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?

**Random sample is when we pick up the subject for an experiment. And each subject that we pick up has the same likelihood of being pick up for the experiment. A randomized experiment is when the subject from a group is being randomized for the experiment, so each subject has equal possibility of being selected for the experiment or for the control, or for experiment A and experiment B if the experiment A is compare to another experiment and no to a control. But that group from where the subject is being randomized for the experiment may not be a random sample. So random sample and randomized experiment are different concepts as explained above.**

1. 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 desired population was a representative population from the US. But the population was taken from subscribers to its magazine, phone number records, and automobile registration records. During that period of time 1936 the subjects who were subscribe to magazine, has a car or has a phone were privilege subject with better economical position. And the subject with better economical position were heavily republican. The sample was heavily weight toward the republican, so the republican had more chances of being selected than the democrats. It was not a sample that represented the US. The democrats and republican had not the same chance of being selected for the pool. The republican had more chance of being selected in the pool for the reason mentioned above.**

1. 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.
   1. 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.

**This is an observational study. There was not randomization of the sample. The farmer decided to participate by sending their survey. There is not randomization of the experiment, the subject decides to pick up the fertilizer that they want. So there is not a cause and effect in this study design, we can only infer association between the increased of bushels and the new fertilizer.**

* 1. 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.

**This is a randomized experiment, so we can infer cause effect for the increased bushels production from the new fertilizer. This cause effect is applicable to the population who called for the order and send back the survey, but it is not applicable to all the farmer because there were not selected randomly.**

* 1. 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.

**This is a randomized experiment, so we can infer cause effect for the increased bushels production from the new fertilizer. Assuming that the sub-selection was done randomly we can apply this cause effect to all the customer who order, not just the subset selected. But this is not applicable to the universe of farmer. It is applicable to the farmer who are customer of the particular store which is located in specific area.**

* 1. 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.

**Assuming the sub-selection is random we have a random sample, so the study result can be applicable to all the farmer from that store and that location. But the experiment is not randomized, actually there is an incentive to use the new fertilizer, so because there is not randomization in the experiment, we cannot infer cause effect. So, we cannot infer that the new fertilizer increased the production of bushels, only we can infer association between the new fertilizer and the increased of bushels production, but this association certainly would be intriguing.**

1. 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

* + - * 1. 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 population from SMU and the population from Seattle U appear different. SMU has a mean of 141.6, Standard Deviation of 304.3, minimum 0, maximum** **1200**

**Seattle U has a mean of 27.0, Standard Deviation of 36.7, minimum 0, maximum 110.**

**But the t-value for method pooled and variance equal was 1.4, with a p value 0.1732. The t-value for a method Satterthwaite variance unequal was 1.49, p value 0.1551. Therefore, according to those mentioned p value we fail to reject the hypothesis that the two population are the same.**

* + - * 1. 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)***

***A screenshot of a cell phone

Description automatically generated***

A screenshot of a cell phone

Description automatically generated

* + - * 1. 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.)