HW 1

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?

   Answer: A ***randomized experiment*** seeks to establish causal relationships by introducing a treatment and studying the responses or outcomes of that treatment on the subjects. The subjects are randomly assigned to treatment groups that are designed to be representative of the sample and as equal as possible except for the treatment applied to the group (Such as in the Creativity study groups.). Because the groups are so similar, a causal inference can be made about the outcome of the experiment. Only experimental studies can establish cause-and-effect relationships because they have a controlled treatment that is applied to the study group. Observational studies can not directly establish causal relations but can show some level of association between outcomes and a subject.

   A ***random sample***, on the other hand, is a selection of a subset of the population that is representative of the mix of the whole population. The method seeks to ensure that all members of the population have an ***equal probability*** of being selected for the study.

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.

   Comment: In the Landon v. Roosevelt presidential race pollsters inaccurately called the election because they did not select a ***random sample***. For the pollsters to select a ***random sample*** they needed to take a subset of the population that was evenly mixed to give all members of the ***voting population*** an equal opportunity to be selected and to respond. Because Time magazine ***did not take its sample at random***, but instead took its sample from readers of its magazine, and owners of phones and cars, they inadvertently ***introduced bias***. At the time of the election the US was still recovering from the depression and many voting Americans could not afford things like phones and cars. Thus, this very important group of voters (the poor who were mostly likely to vote for Roosevelt) was excluded from the poll. ***Had pollsters used a random subset of the larger population of voters so that the sample mix was more reflective of the overall population, their poll may have been more accurate.***

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.

Study A Scope of Inference: This study did not use a ***random sample*** of the entire farming population and thereby introduced bias. Additionally, the study sought volunteers through discounted fertilizer and limited responses to farmers self-selected by sending in the responses to the survey. The scientists also failed to create a ***randomized experiment*** because they did not control the assignment of the fertilizer. ***Without random assignment of the treatment no causal inference can be made***. Despite the statistically significant pvalue, the design was flawed and would have invalidated the results.

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.

Study B Scope of Inference: This study used a ***randomized experiment*** to establish a causal relationship between the use of the new fertilizer and its yield as opposed to that of the old fertilizer, but the survey respondents did not represent a ***random sample***. Because the new and old fertilizers were randomly sent to the farmers, no bias was introduced by allowing them to know which fertilizer they received, nor was there any incentive to report crop yields beyond simply providing the information. However, the response mechanism used allowed farmers to volunteer (self-select) responses, making it very difficult to claim a ***random sample***. The scientist were able to establish a causal relationship as they tracked which farmer received which fertilizer and could directly link it to the yield, but could not make an inferences that represented the population as a whole.

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.

Study C Scope of Inference: The study sought to create a control and treatment group that was randomized. The ***randomized experiment*** sent new and old fertilizer to the subjects at random (***a random sample***) and then sent out teams to record the yields of the crops from a subset of that group. This process greatly reduces the ***probability that bias could be introduced*** and could thus ***induce a causal relationship*** with the outcome.

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.

Study D Scope of Inference: This study did not use a ***random sample*** as the farmers self-selected by taking advantage of the fertilizer discount. The researchers also failed to create a ***randomized experiment*** because the treatment was not applied randomly (self selecting farmers securing a discount) and no control group was established. Despite sending out the teams to randomly collect information on crop yields, there are too many ***compounding variables*** that could play into the study, such as soil quality, water availability, and farming equipment that could also influence outcomes. Without a proper control of the experimental conditions , no causal inference 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.

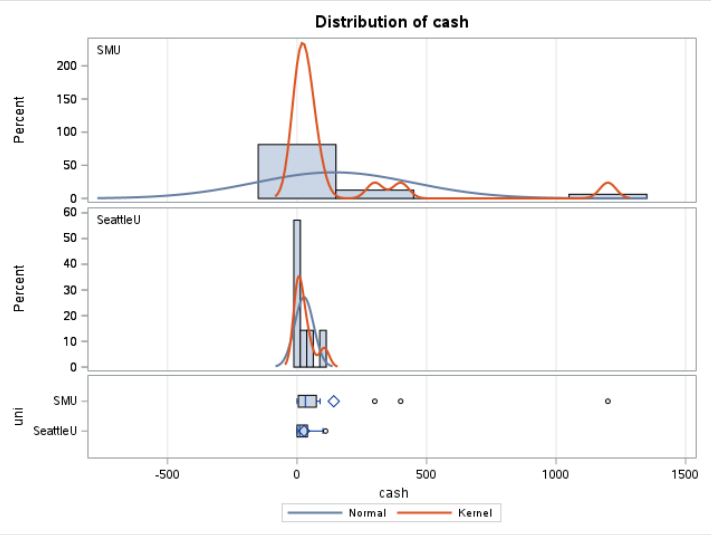
SMU

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

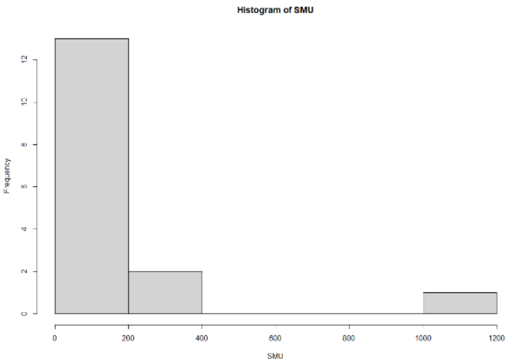
a. Use SAS to make a histogram of the amount of money in a student's pocket from each school. Does it appear there is any difference in ***<u>population</u>*** means? What evidence do you have? Discuss your thoughts.

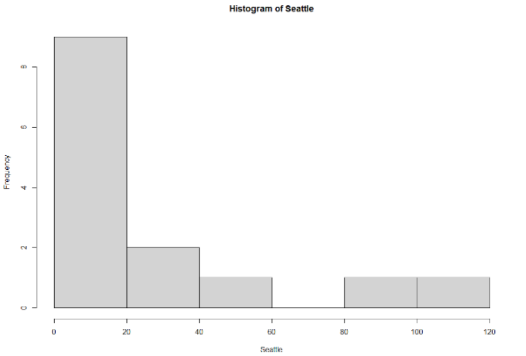


Answer A: In regards to the student's pocket study there is a major difference between the population mean of SMU (141.6) and the Seatle U (27), which gives a difference of (114.6). This difference is likely due to outliers in the data (1400, 400, 300). These outliers may suggest that the population is not a random sample or representative of the whole. The researcher might want to remove the outliers so that they could conduct a better analysis of the population as a whole.

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)
```

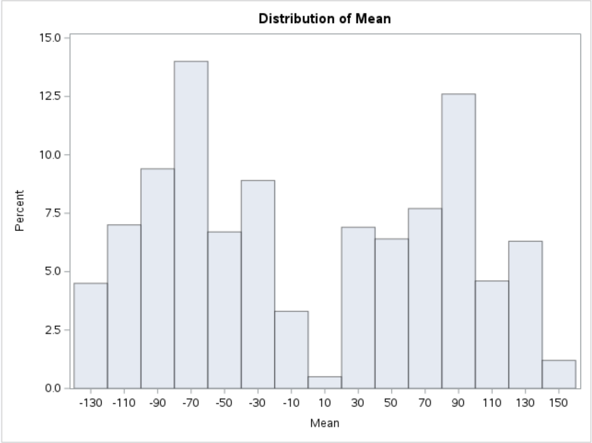




c. Run a permutation test to test if the mean amount of pocket cash from students at SMU is different than that of students from Seattle University. Write up a statistical conclusion and scope of inference (similar to the one from the PowerPoint). (This should include identifying the Ho and Ha as well as the p-value.)

Answer C:

- ***Assumption*** - We assume, for our null hypothesis, that most students will have on average the same amount of cash on hand (H_o: $\mu_1 - \mu_2 = 0$). The alternative hypothesis is that most students will on average not have the same amount of cash on hand (H_a: $\mu_1 - \mu_2 \neq 0$).
- ***Evidence*** – The difference in the mean of the SMU sample (141.6) from the Seattle sample (27) was 114.6.
- ***Probability*** – The estimated probability that a difference of 114.6 or larger in the sample mean scores is high as indicated by the p-value of 0.148.



- ***Conclusion*** – There is not enough evidence to suggest that the mean score of SMU students who have cash on hand ***is different*** (cannot reject the null hypothesis) than the mean score of Seattle students who have cash on hand (p-value = 0.148). Of the 1000 permutations, it was found that 148 out of 1000 had higher scores than our original pooled score. Thus, there is a high probability of observing a score as extreme or more extreme than the original difference.

- ***Scope of Inference*** - This is ***not a randomized experiment*** because the subjects self-selected by participating in the study and there was no randomized treatment among groups, they simply looked in their wallets.
  - No causal relationship between the school and the amount of money can be determined. This test appears to be more an observational study than a randomized experiment and there are many confounding variables that could explain the difference in cash.
  - The subjects are not part of a ***random sample*** of students as they represent a distinct class and not a subset of the total student population.