Problem Set # 11

Econ 103

Lecture Progress

We made it to slide 72 of the Chapter 8 slides.

Homework Checklist

- ☐ **Book Problems (Chapter 9):** 7, 9, 11, 23, 25, 27, 29
- ☐ **Additional Problems:** See below
- ☐ **Ask questions on Piazza**
- ☐ **Review slides**
- ☐ **R Tutorial:** Develop testable hypotheses for your R project and test at least one of them.

Additional Problems

1. 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 Professor DiTraglia's 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
```

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.

2. This problem concerns a dataset comparing the scores of men and women on the Armed Forces Qualifying Test (AFQT). Throughout you may assume that the sample size is large enough for the CLT to apply. As before, the data are available from Professor DiTraglia's 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) For each section of the exam, as well as for overall percentile scores, test the null hypothesis that the population mean scores are equal for men and women at the 5% level. In which cases do you reject, and in which cases do you fail to reject?

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

```
test.men <- subset(test.scores, Gender == 'male')[,-1]
means.men <- apply(test.men, 2, mean)
var.men <- apply(test.men, 2, var)
n.men <- nrow(test.men)
test.women <-subset(test.scores, Gender == 'female')[,-1]
means.women <- apply(test.women, 2, mean)
var.women <- apply(test.women, 2, var)
n.women <- nrow(test.women)
diff.means <- means.men - means.women
SE <- sqrt(var.women/n.women + var.men/n.men)
```

Since we've arranged the differences of means and associated standard errors in *vectors*, we can calculate all of the test statistics at once!

```
test.stats <- diff.means/SE
test.stats
##       Arith        Word       Parag        Math        AFQT
##  7.31241751 -0.07983706 -4.60228958  3.04909284  1.87014777
```

Now, the critical value for a 5%, two-sided test in this case is 1.959964

```
critical.value <- qnorm(1 - 0.05/2)
```

So for which comparisons would we reject the null?

```
abs(test.stats) >= critical.value
## Arith  Word Parag  Math  AFQT
##  TRUE FALSE  TRUE  TRUE FALSE
```

(b) How do your results from part (a) relate to the CIs you constructed using this dataset in an earlier assignment?

Solution: We can construct the confidence intervals from the previous homework assignment as follows:

```
ME <- qnorm(1 - 0.05/2) * SE
LCL <- diff.means - ME
UCL <- diff.means + ME
rbind(LCL, UCL)
##         Arith       Word      Parag      Math        AFQT
## LCL 1.491059 -0.5654962 -0.8119748 0.2686287 -0.09799557
## UCL 2.583052  0.5212295 -0.3269459 1.2354610  4.17891115
```

Because of the relationship between a two-sided test at the 5% level and a 95% confidence interval, we rejected the null in *exactly* the cases where the corresponding confidence interval did not contain zero.

(c) Calculate the two-sided p-values for each test from part (a).

Solution:

```
p.values <- 2 * (1 - pnorm(abs(test.stats)))
round(p.values, 4)
##  Arith   Word  Parag   Math   AFQT
## 0.0000 0.9364 0.0000 0.0023 0.0615
```

We have found extremely strong evidence of a difference between men and women on the Arithmetic and Paragraph portions, and very strong evidence of a difference on the Math portion. We found some evidence suggestive of a difference in overall quantile scores, but no evidence of a difference in the Word knowledge portion of the test.

3. 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 are on my website:
`https://mallickhossain.files.wordpress.com/2016/07/conspiracy.pdf`
In an earlier assignment you used these data to construct confidence intervals. In this question you'll use them to carry out hypothesis tests. Throughout you may assume that the sample size is large enough for the approximate based on the central limit theorem to be valid.

(a) Suppose we wanted to test the null hypothesis that 20% of registered voters believe that a UFO crashed at Roswell, New Mexico in 1947 and the US Government covered it up. There are two possible test statistics we could use. Calculate them both and explain the difference. Which is preferable?

Solution: Overall percentages appear on page 2 of the report, and this question refers to Q3. The sample size is 1247 and $\widehat{p} = 0.21$

```
p.hat <- 0.21
n <- 1247
```

We calculate the numerator of the test statistic as follows

```
p.null <- 0.20
numerator <- p.hat - p.null
```

For the denominator we need the standard error of $\widehat{p}$. There are two possibilities. The first is to use the estimated standard error

```
n <- 1247
SE.est <- sqrt(p.hat * (1 - p.hat)/n)
```

The second option is to use the exact standard error *under the null hypothesis*

```
SE.0 <- sqrt(p.null * (1 - p.null)/n)
```

The two test statistics are as follows:

```
test.stat <- numerator / SE.est
test.stat.refined <- numerator / SE.0
test.stat
## [1] 0.8669819
test.stat.refined
## [1] 0.8828222
```

The refined test statistic is preferable since it *fully imposes* the null hypothesis. This is the test statistic that we will use below.

(b) Suppose that we wanted to test the null hypothesis from the preceding part against the one-sided alternative that more than 20% of registered voters believe in the UFO conspiracy. Calculate the p-value for this test.

Solution:

```
1 - pnorm(test.stat.refined)
## [1] 0.1886662
```

(c) Repeat the preceding part for the *two-sided* alternative.

Solution:

```
2 * (1 - pnorm(test.stat.refined))
## [1] 0.3773324
```

(d) Calculate the p-value for a test of the null hypothesis that equal proportions of Romney and Obama voters believe in the UFO conspiracy against the two-sided alternative. There are two test statistics you could use. Calculate the p-value using each and explain the difference. Which should we prefer?

Solution: Percentages broken down by 2012 vote appear in page 5 of the survey results. Overall percentages of Romney and Obama voters in the sample appear on page 3. Of the 1247 registered voters in the sample, 50% voted for Obama and 44% voted for Romney. We'll call this $n_O = 623$ and $n_R = 547$. The sample proportions are $\widehat{p}_O = 0.16$ for Obama voters versus $\widehat{p}_R = 0.27$ for Romney voters:

```
n.R <- 547
p.R <- 0.27
n.O <- 623
p.O <- 0.16
diff <- p.R - p.O
```

The two statistics correspond to different ways of calculating the standard error of the difference of sample means. The first possibility is to use the estimated standard errors for each population and combine them using the independence of the samples, as we did when constructing confidence intervals:

```
SE.R <- sqrt(p.R * (1 - p.R)/n.R)
SE.O <- sqrt(p.O * (1 - p.O)/n.O)
SE <- sqrt(SE.R^2 + SE.O^2)
```

The second possibility is to construct a *pooled* estimator of the standard error based on a *pooled* sample proportion. This is preferable because it fully imposes the null hypothesis:

```
n.total <- n.O + n.R
p.pooled <- ((n.O * p.O) + (n.R * p.R)) / n.total
SE.pooled <- sqrt(p.pooled * (1 - p.pooled) * (1/n.O + 1/n.R))
```

The resulting test statistics are as follows:

```
test.stat <- diff / SE
test.stat.refined <- diff / SE.pooled
test.stat
## [1] 4.583097
test.stat.refined
## [1] 4.597651
```

The two test statistics are quite similar in this particular example and both p-values are essentially zero:

```
2 * (1 - pnorm(test.stat))
## [1] 4.581394e-06
2 * (1 - pnorm(test.stat.refined))
## [1] 4.272809e-06
```

Using either test statistic, we would find extremely strong evidence against the null hypothesis.

4. Suppose $X_1, \ldots, X_n \sim$ iid $N(\mu_x, \sigma_X^2)$ independently of $Y_1, \ldots, Y_m \sim$ iid $N(\mu_Y, \sigma_Y^2)$.

   (a) Suppose $n = m = 10$ and you know that $\sigma_X^2 = \sigma_Y^2 = 10$. Express the power of a two-sided test of $H_0 \colon \mu_X = \mu_Y$ at the 5% level in terms of the true, unknown difference of population means $\Delta = \mu_X - \mu_Y$.

Solution: Since the population variances are known and both populations are normal, the test statistic

$$T = \frac{\bar{X}_n - \bar{Y}_n}{\sqrt{\sigma_X^2/n + \sigma_Y^2/m}} = \frac{\bar{X}_n - \bar{Y}_m}{\sqrt{2}}$$

follows a standard normal distribution under the null. This is *exact* because we know the population is normal. Under the alternative $H_1 \colon \mu_X \neq \mu_Y$, however, the above test statistic does *not* follow a standard normal distribution. Instead,

$$\frac{(\bar{X}_n - \bar{Y}_m) - (\mu_X - \mu_Y)}{\sqrt{2}} \sim N(0, 1)$$

Hence,

$$\frac{\bar{X}_n - \bar{Y}_m}{\sqrt{2}} = \frac{(\bar{X}_n - \bar{Y}_m) - (\mu_X - \mu_Y)}{\sqrt{2}} + \frac{\mu_X - \mu_Y}{\sqrt{2}} \sim N\left(\frac{\mu_X - \mu_Y}{\sqrt{2}}, 1\right)$$

In other words $T \sim N(\Delta/\sqrt{2}, 1)$. This distribution is normal, but it is only *standard normal* under the null hypothesis that $\mu_X = \mu_Y$, i.e. $\Delta = 0$. Now, at

the 5% level we reject H_0 when $|T| >$ `qnorm(1 - 0.05/2)` ≈ 2. Combining this decision rule with the distribution of the test statistic under the alternative, we calculate power as follows:

$$\begin{aligned}
\text{Power}(\Delta) &= P(\text{Reject } H_0 | H_0 \text{ False}) = P(|T| > 2) \\
&= P(T < -2) + P(T > 2) \\
&= P(Z + \Delta/\sqrt{2} < -2) + P(Z + \Delta/\sqrt{2} > 2) \\
&= P(Z < -2 - \Delta/\sqrt{2}) + P(Z > 2 - \Delta/\sqrt{2}) \\
&= \texttt{pnorm}(-2 - \Delta/\sqrt{2}) + \left[1 - \texttt{pnorm}(2 - \Delta/\sqrt{2})\right]
\end{aligned}$$

(b) Evaluate the power formula you derived in part (a) using R by setting

```
delta <- seq(from = -10, to = 10, by = 0.01)
```

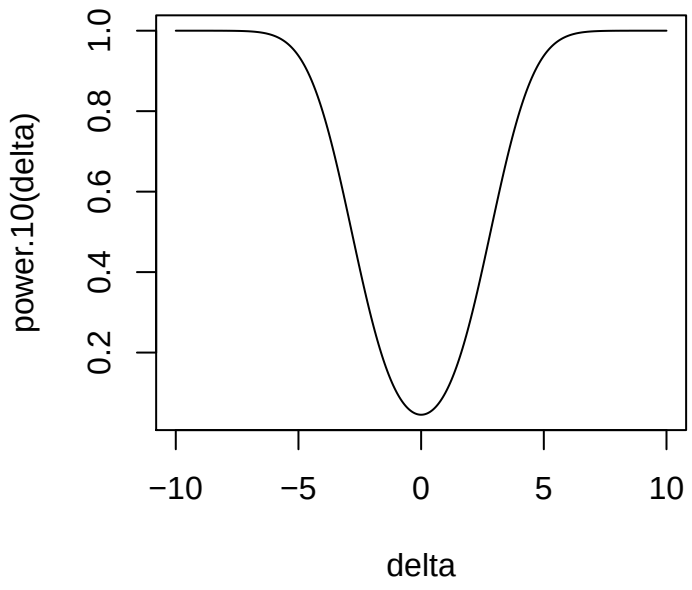
Plot your results. Approximately how large would the true difference of population means have to be for you to have at least a 50% chance of rejecting the null at the 5% level? What is the power when $\Delta = 0$? Why?

Solution: To make things easier, we can write the following function:

```
power.10 <- function(delta){
  greater <- 1 - pnorm(2 - delta/sqrt(2))
  less <- pnorm(-2 - delta/sqrt(2))
  power <- greater + less
  return(power)
}
```

Now, we can make the plot as follows:

```
delta <- seq(from = -10, to = 10, by = 0.01)
plot(delta, power.10(delta), type = 'l')
```



From this plot, it looks like we need Δ around 3 in absolute value for the power to be at least 0.5. (Notice that the curve is symmetric.) Let's try a few values:

```
power.10(3)
## [1] 0.5483002
```

So it looks like we can make Δ even smaller:

```
power.10(2.7)
## [1] 0.4638674
```

Those are too small. Let's try making it a little bigger

```
power.10(2.75)
## [1] 0.4779274
power.10(2.76)
## [1] 0.4807434
power.10(2.77)
## [1] 0.4835604
```

Notice that because of the symmetry:

```
power.10(-2.77)
## [1] 0.4835604
```

Thus, in order to have at least a 50% of rejecting the null with a sample size of 10, the true difference of means needs to be *at least* 2.77 in absolute value. This is quite a large difference relative to the population standard deviations!

Finally, let's evaluate the power when $\Delta = 0$:

```
power.10(0)
## [1] 0.04550026
```

We get 0.05, which is the significance level of the test! (The only reason it is not exactly equal to 0.05 is that we rounded the critical value for our test to 2. The reason this happens is that when $\Delta = 0$ the null is *true*. In hypothesis testing, we set everything up so the probability of rejecting a true null equals α.

(c) Repeat parts (a) and (b) with $n = m = 100$. How do your results change? Explain.

Solution: We can use the same procedure as above. The only difference is the standard error used in the test statistic. Whereas before this was $\sqrt{2}$, now we have:

$$SE = \sqrt{\sigma_X^2/n + \sigma_Y^2/m} = \sqrt{10/100 + 10/100} = \sqrt{1/5} = 1/\sqrt{5}$$

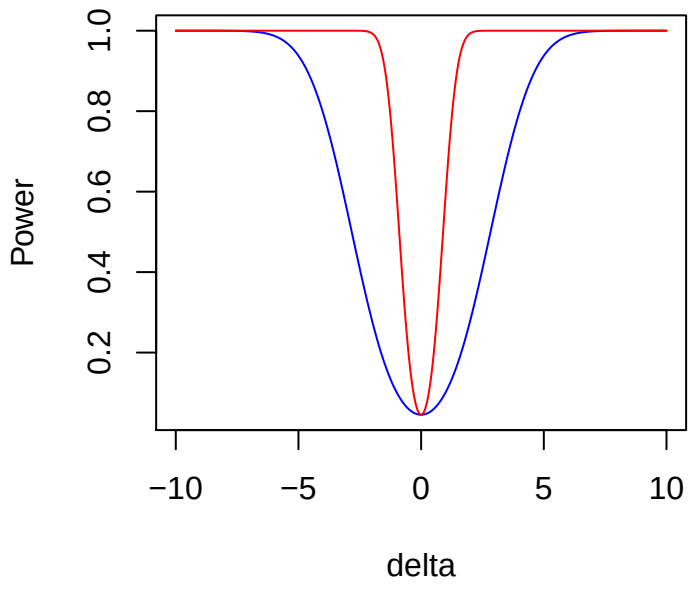
Thus, putting $1/\sqrt{5}$ in place of $\sqrt{2}$, we have

$$\text{Power}(\Delta) = \texttt{pnorm}(-2 - \Delta\sqrt{5}) + \left[1 - \texttt{pnorm}(2 - \Delta\sqrt{5})\right]$$

```
power.100 <- function(delta){
  greater <- 1 - pnorm(2 - delta * sqrt(5))
  less <- pnorm(-2 - delta * sqrt(5))
  power <- greater + less
  return(power)
}
```

We can compare this to the power function when $n = m = 10$ by plotting them on the same graph as follows:

```
y <- cbind(power.10(delta), power.100(delta))
matplot(delta, y, type = 'l', ylab = 'Power', col = c('blue', 'red'), lty = 1)
```



where the red curve is $n = m = 100$ and the blue curve is $n = m = 10$. We see that for any value of Δ, power is higher for the test based on a larger sample size: we are more likely to detect a difference of a given size using large samples. Experimenting with `power.100` as we did above for `power.10`,

```
power.100(1)
## [1] 0.5933214
power.100(0.9)
## [1] 0.5050012
power.100(0.89)
## [1] 0.4960838
power.100(0.88)
## [1] 0.4871686
```

So it looks like we need a difference between the two population means of at least 0.88 (in absolute value) to have at least a 50% chance of rejecting the null of no difference. Notice that this is a much smaller difference than we needed above (2.77) because of the increased sample size.

Finally, as above, if $\Delta = 0$ then the null is true, and the probability of rejecting a *true* null is simply α

```
power.100(0)
## [1] 0.04550026
```

Again, there's only a slight difference from the exact value of 0.05. This is because we rounded the critical value for our test to 2.

5. Professor Neil is interested in determining whether viewing different colors affects subjects' mental states in a way that alters their athletic ability. As a part of her research she carries out the following experiment. Each subject is randomly assigned to wait in one of two rooms: a room in which all of the walls have been painted pink or another in which all of the walls have been painted red. After waiting for five minutes, each subject is taken to a track and asked to run a 5K as fast as possible. Using the data from this experiment, Professor Neil carries out a statistical test of the null hypothesis that the population mean 5K time is equal across groups (those who waited in the pink room versus those who waited in the red room). Testing at the 1% level, she finds a statistically significant difference. For each of the following, answer True or False. If false, explain.

   (a) The p-value for the null hypothesis that population means are equal across groups is greater than 0.01.

   Solution: False: the p-value is *less than or equal to* 0.01.

   (b) Professor Neil would also have found a statistically significant difference had she carried out her test at the 5% level.

   Solution: True.

   (c) If there were really no difference in population means across the two groups, the chance of observing a test statistic at least as extreme as that observed by Professor Neil would be 0.01 or less.

   Solution: True.

   (d) Professor Neil's findings have important practical implications for sports regulatory organizations such as the International Olympic Committee: all locker rooms should be painted exactly the same color to keep from throwing off the outcomes of sporting events.

Solution: False. Professor Neil has found strong evidence of a difference in population means across the two groups. However, none of the information given above provides any indication of whether this difference is large enough to have any practical importance. In the words of the textbook "statistical significance and practical significance are two entirely different matters." For example, the difference of population means could be one second. This is far too small to be likely to change the outcome of a 5K race, but with a large enough sample size we would still be able to detect it. Unlike a confidence interval, which gives us a range of plausible values for the difference in population means, a hypothesis test merely tells us whether we have strong evidence that a difference exists.