**PP249: Problem Set 2**

*DUE: SEPTEMBER 22, 2023 at 5pm*

# Instructions!

*Please do the coding portion of this problem set in Stata or R. Save all code in R or Stata script that can be run from start to finish without errors. Save your script as [LAST NAME]\_[FIRST NAME]\_ps2.[FILE EXTENSION] and email it to Raheem by the deadline. Final answers should be pasted into this document.*

# Question 1

[READI Chicago](https://www.heartlandalliance.org/readi/program/) is an organization that provides men in Chicago who have been most affected by gun violence with access to employment, cognitive behavioral therapy, and other support services in order to “help [their] participants create a viable path and opportunities for a different future.”

Let’s say that another organization attempts to replicate the program. However, instead of evaluating the program using a randomized control trial, they do the following: They start by recruiting participants from a few neighborhoods of another large city. Then, they offer the program to the first *x* individuals who agree to participate, and place the next *x* individuals who agree to participate on a waitlist where they don’t have immediate access to the program. They then compare outcomes of individuals who participate in the program to those on the waitlist. The hypothesis is that individuals who participate will be less likely to be shot or killed (“shooting and homicide victimization”) compared to those on the waitlist. For this question, please do the following:

**Question 1.1**

Specify the treatment, observed outcome, and potential outcomes for this causal claim.

Treatment: A program based on READI Chicago is offered to certain participants.

Observed outcome: Whether or not participants are shot or killed.

Potential outcomes: There could be a reduction, increase, or the number could remain the same vs the control group.

**Question 1.2**

Describe a possible form of selection bias that might affect who ends up participating in the program.

The first people who sign up for the program may be more motivated to participate than those who sign up later and may have a lower likelihood of being shot or killed than those who sign up later.

*(You can read more about a recent experimental evaluation of READI Chicago* [*here*](https://osf.io/preprints/socarxiv/dks29)*.)*

# Question 2

You are given cleaned data from the [Tennessee Star Experiment](https://www.aft.org/sites/default/files/periodicals/STARSummer99.pdf), which aims to address the impact of class size on student outcomes. Your task is to do the following, in either R or Stata (please be sure to share your code!):

1. Load in the dataset, which is stored here on bCourses as “TN\_STAR\_reduced.csv.”
2. Conduct summary statistics as requested in 2.1.
3. Conduct a regression with *robust* standard errors, where “math\_score” is the outcome and our only predictor variable is “treat”.
4. Answer the questions below!

*If you are unsure of what code to use, all of this has been included within the Stata and R videos (or can be yielded through a quick Google search!)*

# Question 2.1

Before we get into the regression, we want to run some summary statistics. Fill in the table below for the outcome variable, “math\_score”.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Treatment Group** | ***n*** | **Minimum** | **Mean** | **Standard Deviation** | **Maximum** |
| Control (treat=0) | 4035 | -3.54 | -0.0493 | 1.00 | 3.01 |
| Treatment (treat=1) | 1733 | -2.81 | 0.130 | 1.06 | 3.01 |

# Question 2.2

Conduct the regression with robust standard errors (using “math\_score” as the outcome and “treat” as the predictor). Please answer the following:

Estimate Std. Error t value Pr(>|t|) CI Lower CI Upper

(Intercept) -0.04933773 0.01575580 -3.131401 1.748448e-03 -0.08022502 -0.01845045

treat 0.17980277 0.02992711 6.008023 1.992814e-09 0.12113439 0.23847114

DF

(Intercept) 5766

treat 5766

* What is control group mean? -0.04933773
* What is the treatment group mean? Add the mean of the control group to the estimated treatment effect: -0.04933773 + 0.17980277 = 0.13046504
* What is the estimated treatment effect? 0.17980277
* What is the estimated standard error on the treatment effect? 0.02992711
* What is the “t statistic”? This measures how many standard errors the point estimate, or coefficient, is from zero. In this case, it is over 6 standard errors away. So the impact of the treatment is highly significant, much larger than statistical significance at the .05 level.
* Is the impact of treat on “math\_score” statistically significant at the .05 level?
  + See above

# Question 2.3

Conduct a two-sample t-test for the difference in means between the treatment group and the control group. What is the t statistic? Does this differ from your answer to Question 2.2? The t statistic is the same. This is consistent with the result from 2.2

> control\_group <- tn\_star$math\_score[tn\_star$treat == 0]

> treatment\_group <- tn\_star$math\_score[tn\_star$treat == 1]

> t\_test\_result <- t.test(treatment\_group, control\_group)

> print(t\_test\_result)

Welch Two Sample t-test

data: treatment\_group and control\_group

t = 6.008, df = 3118.1, p-value = 2.095e-09

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

0.1211239 0.2384816

sample estimates:

mean of x mean of y

0.13046503 -0.04933773

# Question 2.4

Consider the case where we double the sample size. Holding all else equal, we would see the following:

* Standard Error: Increase, Decrease, or Stay the Same? Decrease. The se is inversely proportional to n.
* *t*-statistic: Increase, Decrease, or Stay the Same? Increase. As the se gets smaller, the t statistic would increase so long as the effect size remains equal.
* *p*-value: Increase, Decrease, or Stay the Same? Decrease as the likelihood of the answer happening by chance would be reduced.

# Question 3

You are given cleaned data from the [Gerber and Green experiment](https://isps.yale.edu/sites/default/files/publication/2012/12/ISPS00-001.pdf) on the effects of canvassing, direct mail, and phone calls on voter turnout. Your task is to do the following, in either R or Stata (please be sure to share your code!):

1. Load in the dataset, which is stored here on bCourses as “Gerber\_APSR\_Reduced.csv”
2. Conduct a regression with *robust* standard errors, where “vote” is the outcome and our only predictor variable is “treat”
3. Answer the questions below!

*If you are unsure of what code to use, all of this has been included within the Stata and R videos (or can be yielded through a quick Google search!)*

# Question 3.1

Conduct the regression with robust standard errors (using “vote” as the outcome and “treat” as the predictor). Please answer the following:

* What is control group mean? 0.29663831
* What is the treatment group mean? 0.29663831 + 0.08130991 = 0.3779482
* What is the estimated treatment effect? 0.08130991
* What is the estimated standard error on the treatment effect? What is the “t statistic”? 0.002691753, 30.20705
* Is the impact of “treat” on “vote” statistically significant at the .05 level? Oh yeah. The mean of the treatment group is over 30 standard errors away from the mean of the control group. The p value is 4.739272e-200!

# Question 3.2

Conduct a two-sample t-test for the difference in means between the treatment group and the control group. What is the t statistic? Does this differ from your answer to Question 3.1? The T statistic is the consistent with the regression.

# Question 3.3

Consider the case where we double the sample size. Holding all else equal, we would see the following:

* Standard Error: Increase, Decrease, or Stay the Same? Decrease
* *t*-statistic: Increase, Decrease, or Stay the Same? Increase
* *p*-value: Increase, Decrease, or Stay the Same? Decrease