Causal Inference

Homework 2

2026

1 Problem 1 (Neyman's approach & Regression approach)

A completely randomized experiment was carried out in 12 first-grade classrooms spread across one school district to investigate the effectiveness of a new reading curriculum. Six of the classrooms were randomly assigned to use the new curriculum while the other six used the traditional curriculum. Average scores for each class on a standardized test were measured at the end of the year. The data:

new curriculum: 57.6, 43.8, 66.5, 70.9, 72.7, 80.8;

old curriculum: 29.3, 39.3, 51.1, 59.9, 62.2, 63.5.

a. A colleague of the experimenter notes that the sample size would have been larger if she had used individual student exam scores as the response. Explain why the original design is superior.

b. Use Neyman's procedure to estimate the average treatment effect in this set of classrooms. Give an approximate 95% confidence interval for the average treatment effect. Is there evidence to reject the null hypothesis of no treatment effect?

c. Students in the district were all given a pretest at the start of the year. For the new curriculum group the average pretest scores (in the same order as the classes listed above) were 46.6, 63.2, 65.4, 66.0, 73.4, 84.8. For the old curriculum group the average pretest scores were: 50.0, 55.9, 55.8, 69.8, 81.1, 91.5. Based on these pretest scores do you believe that treatments were randomly assigned? Why or why not?

d. One way to use the pretest scores is to focus on changes (post - pre) rather than just looking at the end of year scores. Repeat part (b) using the change in score. How does your answer change?

e. Suppose these 12 classrooms can be viewed as i.i.d. samples from all the classrooms in this district. From now on, let us change our interested estimand to the average treatment effect for all the classrooms in this district (that is τ_{SP}). Use Neyman's procedure to estimate τ_{SP} and provide an approximate 95% confidence interval for τ_{SP}. Compare your results with those in (b).

f. Continue to focus on the estimand in (e) (that is τ_{SP}). Assume that the potential outcomes $Y(1)$ and $Y(0)$ have equal variances, that is $\sigma_t^2 = \sigma_c^2$, where t and c denote the treated (new curriculum) and control (traditional

curriculum) groups, respectively. Please apply linear regression without adjusting for pretest scores to estimate τ_{SP}, and provide an approximate 95% confidence interval for τ_{SP}.

g. (optional) Assume that the potential outcomes $Y(1)$ and $Y(0)$ have different variances, that is $\sigma_t^2 \neq \sigma_c^2$. Please apply linear regression without adjusting for pretest scores to estimate τ_{SP}, and provide an approximate 95% confidence interval for τ_{SP}.

h. Compare the results from (b), (e), and (f). (If you skip (f), compare (b) and (e).) Comment on your findings.

2 Problem 2 (Theory for Neyman's approach)

(1). Consider the education experiment in the previous problem:

a. Write the test statistic for part (b) (the difference in sample means) as an explicit function of the potential outcomes and the assignment indicators for the 12 units.

b. Show that under complete randomization the statistic is unbiased for the average treatment effect in the 12 units.

c. Write the test statistic for part (d) (the difference in the sample mean gain (post - pre) scores) as an explicit function of the potential outcomes, the covariate (pretest scores) and the assignment indicators for the 12 units.

d. Show that under complete randomization this statistic is unbiased for the average treatment effect in the 12 units.

e. If both are unbiased (they should be), then why are the two statistics different in part (b) and (d) of the previous problem?

(2). Please derive an expression for $\mathbb{E}_{\text{SP}}\{Y_i^{obs}(X_i - \mu_X)\}$. Note that this is for the super population. The resulting expression may involve only the sample size and the expectations, variances, and covariances of $X_i, Y_i(0), Y_i(1)$.

3 Problem 3 (Theory for Regression approach)

Consider the model $Y_i^{obs} = \alpha + \tau W_i + \varepsilon_i$ and use it to estimate the super-population average treatment effect τ_{SP}. Please use the following derivations to understand the second and third views discussed in class.

a. Imputation view (Lecture 3 handout, p. 20): Use $\bar{Y}_t$ from data to impute unobserved $Y(1)$, and $\bar{Y}_c$ from data to impute unobserved $Y(0)$. Please prove that $\frac{1}{N}\sum_{i=1}^{N} \hat{\tau}_i = \hat{\tau}^{dif}$.

b. For the result on p. 21 of the Lecture 3 handout, please derive the minimizers τ^* and α^* of $Q(\alpha, \tau) = \mathbb{E}_{SP}[(Y_i^{obs} - \alpha - \tau \cdot W_i)^2]$. (Hint: (1) This is a completely randomized experiment, so W_i has special properties, (2) $(Y_i(0), Y_i(1))$ is i.i.d. sampled from a super-population. (3) You may obtain $\tau^* = \tau_{sp}$ directly, without the intermediate step $\tau^* = \mathbb{E}_{SP}[Y_i^{obs}|W_i = 1] - \mathbb{E}_{SP}[Y_i^{obs}|W_i = 0]$.)