

Causal Inference Assignment 1

Due noon Saturday 6/21

ICPSR Session 1 (June 20, 2025)

1. Refer to the the **acorn** data set that accompanies this assignment. Using the mean of turnout proportions in treatment group precincts, $n_1^{-1}\mathbf{Z}'\mathbf{y}$, as test statistic, simulate its rerandomization distribution under the null hypothesis of strictly no effect, reporting:

- (a) your simulation p -value;
- (b) your simulation approximation of $E[n_1^{-1}\mathbf{Z}'\mathbf{y}]$, the null expected value of the test statistic;
- (c) your simulation approximation of $\text{Var}[n_1^{-1}\mathbf{Z}'\mathbf{y}]$, this test statistic's variance under the null.

Note that n_1 is the number of treated units, n is the total number of units, \mathbf{Z} is the random assignment variable and \mathbf{y} is the observed outcome variable

2. Calculate $E[n_1^{-1}\mathbf{Z}'\mathbf{y}]$, the expected value of the sample mean under the strict null hypothesis of no effect, from first principles — i.e., without simulations — using data in \mathbf{y} .
3. In this setup, $\text{Var}[\bar{y}_1] = \frac{1}{n_1} \frac{n_0}{n} \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n-1}$, where $\bar{y}_1 = n_1^{-1}\mathbf{Z}'\mathbf{y}$ is the mean of y s among the group assigned to treatment ($\{i : Z_i = 1\}$) while $\bar{y} = n^{-1}\mathbf{1}'\mathbf{y}$ is the mean of y over the full study population. If you've seen formulas for the sampling variance of the mean in earlier stats courses, you probably saw

$$\widehat{\text{Var}}[\bar{y}_1] = \frac{s_y^2}{n_1} = \frac{1}{n_1} \frac{\sum_{i=1}^{n_1} (y_i - \bar{y}_1)^2}{n_1 - 1}, \text{ and/or } \text{Var}[\bar{y}_1] = \frac{\sigma_y^2}{n_1} = \frac{1}{n_1} \frac{\sum_{i=1}^N (y_i - \bar{y})^2}{N}.$$

Which gives a larger value, our new $\text{Var}[\bar{y}_1]$ -formula or the $\text{Var}[\bar{y}_1]$ formula immediately above? Explain why it makes sense that the formula we've given would differ in the direction it does from this other formula.

4. Calculate $\text{Var}[n_1^{-1}\mathbf{Z}'\mathbf{y}]$ using the appropriate formula. Determine the error of the simulation-based approximation to this quantity that you reported in question 1, expressing it as a percentage of $\text{Var}[n_1^{-1}\mathbf{Z}'\mathbf{y}]$.
5. Determine the Normal theory approximation to $\Pr(n_1^{-1}\mathbf{Z}'\mathbf{y} \geq n_1^{-1}\mathbf{z}'\mathbf{y})$. (Hints: Use the variance and expected values calculated in 4 and 2 to transform your observed treatment group mean into a corresponding “z-score.” To determine Normal quantiles corresponding to z-scores in R, use `pnorm()`; type `?pnorm` for help. And \mathbf{z}' refers to the observed Acorn data treatment assignment.)
6. A researcher plans to ask six subjects to donate time to an adult literacy program. Each subject will be asked to donate either 30 ($Z = 0$) or 60 ($Z = 1$) minutes. The researcher is considering three methods for randomizing the treatment. Method I is to make independent decisions for each subject, tossing a coin each time. Method C is to write “30” and “60” on three playing cards each, and then shuffle the six cards. Method P tosses one coin for each of the 3 pairs (1, 2), (3, 4), (5, 6), asking for 30 (60) minutes from exactly one member of each pair.
 - a Discuss strengths & weaknesses of each method.
 - b How would your answers to (a) change if $n : 6 \mapsto 600$?
 - c Determine $E[Z_1]$ under each method.
 - d Determine $E[Z_1 + Z_2 + \dots + Z_6]$ under each method.
 - e Calculate $E[\mathbf{Z}'\mathbf{1}]$ under each of the three methods.
 - f For which of the methods does $E[(\mathbf{Z}'\mathbf{1} - E[\mathbf{Z}'\mathbf{1}])^2] = 0$?¹
 - h For two of the three methods, the linearity property of expected value² entails that $E[\frac{\mathbf{Z}'\mathbf{x}}{\mathbf{Z}'\mathbf{1}}] = (x_1 + x_2 + \dots + x_6)/6$. Which two are these, and why doesn't the argument work for the third?

¹I.e., for which does $\text{Var}[\mathbf{Z}'\mathbf{1}] = 0$? (In general, $\text{Var}[V] = E[V - E[V]]^2$.)

²For constant α and β , $E(\alpha + \beta Y) = \alpha + \beta E(Y)$. Also $E(\alpha X + \beta Y) = \alpha E(X) + \beta E(Y)$.