Problem Set # 10

Econ 103

Lecture Progress

We made it to slide 15 of the Chapter 8 slides.

Homework Checklist

- ☐ **Book Problems (Chapter 6):** 15, 17, 19ab, 21
- ☐ **Book Problems (Chapter 8):** 11, 17(c), 17(d), 19, 21
- ☐ **Assigned Reading:** *Extracting the Signal from the Noise: 7 Tips for Interpreting Macroeconomic Data*. It will be helpful for your R project and is good for econ in general.
- ☐ **Additional Problems:** See below
- ☐ **Ask questions on Piazza**
- ☐ **Review slides**
- ☐ **R Tutorial:** Use RMarkdown to write a description of your data (will be a section of your R Project) as well as load your data and summarize some statistics of interest. Will take up next week

I'll provide full solutions to 6-17 and 8-21.

Problems from the Textbook

Solution: (**6-17**) There are a number of different ways to solve this question. The "exact" solution, which is not the one in the book, directly uses the fact that this is a Binomial sampling model: given that 20% of the cars in the population are defective, what is the probability that no more than 5 of the cars in a sample of size 50 are defective? We can calculate the answer in R as follows:

```
pbinom(5, size = 50, prob = 0.2)
## [1] 0.04803
```

The "point" of this question, however, is to get an *approximate* answer using what we know about the Central Limit Theorem. Since this is a result about the sampling distribution of sample means, we need to re-express the desired probability in these terms. The mechanic wants to know the probability that no more than 5 out of 50 cars are defective. This is the same thing as saying that the *sample mean*, which is just the sample proportion, is no greater than $5/50 = 0.1$. Now, by the CLT, if we center the sample mean at the population mean and scale it by its standard error, the result is approximately standard normal:

$$P(\bar{X}_n \leq 0.1) = P\left(\frac{\bar{X}_n - \mu}{SE(\bar{X}_n)} \leq \frac{0.1 - \mu}{SE(\bar{X}_n)}\right) \approx \texttt{pnorm}((0.1 - \mu)/SE(\bar{X}_n))$$

In this example the standard error is: $\sqrt{p(1-p)/n}$ where n is the sample size, 50, and p is the population proportion: 0.2. The population mean for this problem is simply the population proportion: 0.2. Thus, we have

```
p <- 0.2
n <- 50
SE <- sqrt((p * (1 - p)) / n)
pnorm((0.1 - 0.2) / SE)
## [1] 0.03855
```

which agrees with the answer in the back of the book. Notice that this is *slightly different* from the "exact" answer given above. This is because the CLT is an *approximation*.

Solution: (**8-21**) Although the answer is in the back of the book, there has been some confusion about part (b) in past semesters. Here are my calculations using R.

1. I enter the data and calculate the rates as follows:

```
group <- c('treatment', 'control', 'refused')
n.children <- 1000 * c(200, 200, 340)
n.polio <- c(57, 142, 157)
rate <- n.polio/n.children
polio.data <- data.frame(group, n.children, n.polio, rate)
polio.data
##       group n.children n.polio      rate
## 1 treatment     200000      57 0.0002850
## 2   control     200000     142 0.0007100
## 3   refused     340000     157 0.0004618
```

2. Now I construct the confidence interval for the *rates*:

```
  treatment <- subset(polio.data, group == 'treatment')
control <- subset(polio.data, group == 'control')

estimate <- control$rate - treatment$rate

SE <- sqrt(
  control$rate * (1 - control$rate)/control$n.children
  + treatment$rate * (1 - treatment$rate)/control$n.children
  )

ME <- qnorm(1 - 0.05/2) * SE
CI <- c(estimate - ME, estimate + ME)
```

   before converting them to cases per 100,000 children:

```
estimate * 10^5
## [1] 42.5

ME * 10^5
## [1] 13.82

CI * 10^5
## [1] 28.68 56.32
```

   After rounding, this agrees with the answer in the book.

Additional Problems

1. In April of 2013, Public Policy Polling carried out a survey of 1247 registered voters to determine whether Republicans and Democrats differ in their beliefs about various conspiracy theories. To answer this question, you'll need to download the full results of their survey which I've posted on my website for convenience: `http://www.ditraglia.com/econ103/conspiracy.pdf`. Note that this is a *pdf file* so you can't import it into R. You'll need to go read through the document to find the data from the poll.

   (a) Construct a 99% confidence interval for the proportion of registered voters who belive that a UFO crashed at Roswell, New Mexico in 1947 and the US Government covered it up.

   Solution: Overall percentages appear on page 2 of the report, and this question refers to Q3. The sample size is 1247 and $\widehat{p} = 0.21$. We can carry out the calculations in R as follows:

```
p.hat <- 0.21
n <- 1247
SE <- sqrt(p.hat * (1 - p.hat)/n)
ME <- qnorm(1 - 0.01/2) * SE
LCL <- p.hat - ME
UCL <- p.hat + ME
c(LCL, UCL)
## [1] 0.1803 0.2397
```

   (b) Is there evidence that male and female voters differ in their beliefs about Roswell and UFOs?

   Solution: Percentages broken down by sex appear on page 15, while overall percentages of men and women appear on page 3. Of the 1247 registered voters in the poll, about 50% were women and 50% were men. We'll call that $n = 623$ for each. The sample proportions are $\widehat{p}_W = 0.19$ for women versus $\widehat{p}_M = 0.24$ for men. Using R, we find:

```
n <- 623
p.M <- 0.24
p.W <- 0.19
SE.M <- sqrt(p.M * (1 - p.M)/n)
SE.W <- sqrt(p.W * (1 - p.W)/n)
SE <- sqrt(SE.M^2 + SE.W^2)
ME <- qnorm(1 - 0.01/2) * SE
```

```
diff <- p.M - p.W
LCL <- diff - ME
UCL <- diff + ME
c(LCL, UCL)
## [1] -0.009846  0.109846
```

This 99% CI just barely includes zero. A 95% wouldn't (try this out for yourself). We have found evidence suggesting that a higher proportion of men believe in the Roswell conspiracy compared to women.

(c) Is there evidence that Romney voters differ from Obama voters in their beliefs about Roswell and UFOs?

Solution: Percentages broken down by 2012 vote appear in page 5. Overall percentages of Romney and Obama voters in the sample appear on page 3. Of the 1247 registered voters in the sample, 50% voted for Obama and 44% voted for Romney. We'll call this $n_O = 623$ and $n_R = 547$. The sample proportions are $\widehat{p}_O = 0.16$ for Obama voters versus $\widehat{p}_R = 0.27$ for Romney voters. Using R, we find:

```
n.R <- 547
p.R <- 0.27
SE.R <- sqrt(p.R * (1 - p.R)/n.R)
n.O <- 623
p.O <- 0.16
SE.O <- sqrt(p.O * (1 - p.O)/n.O)
SE <- sqrt(SE.R^2 + SE.O^2)
ME <- qnorm(1 - 0.01/2) * SE
diff <- p.R - p.O
UCL <- diff + ME
LCL <- diff - ME
c(LCL, UCL)
## [1] 0.04818 0.17182
```

We have found strong evidence that a substantially greater proportion of Romney voters believe in the Roswell conspiracy.

(d) How should we interpret the results of the preceding two parts?

Solution: Since we know the men are more likely to vote for Republican candidates than women, it's difficult to tell whether the effect has to do with sex or

political affiliation. To learn more, we'd need to compare *female* Romney voters to *female* Obama voters and then *separately* compare male Obama voters to male Romney voters.

2. This problem concerns a dataset comparing the scores of men and women on the Armed Forces Qualifying Test (AFQT). The data are available from Professor Ditraglia's website:

```
data.url <- "http://www.ditraglia.com/econ103/ex0222.csv"
test.scores <- read.csv(data.url)
head(test.scores)
##   Gender Arith Word Parag Math AFQT
## 1   male    19   27    14   14 70.3
## 2 female    23   34    11   20 60.4
## 3   male    30   35    14   25 98.3
## 4 female    30   35    13   21 84.7
## 5 female    13   30    11   12 44.5
## 6 female     8   15     6    4  4.0
```

Each row is an individual who took the test. The first column gives that individual's sex, while the second through fifth columns give the individual's score on four parts of the test. The final column is an overall percentile score for the test.

(a) Suppose we want to compare the scores of men and women. Is this a problem based on two independent samples or matched pairs data?

Solution: Independent samples: each person's score on the exam is independent of every other person's score. There is no sensible way to match up pairs of observations here. Indeed, there are different numbers of men and women!

(b) For each of the four parts of the test, as well as for the overall percentile score, construct an approximate 95% CI for the difference of population means (men - women) based on the CLT. To make the calculations easier, notice that we can use the function `apply` to calculate the mean and variance of *each column at once*. For example, extracting the data for men:

```
test.men <- subset(test.scores, Gender == 'male')[,-1]
means.men <- apply(test.men, 2, mean)
var.men <- apply(test.men, 2, var)
```

Setting the second argument equal to 2 tells R to apply the function in the third argument to the *columns* of `test.men`.

Solution:

```
test.men <- subset(test.scores, Gender == 'male')[,-1]
test.women <-subset(test.scores, Gender == 'female')[,-1]
means.men <- apply(test.men, 2, mean)
var.men <- apply(test.men, 2, var)
n.men <- nrow(test.men)
means.women <- apply(test.women, 2, mean)
var.women <- apply(test.women, 2, var)
n.women <- nrow(test.women)
diff.means <- means.men - means.women
SE <- sqrt(var.women/n.women + var.men/n.men)
ME <- qnorm(1 - 0.05/2) * SE
LCL <- diff.means - ME
UCL <- diff.means + ME
CI <- rbind(LCL, UCL)
round(diff.means, 2)
## Arith  Word Parag  Math  AFQT
##  2.04 -0.02 -0.57  0.75  2.04
round(CI, 2)
##     Arith  Word Parag Math  AFQT
## LCL  1.49 -0.57 -0.81 0.27 -0.10
## UCL  2.58  0.52 -0.33 1.24  4.18
```

(c) Interpret your results.

Solution: Men score, on average, higher on the Arithmetic Reasoning and Mathematical Knowledge portions of the test. Women score higher, on average, on the Paragraph Comprehension portion of the test, while men and women appear to score about the same on the Word Knowledge portion. In terms of overall results, men seem to score higher than women, although the 95% CI does include zero.

3. This problem uses a dataset that investigates the relationship between schizophrenia and the volume (in cm^3) of a particular region of the brain (the left hippocampus) measured using an MRI machine. The dataset contains 15 sets of monozygotic (i.e. identical) twins, one of whom has schizophrenia ("Affected") and the other who does not ("Unaffected"). The idea of using identical twins is to hold constant unobserved

genetic and socioeconomic confounding variables that might influence whether someone develops schizophrenia. You can download the data from Professor Ditraglia's website as follows:

```
data.url <- "http://www.ditraglia.com/econ103/case0202.csv"
twins <- read.csv(data.url)
head(twins)
##   Unaffected Affected
## 1       1.94     1.27
## 2       1.44     1.63
## 3       1.56     1.47
## 4       1.58     1.39
## 5       2.06     1.93
## 6       1.66     1.26
```

(a) Should these data be analyzed as independent samples or matched pairs?

> **Solution:** This is matched pairs data. We would expect the size of the left hippocampus to be very similar for identical twins!

(b) Construct an approximate 95% confidence interval for the difference of means using the CLT and treating the data as two independent samples.

> **Solution:**
>
> ```
> mean.affected <- mean(twins$Affected)
> var.affected <- var(twins$Affected)
> n.affected <- length(twins$Affected)
> mean.unaffected <- mean(twins$Unaffected)
> var.unaffected <- var(twins$Unaffected)
> n.unaffected <- length(twins$Unaffected)
> diff.means <- mean.unaffected - mean.affected
> SE.indep <- sqrt(
>     var.affected/n.affected
>     + var.unaffected/n.unaffected)
> ME.indep <- qnorm(1 - 0.05/2) * SE.indep
> CI.indep <- c(diff.means - ME.indep, diff.means + ME.indep)
> round(CI.indep, 3)
> ## [1] 0.003 0.394
> ```

(c) Construct an approximate 95% confidence interval for the difference of means using the CLT and treating the data as matched pairs.

Solution:

```
twin.diff <- twins$Unaffected - twins$Affected
n.twins <- length(twin.diff)
SE.paired <- sqrt(var(twin.diff)/n.twins)
ME.paired <- qnorm(1 - 0.05/2) * SE.paired
CI.paired <- c(diff.means - ME.paired, diff.means + ME.paired)
round(CI.paired, 3)
## [1] 0.078 0.319
```

(d) The dataset only contains 15 pairs, a fairly small sample. Since the CLT is a large sample approximation, it may not work well in this situation. Suppose we were willing to assume that the within-twin differences came from a normal population. Construct an *exact* 95% confidence interval for the difference of means (again treating the data as matched pairs) under this assumption.

Solution:

```
ME.t <- qt(1 - 0.05/2, df = n.twins - 1) * SE.paired
CI.paired.t <- c(diff.means - ME.t, diff.means + ME.t)
round(CI.paired.t, 3)
## [1] 0.067 0.331
```

(e) Compare each of the intervals you have constructed. Why and how do they differ? What should we conclude?

Solution: The shortest interval is the one based on matched pairs using the CLT (`qnorm`). The widest is the one that assumes the samples are independent, which they are not. This interval is wider because the measurements are correlated across twins so that the sample variance of the differences is less than the sum of the sample variances of the affected and unaffected twins.

The interval based on the assumption that the differences come from a normal distribution is narrower than that based on assuming independent samples for the same reason, but wider than the equivalent interval based on the CLT. This is because each of them uses the same standard error estimate but `qt(0.975, df = 14)` is larger than `qnorm(0.975)`.

Although we may doubt that 15 is large enough for the approximation based on the CLT to work well, we may equally well doubt that the differences come from a normal population. Fortunately, both of the intervals based on differences give the same basic result: the twin with schizophrenia has, on average, a smaller

left hippocampus. If we wanted to be conservative, we could report the wider of the two intervals.

4. This question examines a situation in which the textbook confidence interval for a population proportion, based on the CLT, performs poorly but the refined interval works well. Recall that the refined CI is based on the quantity

$$\widetilde{p} = \frac{1}{n+4}\left(2 + \sum_{i=1}^{n} X_i\right)$$

while the textbook CI is based on $\widehat{p} = (\sum_{i=1}^{n} X_i)/n$.

(a) Show that $\widetilde{p} = (n\widehat{p} + 2)/(n+4)$

Solution:

$$\begin{aligned}
\widetilde{p} &= \frac{1}{n+4}\left(2 + \sum_{i=1}^{n} X_i\right) \\
&= \frac{2}{n+4} + \frac{n}{n+4}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\
&= \frac{2}{n+4} + \left(\frac{n}{n+4}\right)\widehat{p} \\
&= \frac{n\widehat{p}+2}{n+4}
\end{aligned}$$

(b) Suppose the true population proportion is $p = 0.5$ and we draw an iid sample of size 50, that is $X_1, \ldots, X_{50} \sim$ iid Bernoulli(0.5). We want to examine how often the textbook CI contains the true population proportion (0.5) in a large number of repeated samples. Since $\widehat{p}$ does not use the *individual* X_i, but only their sum, we can simulate $\widehat{p}$ based on an iid sample of size 50 by drawing a *single* Binomial(50, 0.5) random variable and dividing it by 50. In R,

```
rbinom(1, size = 50, prob = 0.5)/50
## [1] 0.58
```

Note that you may get a different answer from me since this is *random*. Indeed, if you run it repeatedly, you will typically get a different answer. The idea is to run this *many times*, and construct a confidence interval based on each result and see how many of them contain 0.5. Here is some code that does exactly that. Explain, step-by-step, how it works and what the result means. Then try running it yourself.

```
n <- 50
p <- 0.5
N.reps <- 100
p.hat <- rbinom(N.reps, size = n, prob = p)/n
ME.hat <- qnorm(0.975) * sqrt(p.hat * (1 - p.hat) / n)
LCL.hat <- p.hat - ME.hat
UCL.hat <- p.hat + ME.hat
CI.hat <- cbind(LCL.hat, UCL.hat)
Coverage <- (LCL.hat <= p) & (p <= UCL.hat)
Coverage <- sum(Coverage)/N.reps
Coverage
## [1] 0.96
```

Solution: The first four lines calculate 100 values of $\widehat{p}$ from 100 repeated samples of size 50 from a Bernoulli population with probability of success 0.5. These values are stored in the vector `p.hat`. The next four lines construct the approximate textbook 95% confidence interval for a population proportion corresponding to *each* of the 100 values for $\widehat{p}$ from the repeated samples. The third to last command checks each of these intervals to make sure that it contains the true value: 0.5. If so, it stores the value `TRUE` otherwise it stores the value `FALSE`. The second to last command uses a clever trick: if you sum a vector of `TRUE` and `FALSE` in R, it will treat the `TRUE` values as 1 and the `FALSE` values as 0. Thus, the sum *counts* how many of the intervals contain the true population parameter before dividing it by 100 to get the *fraction* of intervals that contain the truth. The result is close to what it should be: 0.95.

(c) How would the results change if you re-ran the above code with `N.reps <- 10000`? Try making the change and re-running the code.

Solution: This just changes *how many times we repeat the sampling*. It does *not change the sample size*. If we increase this number, we get closer to what we actually mean by repeated sampling, namely an *infinite number* of replications. In practical terms, the answer is much more precise and once again is close to 0.95 which is what we would expect.

(d) From here on, use `N.reps <- 10000`. What happens if you re-run the above code with `p <- 0.1` and `n <- 10`?

Solution: This changes the population from which we are sampling as well as

the sample size. Formerly the population proportion was 0.5 and the sample size 50 whereas now the population proportion is 0.1 and the sample size is 0.1. The result is as follows:

```
n <- 10
p <- 0.1
N.reps <- 10000
p.hat <- rbinom(N.reps, size = n, prob = p)/n
ME.hat <- qnorm(0.975) * sqrt(p.hat * (1 - p.hat) / n)
LCL.hat <- p.hat - ME.hat
UCL.hat <- p.hat + ME.hat
CI.hat <- cbind(LCL.hat, UCL.hat)
Coverage <- (LCL.hat <= p) & (p <= UCL.hat)
Coverage <- sum(Coverage)/N.reps
Coverage
## [1] 0.6399
```

As we talked about in class, the textbook CI for a population proportion can work poorly if p is close to zero or one and n is small.

(e) Adapt the above code to examine the performance of the refined CI when $p = 0.1$ and $n = 10$. Use `N.reps <- 10000` as above. Hint: you can reuse the `p.hat` vector from part (c) by using the formula from part (a).

Solution:

```
p.tilde <- (n * p.hat + 2) / (n + 4)
ME.tilde <- qnorm(0.975) * sqrt(p.tilde * (1 - p.tilde) / (n + 4))
LCL.tilde <- p.tilde - ME.tilde
UCL.tilde <- p.tilde + ME.tilde
CI.tilde <- cbind(LCL.tilde, UCL.tilde)
Cover.tilde <- (LCL.tilde <= p) & (p <= UCL.tilde)
Cover.tilde <- sum(Cover.tilde)/N.reps
Cover.tilde
## [1] 0.9352
```