Problem Set #9

Econ 103

For all of the questions on this assignment, assume the data are a random sample from a normal population, that is $X_1, \ldots, X_n \sim$ iid $N(\mu, \sigma^2)$. Depending on the specifics of the question, σ may be known or unknown. I'd recommend that you use R for any problems in which you need to calculate sample means or variances from a dataset: in an exam, I wouldn't make you do this by hand.

Part I − Problems from the Textbook

Chapter 8: 1, 3, 5, 7, 9

To avoid typing, you can download the data for question 8-9 from my website and import it directly into R with the following commands:

```
data.url <- "http://www.ditraglia.com/econ103/speed.csv"
speed <- read.csv(data.url, header = FALSE)
speed <- speed[,1]
```

The second command is simply a quick way to tell R to store `speed` as a vector rather than a dataframe since it only contains one column.

Part II − Additional Problems

1. Oranges sold at Iovine Brothers Produce in Reading Terminal Market have weights that follow a normal distribution with a mean of 12 ounces and standard deviation of 2 ounces.

   (a) If we choose an orange at random, what is the probability that it will weigh less than 10 ounces?

   > **Solution:**
   > ```
   > pnorm(10, mean = 12, sd = 2)
   > ## [1] 0.1587
   > ```

(b) If we choose 25 oranges at random, what is the probability that they will have a total weight of less than 250 ounces?

> **Solution:** Since the weight of any individual orange is an independent draw from a normal distribution with mean 12 ounces and standard deviation 2 ounces, the weight of *25* oranges can be represented as a draw from a normal distribution with mean $25 \times 12 = 300$ and variance $25 \times 4 = 100$. Hence, the standard deviation is 10. Plugging this into R:
>
> ```
> pnorm(250, mean = 300, sd = 10)
> ## [1] 2.867e-07
> ```

2. All other things equal, how would the following change the width of a confidence interval for the mean of a normal population? Explain.

> **Solution:** All answers below are based on the following expression for a $(1 - \alpha) \times 100\%$ confidence interval for the mean of a normal population when the population standard deviation is unknown:
>
> $$\bar{X}_n \pm \mathtt{qt}(1 - \alpha/2, df = n - 1) \times \frac{S}{\sqrt{n}}$$
>
> The width of this interval is
>
> $$\text{Width} = 2 \times \mathtt{qt}(1 - \alpha/2, df = n - 1) \times \frac{S}{\sqrt{n}}$$

(a) The sample mean is smaller.

> **Solution:** No effect: width doesn't involve the sample mean.

(b) The population mean is smaller.

> **Solution:** No effect: width doesn't involve the population mean.

(c) The sample standard deviation is smaller.

> **Solution:** If S decreases, all other things constant, the width decreases.

(d) The sample size is smaller.

> **Solution:** Changing sample size has two effects but they both go in the same direction. First, if n gets smaller, $\mathtt{qt}(1 - \alpha/2, df = n - 1)$ gets larger as we can see from the table presented in the lecture slides. Second, as n gets smaller holding all other things fixed, $S/\sqrt{n}$ gets larger. Hence, decreasing sample size, all other things equal, increases the width.

3. Do you agree or disagree with the following statement: "the household unemployment survey is hardly flawless; its 60,000 families constitute less than 0.1% of the workforce." Explain your answer.

> **Solution:** There may be flaws in the household unemployment survey, but the fact it is based on a sample of less than 0.1% of the workforce is not one of them. Sampling distributions only depend on the *distribution* of the population, not on the *number of individuals* that make up that population. As we saw in class, the variance of the sampling distribution of the sample mean is σ^2/n. Thus, the accuracy of the sample mean, in probabilistic terms, depends on the relative size of the *population variance* versus the *sample size*. It does *not* depend on the relative size of the sample versus the population.

4. Suppose you want to construct a 99% confidence interval for the average height of US males above the age of 20. Based on past studies you think the standard deviation of heights for this population is around 6 inches. How large a sample should you gather to ensure that your confidence interval has a width no greater than 1 inch?

> **Solution:** Assuming the population is normal and σ is known, our confidence interval takes the form:
> $$\bar{X}_n \pm \mathtt{qnorm}(1 - \alpha/2) \times \sigma/\sqrt{n}$$
> Thus, the width equals $2(1 - \alpha/2) \times \sigma/\sqrt{n}$. From the problem statement $\sigma = 6$. For a 99% confidence interval we set $\alpha = 0.005$. Plugging this into R, we find $\mathtt{qnorm}(1 - 0.01/2) \approx 2.58$. Thus, in terms of n, the width of our interval is approximately
> $$2 \times 2.58 \times 6/\sqrt{n} \approx 31/\sqrt{n}$$
> Solving $1 = 31/\sqrt{n}$ for n gives $n = 961$. Double checking this in R:

```
n <- 950:961
width <- 2 * qnorm(1 - 0.01/2) * 6/sqrt(n)
cbind(n, width)

##          n  width
##  [1,] 950 1.0029
##  [2,] 951 1.0023
##  [3,] 952 1.0018
##  [4,] 953 1.0013
##  [5,] 954 1.0007
##  [6,] 955 1.0002
##  [7,] 956 0.9997
##  [8,] 957 0.9992
##  [9,] 958 0.9987
## [10,] 959 0.9981
## [11,] 960 0.9976
## [12,] 961 0.9971
```

so the exact answer is 956, which is pretty close to what we got using a rounded value for $2 \times$ `qnorm(1 - 0.01/2)` $\times 6$ as we did above.

5. A well-known weekly news magazine once wrote that the width of a confidence interval is inversely related to sample size: for example, if a sample size of 500 gives a confidence interval of plus or minus 5, then a sample of 2500 would give a confidence interval of plus or minus 1. Explain the error in this argument.

Solution: This question involves symmetric confidence intervals, i.e. intervals of the form $\widehat{\theta} \pm ME$. As we have seen in class, the width of such intervals, whether based on the normal or t distributions, depends on $\sqrt{n}$ rather than n:

$$\sigma \text{ Known:} \quad \bar{X}_n \quad \pm \quad \texttt{qnorm}(1 - \alpha/2) \times \frac{\sigma}{\sqrt{n}}$$

$$\sigma \text{ Unknown:} \quad \bar{X}_n \quad \pm \quad \texttt{qt}(1 - \alpha/2, df = n - 1) \times \frac{S}{\sqrt{n}}$$

Thus, all other things equal, we would have to quadruple the sample size to cut the width of the interval in half. (There is a slight complication that arises from the fact that the quantile of the t interval also involves n, but as discussed in class, this only makes a practical difference in the confidence interval when n is very small.)

6. This question uses the same data as question 8-9 from the textbook. (See above for instructions on how to download the data.) You may assume that the data constitute a random sample from a normal population.

> **Solution:** First, read in the data and make sure they look reasonable:
>
> ```r
> speed.data <- read.csv("http://www.ditraglia.com/econ103/speed.csv", header = FALSE
> speed.data <- speed.data[,1]
> head(speed.data)
>
> ## [1] 49 71 78 58 83 74
> ```

(a) Construct a 90% confidence interval for population variance.

> **Solution:**
>
> ```r
> n.speed <- length(speed.data)
> a.speed <- qchisq(0.1/2, df = n.speed - 1)
> b.speed <- qchisq(1 - 0.1/2, df = n.speed - 1)
> LCL <- (n.speed - 1) * var(speed.data)/b.speed
> UCL <- (n.speed - 1) * var(speed.data)/a.speed
> c(LCL, UCL)
> ## [1] 100.0 212.4
> ```

(b) The data for this problem are given in kilometers per hour. How would the confidence interval change if the data were expressed in miles per hour? Note that 1 km ≈ 0.62 miles.

> **Solution:** One way to solve this would be to convert the data to miles per hour and then repeat the above, noting that this will only change the sample variance so we can re-used everything else:
>
> ```r
> LCL.mhr <- (n.speed - 1) * var(0.62 * speed.data)/b.speed
> UCL.mhr <- (n.speed - 1) * var(0.62 * speed.data)/a.speed
> c(LCL.mhr, UCL.mhr)
> ## [1] 38.45 81.65
> ```
>
> Another way would be to transform the confidence interval we have already created. From above:
> $$P(100.01 < \sigma^2 < 212.41) = 0.9$$

where σ^2 has units $(\text{km/hr})^2$. To transform to units of $(\text{miles/hr})^2$ we need to multiply by $(0.62 \text{ miles/km})^2$. Notice that this leaves the meaning of the probability statement unchanged:

$$P(38.45 < 0.62^2 \times \sigma^2 < 81.65) = 0.9$$

Notice that this agrees with the alternative solution that we found by transforming the *data* rather than the interval.

(c) Construct a 90% confidence interval for the population *standard deviation* in kilometers per hour.

Solution: To solve this, we use the fact that we can carry out algebraic manipulations within a probability statement without changing the meaning of that statement and hence the associated probability. (This is the logical equivalence rule). From above, we found a confidence interval

$$P(100.01 < \sigma^2 < 212.41) = 0.9$$

where the data are measured in km/hour. Taking square roots of everything:

$$P(10 < \sigma < 14.57) = 0.9$$

7. Researchers asked a random sample of college students how many hours they sleep every night. The data are stored on my website and can be imported directly into R as follows:

```
data.url <- "http://www.ditraglia.com/econ103/sleep.csv"
sleep <- read.csv(data.url, header = FALSE)
sleep <- sleep[,1]
```

The second command is simply a quick way to tell R to store `sleep` as a vector rather than a dataframe since it only contains one column. You may assume that the data represent a random sample from a normal population.

Solution: First, read in the data and make sure they look reasonable:

```
data.url <- "http://www.ditraglia.com/econ103/sleep.csv"
sleep.data <- read.csv(data.url, header = FALSE)
sleep.data <- sleep.data[,1]
head(sleep.data)

## [1] 5.0 4.0 6.5 5.0 5.0 6.0
```

(a) Construct a 95% confidence interval for the population mean.

Solution:

```
n.sleep <- length(sleep.data)
SE.sleep <- sd(sleep.data)/sqrt(n.sleep)
ME.sleep <- qt(1 - 0.05/2, df = n.sleep - 1) * SE.sleep
LCL <- mean(sleep.data) - ME.sleep
UCL <- mean(sleep.data) + ME.sleep
c(LCL, UCL)
## [1] 5.528 6.222
```

(b) Construct a 95% confidence interval for the population variance.

Solution:

```
a.sleep <- qchisq(0.05/2, df = n.sleep - 1)
b.sleep <- qchisq(1 - 0.05/2, df = n.sleep - 1)
LCL <- (n.sleep - 1) * var(sleep.data)/b.sleep
UCL <- (n.sleep - 1) * var(sleep.data)/a.sleep
c(LCL, UCL)
## [1] 0.7893 1.9394
```