

Jonathan A. Rocha

Dr. Monnie McGee & Dr. Bivin Sadler

DS-6371 Statistical Foundations for Data Science

January 9, 2025



HW 1



Question 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?

Key differences:

-  A random sample is about how participants are selected from a population. Each member of the population has an equal chance of being selected for the study. This allows generalization to the population.
-  A randomized experiment involves assigning participants to treatments once they're in the study. Each participant has an equal chance of receiving each treatment, which allows causal inference.



Causal inference can only be made with a randomized experiment because:

1. Random assignment helps balance confounding variables between groups
2. It ensures any systematic differences between groups are due to the treatment
3. It minimizes selection bias in treatment assignment



Question 2: Regarding the 1936 Literary Digest poll:

Population of interest: All eligible American voters who would vote in the 1936 presidential election

Actual population sampled: A much wealthier subset of Americans, specifically:

1. Magazine subscribers (higher income and education level)
2. Phone owners (during the Great Depression, phones were luxury items)
3. Car owners (automobiles were also indicators of wealth in 1936)

This created significant sampling bias because:

- The sample was not representative of the general voting population
- Wealthier Americans tended to favor Republican candidate Landon
- Poll missed many lower-income voters who supported Roosevelt
- The large sample size (1 in 4 Americans) didn't correct for a biased sampling frame

The 1936 Literary Digest poll represents a classic example of how a large sample size cannot compensate for systematic sampling bias. Although it surveyed an impressive 2.4 million Americans (about one-quarter of the electorate), the poll disastrously failed to predict Roosevelt's landslide victory because it sampled from an unrepresentative subset of the population. By drawing its sample from magazine subscribers, telephone owners, and automobile registrations during the Great Depression, the Literary Digest inadvertently targeted primarily wealthy Americans who were more likely to support the Republican candidate. This sampling strategy completely overlooked American voters' economic and demographic diversity, particularly lower-income citizens who overwhelmingly backed Roosevelt's New Deal policies. The poll's

failure illustrates that selecting participants is far more crucial than the number of participants—even a massive sample size cannot correct for a biased sampling frame. This case has become a foundational example in statistics of how seemingly impressive "big numbers" can lead to dramatically incorrect conclusions when the sample is not representative of the population of interest.

Question 3: Let's analyze each fertilizer study scenario:

a) Survey sent to previous customers:

- No random sampling (self-selected participants)
- No random assignment (farmers chose their fertilizer)
- Scope: Cannot generalize to all farmers or make causal claims
- High risk of response bias and self-selection bias
- Results only apply to survey respondents

b) Random assignment of fertilizer type:

- No random sampling (existing customers only)
- Has random assignment (company randomly chose fertilizer type)
- Scope: Can make causal claims about fertilizer effect but only for responding customers
- Response bias is still present, but treatment assignment was controlled
- Can determine causation for this specific group

c) Random assignment plus team measurement:

- ♥• No random sampling (existing customers only)
 - Has random assignment (company randomly chose fertilizer type)
 - No response bias (team measured all selected fields)
 - Scope: Can make causal claims about fertilizer effect on customer population
 - The most rigorous design of the four options
 - Eliminates both response bias and measurement error

d) Self-selected fertilizer with team measurement:

- ♥• No random sampling (existing customers only)
 - No random assignment (farmers chose their fertilizer)
 - No response bias (team measured all selected fields)
 - Scope: Cannot generalize to all farmers or make causal claims
 - Eliminates response bias, but selection bias remains
 - More reliable measurements but still can't determine causation

In all four scenarios, the strong statistical significance ($p = 0.0001$) suggests a real difference in yield, but the scope of inference varies dramatically based on the study design. Only scenarios b and c, which use random assignment, allow for causal conclusions about the fertilizer's effect.

♥Question 4. Based on the analysis, here's my answer to Question 4:

a) The histograms show that both distributions are right-skewed, with several students having \$0 in cash. The SMU distribution shows more extreme values and significant variability, ranging from \$0 to \$400, while Seattle U's distribution is more compressed, ranging from \$0 to \$110. The SMU mean (\$79.13) is higher than Seattle U's mean (\$27.00), but this difference appears to be influenced by a few large values in the SMU sample.

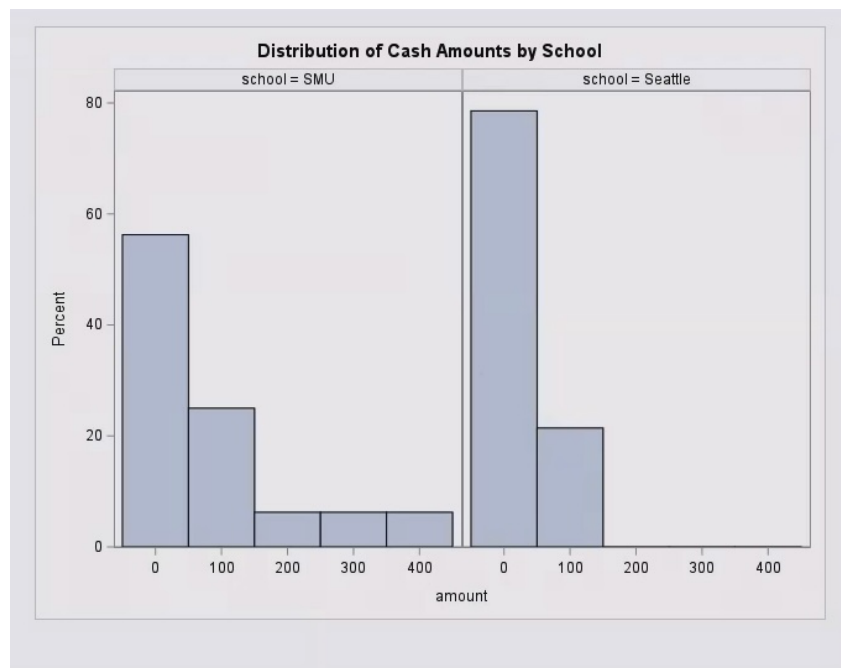
b) As requested, the histograms have been created and visualized in SAS and R formats.

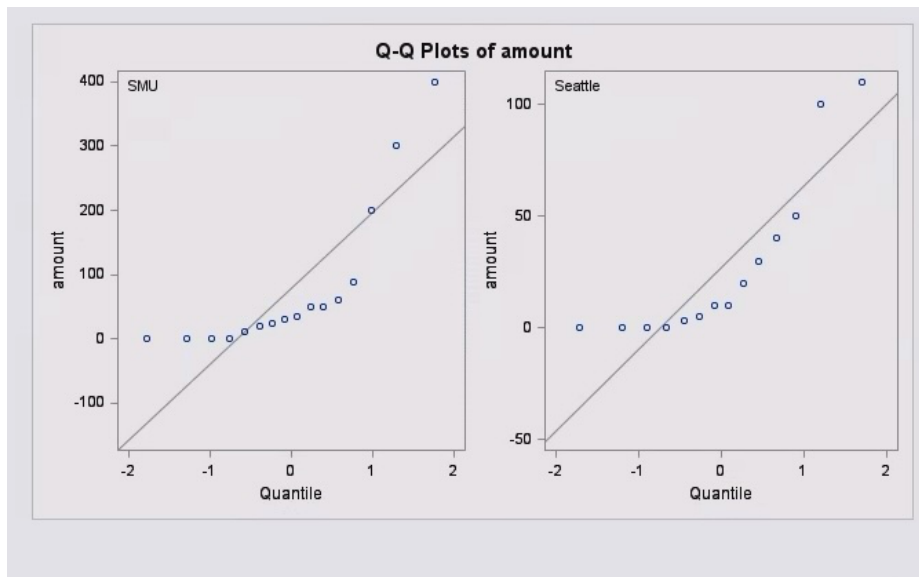
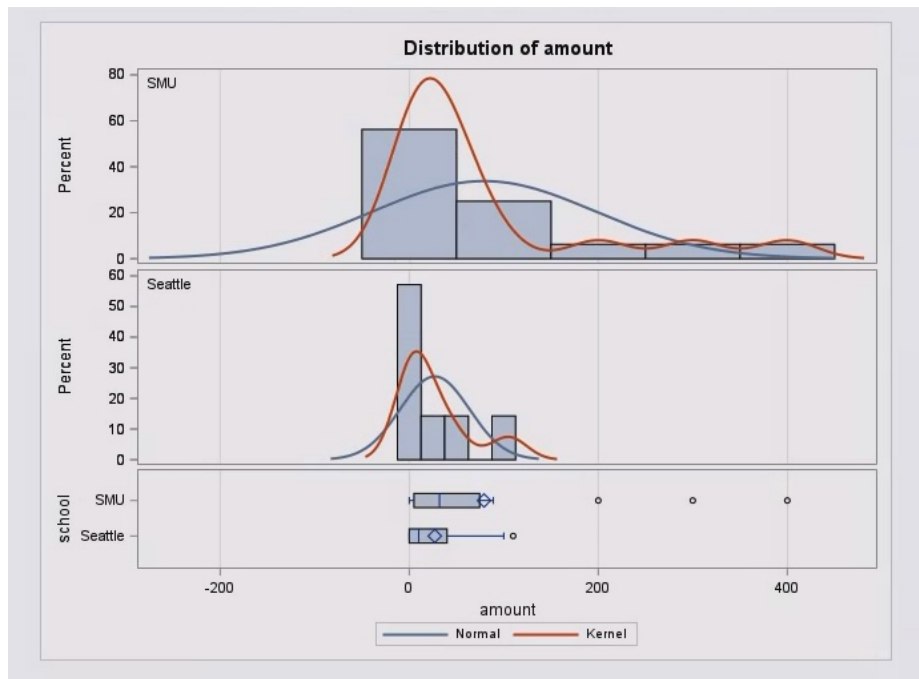
SAS outputs:

Summary Statistics of Cash Amounts by School

The MEANS Procedure

Analysis Variable : amount				
school	N Obs	Mean	Std Dev	N
SMU	16	79.13	118.16	16
Seattle	14	27.00	36.72	14





Two-Sample T-Test Results

The TTEST Procedure

Variable: amount

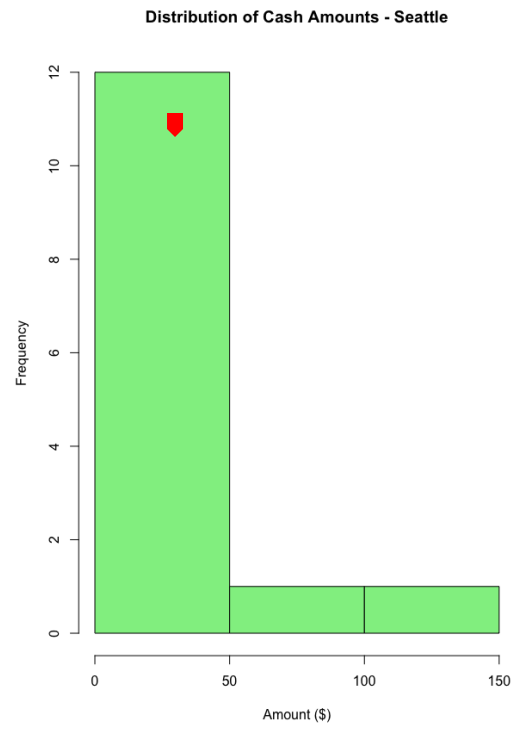
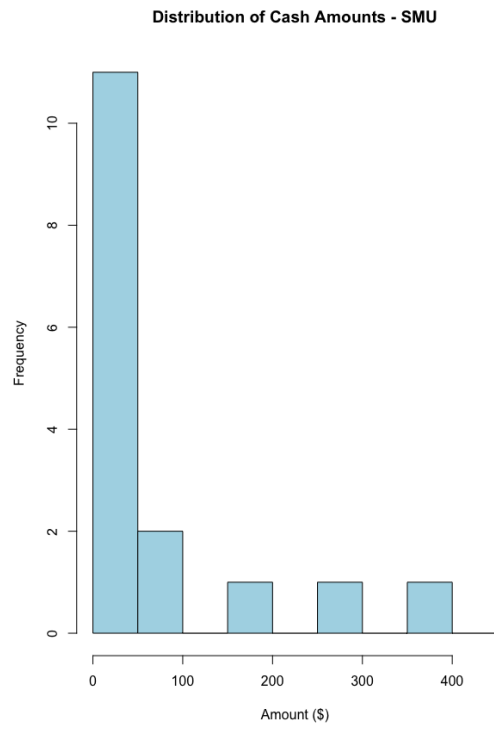
school	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
SMU		16	79.1250	118.2	29.5405	0	400.0
Seattle		14	27.0000	36.7193	9.8136	0	110.0
Diff (1-2)	Pooled		52.1250	90.0321	32.9484		
Diff (1-2)	Satterthwaite		52.1250		31.1279		

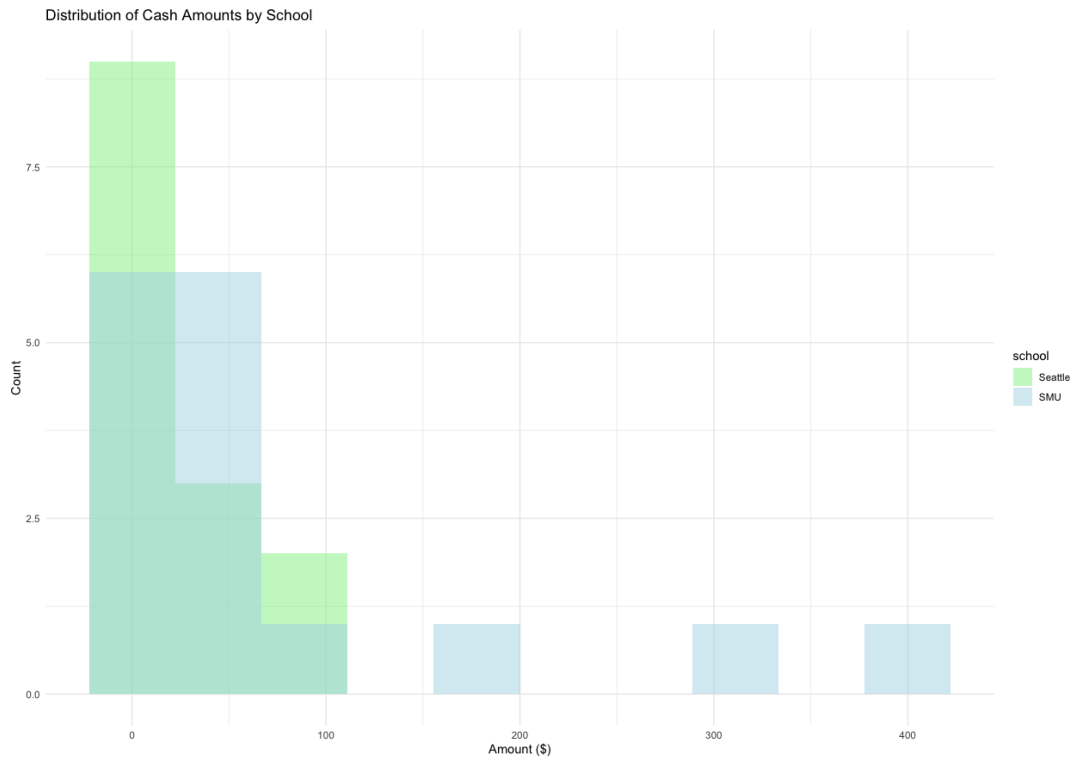
school	Method	Mean	95% CL Mean	Std Dev	95% CL Std Dev
SMU		79.1250	16.1609 142.1	118.2	87.2868 182.9
Seattle		27.0000	5.7989 48.2011	36.7193	26.6198 59.1564
Diff (1-2)	Pooled	52.1250	-15.3667 119.6	90.0321	71.4476 121.8
Diff (1-2)	Satterthwaite	52.1250	-13.2114 117.5		

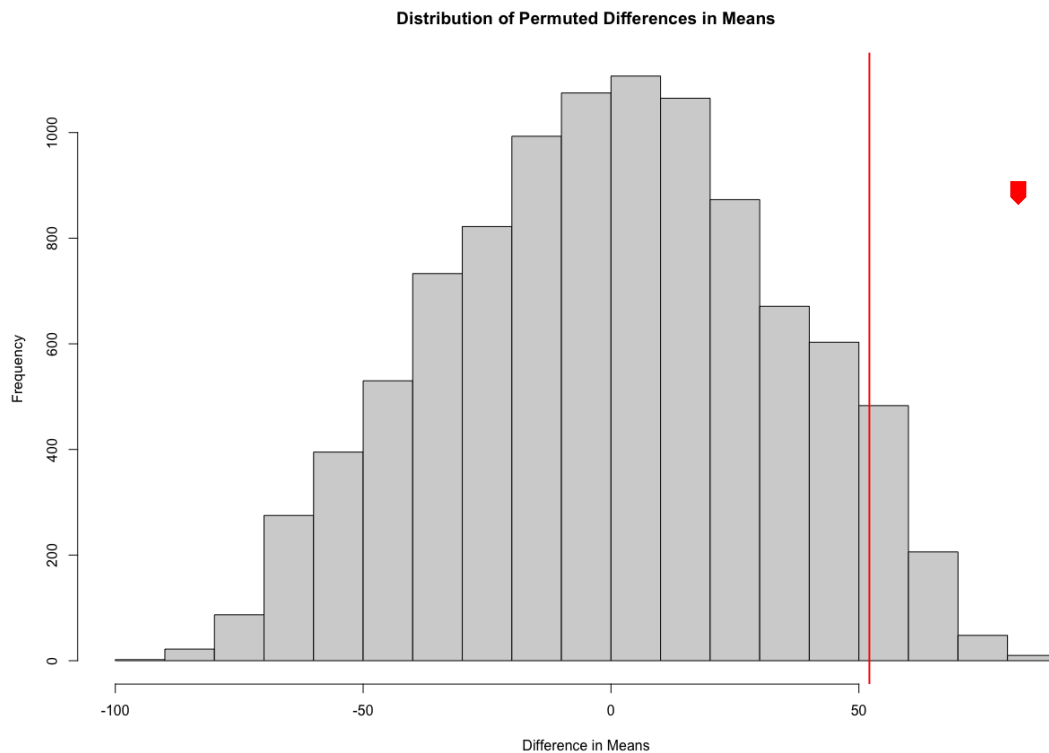
Method	Variances	DF	t Value	Pr > t
Pooled	Equal	28	1.58	0.1249
Satterthwaite	Unequal	18.237	1.67	0.1111

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	15	13	10.36	0.0001

R outputs:







c) Statistical Analysis:

Null Hypothesis (H_0): There is no difference in the mean amount of pocket cash between SMU and Seattle U students ($\mu_{SMU} = \mu_{Seattle}$).

Alternative Hypothesis (H_a): There is a difference in the mean amount of pocket cash between SMU and Seattle U students ($\mu_{SMU} \neq \mu_{Seattle}$)

A permutation test was conducted with 10,000 permutations. The observed mean difference was \$52.13 (SMU mean—Seattle mean). The permutation test yielded a p-value of 0.1371, more significant than the conventional significance level of 0.05.

Statistical Conclusion: Fail to reject the null hypothesis. There is not enough evidence to conclude that there is a significant difference in the mean amount of pocket cash between SMU and Seattle U students ($p = 0.1371$).

Scope of Inference:

- Generalization: The results can only be generalized to the specific classes sampled, not to the broader student populations at either university, as this was not a random sample from the entire student body.
- Causation: No causal claims can be made as this was an observational study, not a randomized experiment.

The significant difference in means (\$52.13) was not statistically significant, likely due to the high variability in the data and small sample sizes (SMU $n=16$, Seattle $n=14$). This suggests that while SMU students in the sample carried more cash on average, this difference could be due to chance rather than an actual population difference.