Lab Handout Week 12

Chapter covered:

- Chapter 9.4 One sample estimate
- Chapter 9.8 Two samples: Estimating the difference between Two means

1. For this problem, load and attach the **Getting to Know You Survey** for Spring 2019 using the following code:

```
NoU<-read.csv ("http://stat.umn.edu/~wuxxx725/data/Getting2NoUS2019.csv", header = TRUE )
#Check names of NoU
names(NoU)
```

We assume that the **Getting To Know You Survey** represents a random sample from the population of all U of M students. Consider the problem of constructing a 98% confidence interval for μ, the population mean hours per day spent on social media for Freshmen students. Use the following R command to define a new variable `social.Fr`. (social network hours for Freshmen only).

```
social.Fr <- NoU$social.networks.hrs[NoU$year == "Freshman"]
```

(a) Explore the distribution of `social.networks.hrs` graphically using either the `hist()` or `boxplot()` command. Describe the shape of the data.

```
hist(social.Fr, xlim=c(0, 10), breaks=20,
   main="Histogram of hours per day spent on social media for Freshmen", xlab="Hours
       per day")
```

In the R command above, option `xlim=c(0, 10)` sets the x-axis limit from 0 to 10 hours. `breaks=20` specifies the number of bars in the histograms. `main` gives the title of the plot and `xlab` gives the x-axis title.

(b) Construct boxplot and identify the interval where the middle 50% of the distribution falls within.

```
boxplot(social.Fr, ylim=c(0, 10))
```

(c) Construct Q-Q plot using `qqnorm()` `qqline()` as well. Q-Q plot compares the data distribution to standard normal distribution. If most of dots are along the straight line, we can conclude that the sample comes from a normal population distribution.

```
qqnorm(social.Fr) ##plot
qqline(social.Fr) ##add a straight line over the existing plot
```

Slightly skewed to the right. Most of points are along the straight line except for one outlier (a student spending 11 hours on social media)

(d) Based on your plots, does `social.Fr` appear to follow a normal distribution? If not, briefly explain why you can still construct the confidence interval for μ using large sample confidence interval method.

Not normal. As the the sample size is large, with the shape of the distribution not too skewed, sampling theory guarantees good results.

(e) Find the point estimate of μ. (Hint: Use `mean()` command)

`mean(social.Fr)=2.83`

(f) Find the standard error of point estimate. (Hint: Use `sd()` and `length()`)

$SE(\bar{X})=\frac{s}{\sqrt{n}} = 0.2179$

(g) Calculate the margin of error to construct a large-sample confidence interval.

moe $= z_{(0.02/2,df=67)}SE(\bar{X}) = z^*_{0.02/2}\frac{s}{\sqrt{n}} =$ `qnorm(1-0.02/2)`$\frac{s}{\sqrt{n}}$
$= 2.326(0.2179) = 0.5068$ **(in hours)**

(h) Construct a 98% confidence interval to estimate μ. Interpret the result.

$\bar{x} \pm z_{\alpha/2}\frac{s}{\sqrt{n}}$
$= 2.83 \pm 0.5068 = (2.324, 3.338)$
We are 98% confident that the interval (2.324, 3.338) (2 hours and 19 minutes, 3 hours and 20 minutes) contains the mean hours spent on social network for U of M Freshmen students.
Or if we repeatedly draw a random sample of size 68 and construct confidence interval using this method, then 98% of intervals constructed contain the true population mean.

(i) Based on your answer above, is 2.5 (2 hours and 30 minutes) a plausible value for μ?

Yes, it falls within the interval.

(j) Use `t.test()` to construct the 98% confidence interval for μ. Compare the result with your answer from (h). Use `?t.test` in console to learn more details.

```
t.test(social.Fr, conf.level = 0.98, alternative="two.sided")
```

(k) Now, we want to compare mean hours spent on social network for Freshmen and for Senior students. Use R command below to construct a side-by-side boxplot. Do you think Freshmen and Senior mean hours spent on social network are different? Do you think two population have equal variance?

```
social.Sr<- NoU$social.networks.hrs[NoU$year == "Senior"]
boxplot(social.Fr, social.Sr, names=c("Freshmen", "Senior"))
```

No, Freshmen and female mean shoe sizes seem different. The middle 50% of hours spent on social network for Freshmen overlap with Seniors. It seems that they have similar population standard deviation. The exact sample standard deviations of each gender can be calculated using the following command.

```
tapply(NoU$social.network.hrs, NoU$year, sd)
## OR
sd(social.Fr)
sd(social.Sr)
```

(l) (From Chapter 9.8, which we haven't covered yet.) Use the following command below to estimate $\mu_{Fr} - \mu_{Sr}$ where μ_{Fr} represents the mean hours spent on social media for Freshmen and μ_{Sr} represents the mean hours spent on social media for Senior students. Interpret the interval.

```
t.test(x=social.Fr, y=social.Sr, conf.level=0.98, alternative="two.sided",
    var.equal=TRUE)
```

```
Two Sample t-test

data:  social.Fr and social.Sr
t = 0.6586, df = 112, p-value = 0.5115
alternative hypothesis: true difference in means is not equal to 0
98 percent confidence interval:
 -0.5740227  1.0183961
sample estimates:
mean of x mean of y
 2.830882  2.608696
```

We are 98% confident that the difference in the mean hour spent on social network between Freshmen and Senior students is between -0.57 hours (about -34 minutes) and 1.018 hours (about 60.6 minutes).