Problem Set #8

Econ 103

Midterm Announcements

- □ Midterm will only cover material **since the first midterm** (Chapter 4-7 slides)
- □ The only R code you are responsible for writing is the commands related to distributions (rnorm, dnorm, qnorm, pnorm, rchisq, etc.). Make sure you know what arguments each of those takes and how to decide when to use "r", "d", "q", or "p".
- □ You are not responsible for any R material outside of what has been covered in the lecture slides. For example, you will not be tested on RMarkdown. I want you to focus on studying the slides and getting comfortable with the material.

Lecture Progress

We made it to slide 70 of the Chapter 7 slides.

Homework Checklist

- □ **Book Problems (Chapter 8):** 1, 3, 5, 7, 9
- □ **Additional Problems:** See below
- □ **Practice Midterms:** Do them!
- □ **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. (Not due until the week after the midterm)

Additional Problems

1. For this question assume that we have a random sample from a normal distribution with unknown mean but *known* variance.

   (a) Suppose that we have 36 observations, the sample mean is 5, and the population variance is 9. Construct a 95% confidence interval for the population mean.

   Solution: Since the population variance is 9, the population standard deviation is 3. Hence, the desired confidence interval is $5 \pm 2 \times 3/\sqrt{36} = 5 \pm 1 = (4, 6)$

   (b) Repeat the preceding with a population variance of 25 rather than 9.

   Solution: The population standard deviation becomes 5 so the confidence interval becomes $5 \pm 2 \times 5/\sqrt{36} = 5 \pm 10/6 \approx (3.3, 6.7)$

   (c) Repeat the preceding with a sample size of 25 rather than 36.

   Solution: $5 \pm 2 \times 5/\sqrt{25} = 5 \pm 2 = (3, 7)$

   (d) Repeat the preceding but construct a 50% rather than 95% confidence interval.

   Solution: Here we need to use R to get the appropriate quantile:

   $$5 \pm qnorm(0.75) \times 5/\sqrt{25} = 5 \pm 0.67 = (4.33, 5.67)$$

   (e) Repeat the preceding but construct a 99% rather than a 50% confidence interval.

   Solution: Again we use R to get the appropriate quantile:

   $$5 \pm qnorm(0.995) \times 5/\sqrt{25} = 5 \pm 2.58 = (2.42, 7.58)$$

2. In this question you will carry out a simulation exercise similar to the one I used to make the plot of twenty confidence intervals from lecture 16.

   (a) Write a function called `my.CI` that calculates a confidence interval for the mean of a normal population when the population standard deviation is known. It should take three arguments: `data` is a vector containing the observed data from which we will calculate the sample mean, `pop.sd` is the population standard deviation, and `alpha` controls the confidence level (e.g. `alpha = 0.1` for a 90% confidence interval). Your function should return a vector whose first element is the lower confidence limit and whose second element is the upper confidence limit. Test out your function on a simple example to make sure it's working properly.

Solution: Your function might look something like this:

```
my.CI <- function(data, pop.sd, alpha) {
x.bar <- mean(data)
n <- length(data)
ME <- qnorm(1 - alpha / 2) * pop.sd / sqrt(n)
lower <- x.bar - ME
upper <- x.bar + ME
ans <- c(lower, upper)
return(ans)
}
```

One simple example would be to work this out for a data set of zeros with a population variance of 1. We could test the function using the following code:

```
data <- rep(0, 25)
my.CI(data, pop.sd = 1, alpha = 0.05)
[1] -0.3919928  0.3919928
```

Checking this by hand, our confidence would be $0 \pm 2 \times 1/5 = (-0.4, 0.4)$ This is basically the same. The discrepancy arises because we're using an approximation when multiplying by 2. R evaluates this expression using the more accurate value of about 1.96, which results in slightly lower values for the upper and lower confidence limits.

(b) Write a function called `CI.sim` that takes a single argument `sample.size`. Your function should carry out the following steps. First generate `sample.size` draws from a standard normal distribution. Second, pass your sample of standard normals to `my.CI` with `alpha` set to 0.05 and `pop.sd` set to 1. Third, return the resulting confidence interval. Test your function on a sample of size 10. (What we're doing here is constructing a 95% confidence interval for the mean of a normal population using simulated data. The population mean is in fact zero, but we want to see how our confidence interval procedure works. To do this we "pretend" that we don't know the population mean and only know the population variance. Think about this carefully and make sure you understand the intuition.)

Solution: Your function might look something like this:

```
CI.sim <- function(sample.size) {
data <- rnorm(sample.size)
CI <- my.CI(data, pop.sd = 1, alpha = 0.05)
return(CI)
}
```

To test it out, we simply run the following. Remember, in this case, because we are simulating our data, we know that the population mean is 0 (because we specified the data is being drawn from a standard normal!). Hence, we should expect for the resulting confidence interval to contain 0.

```
CI.sim(10)
[1] -0.7021254  0.5374647
```

You will probably not get the same confidence interval since R likely generated different draws from a normal on your computer.

(c) Use `replicate` to construct 10000 confidence intervals based on simulated data using the function `CI.sim` with `sample.size` equal to 10. (Note that `replicate` will, in this case, return a matrix with *2* rows and *10000* columns. Each column corresponds to one of the simulated confidence intervals. The first row contains the lower confidence limit while the second row contains the upper confidence limit.) Calculate the proportion of the resulting confidence intervals contain the true population mean. Did you get the answer you were expecting?

Solution: Here, we just use what we constructed above, but we have to replicate it a bunch of times. One way of doing this is the following:

```
    sims <- replicate(10000, CI.sim(10))
```

To see what proportion of these confidence intervals contain the true population mean of 0, we need to find the ones where the lower confidence limit is below 0 and the upper confidence limit is above zero.

```
   lower <- sims[1, ]
   upper <- sims[2, ]
   covers.truth <- (lower < 0 & upper > 0)
   sum(covers.truth) / length(covers.truth)
   [1] 0.9464
```

Great! This is really close to the 0.95 we were expecting!

(d) Repeat the preceding but rather than using `CI.sim` write a new function called `CI.sim2`. This new function should be identical to `CI.sim` except that, when calling `my.CI`, it sets `pop.sd = 1/2` rather than 1. How do your results change? Try to provide some intuition for any differences you find.

Solution: First, let's write our new function CI.sim2.

```
CI.sim2 <- function(sample.size) {
data <- rnorm(sample.size)
CI <- my.CI(data, pop.sd = 0.5, alpha = 0.05)
return(CI)
}
```

Now let's run the same procedure we did in part (c):

```
     sims2 <- replicate(10000, CI.sim2(10))
    lower <- sims2[1, ]
    upper <- sims2[2, ]
    covers.truth <- (lower < 0 & upper > 0)
    sum(covers.truth) / length(covers.truth)
    [1] 0.679
```

In this case the procedure didn't work: many fewer than 95% of the intervals contain the true population mean. The problem is that `CI.sim2` constructs a confidence interval using the *wrong* population standard deviation! Since it uses 1/2 rather than 1, the resulting intervals are too short, so too few of them contain the true population mean.

3. 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: Using R, we get the following

   ```
   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 $25 \times 4 = 100$ (so the standard deviation is 10). In other words, remember that the sum of normal random variables is also a normal random variable, you just have to compute the new mean and variance. Plugging this into R:

```
pnorm(250, mean = 300, sd = 10)
[1] 2.867e-07
```

4. All other things equal, how would the following change the width of a confidence interval for the mean of a normal population? The population standard deviation is unknown. 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:

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

The width of this interval is

$$\text{Width} = 2 \times \texttt{qt}(1-\alpha/2, df = n-1) \times \frac{S}{\sqrt{n}}$$

(a) The sample mean is smaller.

Solution: No effect: width doesn't involve the sample mean. It cancels out when you subtract the upper and the lower confidence limits.

(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, $\texttt{qt}(1-\alpha/2, df = n-1)$ gets larger as we can see from the table presented in the lecture slides (use R to convince yourself as well). 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.

5. 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.

6. 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:

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

Thus, the width equals $2qnorm(1-\alpha/2) \times \sigma/\sqrt{n}$. From the problem statement $\sigma = 6$. For a 99% confidence interval we set $\alpha = 0.005$. Plugging this into R, we find $\texttt{qnorm}(1-0.01/2) \approx 2.58$. 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 <- 950:960
width <- 2 * qnorm(1 - 0.01 / 2) * 6 / sqrt(n)
cbind(n, width)
          n     width
```

```
 [1,] 950 1.0028513
 [2,] 951 1.0023239
 [3,] 952 1.0017973
 [4,] 953 1.0012715
 [5,] 954 1.0007466
 [6,] 955 1.0002225
 [7,] 956 0.9996993
 [8,] 957 0.9991768
 [9,] 958 0.9986552
[10,] 959 0.9981344
[11,] 960 0.9976144
```

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

7. 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:

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

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.)