Econ 270 / GSB 603 **Department of Economics**

Fall 2025 **Stanford University**

Problem Set I

Due: Sunday, September 28th, 2025, 6pm.

Be concise but clear as to what numbers you are reporting, and answer in full sentences. You should also hand in supporting code, but all answers should be in a PDF or Word document.

Use the experimental data from the Riverside GAIN experiment data posted on the course website under `riverside_2025.txt`. There are eight variables, the latter five of which are pre-treatment variables:

- the binary treatment indicator (1 for the individuals who were assigned to the training program, and 0 for those who were not);
- earnings one year after the program, in thousands of dollars;
- earnings four years after the program, in thousands of dollars;
- a binary indicator for high school graduation;
- a binary indicator for being female;
- age, in years;
- a binary indicator for having at least one child under the age of six; and
- a binary indicator for being single.

There are 5,418 observations in the sample. We are interested in analyzing the effect of the program on the two outcome variables (earnings one and four years after the program).

1. (a) Test the null hypothesis that there is no effect of the training program whatsoever on earnings the year after the program, assuming the experiment was based on drawing M units at random from the $N = 5418$ total units to be in the treatment group. Use as the test statistic the difference in averages by treatment status:

$$T = \overline{Y}_T - \overline{Y}_C.$$

Report the value of the test statistic as well as the overall p-value. How many draws from the randomization distribution did you use? How did you decide on that number? Include a histogram of the randomization distribution of the test statistic under the null hypothesis.

(b) Use the statistic based on the difference in medians by treatment status. If the number of observations in a group is even, take the average of the middle two values as the sample median. Do you think that this is a better choice for a test statistic? Why or why not?

(c) Choose another statistic that you think may be a good choice in this setting. The statistic may, but need not, depend on the pre-treatment variables. Explain your choice of test statistic.

(d) For each of the statistics, recalculate the p-value as if the experiment consisted of flipping an independent coin with probability of heads equal to 0.2 for each unit to determine whether it was in the treatment group or in the control group. What differences do you see in the p-values? When would it matter whether you use this randomization distribution or the one in part (a)?

2. Theoretical Calculations

(a) Suppose I do an experiment with 100 units. I flip a fair coin for each unit to determine whether it is in the treatment group or in the control group. I end up with 49 units in the treatment group and 51 units in the control group. I can analyze this experiment as conducted, or as an experiment where I randomly picked 49 units out of 100 to be in the treatment group and 51 to be in the control group. Which way do you recommend for the analysis? Explain.