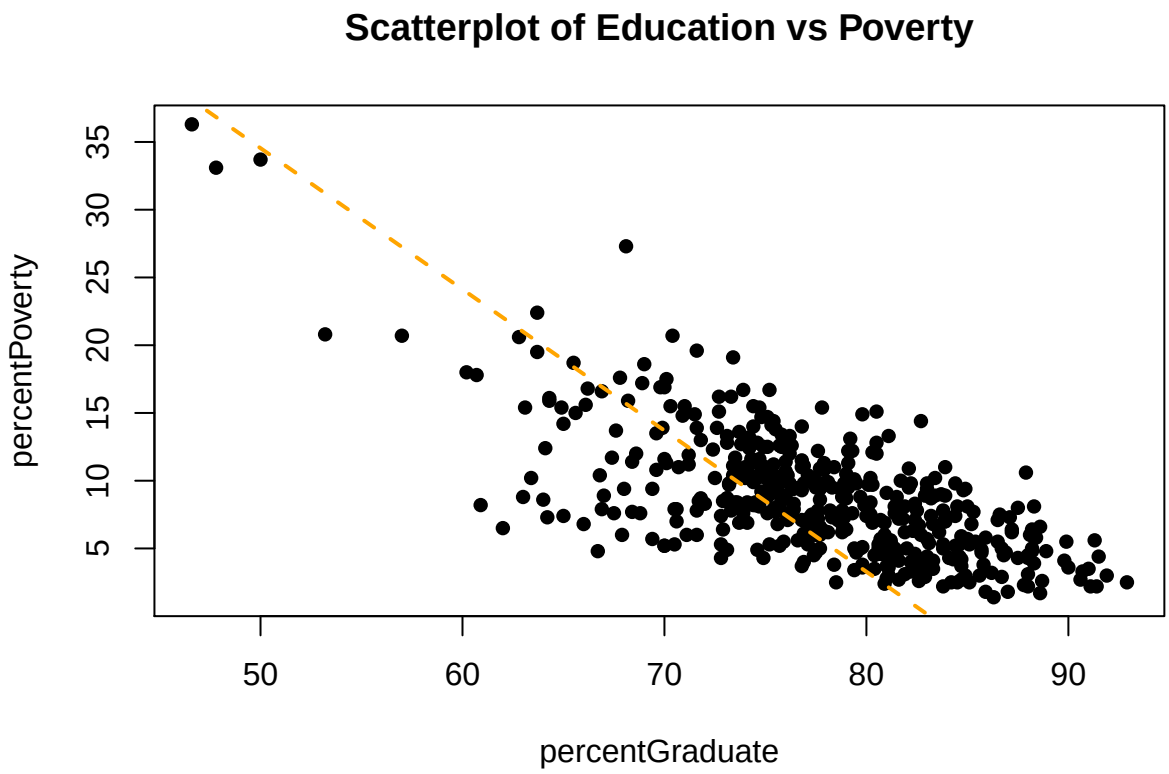
Scatter plots are great tools for quickly visualizing the relationship between two variables. Each pair (x_i, y_i) simply becomes a point on the plot. The `plot()` function in R handles this readily.

```
plot(percentGraduate, percentPoverty, pch=16, main='Scatterplot of Education vs Poverty')
```



Linear Regression is a big topic in Statistics (you can take an entire course on it here at UNM). It deals with understanding the relationship between two or more variables, and often involves using a "covariate" to predict a "response" variable of interest. Although it is not within the scope of this class, we comment that R makes it very easy to add a Linear Regression line to a Scatterplot.

```
plot(percentGraduate, percentPoverty, pch=16, main='Scatterplot of Education vs Poverty')
fit <- lm(percentGraduate ~ percentPoverty)
abline(fit, lwd=2, lty=2, col='orange')
```



It should be apparent from the plot that a linear relationship may not be adequate for describing the relationship between these variables.