Lab 6: Confidence Intervals

August 5, 2016

1 Goals

Today in lab, we hope to

- Look at properties of the T distribution
- Analyze

2 T Distribution

In class, we mentioned that the T distribution arises as a ratio of random variables. Specifically, we use the T distribution when we don't know the population standard deviation and divide by the sample standard deviation in stead. Let's take a deeper look to see.

Recall that QQ plots help us compare two distributions against each other (most often we use the standard normal as a reference distirbution). Specifically, if the QQ-plot is roughly linear, we would say that the two distirbutions are similar. Below for various values of N, we look at the distributions of

$$Z = \sqrt{n}\frac{\bar{x}_n - \mu}{\sigma}$$

and

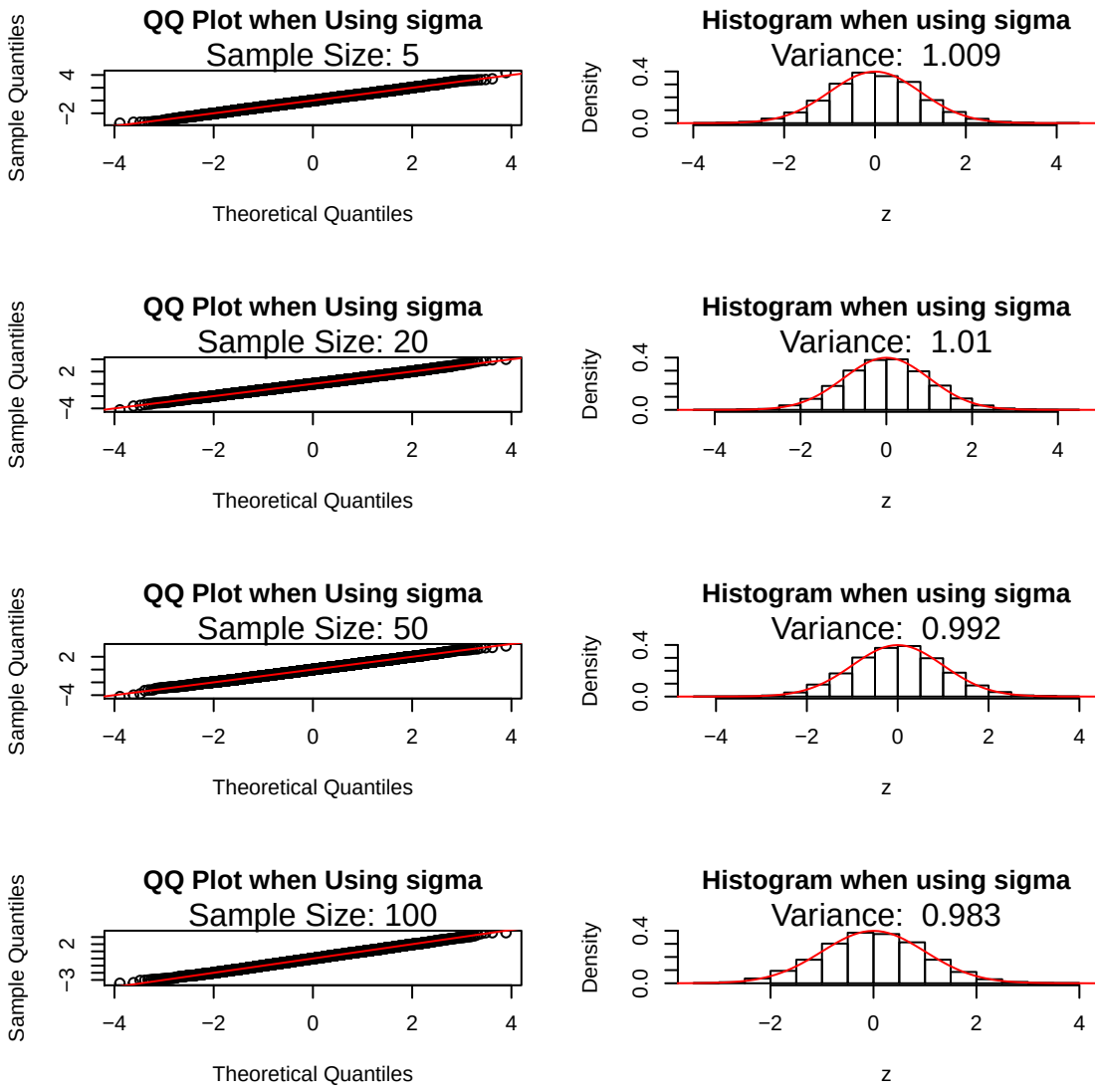
$$T = \sqrt{n}\frac{\bar{x}_n - \mu}{s_x}$$

and comparing them to a standard normal distribution using QQ-Plots.

First, let's consider the case where we know σ. For each setting of n: We take 1000 samples and record $Z = \sqrt{n}\frac{\bar{x}-\mu}{\sigma}$ for each sample. We then take each of the $\sqrt{n}\frac{\bar{x}-\mu}{\sigma}$ values, and form a QQ-plot. We also plot the histogram of the Z values as well as the variance of the Z values. If the Z values actually are $\mathcal{N}(0, 1)$, we would expect the calculated variance of the Z's to be very close to 1. We also plot the density of a standard normal in red on top of the histogram.

```
# Number of samples
sim.size <- 10000
z <- rep(0, sim.size)

# Settings for N
n.list <- c(5, 20, 50, 100)

# Plots should have 4 rows and 2 columns
par(mfrow = c(4, 2))

# Go through each value of n
```

```r
for(n in n.list){

  # Go through each sample
  for(i in 1:sim.size){
    X <- rnorm(n = n, mean = 10, sd = 4)

    # Change this line later
    z[i] <- sqrt(n) * (mean(X) - 10) / 4
  }

  # Plot QQ plot
  qqnorm(z, main = "QQ Plot when Using sigma")
  mtext(paste("Sample Size: ", n, sep=""))
  abline(a = 0, b = 1, col = "red")

  # Plot histogram
  hist(z, main = "Histogram when using sigma", freq = F)
  mtext(paste("Variance: ", round(var(z), 3)))
  lines(seq(-5, 5, by = .1), dnorm(seq(-5, 5, by = .1)),col = "red")
}
```



2.1 Questions

- What do we see? Does the distribution look roughly normal according to the QQ-plots?
- What do we see? Does the distribution look roughly normal according to the histograms?
- What do we see? Are the calculated variances about what we would expect?
- Does this change as the sample size grows larger?

Now repeat the analysis above, but this time, instead of dividing by the known standard deviation, divide by an estimate of the standard deviation s_x. You should only have to change one line of the code above to calculate

$$T = \sqrt{n}\frac{\bar{x}_n - \mu}{s_x}$$

instead of the Z value we used before.

2.2 Questions

- What do we see? Does the distribution look roughly normal according to the QQ-plots?
- What do we see? Does the distribution look roughly normal according to the histograms?
- What do we see? Are the calculated variances about what we would expect?
- Does this change as the sample size grows larger?

3 T Distributions and Confidence intervals

Now let's see how estimating σ affects confidence intervals.

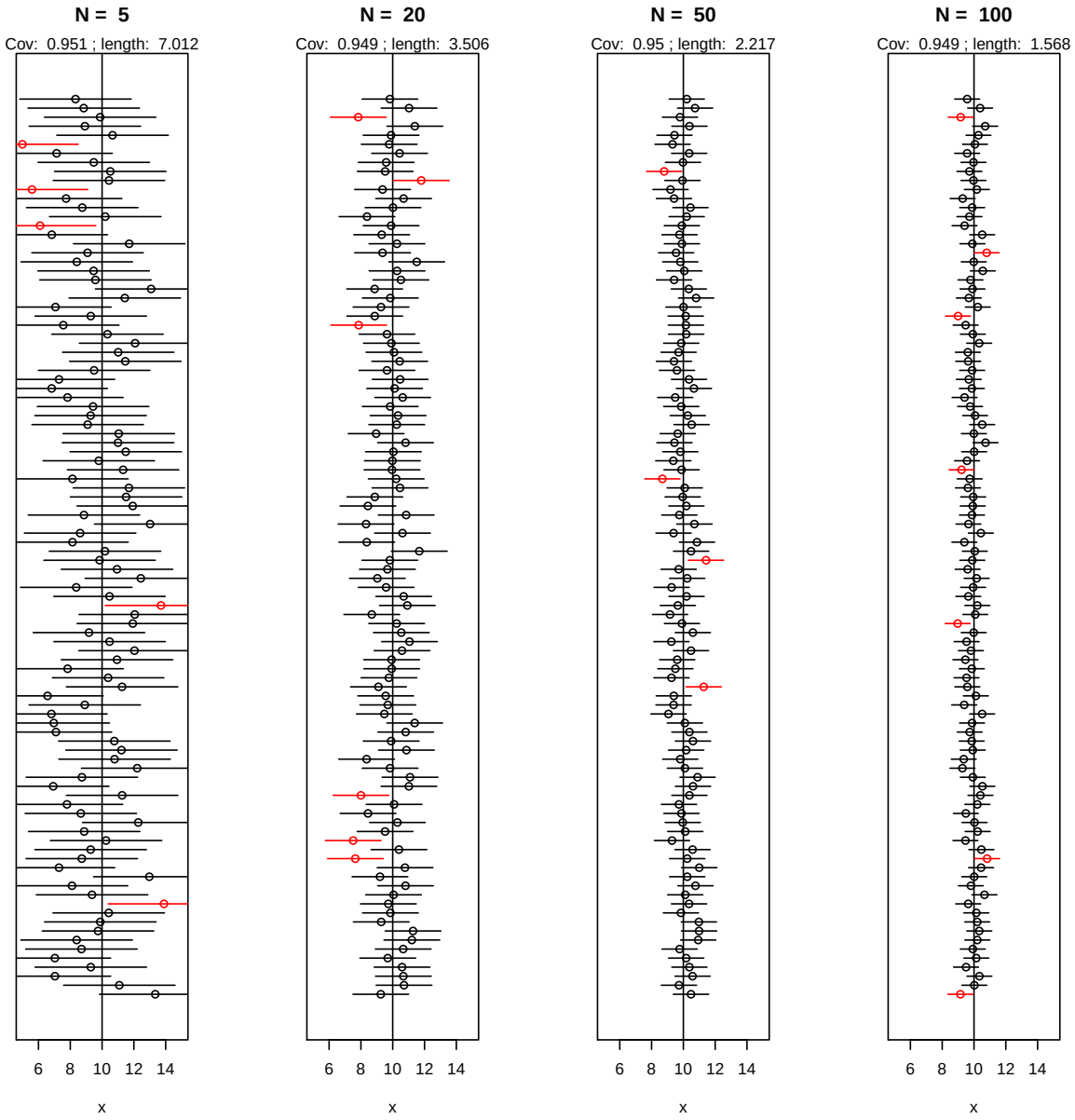
3.1 Questions

- Explain to your neighbor what is meant by a 95% confidence interval. What does that imply about each individual confidence interval we create? What does that imply about when we repeat this procedure many many times?

Ideally, if our assumptions are correct, 95% of the confidence intervals we form will contain the true mean. First, we will simulate data in which each individual comes from a normal distribution. For now, let's assume we know what the true population standard deviation σ is and form a confidence intervals with σ rather than s_x.

```
# Number of samples
sim.size <- 10000
x.bar <- rep(0, sim.size)
stnd.err <- rep(0, sim.size)
# Settings for N
n.list <- c(5, 20, 50, 100)



# Plots should have 4 rows and 2 columns
par(mfrow = c(1, 4))

# Go through each value of n
for(n in n.list){

  # Multiplier
  multiplier <- -qnorm(.025)

  # Go through each sample
  for(i in 1:sim.size){
    X <- rnorm(n = n, mean = 10, sd = 4)

    # Point Estimate
    x.bar[i] <- mean(X)

    # Standard Error
    stnd.err[i] <- 4 / sqrt(n)
  }
```

```r
  # Create the Confidence Interval
  lower.bound <- x.bar - multiplier * stnd.err
  upper.bound <- x.bar + multiplier * stnd.err

  # Check whether or not it contains the true mean
  contains.mu <- lower.bound < 10 & upper.bound > 10

  coverage <- mean(contains.mu)
  avg.length <- mean(stnd.err * multiplier * 2)


  # Plot 100 randomly selected samples from the set of 1000
  samples.to.be.plotted <- sample(10000, 100)
  plot.col <- ifelse(contains.mu[samples.to.be.plotted], "black", "red")
  plot(-1, -1, xlim = c(5, 15), ylim = c(0, 101),
       xlab = "x", main = paste("N =  ", n, sep = ""), ylab = "", yaxt = "n")
  mtext(paste("Cov: ", round(coverage, 3), "; length: ", round(avg.length,3)), cex = .7)
  points(x.bar[samples.to.be.plotted], c(1:100), col = plot.col)
  abline(v = 10)
  segments( lower.bound[samples.to.be.plotted], c(1:100),
            upper.bound[samples.to.be.plotted], c(1:100), col = plot.col)
}
```



3.2 Questions

- How does the simulated coverage rates compare to what we would expect?
- How does the average length of the confidence interval change as n increases?

Now try the procedure again, using the normal multiplier, but for the standard error, use the sample standard deviation (s_x) instead of the truth.

3.3 Questions

- How does the simulated coverage rates compare to what we would expect?

- How do the simulated coverage rates compare to what we saw before?
- How does the average length of the confidence interval change as n increases?
- How do the average interval lengths compare to what we saw before?

Finally try the procedure again, using the T distribution multiplier, and for the standard error, use the sample standard deviation (s_x) instead of the truth.

3.4 Questions

- How does the simulated coverage rates compare to what we would expect?
- How do the simulated coverage rates compare to what we saw before?
- How does the average length of the confidence interval change as n increases?
- How do the average interval lengths compare to what we saw before?

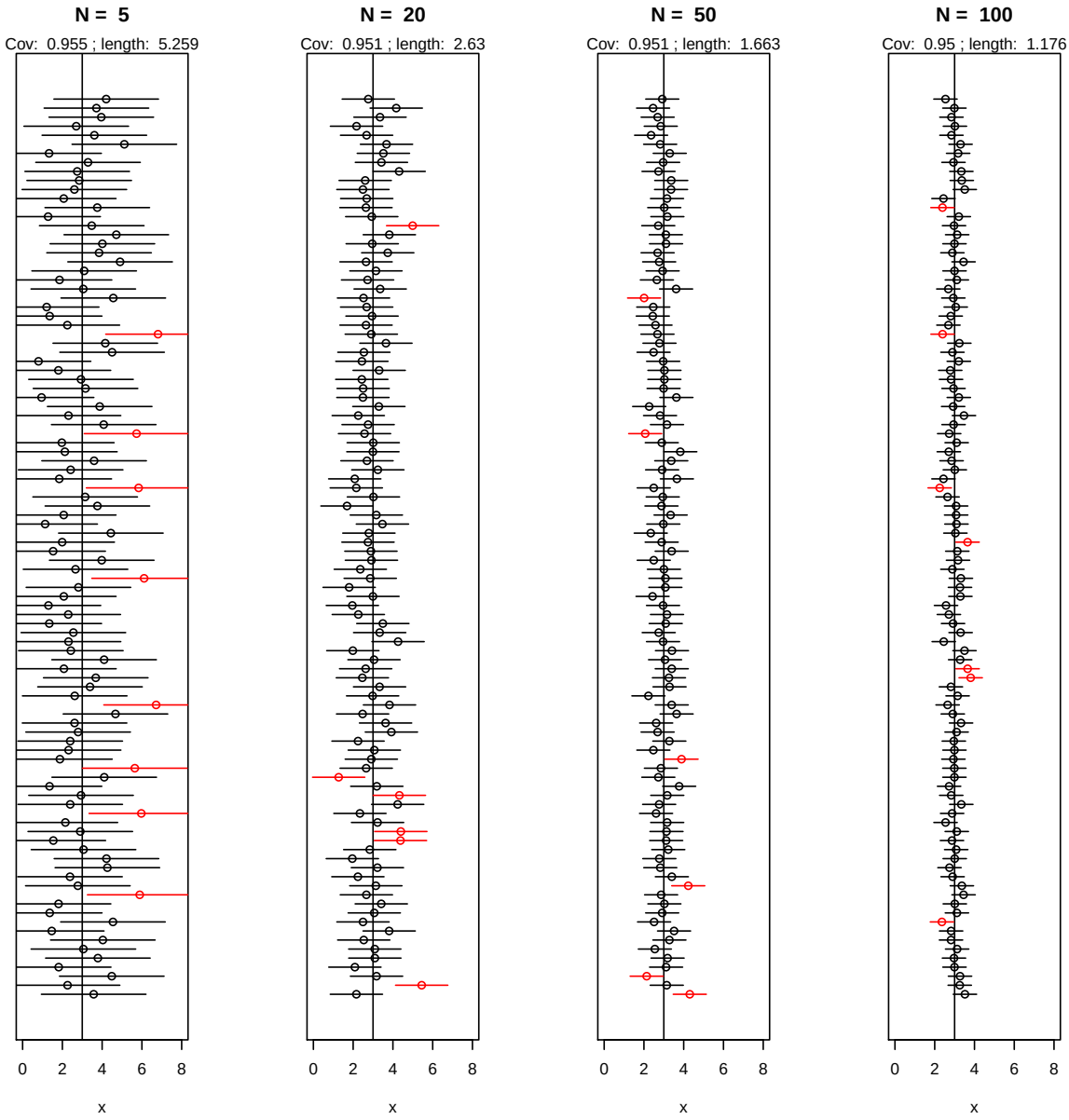
3.5 Questions

- How does the simulated coverage rates compare to what we would expect?
- How do the simulated coverage rates compare to what we saw before?
- How does the average length of the confidence interval change as n increases?
- How do the average interval lengths compare to what we saw before?

Now let's look at the effect when the underlying population is not normally distributed, but exponential instead. When we first gave the motivation for confidence intervals, we had to appeal to the Central Limit Theorem. In this case, if the underlying population is not normally distributed, we actually need n to grow for the normal distribution approximation to be good. Let's take a look at how the confidence intervals properties change.

```
# Number of samples
sim.size <- 10000
x.bar <- rep(0, sim.size)
stnd.err <- rep(0, sim.size)
# Settings for N
n.list <- c(5, 20, 50, 100)



# Plots should have 4 rows and 2 columns
par(mfrow = c(1, 4))

# Go through each value of n
for(n in n.list){

  # Multiplier
  multiplier <- -qnorm(.025)

  # Go through each sample
  for(i in 1:sim.size){
    X <- rexp(n = n, rate = 1/3)

    # Point Estimate
```

```
    x.bar[i] <- mean(X)

    # Standard Error
    stnd.err[i] <- 3 / sqrt(n)
  }


  # Create the Confidence Interval
  lower.bound <- x.bar - multiplier * stnd.err
  upper.bound <- x.bar + multiplier * stnd.err

  # Check whether or not it contains the true mean
  contains.mu <- lower.bound < 3 & upper.bound > 3

  coverage <- mean(contains.mu)
  avg.length <- mean(stnd.err * multiplier * 2)


  # Plot 100 randomly selected samples from the set of 1000
  samples.to.be.plotted <- sample(10000, 100)
  plot.col <- ifelse(contains.mu[samples.to.be.plotted], "black", "red")
  plot(-1, -1, xlim = c(0, 8), ylim = c(0, 101),
       xlab = "x", main = paste("N =  ", n, sep = ""), ylab = "", yaxt = "n")
  mtext(paste("Cov: ", round(coverage, 3), "; length: ", round(avg.length,3)), cex = .7)
  points(x.bar[samples.to.be.plotted], c(1:100), col = plot.col)
  abline(v = 3)
  segments( lower.bound[samples.to.be.plotted], c(1:100),
            upper.bound[samples.to.be.plotted], c(1:100), col = plot.col)
}
```



Now try the procedure again, using the normal multiplier, but for the standard error, use the sample standard deviation (s_x) instead of the truth.

3.6 Questions

- How does the simulated coverage rates compare to what we would expect?
- How do the simulated coverage rates compare to what we saw before?
- How does the average length of the confidence interval change as n increases?
- How do the average interval lengths compare to what we saw before?
- How does this compare to when the underlying population was normally distributed?

Finally try the procedure again, using the T distribution multiplier, and for the standard error, use the sample standard deviation (s_x) instead of the truth.

3.7 Questions

- How does the simulated coverage rates compare to what we would expect?
- How do the simulated coverage rates compare to what we saw before?
- How does the average length of the confidence interval change as n increases?
- How do the average interval lengths compare to what we saw before?
- How does this compare to when the underlying population was normally distributed?

4 IMDB Movie Ratings

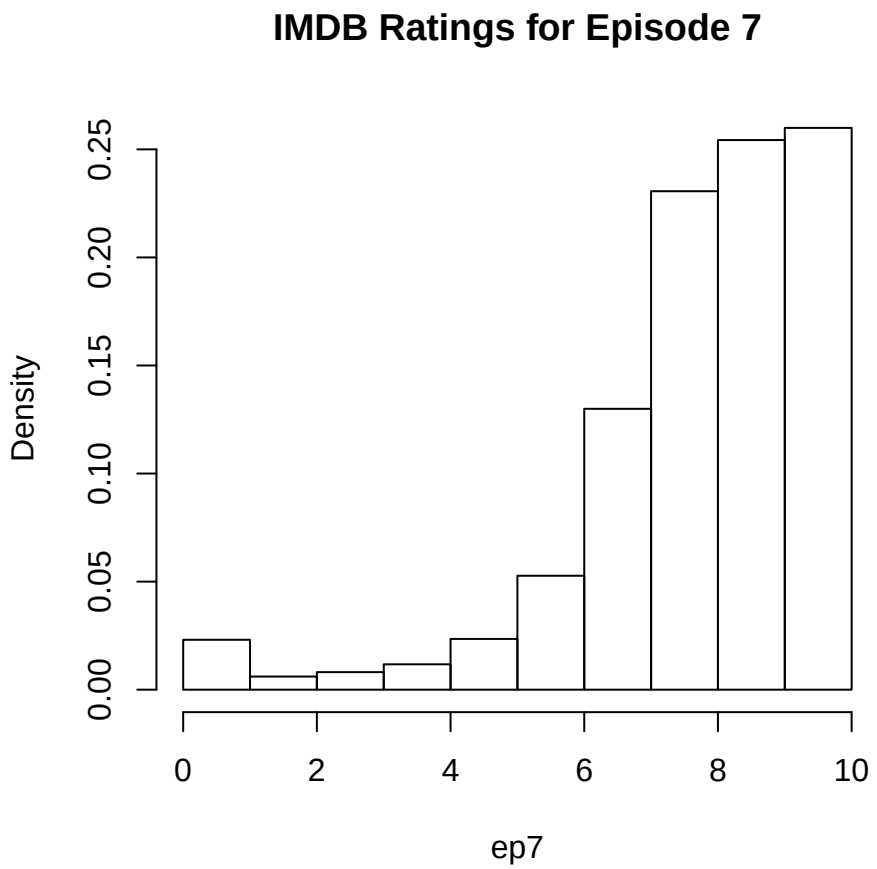
The Internet Movie Database (IMDB) which provides data about a wide variety of movies. In addition to providing plot synopsis and information on the actors/actresses in the movie, the website also allows individuals provide a 1-10 rating on each movie. In particular for the lab today, we will be looking at the ratings for the Star Wars movies. Although this is a voluntary sample, for the purpose of this lab, let's suppose it's representative of the overall population.

5 Questions

- What would you specify as the population of interest? (There's no exat right answer)
- What is the parameter of interest?

We will first look at the ratings for the newest Star Wars movie, Episode VII: The Force Awakens.

```
ep7 <- unlist(read.csv("http://www.stat.washington.edu/~ysamwang/notes/ep7.csv", header = F))

hist(ep7, freq = F, main = "IMDB Ratings for Episode 7", breaks = c(0:10))
```



The sample size on IMDB is really big, so to make things a bit more interesting, let's suppose we only have 500 ratings. We'll take a subsample of 500 of the actual ratings.

```
set.seed(111)
ep7.subsample <- sample(ep7, 500)
```

Let's form a 90% confidence interval for the true population rating of Episode 7.

```
x.bar <- mean(ep7.subsample)
n <- length(ep7.subsample)
se <- sd(ep7.subsample) / sqrt(n)
multiplier <- -qt(.05, df = n-1)
```

Notice that the multiplier here using a t distibution is -1.647913 while using a normal distribution would give us a multiplier of -1.6448536. Thus, there is barely any difference between the normal and t distribution with 499 many degrees of freedom.

So to form the confidence interval, we have

```
x.bar - multiplier * se
```

```
## [1] 8.048887
```

```
x.bar + multiplier * se
```

```
## [1] 8.303113
```

5.1 Question

- Make a "confidence statement" regarding the confidence interval we created
- Give an interpretation of the confidence statement in plain English