Solution Set for Problem Set #2

W241: Field Experiments

1. *FE exercise 3.6*

Note: due to random noise in the simulations, your exact answers will likely differ slightly.

```
> ## FE 3.6 ##

> #DATA PROCESSING
> #Read in data
> data <- read.csv('ex3_6_data.csv')

> #ESTIMATION
> #Estimate the ATE
> est.ate <- function(outcomes, treatment) mean(outcomes[treatment==1]) -
mean(outcomes[treatment==0])
>
> ate <- est.ate(data$rating, data$treatment)
>
>
> #Functions to conduct randomization inference
> #Full random assignment
> assign.treatment <- function(n) sample(0:1, n, replace=TRUE)
> ri.sim.once <- function(outcomes) est.ate(outcomes,
assign.treatment(length(outcomes)))
>
> distribution.under.sharp.null <- replicate(10000, ri.sim.once(data$rating))
>
> # how many bigger?
> n.bigger <- sum(distribution.under.sharp.null >= ate)
> n.bigger
[1] 17
>
> # one-tailed p-value estimate
> n.bigger / length(distribution.under.sharp.null)
[1] 0.0017
>
> # how many are bigger in absolute value terms?
> n.bigger.abs <- sum(abs(distribution.under.sharp.null) >= abs(ate))
> n.bigger.abs
[1] 26
>
> # two-tailed p-value estimate
> n.bigger.abs / length(distribution.under.sharp.null)
```

```
[1] 0.0026
```

2. *FE exercise 3.7*

```
> y0 <- c(1,0,0,4,3)
> y1 <- c(2,11,14,0,3) - 2
> all.outcomes <- c(y0, y1)
> n <- length(all.outcomes)
> all.assignments <- combn(1:n, length(y1))
> all.assignments <- apply(all.assignments, 2, function(x){
+     zs <- rep(0,n)
+     zs[x] <- 1
+     return(zs)
+     })
> est.ate <- function(treatment, outcomes) mean(outcomes[treatment==1]) -
mean(outcomes[treatment==0])
> dist.under.sharp.null <- apply(all.assignments, 2, est.ate, all.outcomes)
> real.ate <- est.ate( c(rep(0,length(y0)), rep(1,length(y1))),all.outcomes)
> real.ate
[1] 2.4
> mean(dist.under.sharp.null >= real.ate)
[1] 0.2380952
```

We want to test the positive claim that the treatment program causes participants to lose at least two pounds more than they otherwise would have during the first two weeks. How likely is it that we would have received an answer as large or larger than the estimated answer? 23.8%. Therefore, this experiment constitues weak evidence that weight loss is 2 pounds or more.

Note that this is a one-tailed test, as we are only interested in the claim that participants *lose* at least seven pounds, and not how likely it is we would have observed that their weight changes by at least 2 pounds in an absolute sense, either + or - (gain *or* loss of 2).

3. *FE exercise 3.8*

```
> #Helper function.
> calc_est_ate <- function(group, outcome){
+    return(mean(outcome[group==1]) - mean(outcome[group==0]))
+ }
>
> # (a)
> est_ate <- c(
+   ar = with(subset(p38_df, st == 'ar'), calc_est_ate(group, bills)),
+   tx = with(subset(p38_df, st == 'tx'), calc_est_ate(group, bills))
+   )
> est_ate
```

```
       ar        tx
-10.95139 -16.74167

> # (b)
> calc_est_se <- function(group, outcome){
+   y0 <- outcome[group==0]
+   y1 <- outcome[group==1]
+   m <- length(y1)
+   N <- m + length(y0)
+   v0 <- var(y0)
+   v1 <- var(y1)
+   return(sqrt(v0/(N-m) + v1/m))
+ }
> est_se <- c(
+   ar = with(subset(p38_df, st == 'ar'), calc_est_se(group, bills)),
+   tx = with(subset(p38_df, st == 'tx'), calc_est_se(group, bills))
+   )
> est_se
      ar       tx
3.395979 9.345871

> # (c)
> n_st <- c(
+   ar=sum(p38_df$st == 'ar'),
+   tx=sum(p38_df$st == 'tx')
+   )
>
> calc_block_ate <- function(ates, ns){
+   return(sum(ates * ns) / sum(ns))
+ }
>
> block_ate <- calc_block_ate(est_ate, n_st)
> block_ate
[1] -13.2168
```

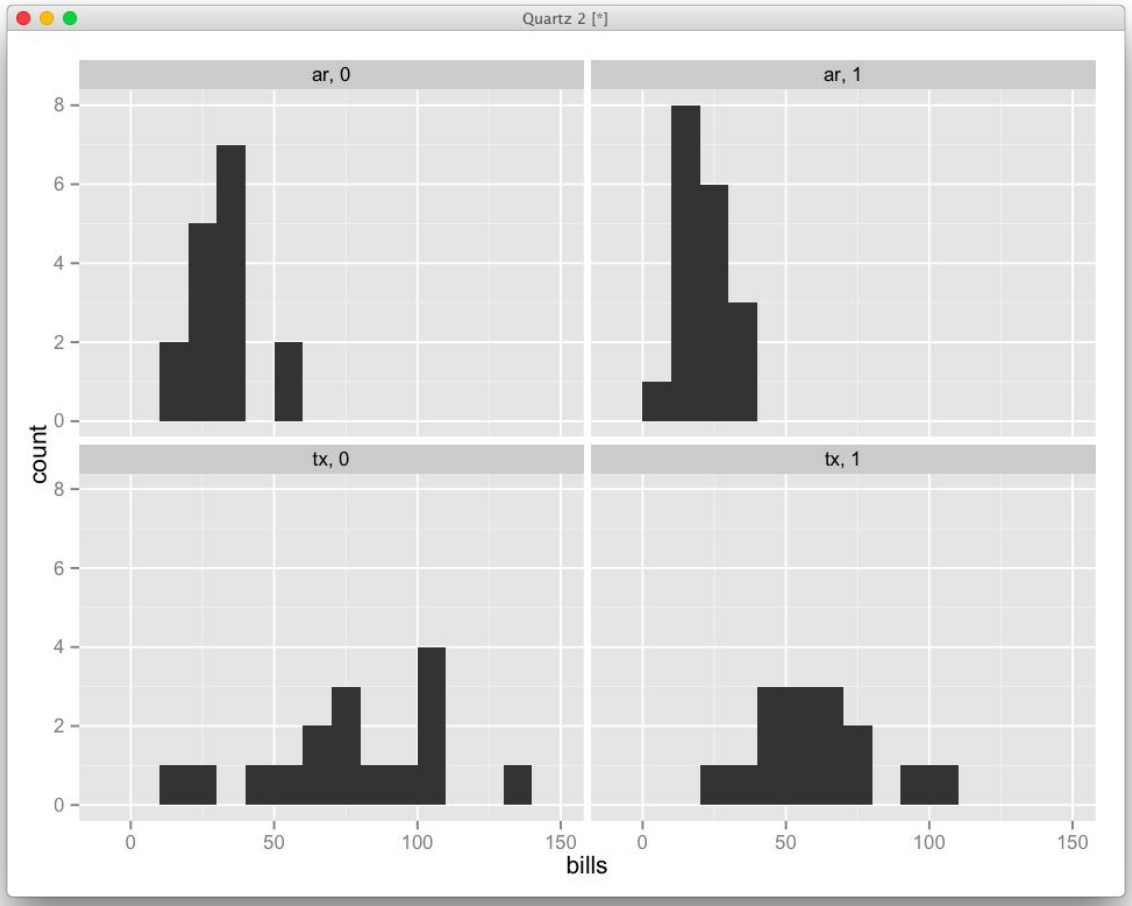
d) The treatment assignment probabilities are different in Arkansas and Texas, such that the treatment group overrepresents Arkansas. Therefore, if outcomes were higher in the treatment group, it might reflect differences between Texas and Arkansas rather than an effect of the treatment.

```
> #(e)
> total_ate_se <- sqrt(
+   est_se['ar']^2 * n_st['ar'] / nrow(p38_df) +
+     est_se['tx']^2 * n_st['tx'] / nrow(p38_df)
+   )
> as.numeric(total_ate_se)
[1] 4.74478
```

```
> # (f)
> # Need to randomize by block.
>
> # Function to assign treatment within a block, given a number in control and
a number in treatment.
> assign_treat <- function(n0, n1){
+   return(sample(c(rep(0,n0),rep(1,n1))))
+ }
>
> # Run one simulated random assignment and calculate the ate.
> run_block_sim <- function(block, group, outcome){
+     n_bg = table(block, group) #count number in treatment and control by
block
+     ates <- c()
+     ns <- c()
+     # Calculate the ATE within each block.
+     for(blocki in unique(block)){
+           assign <- assign_treat(n_bg[blocki,'0'], n_bg[blocki,'1'])
+           ates[blocki] <- calc_est_ate(assign, outcome[block == blocki])
+           ns[blocki] <- sum(n_bg[blocki,])
+     }
+     # Pool the ATEs, with the function made earlier.
+     ate <- calc_block_ate(ates, ns)
+     return(ate)
+ }
>
> block <- p38_df$st
> group <- p38_df$group
> outcome <- p38_df$bills
>
> iter = 10000
> rand_ate <- replicate(iter, run_block_sim(block, group, outcome))
> sum(abs(rand_ate >) abs(block_ate)) / iter
[1] 0.0038


> # Histograms
> library(ggplot2)
> ggplot(p38_df, aes(x=bills)) + geom_histogram(binwidth=10) +
facet_wrap(st~group)
```

In addition to the problems in the book, plot histograms for both the treatment and control groups in each state (for 4 histograms in total).



4. *FE exercise 3.11*
    a. *Note: assume 3 clusters in treatment and 4 in control.*
    b. *Note: when Gerber and Green say "simulate", they do not mean "run simulations with R code", but rather, in a casual sense "take a look at what happens if you do this this way." There is no randomization inference necessary to complete this problem.*

```
> ## FE 3.8 ##
>
> calc_cluster_se <-function(y0,y1,m,cluster){
+   k <- length(unique(cluster))
+   N <- length(y0)
+
+   cluster_means <- aggregate(data.frame(y0=y0,
y1=y1),list(cluster=cluster),mean)
+
+   sigma <- cov(cluster_means[,c('y0','y1')])
```

```
+   vary0 <- sigma[1,1]
+   vary1 <- sigma[2,2]
+   cov01 <- sigma[1,2]
+
+   # Equation 3.22.
+   return(sqrt( 1/(k-1) * ( m/(N-m)*vary0 + (N-m)/m*vary1 + 2*cov01)) )
+ }
>
> y0 <- c(0,1,2,4,4,6,6,9,14,15,16,16,17,18)
> y1 <- c(0,0,1,2,0,0,2,3,12,9,8,15,5,17)
>
> # (a)
> cluster <- rep(1:7,each=2)
> calc_cluster_se(y0,y1,m=3*2,cluster)
[1] 4.918953

> # (b)
> cluster <- c(1:7,7:1)
> calc_cluster_se(y0,y1,m=3*2,cluster)
[1] 1.264924
```

Note: some of you realized that R calculates variance and covariance slightly differently than the formulas in Gerber and Green (R uses n-1 instead of n since it is assuming a sample). These answers are perhaps *more* correct, so we have also accepted them.

c) The variation between clusters was extremely high in the first clustering scheme. In other words, in the first scheme, the cluster was highly predictive of the level of potential outcomes. This means that one very high or low cluster being included in the treatment or control group had a large effect on the ATE estimate, resulting in a high standard error because the ATE is likely to differ by a great deal from assignment to assignment. In the second scheme, the average composition of each cluster was more similar across clusters, meaning there will be less variation in the ATE from assignment to assignment. The lesson for experimental design is that you ideally want your cluster means to be similar, and that experiments where clusters differ dramatically will have much higher standard errors.

5. *You are an employee of a newspaper and are planning an experiment to demonstrate to Apple that online advertising on your site causes people to buy iPhones. Each user shown the ad campaign is exposed to $0.10 worth of advertising for iPhones. There are 1,000,000 users available to be shown ads on your newspaper's website during the one week campaign. Apple indicates that they make a profit of $100 every time an iPhone sells and that 0.5% of visitors to your newspaper's website buy an iPhone in a given week in general.*

a. *By how much does the ad campaign need to increase the probability of purchase in order to be "worth it" and a positive ROI (supposing there are no long-run effects and all the effects are measured within that week)?*

A user seeing an ad needs to increase profitability by at least $0.10 for the ad campaign to be worth the cost. If Apple makes $100 on each iPhone, an increased probability of sale of 0.1% is worth an increase in $0.10. So each ad needs to increase the probability of purchase by 0.1% on average for the ad campaign to be "worth it."

b. *Assume the measured effect is 0.2%. If users are split 50:50 between the treatment and control groups, what will be the confidence interval of your estimate on whether people purchase the phone? (Use the standard formula for two-sample proportion tests, which you can find here.)*

```
> ### (b)
> N = 1000000
> p1 = .005
> p2 = .007  # Baseline of .005 + .002 measured treatment effect.
>
> calc_ci <- function(p1,p2,n1,n2){
+   x1 = p1*n1
+   x2 = p2*n2
+   xbar = p1-p2
+   p = (x1+x2)/(n1+n2)
+   se = sqrt(p*(1-p)*(1/n1 + 1/n2))
+   ci = c(xbar - se*1.96, xbar + se*1.96)
+   return(ci)
+ }
>
> n1 = N*.5
> n2 = N*.5
>
> ci1 <- calc_ci(p1,p2,n1,n2)
> ci1
[1] 0.00169727 0.00230273
```

The 95% confidence interval is (0.1697%, 0.2303%).

c. *Is this confidence interval precise enough that you would recommend running this experiment? Why or why not?*

We noted in part (a) that an effect of 0.1% would be sufficient for the campaign to be profitable. The lower tail of the 95% confidence interval in part (b) is larger than this value, meaning that Apple should look at this result and be very confident that the ad campaign is profitable.

Therefore, as the newspaper attempting to argue that our ads work, we would recommend running this experiment.

d. *Your boss at the newspaper, worried about potential loss of revenue, says he is not willing to hold back a control group any larger than 1% of users. What would be the width of the confidence interval for this experiment if only 1% of users were placed in the control group?*

```
> # (d)
> n1 = 0.99*N
> n2 = 0.01*N
> ci2 <- calc_ci(p1,p2,n1,n2)
> ci2
[1] 0.000359995 0.003640005
> ci2[2] - ci2[1]
[1] 0.00328001
```

The 95% confidence interval nearly includes 0 (the ads don't do anything) and certainly includes the break-even point of 0.1% or 0.001. We cannot reliably distinguish between a campaign that is very unprofitable and a campaign that is wildly profitable.

6. *Here you will find a set of data from an auction experiment by John List and David Lucking-Reiley (2000). In this experiment, the experimenters invited consumers at a sportscard trading show to bid against one other bidder for a pair trading cards. We abstract from the multi-unit-auction details here, and simply state that the treatment auction format was theoretically predicted to produce lower bids than the control auction format. We provide you a relevant subset of data from the experiment.*
    a. *Compute a 95% confidence interval for a difference between two means.*

```
> # a
> treat.outcomes <- data$bid[data$uniform_price_auction==1]
> control.outcomes <- data$bid[data$uniform_price_auction==0]
> var_treat <- var(treat.outcomes)
> var_control <- var(control.outcomes)
> n_treat <- length(treat.outcomes)
> n_control <- length(control.outcomes)
> SE <- sqrt(var_treat/n_treat + var_control/n_control)
> SE
[1] 4.326572
> ATE <- mean(treat.outcomes) - mean(control.outcomes)
>
> lower <- ATE - 1.96 * SE
> upper <- ATE + 1.96 * SE
>
```

```
> c(lower,upper)
[1] -20.685963  -3.725801
```

b. In plain language, what does this confidence interval mean?

This confidence interval indicates there is a 95% probability that the true ATE falls within this range.

c. Compute the confidence interval using regression.

```
> reg_summary <- summary(lm(bid~uniform_price_auction, data))
> reg_summary

Call:
lm(formula = bid ~ uniform_price_auction, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-28.824 -11.618  -3.221   8.382  58.382

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)            28.824      3.059   9.421 7.81e-14 ***
uniform_price_auction  -12.206      4.327  -2.821  0.00631 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 17.84 on 66 degrees of freedom
Multiple R-squared:  0.1076,   Adjusted R-squared:  0.09409
F-statistic: 7.959 on 1 and 66 DF,  p-value: 0.006315
> ate <- reg_summary$coefficients[2,1]
> se <- reg_summary$coefficients[2,2]
>
> ci <- c(ate - 1.96 * se, ate + 1.96 * se)
> ci
[1] -20.685963  -3.725801
```

d. Compute the p-value using regression.

```
> # p-value with regression
> p.reg.two.tailed <- reg_summary$coef[2,4]
> p.reg.two.tailed
[1] 0.006314796
```

e. Compute the p-value using randomization inference.

```
> # p-value with randomization inference
> assign.treatment <- function(n0, n1) return(sample(c(rep(0,n0),rep(1,n1))))
> est.ate <- function(outcomes, treatment) mean(outcomes[treatment==1]) -
mean(outcomes[treatment==0])
> ri.sim.once <- function(outcomes){
+     simulated.treatment <-
assign.treatment(sum(auction_df$uniform_price_auction==0),
sum(auction_df$uniform_price_auction==1))
+     return(est.ate(outcomes, simulated.treatment))
+ }
> distribution.under.sharp.null <- replicate(10000,
ri.sim.once(auction_df$bid))
>
> actual.ate <- est.ate(auction_df$bid, auction_df$uniform_price_auction)
>
> p.rand.two.tailed <- mean(abs(actual.ate) <=
abs(distribution.under.sharp.null))

> p.rand.two.tailed
[1] 0.0059
```

f. Compute the same p-value again using analytic formulas for a two-sample t-test from your earlier statistics course. (Also see part (a).)

```
> # f
> t <- ATE / SE
> pnorm(t)
[1] 0.002392636
```

g. Compare the two p-values. Are they much different? Why or why not? How might your answer to this question change if the sample size were different?

The p-values were not much different, as is usually the case when samples are larger than about 30 observations. If the sample size were smaller, then we might expect a bigger difference in the two methods of calculating p-values.