HW 1

Raag Patel DS 6371 Summer 22

Questions 1,2,3a and 3b covered in Live Session. You may use your answers from live session but add them to you HW for completeness. Also, put your answer in your own words (do not simply copy and paste the solution given in live session.) There are many ways to write responses to the questions below.

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

Only in randomized experiments can you make causal inferences, you can not justifiably make causal inferences in random samples. Also in a randomized experiment, people in the study are moved randomly to treatments. While in a random sample, groups from a population is moved to different treatments, randomly.

2. **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 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 study was conducted during the great depression, in which a lot of the poorer population did not have the money to spend on magazine subscriptions. This caused a demographic shift from the desired average income readers to higher income readers.

3. **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.**
    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 the 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.**

Since the new fertilizer was offered as an opt in, it was not random people from the population, nor were the groups randomized. Therefore no causal inferences can be made.

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.**

We do not know how many farmers sent in their end of year reports, so there can be a bias/skew in a certain direction. But since the placebo was sent randomly, this is a randomized experiment, therefore causal inferences can be made, but only to the farmers who sent in their reports.

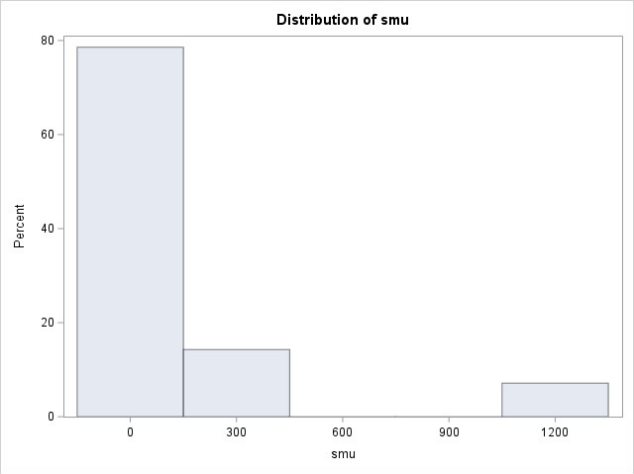
c. **When a customer makes an order, we randomly send them either the old or new fertilizer. At the end of the season, we sub-select from the fertilizer orders and send a team out to count those farmers' crop yields.**

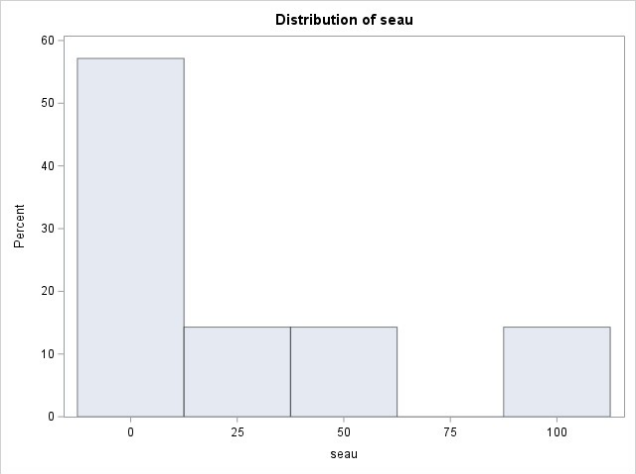
So now random fertilizer is sent to every consumer, it's still a random experiment where you can make causal inferences. It seems like only those farmers that you've 'sub-selected' will have crop yields recorded, it cannot be a random sample, since it doesn't seem like they'll be selected randomly.

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 farmers' crop yields. From our records, we know which of these farmers used the new fertilizer and which used the old one.**

Once again, the fertilizer is offered as an opt in, already ruling out the random sample. On top of farmers being sub selected, assuming that means not randomly selected. But since farmers are opting in and they aren't getting random fertilizer, this is not a randomized experiment either. No causal inferences can be made.

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





The SAS System

The TTEST Procedure

Variable: smu

N	Mean	Std Dev	Std Err	Minimum	Maximum
14	154.0	324.8	86.8079	0	1200.0

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
154.0	-33.5371	341.5	324.8	235.5	523.3

DF	t Value	Pr > |t|
13	1.77	0.0995

The SAS System

The TTEST Procedure

Variable: seau

N	Mean	Std Dev	Std Err	Minimum	Maximum
14	27.0000	36.7193	9.8136	0	110.0

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
27.0000	5.7989	48.2011	36.7193	26.6198	59.1564

DF	t Value	Pr > |t|
13	2.75	0.0165

This tells us with overwhelming evidence that there is a difference in the mean. Since the P-Values are really small, we can confidently say that there is enough evidence to suggest that the mean difference is != 0.

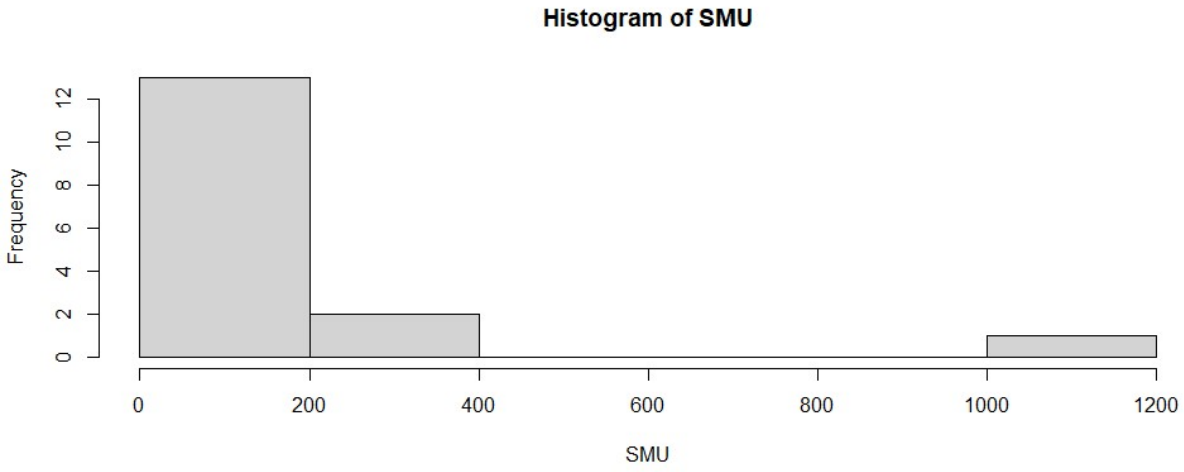
b. Use the following R code to reproduce your histograms. Simply cut and paste the histograms into your HW.

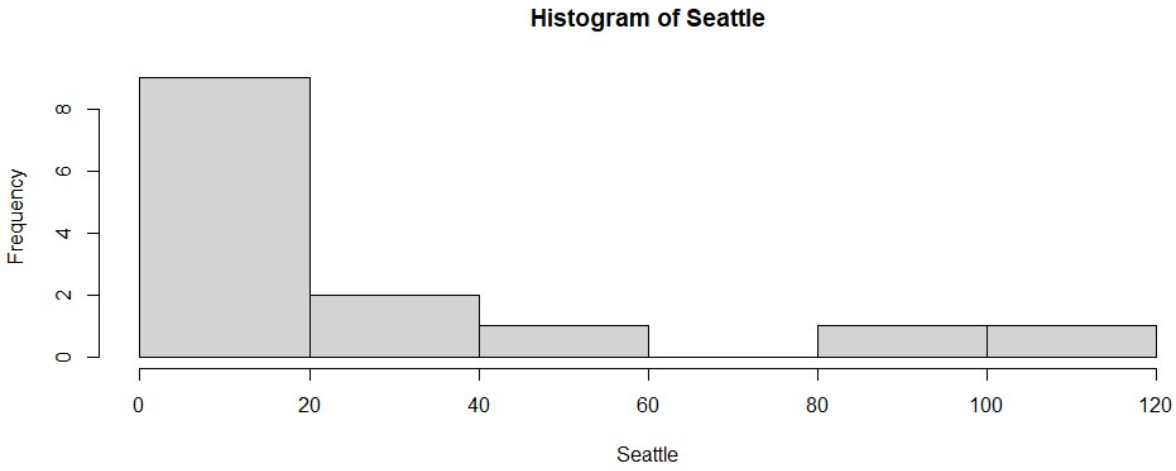
SMU = c(34, 1200, 23, 50, 60, 50, 0, 0, 30, 89, 0, 300, 400, 20, 10, 0)

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

hist(SMU)

hist(Seattle)





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.)**

Ha : true difference in means != 0 || Ho : true difference in means = 0

P Value of .1551, tells us that we do not have enough evidence to suggest reject Ho. I believe if we were to have more samples, we would get a different result. That also helps get a wide range of different values, giving us a more accurate reading of the school. Since significantly less than 1% of the school was polled, you cannot make any causal inferences.

Link to RCode