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?

**In a randomized experiment, the investigator selects subjects to receive a particular treatment in a random fashion so as to isolate the effects of the treatment from other variables of influence. In a random sample, the experimenter selects a random slice from a population for observation and attempts to draw conclusions based on statistical inference of the results. Only a randomized experiment can lead to causal inference since a random sample can never fully decouple the effects of covariance amongst the different variables.**

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.

* **Desired population: voting age americans**
* **Sample population: magazine subscribers (disposable income, pollical leaning?), phone number records (in 1936 post great depression people had no phones), automobile records (rich owned cars)**
* **Sample bias towards the wealthy leads to a prediction of a Landon victor. Exclusion of the poor from the sample set (or failing to gather income data and adjust accordingly) leads to an underestimation of the population looking to vote for FDR (New Deal etc?)**

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.
   * 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.

* **Not a random sample. Only pre-existing customers are selected. Only some farmers send in responses to the survey. Bias introduced by selling at a discount and allowing the control variable to be selected by the subjects.**
  + 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.   
     **Random sample – but not requiring all to respond creates bias in the sample set. Variability in measurement methods could lead to error.**
  + 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.
  + 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.

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.
        2. 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)***

* + - * 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.)