UNIT 1 HW

1. What is the difference between a randomized experiment and a random sample?

A randomized experiment is when you have groupings that all have an equal chance of receiving a treatment. A random sample is a sample (subset) of a population is chosen to be examined.

Under what type of study/sample can a causal inference be made?

A causal inference can only be made with a randomized experiment and not an observation study.

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 population of interest should have been people who vote. For instance, they could have sampled randomly from being who were registered to vote. However, since they drew from the following three sources: subscribers to a magazine, phone records, and automobile registration records, they likely sampled from only the affluent population. In 1936, it was not common to have phones in the household nor own a vehicle as these items were luxury items.

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.

Inference cannot be made since the fertilizer was not assigned randomly and was assigned by existing customers who already likely used and applied old fertilizer. Since farmers are sending in the survey, inference cannot be made to the entire population.

* 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.

Here causal inference can be made. However, the population is all the customers who ordered fertilizer. Now the sample that we have to collect data on is not random and is voluntary. If the sample that we get back is representative of the population meaning if the sample is a large percentage of the population, then we can say that the sample is representative of the population because we have more power. However, if the sample is low, then we can’t infer causality on the entire population and only on the volunteers.

* 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 would be the ideal experiment. There is randomization in the fertilizer and randomization in selecting the sample to be studied. Causal inference can be made here to the entire population.

* 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.

Causal inference cannot be made since the treatment was not randomized. Though it is randomized to see crop yields, this would be observational in nature and we cannot infer on the population.

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.

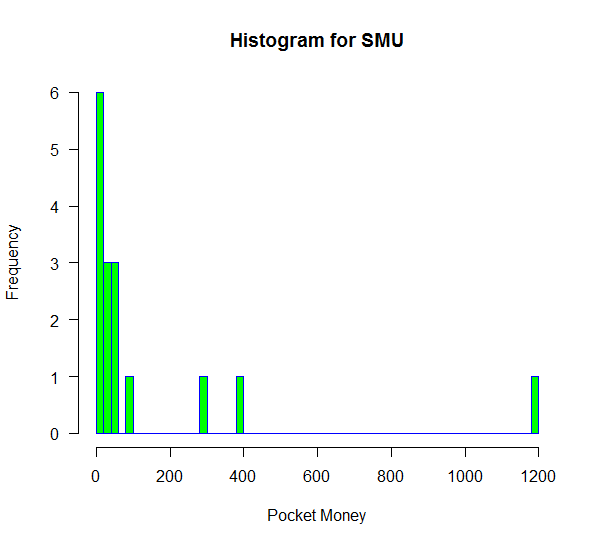
**SMU**

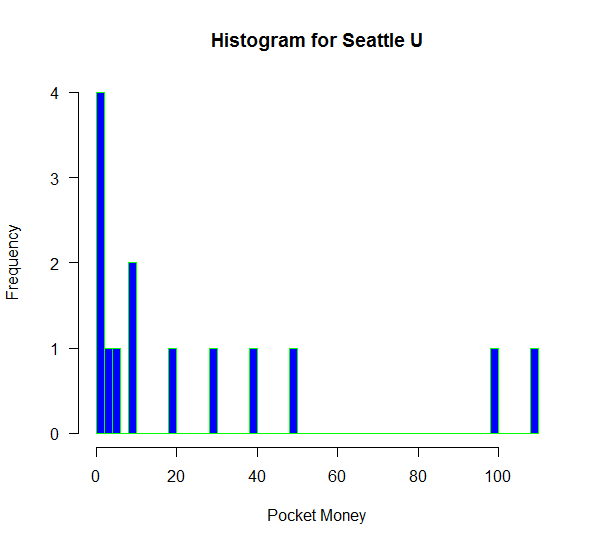
34, 1200, 23, 50, 60, 50, 0, 0, 30, 89, 0, 300, 400, 20, 10, 0

**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. (? Same as C)





* + - * 1. Does it appear there is any difference in ***population*** means? What evidence do you have? Discuss your thoughts.

So the population would be everyone who has access to the vending machine. The sample that was selected was the students of the business cIass. I think there may be a difference in population means unless the vending machine is only accessible by the Business Stats class which I doubt. Also, this vending machine won’t be accessible by Seattle, so I’m missing something here. The sample here may not be representative of the entire population that can access the vending machine therefore the population mean may be different than the sample mean.

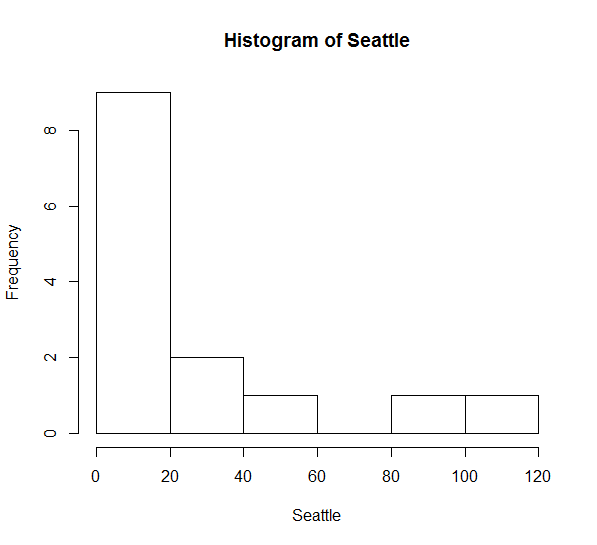
* + - * 1. Use the following R code to reproduce your histograms. Simply cut and paste the histograms into your HW.

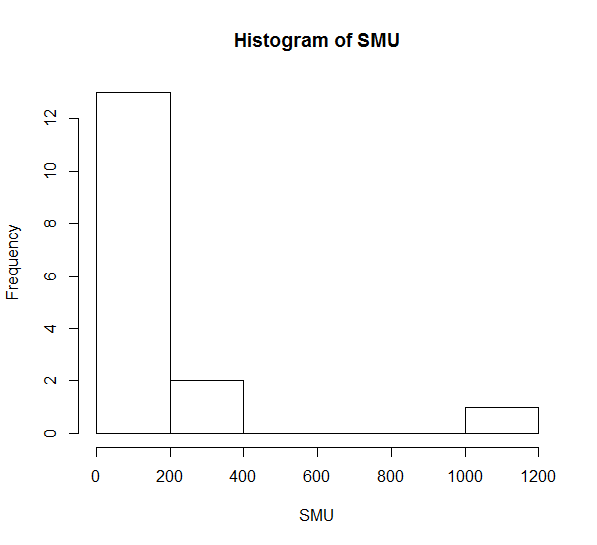
***SMU = c(34, 1200, 23, 50, 60, 50, 0, 0, 30, 89, 0, 300, 400, 20, 10, 0)***

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

***hist(SMU)***

***hist(Seattle)***





* + - * 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.)

Alpha was not provided. Going to assume .05

#test stat: difference in means

#H0: ui- ue = 0

#HA: ui-ue <> 0

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)

df1 <- cbind("SMU", SMU)

df2 <- cbind("Seattle", Seattle)

df <- rbind(df1, df2)

df <- as.data.frame(df)

df <- rename(df, c(V1 = "School", "SMU" = "PocketMoney" ))

df$PocketMoney <- as.numeric(df$PocketMoney)

t.test(df$PocketMoney ~ df$School)

number\_of\_permuations = 10000;

xbarholder = c();

observed\_diff = mean(subset(df, School == "SMU")$PocketMoney)-mean(subset(df, School == "Seattle")$PocketMoney)

observed\_diff

counter = 0;

summary(df)

for(i in 1:number\_of\_permuations)

{

scramble = sample(df$PocketMoney,30);

SMU = scramble[1:16];

Seattle = scramble[17:30];

diff = mean(SMU)-mean(Seattle);

xbarholder[i] = diff;

if(diff > observed\_diff)

counter = counter + 1;

}

hist(xbarholder)

pvalue = counter/ number\_of\_permuations;

pvalue

> pvalue

[1] 0.1074

Alpha = .05

Pvalue = .1074

Going to fail to reject the null hypothesis.