Problem Set (Week 5)

Econ 103

Lecture 16 - 17

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] 0.1586553
   ```

   (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] 2.866516e-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:

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

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.

(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. 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. Researchers asked a random sample of college students how many hours they sleep every night. The data can be imported directly into R using the following codes:

```
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] 5.528139 6.221861
```

(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] 0.7893144 1.9393917
```

Lecture 18

4. 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 using the following codes:

```
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.

Lecture 19 - 20

5. Let $X_1, \ldots, X_n \sim$ iid $N(\mu, \sigma^2)$ and suppose we know that $\sigma^2 = 1$.

   (a) Write down the sampling distribution of $\sqrt{n}(\bar{X}_n - \mu)$.

   Solution: $N(0, 1)$

   (b) Suppose we wanted to test the null hypothesis that $\mu = 0$ against the two-sided alternative at the 5% level. What test statistic should we use and what should our decision rule be?

   Solution: Under the null,

   $$T = \sqrt{n}\bar{X}_n \sim N(0, 1)$$

   and we reject if $|T| \geq$ `qnorm(0.975)` ≈ 2.

   (c) How would your answer to (b) change if we tested instead against the one-sided alternative that $\mu > 0$?

   Solution: We still examine the test statistic $T = \sqrt{n}\bar{X}_n$ except in this case the decision rule is: reject if $T \geq$ `qnorm`(0.95). This critical value is *less* than 2.

6. (You will be able to solve this part after lecture 21) This problem uses the same dataset as in question 4. In the previous question you used this dataset to construct confidence intervals. In this question you will carry out hypothesis tests. For this question you may assume that the sample differences between the left hippocampus volume of the "Affected" and "Unaffected" twins are drawn from a normal population with unknown variance.

   (a) Carry out a one-sided test at the 5% level of the null hypothesis of no difference against the alternative that the affected twin has a larger left hippocampus, on

average. What is your test statistic? What is your critical value? What is your decision rule? What is your decision?

Solution: First load the data and calculate the quantities we'll need to carry out all of the tests below.

```
twin.diff <- twins$Unaffected - twins$Affected
mean.diff <- mean(twin.diff)
n.twins <- length(twin.diff)
SE.paired <- sqrt(var(twin.diff)/n.twins)
```

Notice that we calculated the differences as Unaffected twin minus Affected twin. We need to be careful about the sign here to see when we should reject the null. In this part, we are testing against the one-sided alternative that the *Affected* twin has a larger left hippocampus. Thus, we should reject when `twin.diff` is *sufficiently negative*. Now we calculate the test statistic and critical value for the one-sided test:

```
test.stat <- mean.diff/SE.paired
test.stat
## [1] 3.228928
critical.value <- qt(0.05, df = n.twins - 1)
critical.value
## [1] -1.76131
test.stat <= critical.value
## [1] FALSE
```

In this case, our decision rule is to reject the null if the test statistic is *less than* -1.7613101. (Remember, we have to keep track of the sign.) We see that this is not the case, so we fail to reject the null. We have not found evidence that the Affected twin has a larger left hippocampus.

(b) Repeat part (a) for a test against the *opposite* one-sided alternative.

Solution: The test statistic remains the same in this case, but our decision rule and critical value have changed. Again, note that we calculated the differences as Unaffected twin minus Affected twin. Thus, a large *positive* value of `mean.diff` would provide evidence that we should reject the null in favor of the one-sided alternative that the Unaffected twin has the larger left hippocampus, on average. The critical value simply changes sign to reflect this:

```
critical.value <- qt(1 - 0.05, df = n.twins - 1)
critical.value
## [1] 1.76131
test.stat >= critical.value
## [1] TRUE
```

Our decision rule is to reject when the test statistic is greater than or equal to 1.7613101. Since this is the case, we reject the null hypothesis that the difference is zero at the 5% significance level. We have found evidence that schizophrenia is associated with a smaller left hippocampus based on the twin data.

(c) Repeat part (a) but test against the *two-sided* alternative.

Solution: Again, the test statistic remains the same. Since we're testing against the two-sided alternative, however, we reject if it is too large *or* two small. This is equivalent to asking whether the *absolute value* of the test statistic is larger than the appropriate (positive) two-sided critical value:

```
critical.value  <- qt(1 - 0.05/2, df = n.twins - 1)
critical.value
## [1] 2.144787
abs(test.stat) >= critical.value
## [1] TRUE
```

We see that this is indeed the case, so we would reject that null hypothesis that the difference is zero against the two-sided alternative at the 5% significance level.

(d) Explain the differences between your results in parts (a), (b), and (c).

Solution: Both parts (b) and parts (c) give the same result: reject the null. Parts (a) and (b) are mutually exclusive: if we reject in favor of one of the one-sided alternatives, we can't reject in favor of the other. This is because the `mean.diff` is either positive or negative: if positive, it suggests that the Unaffected twin has a larger left hippocampus; if negative, that the Affected twin does. We saw in class that, in borderline cases, it is possible to reject against a one-sided alternative without rejecting against the two-sided alternative. We are not in such a situation here.

(e) Calculate the p-values corresponding to parts (b) and (c).

Solution:

```
1 - pt(test.stat, df = n.twins - 1)
## [1] 0.003030772
2 * (1 - pt(abs(test.stat), df = n.twins - 1))
## [1] 0.006061544
```

We see that both the one-sided p-value for part (b) and the two-sided p-value for (c) are quite small: there is extremely strong evidence against the null hypothesis of no difference. From these values, for example, we see that we would have still rejected if we had carried out these two tests at the 1% rather than the 5% level.