P-Set II

POLI 784 - Spring 2026

2026-02-04

Complete all of the questions below together with your group members. Please submit both (1) a compiled PDF or HTML file with your responses and (2) the `R Markdown`, `Quarto`, etc. file containing both your responses and code as a group. You are encouraged to use `R Studio` projects and virtual environments to complete this p-set, although it is not required. Similarly, you are encouraged to use consistent style and linters when writing code.

Question 1

a. (1 point) Prove that the Horvitz-Thompson estimator reduces to the difference-in-means estimator when a fixed number n of N units is assigned to treatment, with $p_i = p = \frac{n}{N}$. Under which treatment assignment mechanism is this the case? *(Note: For the proof, begin by stating the canonical form of the Horvitz-Thompson estimator and simplify it until you arrive at the difference-in-means estimator. Make sure to show your work!)*

b. (1 point) Prove that the Hajek estimator reduces to the difference-in-means estimator when $p_i = p, \forall i$. Under which treatment assignment mechanisms is this the case? *(Note: For the proof, begin by stating the canonical form of the Hajek estimator and simplify it until you arrive at the difference-in-means estimator. Make sure to show your work!)*

Question 2

a. *(0.5 points)* Simulate the following data generating process in `R`, save the results to a dataframe called `df`, and compute the *true* (sample) ATE. Save the true ATE to an object called `SATE` and print it to console.

1. set seed to `42` to make your results reproducible
2. set sample size to $N = 100$
3. simulate X_i as $X_i \sim U(0, 1)$

4. simulate $Y_i(0)$ as a quadratic function of X_i, where $Y_i(0) = 3 + 2 * X_i + X_i^2 + \varepsilon_i$ and $\varepsilon_i \sim N(0, 1)$
5. simulate the individualistic treatment effect as $\tau_i \sim N(5 * \sin(X_i), 1)$
6. simulate the outcome under treatment as $Y_i(1) = Y_i(0) + \tau_i$
7. assign treatment status $D_i = 1$ randomly for each unit using a Bernoulli trial with probability of success $p = 0.4$
8. reveal the observed outcome Y_i for each unit by saving it in a new column `Y`

b. *(1 point)* Compute the Neyman, Horvitz-Thompson, and Hajek estimator (1) ATE point estimates and (2) respective analytical Neyman variance estimates. Plot these ATE point estimates, together with their asymptotic 95% CIs (derived from the Neyman variance estimates and corresponding critical value from the standard normal distribution for a two-tailed $\alpha = 0.05$ test), next to one another and add horizontal lines for the true SATE and no effect.

c. *(1 point)* Using the data in `df`, simulate additional experiments by randomly reassigning treatment with $p = 0.4$ for all $N = 100$ units and revealing their new observed outcomes 1,000 times. During each simulated trial, compute and save the point estimates and variance estimates from the three estimators in a new dataframe called `df_randomization`. Plot the distributions of the three estimators' point estimates (using density curves) and their averages (using vertical dashed lines) against the true SATE (using a vertical solid line). Distinguish the density curves and lines for the different estimators using different colors. *(Note: You may also want to use one dashed and one solid density curve for overlapping distributions.)*

d. *(0.5 points)* Compute (1) the *true* (based on this simulation) variance (i.e., design-based uncertainty) and (2) the average estimated variance of the three estimators' point estimates. Plot the distributions of the three estimators' variance estimates (using density curves) and their averages (using vertical dashed lines) against the respective true variances (using vertical solid lines). Distinguish the density curves and lines for the different estimators using different colors. *(Note: You may also want to use one dashed and one solid density curve for overlapping distributions.)*

e. *(0.5 points)* Calculate the coverage rates of the analytical 95% confidence intervals derived from the three estimators. To do so, first compute the asymptotic 95% confidence interval around the point estimate for each estimator-trial dyad and then compute the proportion of confidence intervals that contain the true ATE for each estimator. Print the results to console.

f. *(1 point)* Based on the above results, briefly discuss the three estimators: Which estimators are unbiased under the assignment mechanism employed above? Which estimators are the most efficient (i.e., have the lowest average estimated variance)? What does the combination of these two qualities tell us about the likely size of the bias of their respective *estimates in a single trial* relative to one another? Which of the three estimators are conservative for their *true* variance and have appropriate coverage rates? Given the

respective bias, efficiency, conservativeness, and coverage rates of the three estimators, which estimator would statisticians prefer for a single trial (like *2b*) and why?

Question 3

In this exercise, you will be working with data from Santoro and Broockman (2022). In their experiment, the researchers invited subjects to participate in an online video chat with another respondent, who was an out-partisan. During their conversation, subjects talked about what their perfect day would be like. Subjects were then randomly assigned into either the treatment group ($D_i = 1$), where they were informed that the other subject was an out-partisan, or the control group ($D_i = 0$), where they received no additional information. The outcome variable is post-treatment warmth toward out-partisans. You can find the data required for this exercise in `data.csv`.

a. *(0.5 points)* Compute an ATE point estimate using the Neyman estimator, the corresponding analytical Neyman variance estimate, and its 95% confidence interval.

b. *(0.5 points)* Compute bootstrapped ($B = 1,000$, with same N for each b as in the original data) 95% confidence intervals for the ATE point estimate from the Neyman estimator using the percentile method and the normal approximation. Plot the ATE point estimate from the Neyman estimator in a coefficient plot three times: once with the analytical CI, once with the percentile method CI, and once with the normal approximation CI. Is the estimate statistically significant at the $\alpha = 0.05$ level? (*Note: Make sure to set the seed to 42.*)

c. *(0.5 points)* Apply Fisher's randomization test (FRT) ($B = 1,000$, with p equal to the proportion of treated units in the data) to construct the distribution of ATE estimates under the sharp null. Plot this distribution (using a density curve) and its 97.5th percentile (using a dashed vertical line) against the ATE estimate from *3a* (using a solid vertical line). Can you reject the sharp null hypothesis? Is this consistent with the results of your test pertaining to the weak null above?

Question 4

Imagine an experiment in which you are randomly assigning participants to either a condition where they receive a universal basic income (UBI) of 1,000 dollars per month for a year on top of the support they are already receiving from the state ($D_i = 1$) or a condition where they continue to receive monetary support from the state as usual ($D_i = 0$). Your outcome of interest is participants' trust in their country's political institutions (on a scale of 0 to 10), and your theory is that receiving a UBI increases trust. The below exercise revolves around simulating data for this experiment under various conditions and analyzing it.

a. *(0.5 points)* Simulate the following dataset and compute the *true* SATE:

   1. set seed to `42` to make your results reproducible
   2. set sample size to $N = 1,000$
   3. simulate unemployment status as $X_i \sim Binom(N, 0.35)$
   4. simulate the outcome under control for all i as $Y_i(0) = 5 - 2 * X_i + \varepsilon_i$, where $\varepsilon_i \sim N(0, 1)$
   5. simulate the individualistic treatment effect as $\tau_i \sim N(0.75, 0.5)$
   6. simulate the outcome under treatment as $Y_i(1) = Y_i(0) + \tau_i$
   7. make sure to truncate $Y_i(0)$ and $Y_i(1)$ to be within the range of 0 to 10

b. *(0.5 points)* You know from prior studies on trust in political institutions that unemployment status, X_i, is a strong predictor of the outcome. To ensure that your treatment and control groups are balanced on this covariate, you decide to employ stratified random assignment with two strata defined by unemployment status. Assign treatment status using complete random assignment with $p_b = 0.4, \forall B$ within each stratum b (i.e., *exactly* 40% of units within each stratum receive treatment; round up the number of units receiving treatment to the nearest integer within strata) and reveal the observed outcome for each unit. Compute one ATE point estimate using the Neyman estimator and another estimate using the Hajek estimator. Why do we use complete random assignment instead of simple random assignment within strata here?

c. *(0.5 points)* Although you oversampled unemployed individuals on purpose, you still have twice as many employed as unemployed participants in your sample. You realize that by assigning participants within the employed and unemployed strata to treatment with equal p_b, you will end up treating many more employed than unemployed respondents. Your preference, however, would be to treat a disproportionately large share of unemployed participants, both because they stand to gain more from the treatment in their everyday lives (meaning your funding doesn't go toward paying participants who didn't need the money in the first place) and because you suspect that the treatment effect is heterogenous, with individuals in the unemployed stratum seeing larger effects that you would like to statistically detect. Create a new treatment variable using complete random assignment within each stratum, where $p_{employed} = 0.35$ and $p_{unemployed} = 0.7$, and reveal the new observed outcome for each unit. Compute an ATE point estimate using the Neyman, the Hajek, and the weighted DIM estimator, respectively.

d. *(0.5 points)* Plot the five ATE estimates you obtained above in a coefficient plot (without CIs) next to the *true* SATE. Briefly discuss: Based on your theoretical knowledge of ignorable treatment assignment and the above estimators, which of the above combinations of assignment mechanism and estimator are biased and which are unbiased for the (S)ATE and why?