Problem Set # 10

Econ 103

Part I – Problems from the Textbook

Chapter 6: 15, 17, 19(b), 21
Chapter 8: 17(c), 17(d), 19, 21

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

Part II – Additional Problems

1. Write R code to carry out the simulation experiments presented on slides 17–19 of Lecture 18 illustrating the central limit theorem. The R command for drawing from a Uniform$(0, 1)$ distribution is `runif` and the corresponding density is `dunif`. In each case, plot the density or mass function of the population and compare it to the histograms of the sample mean computed for random samples drawn from that population. In each simulation, use 10000 replications.

2. 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.

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

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

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

3. Construct an approximate 95% confidence interval for the Anchoring Experiment based on the CLT using *this semester's* data, following the details in Lecture 19. Be sure to properly account for missing values. How does it compare to the interval based on the data from lecture?

4. This problem concerns a dataset comparing the scores of men and women on the Armed Forces Qualifying Test (AFQT). The data are available from my 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?

(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`.

(c) Interpret your results.

5. 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 my 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?

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

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

   (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.

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

6. 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 $\widetilde{p} = (n\widehat{p} + 2)/(n + 4)$

(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
```

(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.

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

(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).