Problem Set #7

Econ 103

Lecture Progress

We made it to the end of the Chapter 6 slides.

Homework Checklist

- ☐ **Book Problems (Chapter 6):** 1, 3, 5, 7, 33
- ☐ **Book Problems (Chapter 7):** 1, 3, 7, 9, 17
- ☐ **Additional Problems:** See below
- ☐ **R Tutorial:** Do the RMarkdown tutorial at `https://www.datacamp.com/courses/reporting-with-r-markdown`. This tutorial uses the package "dplyr" to clean the data, which we have not covered, but you should see how you can do similar cleaning using "data.table". If you have questions about what the tutorial is doing, please post on Piazza!
- ☐ **Ask questions on Piazza**
- ☐ **Review slides**
- ☐ **Work on R Project:** Use RMarkdown to report your summary statistics!
- ☐ **RMarkdown Cheat Sheet:** `http://www.rstudio.com/wp-content/uploads/2015/03/rmarkdown-reference.pdf`

Additional Problems

1. James is running a polling operation. He interviews 1,000 people and asks who they are voting for in the upcoming election (for simplicity, assume only two candidates). Assume each interview is an iid draw from the voting population and the share of the population voting for Hillary Clinton is μ, with a variance of σ^2.

(a) Is the sample share of Hillary Clinton voters (percent of voters polled that said they would vote for Hillary Clinton) an unbiased estimator of the true μ?

Solution: Yes. We can think of each interview as being either a 1 if the voter is a Hillary Clinton supporter or a 0 if they are a Donald Trump supporter. Then, the sample share is $\bar{X} = \frac{1}{1000}\sum_{i=1}^{1000} X_i$. To see whether this is biased or not, we look at the expectation $E[\bar{X}] = E[\frac{1}{1000}\sum_{i=1}^{1000} X_i] = \frac{1}{1000}\sum_{i=1}^{1000} E[X_i] = \mu$. Hence, this is an unbiased estimator of the true share of the population that will vote for Hillary Clinton, μ

(b) What is the variance of this estimate?

Solution: The variance can be computed as

$$Var\left(\frac{1}{1000}\sum_{i=1}^{1000} X_i\right) = \frac{1}{1000^2}\sum_{i=1}^{1000} Var(X_i) = \frac{\sigma^2}{1000}$$

(c) James decides to only report the results of the first person he interviews and ignore the other 999 people he interviews (so his estimator is $X = X_1$). Is this an unbiased estimator of the true population share μ?

Solution: Yes! Taking expectations, we get $E[X] = E[X_1] = \mu$. While this seems like a silly way to do polling, it is still an unbiased estimator!

(d) What is the variance of this new estimator?

Solution: It is simply $Var(X) = Var(X_1) = \sigma^2$

(e) Is the first estimator consistent? What about the second?

Solution: Recall that we are interested in MSE consistency, which means that

$$\lim_{n\to\infty} MSE(\hat{\theta_n}) = 0$$

We know that $MSE = variance + bias^2$ and we established that both estimators are unbiased, so we only have to examine how their variance behaves as $n \to \infty$. The variance of the first estimator is $\frac{\sigma^2}{n}$ while for the second estimator it is σ. Hence, only the first estimator estimator is consistent since its MSE error goes to 0 in the limit while the second estimator does not.

(f) Which estimator do you prefer? Why?

Solution: I would prefer the first estimator because even though both estimators are unbiased, the first has a smaller variance and is hence, more *efficient* than the second one. This means we can be more confident that the resulting estimate is somewhere close to the actual truth while the estimator with the higher variance could be farther off (and most definitely will be).

2. In this question you will replicate some of the density plots of normal, t-, F-, and chi-squared distributions. Feel free to experiment with the parameters and see how the different distributions behave!

(a) Plot a standard normal pdf on the same graph as a $t(1)$ pdf on the interval $[-5, 5]$. How do the pdfs compare? Explain.

Solution:

```
library(plotly)
x <- seq(from = -5, to = 5, by = 0.01)
normal <- dnorm(x)
tdist1 <- dt(x, df = 1)
plot_ly(x = x, y = normal, type = "scatter", mode = "lines")
%>% add_trace(x = x, y = tdist1, type = "scatter")
```

The $t(1)$ has much fatter tails than the normal: it's much more spread out. Although both are centered at zero, the t is much more likely to take on very large postive or negative values.

(b) Plot a standard normal pdf on the same graph as a $t(100)$ pdf on the interval $[-5, 5]$. How do the pdfs compare? Explain.

Solution: Carrying on from the code in the previous part:

```
tdist100 <- dt(x, df = 100)
plot_ly(x = x, y = normal, type = "scatter", mode = "lines")
%>% add_trace(x = x, y = tdist100, type = "scatter")
```

The two plots overlap almost perfectly: it looks like we've only plotted one curve! This is because the t distribution gets closer and closer to the standard normal as its degrees of freedom increase.

(c) Plot a χ^2 pdf with degrees of freedom equal to 4 on the interval $[0, 20]$.

Solution:

```
x <- seq(from = 0, to = 20, by = 0.01)
y <- dchisq(x, df = 4)
plot_ly(x = x, y = y, type = "scatter", mode = "lines")
```

(d) Plot an F pdf with numerator degrees of freedom equal to 4 and denominator degrees of freedom equal to 40 on the interval $[0, 5]$.

Solution:

```
x <- seq(from = 0, to = 5, by = 0.01)
y <- df(x, df1 = 4, df2 = 40)
plot_ly(x = x, y = y, type = "scatter", mode = "lines")
```

3. In this question you will verify the empirical rule both directly using `pnorm` and by simulation using `rnorm`.

(a) Draw 100000 iid observations from a standard normal distribution and store your results in a vector called `sims`.

Solution:

```
sims <- rnorm(100000)
```

(b) What proportion of the observations in `sims` lie in the range $[-1, 1]$?

Solution:

```
sum((sims >= -1) & (sims <= 1))/length(sims)
```

(c) What proportion of the observations in `sims` lie in the range $[-2, 2]$?

Solution:

```
sum((sims >= -2) & (sims <= 2))/length(sims)
```

(d) What proportion of the observations in `sims` lie in the range $[-3, 3]$?

Solution:

```
sum((sims >= -3) & (sims <= 3))/length(sims)
```

(e) Use `pnorm` to calculate the exact probabilities for a standard normal pdf that correspond to the above simulation experiments. How accurate were your simulations?

Solution:

```
pnorm(1) - pnorm(-1)
pnorm(2) - pnorm(-2)
pnorm(3) - pnorm(-3)
```

4. In this question you will replicate the Law of Large Numbers (LLN) visualization from lecture, in which we plotted "running" sample means as we kept adding more and more simulations from a $N(\mu = 0, \sigma^2 = 100)$ distribution. Your plot won't look exactly like the one from class since this is a random experiment, but it will show the same qualitative behavior.

   (a) The R command for a "running" or "cumulative" sum is `cumsum`. Look at the help file for this command and test it out on a vector of ten ones and another containing the integers from one to ten to make sure you understand what it does.

   Solution:

   ```
   ones <- rep(1, 10)
   ones
   cumsum(ones)
   cumsum(1:10)
   ```

   (b) Replicate the plot from the end of the lecture. First you'll need to draw 10,000 iid samples from a $N(\mu = 0, \sigma^2 = 100)$ distribution. Then you'll need to calculate the running means. You'll need to figure out how `cumsum` can be used to accomplish this. Finally, plot your results along with a dashed red line at the value to which the sample mean is converging. Make sure to label your axes.

   Solution:

   ```
   n <- 10000
   sims <- rnorm(n, mean = 0, sd = 10)
   running.mean <- cumsum(sims)/(1:n)
   plot_ly(x = 1:n, y = running.mean, type = "scatter", mode = "lines")
   ```

   (c) Repeat the previous part but, rather than drawing $N(\mu = 0, \sigma^2 = 100)$ simulations, draw from a Student-t distribution with one degree of freedom. How do your results differ? Use what you know about the Student-t distribution to guess why our proof that the sample mean is consistent for the population mean doesn't work here.

Solution:

```
n <- 10000
sims <- rt(n, df = 1)
running.mean <- cumsum(sims)/(1:n)
plot_ly(x = 1:n, y = running.mean, type = "scatter", mode = "lines")
```

This plot looks totally different from the previous one: the "running means" never settle down in this case. To prove that the sample mean is consistent for the population mean, we tacitly assumed that both the mean and variance of X_i exist and are finite. (Remember that both quantities are defined as improper integrals, so they could diverge or may be undefined.) It turns out that the Student-t distribution with one degree of freedom has an *infinite variance and a mean that does not exist*. Essentially its mean is $\infty - \infty$ which does *not* equal zero: it's simply undefined. To get the LLN to work for a Student-t, we need a finite mean and variance. Both conditions turn out to hold as long as the degrees of freedom are ≥ 3. For example:

```
n <- 10000
sims <- rt(n, df = 3)
running.mean <- cumsum(sims)/(1:n)
plot_ly(x = 1:n, y = running.mean, type = "scatter", mode = "lines")
```