Homework #1 Solutions

***Note: You earned 8 points just for turning in the assignment!***

## Question 1 (20 points total)

What is the difference between a randomized experiment and a random sample? Under what type of study/sample can a causal inference be made?

**In a randomized experiment, the groups in the study are allocated to each treatment via a chance mechanism. In a random sample, the units/subjects in a study are included in the study from some population via a chance mechanism.**

**Causal inferences can be made only with a randomized experiment.**

*Note: your solution does not need to be word-for-word the same as the above. For example, for the first question you could have phrased the first part as “in a randomized experiment, subjects in the study are randomly assigned to groups.”*

## Question 2 (20 points total)

In 1936, the *Literary Digest* polled 1 out of every 4 Americans and concluded that Alfred Landon would win the presidential election in a landon-slide. Of course, history turned out dramatically different (see [link](http://historymatters.gmu.edu/d/5168/) for further details). The magazine combined three sampling sources: subscribers to its magazine, phone number records, and automobile registration records. Comment on the desired population of interest of the survey and what population the magazine actually drew from.

**The population of interest for this study was intended to be all the people who were going to vote in the 1936 election (not even necessarily all people in the country, as many people don’t vote). The population that the study actually referenced consisted of more affluent people (based on appealing to those with magazine subscriptions and cars during one of the lowest points in the great depression). As FDR’s “New Deal” and other social programs were much more attractive to lower income people, they were not included in anything resembling representative numbers.**

*Note: To get full credit, you must correctly identify the desired population AND what they actually ended up with. If you miss one, around 5 points should be subtracted. There is no one correct answer for the discussion - the one given is an example.*

## Question 3 (30 points total)

Suppose we have developed a new fertilizer that is supposed to help corn yields. This fertilizer is so potent that a small vial of it sprayed over an entire field is a sufficient dose. We find that the new fertilizer results in an average yield of 60 more bushels over the old fertilizer with a p-value of 0.0001. Write up a scope of inference under the following study designs that generated this data.

*Note: To get full credit, you must correctly discuss whether the results can be extended to any broader population (4 points) and whether causal inferences can be made (4 points). This makes each sub-question worth a total of 8 points.*

### Part A

We offer the new fertilizer at a discount to customers who have purchased the old fertilizer along with a survey for them to fill out. Some farmers send in a survey after the growing season, reporting their crop yield. From our records, we know which of these farmers used the new fertilizer and which used the old one.

**As the farmers self-select into the study (the ones who return the survey are considered the sample population), extending the results to any other group or population (such as all the customers of the company who bought fertilizer) is speculative. Also, we did not randomize the assignment of “old” and “new” fertilizer, so no causal inference can be made.**

### Part B

When a customer makes an order, we randomly send them either the old or new fertilizer. At the end of the season, some of the farmers send us a report of their yield. Again, from our records, we know which of these farmers used the new fertilizer and which used the old.

**As the farmers self-select into the study (we have no idea why some farmers send in their report and some did not, making a random sample of fertilizer customers highly unlikely, it was most likely self-selection), extending the results to any other group or population (such as all the customers of the company who bought fertilizer) is speculative and is not appropriate. This is a randomized experiment, as we randomly assign the farmers to the two fertilizers. Hence, causal inferences can be made - but again, only to the particular farmers that sent us their yields. However, because only some of the farmers send their yields (after treatment has been applied), this could introduce a confounding variable into the study via nonresponse bias (a form of selection bias). This could be a cause for concern for the validity of the entire study.**

### Part C

When a customer makes an order, we randomly send them either the old or new fertilizer. At the end of the season, we sub-select from the fertilizer orders and send a team out to count those farmers’ crop yields.

**It’s unclear whether the “sub-selection” is done randomly. If it is assumed that the sub-selection process from all fertilizer customers was random, then we can extend our results to all farmers that ordered fertilizers this year. If the sub-selection was not random (e.g., perhaps the people travelling to the fields wanted to minimize travel time and chose only farmers with fields nearby), we cannot generalize the results to a population any broader than those whose yields were collected. This is a randomized experiment as we randomly assigned the farmers to the two fertilizers. Hence, causal inferences can be made.**

### Part D

We offer the new fertilizer at a discount to customers who have purchased the old fertilizer. At the end of the season, we sub-select from the fertilizer orders and send a team out to count those farmer’s crop yields. From our records, we know which of these farmers used the new fertilizer and which used the old one.

**It’s unclear whether the “sub-selection” is done randomly. If it is assumed that the sub-selection process from all fertilizer customers was random, then we can extend our results to all farmers that ordered fertilizers this year. If the sub-selection was not random (e.g., perhaps the people travelling to the fields wanted to minimize travel time and chose only farmers with fields nearby), we cannot generalize the results to a population any broader than those whose yields were collected. We did not randomize the assignment to “old” and “new” fertilizer, though, so no causal inference can be made.**

## Question 4 (30 points total)

A Business Stats class here at SMU was polled and student were asked them how much money (cash) they had in their pocket at that very moment. The idea was that we wanted to see if there was evidence that those in charge of the vending machines should include the expensive bill/coin acceptor or if it should just have the credit card reader. We asked a professor from Seattle University last year to poll her class with the same question. Below are the results of our polls.

**SMU** > 34, 1200, 23, 50, 60, 50, 0, 0, 30, 89, 0, 300, 400, 20, 10, 0 **Seattle U** > 20, 10, 5, 0, 30, 50, 0, 100, 110, 0, 40, 10, 3, 0

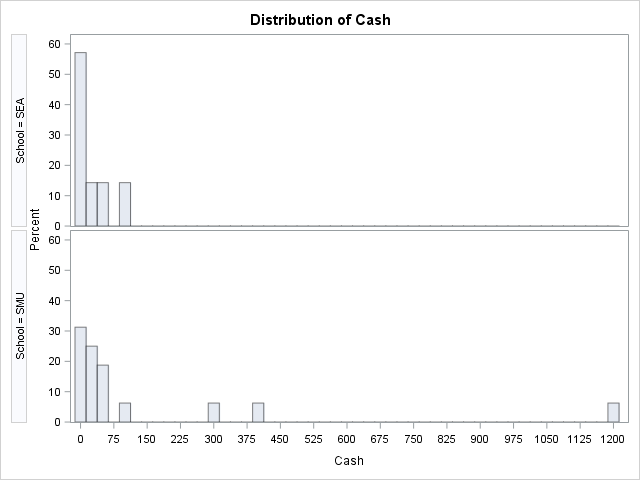
*Note: The cash in student pockets may be a poor proxy for whether they need to install the coin accepter. A more rigorous study might have also noted whether these students had credit cards with them. However, you have just started your statistics education and we are starting with easy examples (comparing pocket change for two groups).*

### Part A

Use SAS to make a histogram of the amount of money in a student’s pocket from each school. Does it appear there is any difference in **population means**? What evidence do you have? Discuss your thoughts. Note, proc univariate in SAS produces histograms. The code is given below. You will need to replace the code with the variables you input in a data step.

*There are multiple ways to get these histograms in SAS. If you used a different method, that is fine… provided that the shapes are the same. As long as the plot is correct, you should receive nearly full credit for the discussion.*

My SAS code:  
proc univariate data=problem4;  
class school;  
histogram cash;  
run;



**It is tough to tell given the scale of the histograms. There may be some evidence of a difference in population means. There appears to be an extreme value at SMU, $1200. Realistically, we should check if this is an error (if we could). Perhaps someone wrote 1200 instead of 12.00. It also may not be relevant to our question of interest in some way. We may wish to perform analyses with and without the extreme value.**

**We conduct a formal permutation test next to test for a difference in means.**

### Part B

Use the following R code to reproduce your histograms. Simply cut and paste the histograms into your HW.

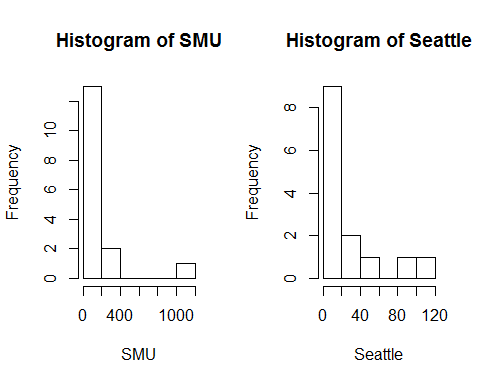
***SMU = c(34, 1200, 23, 50, 60, 50, 0, 0, 30, 89, 0, 300, 400, 20, 10, 0)***

***Seattle = c(20, 10, 5, 0, 30, 50, 0, 100, 110, 0, 40, 10, 3, 0)***

***hist(SMU)***

***hist(Seattle)***

SMU = c(34, 1200, 23, 50, 60, 50, 0, 0, 30, 89, 0, 300, 400, 20, 10, 0)  
Seattle = c(20, 10, 5, 0, 30, 50, 0, 100, 110, 0, 40, 10, 3, 0)  
par(mfrow=c(1,2))  
hist(SMU)  
hist(Seattle)

The added code line par(mfrow=c(1,2)) puts the graphs next to each other. It essentially creates a matrix of graphs using matrix notation. Keep this in mind if you want to see a lot of graphs at once. Also note that the scales are not the same (unlike the SAS plots); the SMU histogram spans a much larger range than that of Seattle.

### Part C

Run a permutation test to test if the mean amount of pocket cash from students at SMU is different than that of students from Seattle University. Write up a statistical conclusion and scope of inference (similar to the one from the PowerPoint). (This should include identifying the Ho and Ha as well as the p-value.) (Note, you will have to replace the data in the SAS code given for the live session questions. Make sure the group variable is coded as 0,1)

*Note: Your answers will vary. As long as you’ve used code similar to that below, your answer should be within a few decimal points. This can be done in SAS or R.*

*Note, if a student runs ONLY a t-test, instead of a permutation test,* ***NO*** *credit should be awarded.*

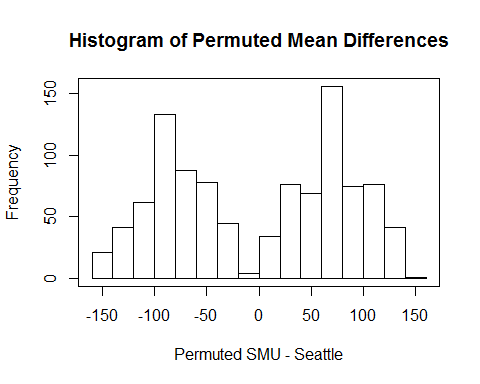
*Note: The point breakdown is as follows: 5 points for running the test, 5 points for getting the correct p-value (which, in all likelihood, should be between 0.1 and 0.2), and 5 points for writing up a conclusion that includes stating the null and alternative hypothesis, as well as drawing the correct conclusion (i.e., there is not sufficient evidence that a difference exists).*

*Note: The t-tests are not truly necessary here. In SAS, it is easy to use a t-test to find the difference in sample means, and then use that value for the permutation test code. In the R code below, a few lines of code were used to find the difference in sample means, create a variable holding that value, and then use that value in the permutation test code. By performing a t-test in R (showing the two-sample means), the difference can be calculated to make sure it matches up with the “observed difference” variable we calculated via coding. (It’s always nice to have a way to check that our code works the way we intend!)*

school1 <- rep('SMU', 16)  
school2 <- rep('Seattle', 14)  
school <- as.factor(c(school1, school2))  
all.money <- data.frame(name=school, money=c(SMU, Seattle))  
  
t.test(money ~ name, data=all.money)

##   
## Welch Two Sample t-test  
##   
## data: money by name  
## t = -1.4945, df = 15.499, p-value = 0.1551  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -277.64481 48.39481  
## sample estimates:  
## mean in group Seattle mean in group SMU   
## 27.000 141.625

number\_of\_permutations <- 1000  
xbarholder <- numeric(0)  
counter <- 0  
observed\_diff <- mean(subset(all.money, name == "SMU")$money)-mean(subset(all.money, name == "Seattle")$money)  
  
set.seed(123)  
for(i in 1:number\_of\_permutations)  
{  
scramble <- sample(all.money$money, 30)  
smu <- scramble[1:16]  
seattle <- scramble[17:30]  
diff <- mean(smu)-mean(seattle)  
xbarholder[i] <- diff  
if(abs(diff) > abs(observed\_diff))  
counter <- counter + 1  
}  
hist(xbarholder, xlab='Permuted SMU - Seattle', main='Histogram of Permuted Mean Differences')  
box()



pvalue <- counter / number\_of\_permutations  
pvalue

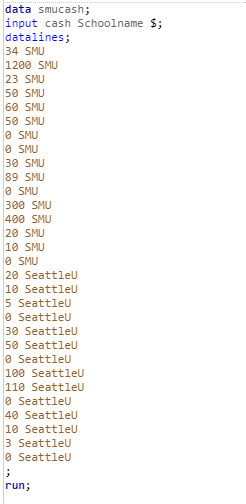
## [1] 0.135

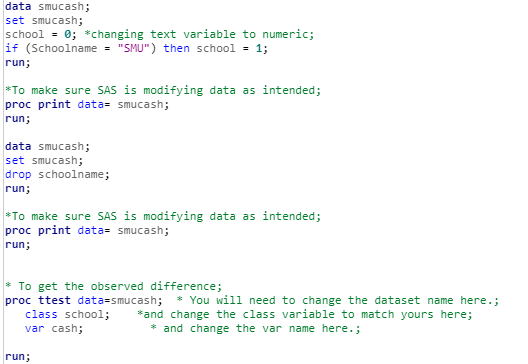
observed\_diff

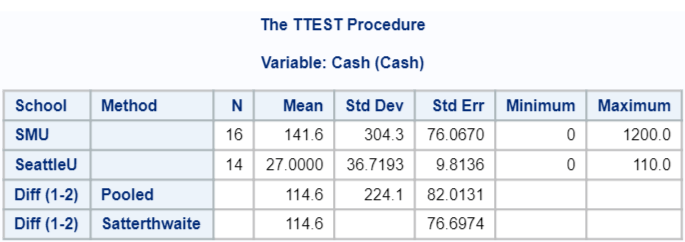
## [1] 114.625

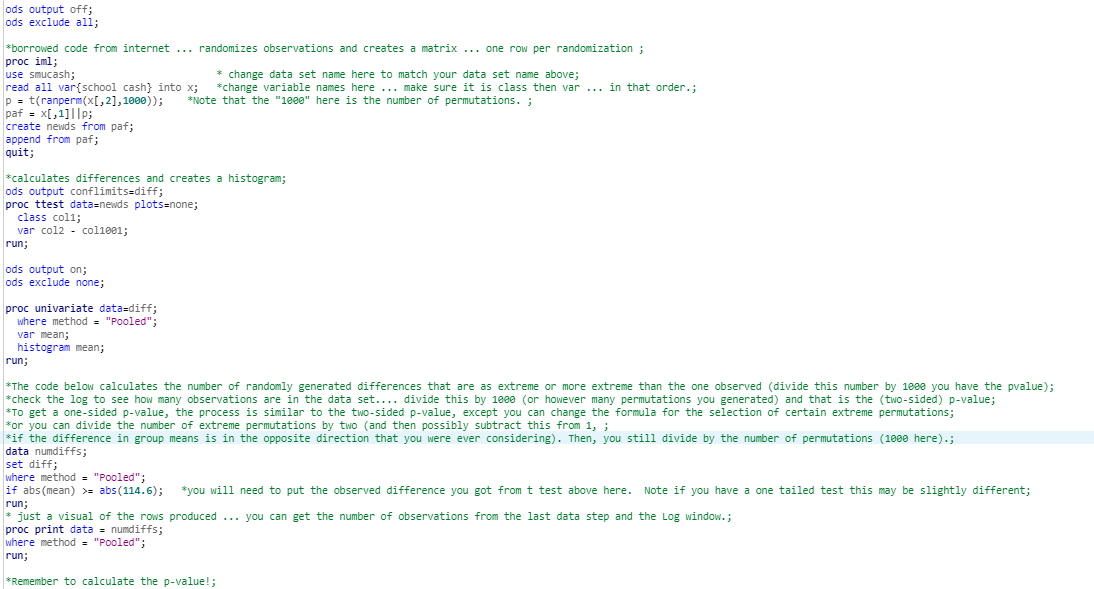
**To test for a difference of population means between the SMU and Seattle groups, a permutation test was conducted on 1,000 random permutations of the data. A histogram of the 1,000 differences of sample means from the 1,000 permutations can be seen above. The observed difference was $114 and 135 of the 1000 permutations yielded a difference in sample means that was as extreme or more extreme than this observed difference (p-value = 135/1000 = 0.135). This does not provide sufficient evidence against the null hypothesis that the mean pocket cash of SMU students is equal to that of Seattle University students at a significance level of 0.5 or even 0.10. These 30 students were not from a random sample; therefore, inference cannot be drawn beyond the 30 subjects in the sample. (Since we failed to reject Ho, there is no need to write up whether casual inference can be drawn. However, had the means been found to be significantly different, no causal inferences can be made, as these subjects were not randomly assigned to these schools.) Note that individual p-values can vary, but they should be somewhere near the p-value in this solution.**

SAS Results from a permutation test:









There were 146 permutations with extreme values, giving a p-value of 146/1000=0.146. The interpretation above applies here as well.

Using the program HW1Session1.sas produced the output table below:

C:\Users\Martin\AppData\Local\Microsoft\Windows\INetCache\Content.Word\New Picture (6).bmp

In this case, There were 164 permutations with extreme values, giving a p-value of 164/1000=0.164.