STAT 311: Homework 6
Due: Aug 15, in class

Name:

There are two versions posted on the catalyst site. This version has all the homework questions, as well as the complete code required for the homework. The other version only contains the questions and is probably want you want to print out for turning in the homework

1 Started at the bottom

We will be using `R` to "scrape" lyrics off of the site `http://www.azlyrics.com/`. This procedure is a little complicated, so you don't necessarily need to follow along with this code exactly, but it could be helpful if you want to use this for other analyses. Note, you will need internet access to complete this assignment.

First we scrape the azlyrics Drake main page to get a list of all the links to Drake's song lyrics Understanding the code here isn't critical, but if you are interested in using `R` for scraping the web, using the `rvest` and `RCurl` packages along with the SelectorGadget tool can be very useful.

```r
### Run this code the first time to install the rvest and RCurl packages
install.packages("rvest")
install.packages("RCurl")

library(rvest, RCurl)

# The site that lists out all of drake's songs
url.drake <- "http://www.azlyrics.com/d/drake.html"

# Pull the site's source code into R
song.list.drake <- read_html(url.drake)

# Only find the links which go to Drake's songs
links.drake <- html_attr(html_nodes(html_nodes(song.list.drake, "#listAlbum"), "a"), "href")
links.drake <- links.drake[which(sapply(links.drake, substr, 0, 2) == "..")]

# Format links correctly
add.url <- function(link){
  gsub("..", "http://www.azlyrics.com", link, fixed = T)
}

# append the azylyrics beginning back onto the urls so you can access pages directly
full.links.drake <- unname(sapply(links.drake, add.url))
```

The object `full.links.drake` now contains all the urls to the song lyrics that azlyrics has for Drake. Now for any given song, we can parse the song lyrics using the following set of `site.to.lyrics` function. You can take any link to a azlyrics lyric site (contained in the `full.links.drake` object we just created), feed it into the `site.to.lyrics` function, and it will return a vector with each of the lyrics to the song. Run the code below to make sure `R` has defined the function, so we can use it later on.

```r
# Takes a string url that points to a azlyircs lyric site
# returns a vector of the individual words in the song
site.to.lyrics <- function(url){

  # read the source code for the lyrics website
  song <- read_html(url)

  # grab the raw lyrics from the source code
  lyrics.raw <- html_text(html_nodes(song, "div")[23])

  # remove the line breaks
  lyrics.intermediate <- gsub("\n", " ", lyrics.raw, fixed = T )
  lyrics.intermediate <- gsub("\r", " ", lyrics.intermediate, fixed = T )

  # Remove all parenthesis
  lyrics.intermediate <- gsub("(", "", lyrics.intermediate, fixed = T )
  lyrics.intermediate <- gsub(")", "", lyrics.intermediate, fixed = T )

  # remove punctuation marks
  lyrics.intermediate <- gsub("[[:punct:]]", "", lyrics.intermediate)

  # Remove verse/chorus tags
  lyrics.intermediate <- gsub("\\[[^\\]]*\\]", "", lyrics.intermediate, perl=TRUE)

  # split into single words based on spaces
  word.tokens <- unlist(strsplit(lyrics.intermediate, " "))

  # remove empty space tokens
  word.tokens <- word.tokens[word.tokens != ""]

  # make all lower case
  word.tokens <- tolower(word.tokens)

  return(word.tokens)
}
```

For instance, the 177 link in the `full.links.drake` object is for Drake's "Hotline Bling" song. So we can feed the url into the `site.to.links` function, and get the lyrics in vector form. The first few words of the song are "You used to call me on...".

```r
head(site.to.lyrics(full.links.drake[177]))
```

So now we have a setup which allows us to select a specific artist and get all their lyrics into `R` so we can analyze them.

2 Started at the bottom

Let's first do just a bit of descriptive analysis on Drake's lyrics. We can load all of the 248 songs that Drake has on azlyircs into `R`. Note that this may take a little bit of time, so instead of getting the lyrics for all 248 songs, we will only get the lyrics for 50 of his songs. Since we are taking a random sample of 50 songs, we will get a different result each time. For grading purposes, **make sure you run the set.seed command**

so you ensure you get the same 50 songs as the solution key.

```
# Make sure you run this line so you get the same songs
set.seed(10101)

# pick 50 out of the 248 song
subsample.drake <- sample(full.links.drake, 50)
```

Now we can pull the lyrics for those 50 songs into one single vector, `complete.list`. Note that this will take at least 2.5 minutes, so be patient.

```
# cycle over each link in our list
# pull the lyrics off the web
# parse the lyrics
# put them into one big vector

complete.list.drake <- c()
for(url in subsample.drake){

  complete.list.drake <- c(complete.list.drake , site.to.lyrics(url))

  # Needed so AZlyrics doesn't block us for accessing their site too many times.
  # To be safe, you can change it to a larger number (in seconds).
  Sys.sleep(3)
}

# You could equivalently do this in one line with the sapply function
# complete.list <- unlist(sapply(subsample.drake, site.to.lyrics)
```

Using the `table` command, we can count the number of times Drake uses each word, and see which words he uses most often.

```
word.counts.drake <- table(complete.list.drake)
head(sort(word.counts.drake, decreasing = T))
```

Using the `unique` command, we can also filter out all the duplicated words and just look at the unique number of words Drake uses. Note that this is not exactly correct, because there may be some punctuation and abbreviations which mess up our count slightly, but this gets us pretty close to what we actually want. When we have the list of unique words, we can use the `length` command to get the number of elements in the vector of unique words.

```
unique(complete.list.drake)
```

Finally, we can use the `%in%` command, to see whether a specific word is in a vector of words. For instance, in class, we have talked a lot about *joint distributions* which concern the probability of two (or more) events occuring. I'm not sure whether Drake has taken a statistics course, but let's see if Drake has ever used the word "joint" in his songs.

```
"joint" %in% complete.list.drake
```

Alternatively, we could look at each song by itself, and analyze specific properties of the song. For instance, we could count how many unique words are in the song, or we could see whether or not a specific word is used in the song.

```
# vector which will record summary statistic
summary.statistics.drake <- rep(0, length(subsample.drake))

# cycle over each link in our list
# pull the lyrics off the web
# parse the lyrics
# put them into one big vector


for(i in 1:length(subsample.drake)){

  # We need to pause a bit before accessing azlyrics too many times
  # or else you will get locked out and need to wait a few hours
  Sys.sleep(3)

  song.lyrics <- site.to.lyrics(subsample.drake[i])

  ### Analyze Song ####
  ## Insert code here
  ## save relevant statistic in summary.statistics[i]
  ##
  ## in order to see if a word "XYZ" is used in a song we could use-
  ## summary.statistics.drake[i] <- "XYZ" %in% song.lyrics
  ##
  ## in order to see the unique words in a song, we could use-
  ## summary.statistics.drake[i] <- length(unique(song.lyrics))
  ##

  }
```

2.1 General Questions

1. *What are the three words that Drake uses most often in the 50 songs we pulled?*

   ```
   head(sort(table(complete.list.drake), decreasing = T))
   ```

   ”i” 1146 times

   ”you” 882 times

   ”the” 791 times

2. *How many total unique words does Drake use in the 50 songs we pulled?*

   ```
   length(unique(complete.list.drake))
   ```

   3416

3. *Of the 50 songs we pulled, in how many songs does Drake use the word “money”?*

   ```
   # vector which will record summary statistic
   summary.statistics.drake <- rep(0, length(subsample.drake))
   ```

```
# cycle over each link in our list
# pull the lyrics off the web
# parse the lyrics
# put them into one big vector


for(i in 1:length(subsample.drake)){
  Sys.sleep(3)

  song.lyrics <- site.to.lyrics(subsample.drake[i])

  summary.statistics.drake[i] <- "money" %in% song.lyrics

  }

sum(summary.statistics.drake)
```

```
23
```

2.2 Confidence Intervals for single proportions

1. *Suppose we are interested in estimating the true proportion of songs for which Drake uses the word "money". Based on our sample of 50 songs, give an estimate of the true proportion.*

   $$\hat{p} = 23/50 = .46$$

2. *If we wanted to form a confidence interval for the true proportion, what standard error would we use. Show the standard error in notation and give a numerical value as well*

   $$se(\hat{p}) = \sqrt{\hat{p}(1-\hat{p})/n} = \sqrt{(.46 \times .54)/50}$$

3. *If we wanted to form a 85% confidence interval for the true proportion, we need to get the appropriate multiplier. What distribution would I use to get the multiplier? Make sure to specify any additional parameters needed for the distribution. If I want 85% of the area under the curve in between ±multiplier, how much area would be in the lower tail (which we would not want to include)?*

   For proportions, we use the standard normal distribution (mean = 0, sd = 1). For an 85% confidence interval, there will be 7.5% in each tail.

4. *Give a numerical value for the multiplier for a 85% confidence by either using* `R` *or looking it up in a table. Give the multiplier for 68% and 92% as well. Hint: For whatever distribution you use, in R, we want the* `qDIST` *function where you replace DIST with the actual distribution.*

```
qnorm(.075)

## [1] -1.439531

qnorm(.925)

## [1] 1.439531

qnorm(.16)

## [1] -0.9944579

qnorm(.84)
```

```
## [1] 0.9944579

qnorm(.04)

## [1] -1.750686

qnorm(.96)

## [1] 1.750686
```

5. *Form an 85% confidence interval for the true proportion of the songs in which Drake uses the word "money"*

```
p.hat <- 23/50

p.hat - qnorm(.925) * sqrt(p.hat * (1-p.hat)/50)

## [1] 0.358536

p.hat + qnorm(.925) * sqrt(p.hat * (1-p.hat)/50)

## [1] 0.561464
```

6. *Explain the confidence interval in plain English*

   If we repeat this procedure and form a confidence interval many times, 90% of the time, the confidence interval we produce will include the true proportion of Drake songs which contain the word money.

7. *Instead of selecting 50 songs randomly (as we actually did), we instead just select 50 of Drake's most popular songs. When forming a confidence interval, would the interpretation still be valid? Explain why or why not.*

   No, the confidence interval would not be valid, because the 50 most popular songs may not be representative of the overall population.

2.3 Confidence Intervals for means

1. *Suppose we are interested in estimating the average number of unique words used in each song. Based on our sample of 50 songs, give an estimate of the true mean.*

```
# vector which will record summary statistic
summary.statistics.drake <- rep(0, length(subsample.drake))

# cycle over each link in our list
# pull the lyrics off the web
# parse the lyrics
# put them into one big vector


for(i in 1:length(subsample.drake)){
  Sys.sleep(3)

  song.lyrics <- site.to.lyrics(subsample.drake[i])

  summary.statistics.drake[i] <- length(unique(song.lyrics))

  }

mean(summary.statistics.drake)
```

212.32

2. *If we wanted to form a confidence interval for the true mean, what standard error would we use. Show the standard error in notation and give a numerical value as well*

   ```
   sd(summary.statistics.drake)
   ```

   $$s_x/sqrt(n) = 85.86/\sqrt{}(50) = 12.14$$

3. *If we wanted to form a 98% confidence interval for the true mean, we need to get the appropriate multiplier. What distribution should I use to get the multiplier? Make sure to specify any additional parameters needed for the distribution. If I want 98% of the area under the curve in between ±multiplier, how much area would be in the lower tail (which we would not want to include)?*

   We would use a T distribution with 50 degrees of freedom. We could also use a normal distribution since the degrees of freedom is greater than 30. For a 98% multiplier, we would have 1% in each tail.

4. *Give a numerical value for the multiplier for a 98% confidence interval by either using* `R` *or looking it up in a table. Give the multiplier for 75% and 86% as well.*

   ```
   qnorm(.01)

   ## [1] -2.326348

   qnorm(.99)

   ## [1] 2.326348

   qnorm(.125)

   ## [1] -1.150349

   qnorm(.875)

   ## [1] 1.150349

   qnorm(.07)

   ## [1] -1.475791

   qnorm(.93)

   ## [1] 1.475791
   ```

5. *Form an 98% confidence interval for the true number of unique words which Drake uses in each of his songs*

   ```
   mean(summary.statistics.drake) - qnorm(.99) * sd(summary.statistics.drake) / sqrt(50)

   ## [1] 184.0704

   mean(summary.statistics.drake) + qnorm(.99) * sd(summary.statistics.drake) / sqrt(50)

   ## [1] 240.5696
   ```

6. *Explain the confidence interval in plain English*

   If we repeat this procedure and form a confidence interval many times, 90% of the time, the confidence interval we produce will include the true mean of unique words for each Drake song.

3 Now We Here

The music industry has a long history of fueds between rappers. In late 2015 and early 2016, "beef" started between the rappers Drake and Meek Mill when Meek tweeted about Drake using ghost writers for his raps. The fued consisted of a series of tweets and "diss tracks" in which the rappers insulted each other [1]. We don't know for certain whether Drake actually write his own lines, but we can at least analyze the tracks that they do put out. In particular, we will be analyzing the vocabulary that each rapper uses. Although this is an imperfect measure, we will assume that better rappers have a larger vocabulary (use more unique words in each song). Whether you are Team Drake or Team Meek, let's see what we can objectively decide about the two rappers. In particular, we will be looking at the parameter

$$\mu_{\text{Drake}} - \mu_{\text{Meek}}$$

where μ represents the true average number of unique words used per song. For this analysis, we will sample 35 of Meek Mill songs and use the 50 Drake's songs from before. We can pull all the links for Meek mill using the code (note this is the same procedure we used for Drake, but we need a different initial url)

```r
# url for meek mill's songs
url.meek <- "http://www.azlyrics.com/m/meekmill.html"

# Pull the site's source code into R
song.list.meek <- read_html(url.meek)

# Only find the links which go to Meek's songs
links.meek <- html_attr(html_nodes(html_nodes(song.list.meek, "#listAlbum"), "a"), "href")
links.meek <- links.meek[which(sapply(links.meek, substr, 0, 2) == "..")]

# append the azylyrics beginning back onto the urls so you can access pages directly
full.links.meek <- unname(sapply(links.meek, add.url))
```

Now let's select a subsample of 35 songs for Meek Mill. Again, **make sure you run the `set.seed` command as shown below so you get the exact same subsample as the solution key**.

```r
# Make sure you run this line so you get the same songs
set.seed(5555)

# pick 35 songs to analyze
subsample.meek <- sample(full.links.meek, 35)
```

Now, modifying the code we used for Drake (replace the references to Drake to references to Meek), get an estimate of the number of unique words Meek uses in each song based on our sample of 35 songs.

3.1 Confidence intervals for two sample means

1. *Are the two sample independent or dependent? Explain why.*

   Independent, because the number of unique words used in a Drake song shouldn't have any association with the number of unique words used in a Meek Mill song.

2. *If we wanted to form a confidence interval for the true difference in means, what standard error would we use. Show the standard error in notation and give a numerical value as well*

[1] See `http://www.mtv.com/news/2224700/drake-meek-mill-ghostwriter-beef-timeline/` for the complete story

```r
# vector which will record summary statistic
summary.statistics.meek <- rep(0, 35)

for(i in 1:length(subsample.meek)){
  summary.statistics.meek[i] <- site.to.lyrics(sub.sample.meek[i])
}
```

$$\sqrt{s^2_{drake}/50 + s^2_{meek}/35}$$

```r
sqrt(sd(summary.statistics.drake)^2 / 50  + sd(summary.statistics.meek)^2 / 35)
```

```
## [1] 16.52276
```

3. *If we wanted to form a 95% confidence interval for the true mean, we need to get the appropriate multiplier. What distribution should I use to get the multiplier? Make sure to specify any additional parameters needed for the distribution. If I want 95% of the area under the curve in between ±multiplier, how much area would be in the lower tail (which we would not want to include)?*

   We would use a T distribution with $\min(50-1, 35-1) = 34$ degrees of freedom. But we could also use a normal distribution since the degrees of freedom is greater than 30.

4. *Give a numerical value for the multiplier for a 95% confidence interval by either using* `R` *or looking it up in a table.*

   ```r
   qnorm(.975)
   ```

   ```
   ## [1] 1.959964
   ```

5. *Form an 95% confidence interval for the true difference in the number of unique words which Drake uses in each of his songs vs the number of unique words which Meek uses in each of his songs*

   ```r
   point.estimate <- mean(summary.statistics.drake) - mean(summary.statistics.meek)
   se <- sqrt(sd(summary.statistics.drake)^2 /50 + sd(summary.statistics.meek)^2/ 35 )
   mult <- qnorm(.975)

   point.estimate - mult * se
   ```

   ```
   ## [1] -47.54974
   ```

   ```r
   point.estimate + mult * se
   ```

   ```
   ## [1] 17.21831
   ```

6. *Explain the confidence interval in plain English*

   If we repeat this procedure and form a confidence interval many times, 90% of the time, the confidence interval we produce will include the true difference of the mean of unique words for each Drake song minus the mean of unique words for each Meek Mill song.

7. *What might we reasonably conclude about who is a better rapper? Can we be confident that one rapper is better than the other (based on our measure of unique words)?*

   It appears that Meek is the better rapper (since our point estimate is negative), but we do not have strong evidence that the true mean is different from 0.

3.2 Hypothesis test for two sample means

1. *Suppose we wanted to test the hypothesis that the two rappers are equally talented vs the hypothesis that Drake is more talented then Meek. State the null and alternative hypothesis*

   $H_0 : \mu_{drake} - \mu_{meek} = 0$ $H_A : \mu_{drake} - \mu_{meek} > 0$

2. *What test statistic could we use to test these hypotheses? Write it out in notation as well as give the numerical value from the samples of data we have gathered.*

   $$\frac{\bar{x}_{drake} - \bar{x}_{meek}}{\sqrt{\frac{s^2_{drake}}{n_{drake}} + \frac{s^2_{meek}}{n_{meek}}}}$$

   ```
   numerator <- mean(summary.statistics.drake) - mean(summary.statistics.meek)
   denomenator <- sqrt(sd(summary.statistics.drake)^2/50 + sd(summary.statistics.meek)^2 / 35)
   numerator / denomenator

   ## [1] -0.9178678
   ```

3. *What is the theoretical distribution of the test statistic under the null hypothesis. Specify all parameters needed for the distribution.*

   The test statistic should follow a T distribution with 34 degrees of freedom, but we can also use a normal distribution since the degrees of freedom is larger than 30.

4. *If the null hypothesis is true, what is the probability we would've observed a test statistic as or more extreme than we actually did. That is, calculate the p-value. In R, you should use the* `pDIST` *command, where DIST is replaced by the actual distribution.*

   Note that since we define "extreme" as large positive values, we actually want the area to the right of the observed test statstic. Thus, we use 1- pnorm.

   ```
   1 - pnorm(-.92)

   ## [1] 0.8212136
   ```

5. *Explain in words what the p-value actually means.*

   The p-value is the probability under the null hypothesis of of observing a difference as large or larger as the one whe actually observed.

6. *Using a cut-off of .05, what would you conclude about the null hypothesis? Does this agree with the confidence interval approach?*

   We would fail to reject the null hypothesis. Thus, we have no evidence that one rapper is more talented than the other. This agrees with the Confidence Interval we found earlier.

4 More Hypothesis testing

Suppose we are developing a retrovirus for Zika. We know the best drugs so far prevent Zika in .6 of the people exposed, and we are intrested in whether the new retrovirus we are developing is better than the existing drugs. In particular, we are interested in testing-

$$H_0 : p_{prevent} = .6$$
$$H_A : p_{prevent} > .6$$

We gather a sample size of 200 individuals.

1. *If the null hypothesis is true, what is the distribution of* $\hat{p}$*?*

$$\hat{p} \sim \mathcal{N}(.6, .6(1 - .6)/200)$$

2. *Given a cut-off of .05, what is the largest value of* $\hat{p}$ *for which we will not reject* H_0*?*

   A cut-off of .05 has a corresponding z-score of 1.645. Thus, we will reject the null hypothesis whenever the z-score is greater than 1.645. We can convert this into a p_{cutoff}.

$$1.645 = \frac{p_{cutoff} - p_0}{\sqrt{p_0(1 - p_0)/n}} = \frac{p_{cutoff} - .6}{\sqrt{.6(.4)/200}}$$

   Solving for p_{cutoff} yields $p_{cutoff} = .657$

3. *Given a cut-off of .05, what percentage of the time will we commit a Type I error?*

   By definition, the probability of a Type I error is the level of the test, or .05 in this case

4. *If the the null hypothesis is not actually true, but in reality* $p = .65$*. What is the power of the hypothesis test specified above? What is the probability of committing a Type II error?*

   If the true proportion is actually .65, then $\hat{p} \sim \mathcal{N}(.65, .65(.35)/200)$. However, since we don't know what the true proportion is when we are running our hypothesis test, we will still use the cut-off of .657 based on the null hypothesis. Thus, we will reject the null hypothesis whenever $\hat{p}$ is greater than .657. Thus, to find the power, we are interested in calculating the probability that $\hat{p} > .657$ when $\hat{p} \sim \mathcal{N}(.65, .65(.35)/200)$

   In this case, we get the z-score for .657 of

$$\frac{.657 - .65}{\sqrt{.65(.35)/200}} = .201$$

   To find the probability that $\hat{p} > .657$, we need the area to the right of .201, which is

```
1 - pnorm(.201)

## [1] 0.4203493
```

5. *If the the null hypothesis is not actually true, but in reality* $p = .7$*. What is the power of the hypothesis test specified above? What is the probability of committing a Type II error?*

   If the true proportion is actually .7, then $\hat{p} \sim \mathcal{N}(.7, .7(.3)/200)$. However, since we don't know what the true proportion is when we are running our hypothesis test, we will still use the cut-off of .657 based on the null hypothesis. Thus, we will reject the null hypothesis whenever $\hat{p}$ is greater than .657. Thus, to find the power, we are interested in calculating the probability that $\hat{p} > .657$ when $\hat{p} \sim \mathcal{N}(.7, .7(.3)/200)$

In this case, we get the z-score for .657 of

$$\frac{.657 - .7}{\sqrt{.7(.3)/200}} = -1.323$$

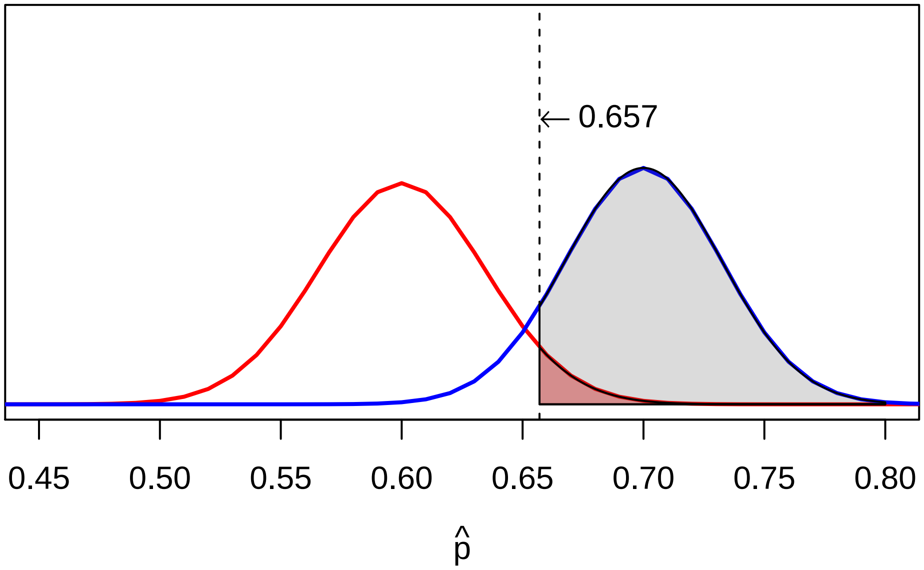
To find the probability that $\hat{p} > .657$, we need the area to the right of -1.323, which is

```r
1 - pnorm(-1.323)
```

```
## [1] 0.9070823
```

6. *Draw the null distribution. Draw a vertical line at the value of $\hat{p}$ which corresponds to the cutoff of .05. Draw the actual distribution of $\hat{p}$ if $p = .7$. Shade the region of the actual distirbution which corresponds to the power ot the hypothesis test.*

   In the plot below, we have the null distribution (what we posit the distribution of $\hat{p}$ to be under the null hypothesis) shown in red and the true distribution (which we do not know when we do the hypothesis test) shown in blue. The area under the red curve to the right of .657 (shaded in red) is .05 (the level of the test). We will reject the null if we observe any value of $\hat{p}$ greater than .657. However, if the true population parameter $p = .7$, we know that $\hat{p} \sim \mathcal{N}(.7, .7(.3)/200)$ (curve shown in blue). Thus, to find the probability of rejecting the null hypothesis when the true $p = .7$, we need to find the area under the blue curve to the right of .657 (shaded in gray and includes the portion shaded in red).


← 0.657
$\hat{p}$