Assignment2

Callin Switzer

September 16, 2014

1. (5 points) Describe a situation in which you had to make a choice, and there was a correct answer that was not available as you made your choice (e.g., "take an umbrella or not", "walk or take a T", etc.). Think about what information you used to help you decide, and then briefly answer the questions below:

(a) What was the decision about? Identify a parameter of interest that you based your decision on.

Once, I had to decide to plan a lesson or go to sleep (thereby not planning my lesson for the next day.) The parameter of interest was my success in teaching the following day.

(b) What were your H0 and H1 about this parameter?

H0: Going to sleep will help me teach better than planning my lesson H1: Planning my lesson will help me teach better than going to sleep

(c) Based on what quantitative or qualitative information were you deciding? *I was basing my my decision qualitative data from my own experience. I generally feel much better when I've had sleep, and am able to think quickly enough to teach acceptably without planning my lesson precisely. However, I was also basing my decision on my qualitative experiences that when my lessons are planned well, I teach better.*

(d) What was your guess of the p-value? Did you reject your H0? *My guess of the p-value was 0.03. I rejected my null hypothesis, and decided to write the lesson plan before going to sleep*

(e) What were the possible consequences of making the wrong decision?

The possible consequences of making the wrong decision were the following:

- I would be embarassed in front of students for doing a bad job at teaching.
- My students wouldn't learn enough material, and my principal would think I was a bad teacher.
- My students wouldn't do well on standardized tests, and would not get scholarships for college.
- I would get fired from my job as a teacher.
- I could make lots of mistakes during my class

2. (10 points) Exercise 17 in Chapter 1 (2nd or 3rd ed.).

```
## [1] "The two-sided p-value is  0.0857"
```

	Group A	Group A	Group A	Group A	Group B	Group B	Group B	DiffBwSamples
1	68	77	82	85	53	64	71	15.33
2	68	77	82	53	85	64	71	-3.33
3	68	77	82	64	85	53	71	3.08
4	68	77	82	71	85	53	64	7.17
5	68	77	85	53	82	64	71	-1.58
6	68	77	85	64	82	53	71	4.83
7	68	77	85	71	82	53	64	8.92
8	68	77	53	64	82	85	71	-13.83
9	68	77	53	71	82	85	64	-9.75
10	68	77	64	71	82	85	53	-3.33
11	68	82	85	53	77	64	71	1.33
12	68	82	85	64	77	53	71	7.75
13	68	82	85	71	77	53	64	11.83
14	68	82	53	64	77	85	71	-10.92
15	68	82	53	71	77	85	64	-6.83
16	68	82	64	71	77	85	53	-0.42
17	68	85	53	64	77	82	71	-9.17
18	68	85	53	71	77	82	64	-5.08
19	68	85	64	71	77	82	53	1.33
20	68	53	64	71	77	82	85	-17.33
21	77	82	85	53	68	64	71	6.58
22	77	82	85	64	68	53	71	13.00
23	77	82	85	71	68	53	64	17.08
24	77	82	53	64	68	85	71	-5.67
25	77	82	53	71	68	85	64	-1.58
26	77	82	64	71	68	85	53	4.83
27	77	85	53	64	68	82	71	-3.92
28	77	85	53	71	68	82	64	0.17
29	77	85	64	71	68	82	53	6.58
30	77	53	64	71	68	82	85	-12.08
31	82	85	53	64	68	77	71	-1.00
32	82	85	53	71	68	77	64	3.08
33	82	85	64	71	68	77	53	9.50
34	82	53	64	71	68	77	85	-9.17
35	85	53	64	71	68	77	82	-7.42

Table 1: Possible ways to randomize

##

3.(20 points) Environmental Voting of Democrats and Republicans in the U.S. House of Representatives. Use the data provided in Problem 26, Chapter 1 (3rd ed.) to answer the following questions. Note that all calculations have to be performed in R. (First two chapters of the 3rd edition have been scanned and uploaded on the course web-site. The data file has been posted online as well.)

(a) Which test is more appropriate to address the research question in this problem, permutation test or randomization test? *Research Question: Do the different parties show differences in the percentage of pro-environment votes?*

A permutation test is more appropriate to address the question, b/c these data are observed

(b) Based on your answer in (a), set up corresponding null and alternative hypotheses.

H0: The difference in the means of PctPro for republicans and democrats is 0

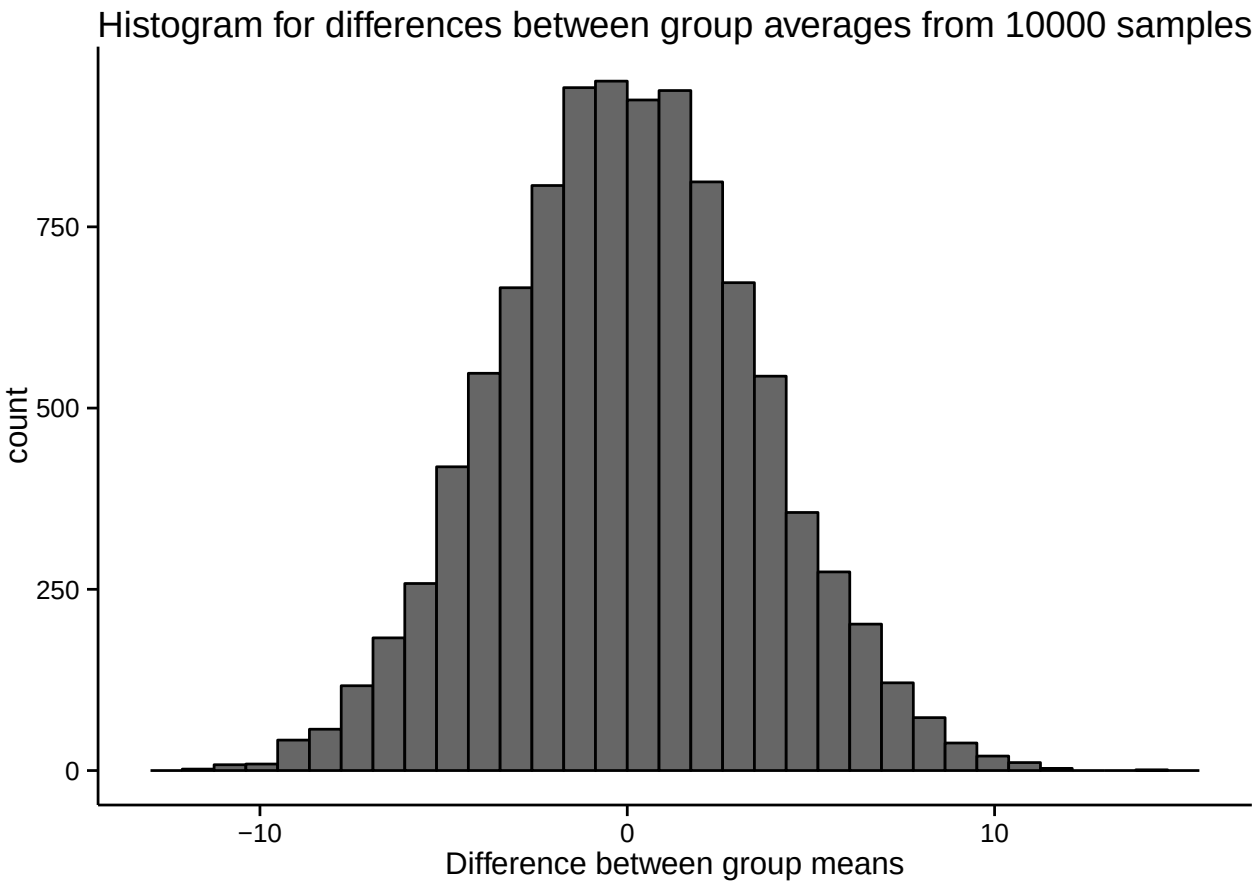
HA: The difference in the means of PctPro for republicans and democrats is NOT 0

(c) Specify the test statistic that you are using and calculate its observed value.

The test statistic is mean(PctPro where Party == "R") - mean(PctPro where Party == "D")

```
The test statistic is -69.73
```

(d) Draw the reference distribution of your test statistic based on 10,000 simulations. You may reuse the R code given here in Section 3.3. by adapting it to this problem.



dist.pdf

(e) What is your p-value?

My approximate p-value is 0. No values from my 10,000 sample are equal or as extreme as the test statistic

(f) Comment on assumptions of the chosen test *The assumption of the permutation test is independence of study units. This may not be valid, because party status is dependent upon where in the US the people live. For instance, if you have two people from the same part of the US, they are probably going to vote the same way.*

(g) Write a brief summary of your findings and comment on the scope of inference

These data suggest that we can reject the null hypothesis – if we reject, that means that we found evidence that party affiliation is associated with the percentage of environmentally friendly votes.

Since the selection of units was not random, and the allocation of units to groups was not random, our scope of inference is quite small. We can neither draw causal inferences, nor can we draw inferences to the population. We're limited to describing the subpopulation we sampled.

CODE

```
################################################
####### Q2
################################################

# setup
require("gtools")
require("mgcv")
require("xtable")

# input data
scores <- c(68, 77, 82, 85, 53, 64, 71)
trt <- c(rep("A", 4), rep("B", 3))
actDiff <- mean(scores[1:4]) - mean(scores[5:7])
choose(7,4) # there are 35 ways to choose 4 out of 7 students to be in group A

# get a combination matrix for 7 choose 4
cho <- t(combn(7,4))

# define empty matrix for putting the values into
possible <- matrix(nrow = 35, ncol = 7)

# variable for calculating the differences
diff <- numeric(35)

# loop over all rows in the matrix
for (i in 1:nrow(cho)){
     nums <- cho[i,]
     possible[i, 1:4] <- scores[nums]
     possible[i, 5:7] <- scores[-nums]
     diff[i] <- mean(scores[nums]) - mean(scores[-nums])
}

# make a nice-looking dataframe to print
posDat <- data.frame(possible, diff)
names(posDat) <- c(rep("Group A", 4), rep("Group B", 3), "DiffBwSamples")

# print the dataframe
options(xtable.comment = FALSE)
print(xtable(posDat, caption = "Possible ways to randomize", digits = c(rep(0, 8), 2)), type = "latex")

# calculate pvalue, and print it
pval = sum(abs(diff) >= actDiff) / length(diff)
print(paste("The two-sided p-value is ", round(pval, 4)))


################################################
####### Q3
################################################

# read in data
ev <- read.csv("data/ex0126.csv")
```

```r
# get rid of independent and no-party data
ev <- ev[ev$Party == "R" | ev$Party == "D", ]

# make separate vectors for each party
rep <- ev$PctPro[ev$Party == "R"]
dem <- ev$PctPro[ev$Party == "D"]

# calculate test statistic
testStat <- round(mean(rep) - mean(dem), 2)

# prepare for histogram and simulations
require(ggplot2)

nsim = 10000 # number of simulations
ndem <- length(dem) # number of democrats
nrep <- length(rep) # of republicans

# create a vector of only the percentagePro
pctP <- ev$PctPro

# resample nsim times, and calculate differences between group means
ts <- numeric(nsim)
for(i in 1:nsim){
     foo <- sample(1:length(pctP), ndem, replace = F)
     democ <- mean(pctP[foo])
     repub <- mean(pctP[-foo])
     ts[i] <- repub - democ
}

# make a data frame for ggplot
ts <- data.frame(ts)

# make the histogram
bb <- ggplot(ts) +
     geom_histogram(aes(ts), fill = "grey40", color = "black") +
     theme_bw() +
     labs(x = "Difference between group means") +
     ggtitle("Histogram for differences between group averages from 10000 samples") +
     theme(plot.background = element_blank()
          ,panel.grid.major = element_blank()
          ,panel.grid.minor = element_blank()
          ,panel.border = element_blank()) +
     #draws x and y axis line
     theme(axis.line = element_line(color = 'black'))
# print the plot
bb

#calculate p-value
pva <- sum(ts <= testStat | ts >= -testStat)
```