Javier Saldana

Sunday, September 2, 2018

# HW 1

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?**
A randomized experiment uses randomization to reduce bias for variables not being measured. However, the subjects are usually self-selected participants, which may create biases in the results. Random sample is when the participants are selected from the population randomly. This presents a greater opportunity for the sample to truly represent the greater population since all participants have the same chance of being selected. Although a random sample may be a strong representative of the population, without an experiment in place it would be difficult to draw any casual inferences from the results. A randomized experiment presents a stronger probability to draw casual inferences from the study.

> **Commented [SJ1]:** 15/20 points. Random sample is clearly defined but randomized experiment was not clearly defined. Final piece touches on experiment and causal inference.

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 magazine was attempting to poll from potential voters in order to determine who the likely candidate would be once the election concluded. The problem was that by sourcing only from phone number records, subscribers, and auto registrations, they excluded the lower class population who isn't able to afford a subscription to a magazine, or a car, or even a landline. Considering lower class population tends to vote Democrat, the results were heavily skewed in the favor of the Republican candidate. As history has shown us, the Democratic candidate won.

> **Commented [SJ2]:** 10/10 points. Answer describes the intended population and the actual population that was sourced.

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.**
   In this design, the scope of inference will not apply to the population or the cause and effect. The fertilizer was only sold to repeat customers and the participants would only be self-selected farmers who submit reports of the yield.

   > **Commented [SJ3]:** 8/8 points. Answer touches on the fact the farmers were self-selected and may not be generalized to broader population. Also touches on lack of causal inference.

   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.**
   In this design, the scope of inference will apply to the population but not to the cause and effect. The farmers were sold the new fertilizer randomly but only those who decide to send the report back would be accounted for, which means the groups are not assigned at random.

   > **Commented [SJ4]:** 0/8 points. Answer indicates population inferences may be made but there is no causal effect. Correct answer is that results may not apply to general population since farmers self-selected. Yet, considering this is a randomized experiment, there lies a potential for causal inferences, but only to the farmers that participated.

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.**

In this design, the scope of inference will apply to the population and the cause and effect. The farmers were sold the new fertilizer at random and the groups would also be assigned randomly.

> **Commented [SJ5]:** 8/8 points. Answer assumes sub-selection was random, which would indicate results could be extended to all farmers that buy the fertilizer. Also confirms there is potential for a causal effect based on random experiment.

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.**

In this design, the scope of inference will apply to the cause and effect but not the population. The fertilizer is limited to repeat customers, yet the groups will be randomized.

> **Commented [SJ6]:** 0/8 points. There was no randomization assignment so no causal inferences can be made. Reason for why it may not be applied to farmers population is incorrect.

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.**

There is a large difference in populations means. The SMU students averaged $141.6, which the Seattle U students averaged $27. The histograms clearly show SMU's students are sitting higher in the mean than their counterpart. However, this data may have been severely skewed by a student who carried $1200 in cash at the time. Yet, even if this individual is recognized as an outlier, the data would still have SMU with a higher mean thanks to the other two students with $300 and $400. The difference in means may have been exaggerated but it is clearly visible.

> **Commented [SJ7]:** 5/5 points. SAS results match and code as well. Excellent identification of 1200 outlier.

b. **Use the following R code to reproduce your histograms. Simply cut and paste the histograms into your HW.**
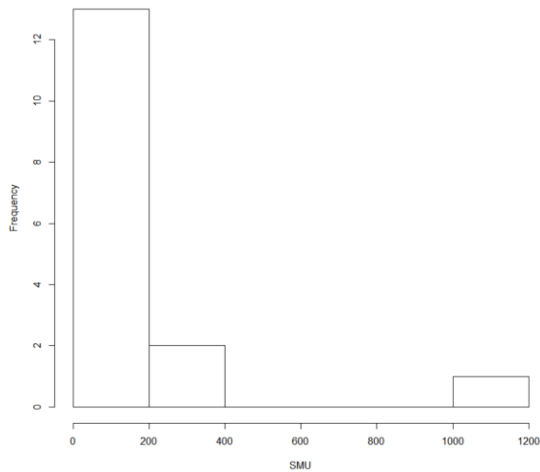*SMU = c(34, 1200, 23, 50, 60, 50, 0, 0, 30, 89, 0, 300, 400, 20, 10, 0)*
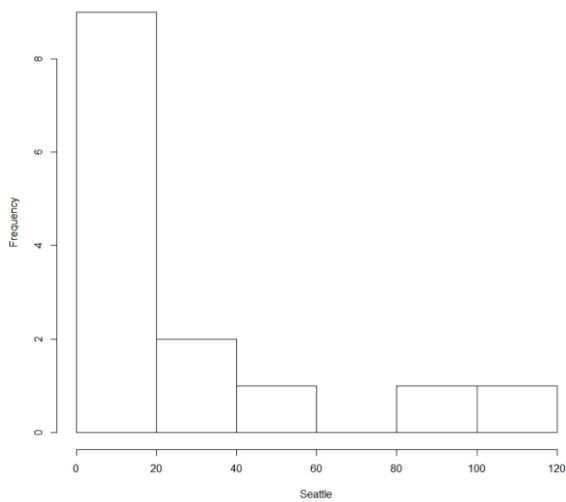*Seattle = c(20, 10, 5, 0, 30, 50, 0, 100, 110, 0, 40, 10, 3, 0)*
*hist(SMU)*
*hist(Seattle)*

**Histogram of SMU**

**Histogram of 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.)**

The goal of study was to determine if students from one school carried more cash then the other school. The null hypothesis would be that SMU and Seattle U students carry the same amount of cash on average. The alternative hypothesis would be that SMU and Seattle U students do not carry the same amount of cash on average. No treatments were placed on

3

the students so this is an observational study. The subjects were a convenience sample and as a result, inference about the student population of either school may not be inferred from this study. However, it can be inferred that the mean cash on-hand by SMU students in this business statistics course is not the same as the mean cash on-hand of a Seattle U students in their business statistics course (p-value = 0.0419).

**Commented [SJ9]:** 5/15 points. Test was run and code was uploaded. However, incorrect p-value was obtained and as a result, the wrong inference was made.