Homework 1 — BST 258: Causal Inference, Theory and Practice

Due February 16th, 2024, 5:00 PM

Christian Testa

Question 1

Done ✓

Find my homework repository at https://github.com/ctesta01/bst258_hw1/

Question 2

Consider a completely randomized experiment (CRE) with $i = 1, \dots, n$ units, m of which are treated, where A_i is 1 when unit i is treated.

a) What is the marginal distribution of the treatment indicator A?

b) What is the joint distribution of A_i and A_j for two units $i \neq j$? (Hint: This amounts to completing a contingency table.)

c) What are $\mathbb{V}(A_i)$ and $\mathrm{Cov}(A_i, A_j)$ for $i \neq j$?

d) The *sample* Average Treatment Effect on the Treated (ATT) is $\theta^{ATT} = \frac{1}{m} \sum_{i=1}^{n} A_i(Y_i(1) - Y_i(0))$. What is the sample ATT in expectation?

Answer to Question 2

a) The marginal distribution is $P(A = 1) = m/n$.
b) The joint distribution is given by the following probabilities, where $i \neq j$:

- $\mathbb{P}(A_i = A_j = 1) = \left(\frac{m}{n}\right) \left(\frac{m-1}{n-1}\right)$,
- $\mathbb{P}(A_i = 1, A_j = 0) = \left(\frac{m}{n}\right) \left(\frac{n-m-1}{n-1}\right)$,
- $\mathbb{P}(A_j = 0, A_j = 1) = \left(\frac{m}{n}\right) \left(\frac{n-m-1}{n-1}\right)$,

- $\mathbb{P}(A_j = A_j = 0) = (\frac{n-m}{n})\left(\frac{n-m-1}{n-1}\right)$.

c) Let $p = m/n$. Then $\mathbb{V}[A_i] = \mathbb{E}[A_i^2] - \mathbb{E}[A_i]^2 = p - p^2 = p(1-p)$. $\text{Cov}(A_i, A_j) = \mathbb{E}[(A_i - \mathbb{E}[A_i])(A_j - \mathbb{E}[A_j])] = (p-p)(p-p) = 0$, where we used that $A_i \perp\!\!\!\perp A_j$ to break apart the expectation.

d) The expectation of the average treatment effect on the treated is given by

$$\begin{aligned}
\mathbb{E}[\theta^{\text{ATT}}] &= \mathbb{E}[\frac{1}{m}\sum_{i=1}^{n} A_i(Y_i(1) - Y_i(0))] \\
&= \frac{1}{m}\sum_{i=1}^{n}\mathbb{E}[A_iY_i(1)] - \frac{1}{m}\sum_{i=1}^{n}\mathbb{E}[A_iY_i(0)] \\
&= \frac{1}{m}\sum_{i=1}^{n} p\mathbb{E}[Y_i(1)] - \frac{1}{m}\sum_{i=1}^{n} p\mathbb{E}[Y_i(0)] \\
&= \frac{1}{n}n(\mathbb{E}[Y_i(1)] - \mathbb{E}[Y_i(0)]) \\
&= \mathbb{E}[Y_i(1) - Y_i(0)] \\
&= \theta^{\text{ATE}}.
\end{aligned}$$

Question 3

Consider an additive treatment effect model, i.e., $\theta = Y_i(1) - Y_i(0)$ for all i, so $Y_i(1) = Y_i(0) + \theta$. Show that $\mathbb{V}(Y_i(1)) = \mathbb{V}(Y_i(0))$ and that the correlation $\rho(Y_i(1), Y_i(0)) = 1$, where expectations are sample expectations, i.e., $\mathbb{E}[Y_i(1)] = \frac{1}{n}\sum_{i=1}^{n} Y_i(1)$.

Answer to Question 3

If $Y_i(1) = Y_i(0) + \theta$, then $\mathbb{V}(Y_i(1)) = \mathbb{V}(Y_i(0) + \theta) = \mathbb{V}(Y_i(0))$. This also implies that $\sigma_{Y_i(1)} = \sigma_{Y_i(0)}$.

The sample correlation will be given by

$$\rho(Y_i(1), Y_i(0)) = \frac{\mathrm{Cov}(Y_i(1), Y_i(0))}{\sigma_{Y_i(1)}\sigma_{Y_i(0)}} = \frac{\mathbb{E}(Y_i(1) - \mathbb{E}[Y_i(1)])(Y_i(1) - \cancel{\theta} - \mathbb{E}[Y_i(1)] + \cancel{\theta})}{\mathbb{V}(Y_i(1))} = \frac{\mathbb{V}(Y_i(1))}{\mathbb{V}(Y_i(1))} = 1.$$

Question 4

It's tea time: (a hologram of) R.A. Fisher places eight cups of tea (with milk) in front of you and asks you to identify which cups had tea poured before milk and vice-versa. Prior to giving you the cups of tea, Fisher poured milk before tea in four of them and tea before milk in the other four. The ordering in which the cups have been served is random. What is the probability that you correctly guess 0, 1, 2, 3, or 4 of all of the cups that had tea poured first?

Answer to Question 4

This scenario can be modelled as a hypergeometrically distributed random variable K representing the number of correct guesses.

Let $\mathbb{M}$ be the set of cups where milk was poured first, and $\mathbb{T}$ be the set of cups where tea was poured first, such that $|\mathbb{M}| = |\mathbb{T}| = 4$.

We are guessing 4 cups for which we speculate milk was poured first. As such, the probability that we guess $K = k$ correctly is given by

$$\mathbb{P}(K = k) = \frac{\binom{\mathbb{M}}{k}\binom{\mathbb{T}}{4-k}}{\binom{\mathbb{M}+\mathbb{T}}{4}}.$$

The exact probabilities are given in the following table:

$K = k$	$\mathbb{P}(K = k)$	$\mathbb{P}(K = k)$ in decimal
0	1/70	0.014
1	16/70	0.229
2	36/70	0.514
3	16/70	0.229
4	1/70	0.014

Question 5

The table below displays the success rates of two distinct, investigational treatments for kidney stones, labeled as A and B. A study enrolls $n = 700$ participants, assigning $n = 350$ individuals to either of the treatment arms, A and B. To summarize data from the study, individuals' kidney stones are categorized as either small or large, and Table 1 is constructed to summarize the success rates of each of the two treatments.

Table 2: Success rates in arms A and B versus kidney stone size

Stone Size	Treatment A	Treatment B
Small Stones	93% (81/87)	87% (234/270)
Large Stones	73% (192/263)	69% (55/80)
Both	78% (273/350)	83% (289/350)

Studying Table 1, your colleague remarks at the discrepancy between the superior overall success rate of treatment B and its relatively lower success (versus treatment A) when stratifying cases by kidney stone size.

a) Describe possible factors that might have contributed to this seemingly contradictory result.

b) Alarmed by this discrepancy, your colleague asks you to further segment the results by reported gender. This newly refined look at the data suggests that for both small and large kidney stones, treatment B is consistently more effective than treatment A across all genders. Construct a hypothetical (i.e., candidate) table that illustrates this (ensure that your candidate table is consistent with the information given in Table 1).

c) What is this phenomenon (it has a name)? What are its broader implications and significance in the interpretation of data?

Answers to Question 5

a) The explanatory factor here is the <u>imbalance within treatment arms</u> between those who had small kidney stones vs. large kidney stones. In treatment A, we see that 263/(263+87)=75% had large kidney stones, while in treatment B we see that only 80/(80+270)=23% had large kidney stones.

Once this is observed, this explains away the apparent "contradictory" results, because the overall success rates should be calculated as weighted-averages of success rates in treating small and large kidney stones with weights according to the number of patients with small/large kidney stones.

E.g., $78\% = (73 \cdot .75) + (93 \cdot .25)$ and $83\% = (69 \cdot .23) + (87 \cdot .77)$.

	Treatment A			Treatment B		
	Men	Women	Total	Men	Women	Total
Small Stones	75.0% (3/4)	94.0% (78/83)	93% (81/87)	78% (100/128)	94.4% (134/142)	87% (234/270)
Large Stones	50.0% (1/2)	73.2% (191/261)	73% (192/263)	53% (10/19)	73.8% (45/61)	69% (55/80)
Both	66.7% (4/6)	78.2% (269/344)	78% (273/350)	75% (110/147)	88.2% (179/203)	83% (289/350)

b) We can come up with cell-sizes for the treatment successes/failures among men and women within each treatment arm and stratified across kidney stone size such that the treatment success rate for treatment B is higher in every sex/gender × kidney stone size strata. See the above table.

c) This is Simpson's paradox, where associations are reversed when looking at data in aggregate vs. within strata.

Question 6

In the middle of a long day, you decide to take a short coffee break. As you meander towards the nearest cafe, a colleague spots you and asks for your help in evaluating the results from a completely randomized experiment (CRE) that they have designed. They've heard that Neyman and Fisher disagreed on the appropriate type of null hypothesis upon which to focus, and they'd like your help in better understanding which—the weak or sharp null—would be most appropriate for the scientific question motivating their experiment. Pressed for time (your day isn't getting shorter), you explain that, in the context of their CRE, the difference-in-means estimator is unbiased for the average treatment effect (ATE):

$$\hat{\gamma}_{ATE} = \frac{1}{n_1} \sum_{i=1}^{n} A_i Y_i - \frac{1}{n_0} \sum_{i=1}^{n} (1 - A_i) Y_i,$$

in which A_i indicates the treatment status for the i^{th} individual and Y_i their outcome; note that $n_1 = \sum_{i=1}^{n} A_i$ and $n_0 = \sum_{i=1}^{n}(1 - A_i)$. Recall the sharp null hypothesis suggested by Fisher:

$$H_0^{sharp} : Y_i(1) - Y_i(0) = 0 \text{ for } i = 1, \dots, n,$$

which states that the potential outcomes do not differ across the treatment conditions $A \in \{0, 1\}$; meanwhile, the weak null hypothesis championed instead by Neyman:

$$H_0^{weak} : \mathbb{E}[Y_i(1)] = \mathbb{E}[Y_i(0)],$$

which states that there is no effect of treatment in expectation, i.e., the mean potential outcomes do not differ.

To illustrate the difference between these null hypotheses for your colleague, you decide to compare the two in a simulation experiment. Note that Fisher's (sharp) null implies Neyman's (weak) null—no difference in the potential outcomes for all $i = 1, \dots, n$ means that there must be no difference in expectation. While Fisher's null implies Neyman's null, the implication need not run in reverse. In your simulation experiment, you will evaluate this—that if there is evidence against Neyman's null, then there should also be evidence against Fisher's. Conduct a simulation study with the following specifications:

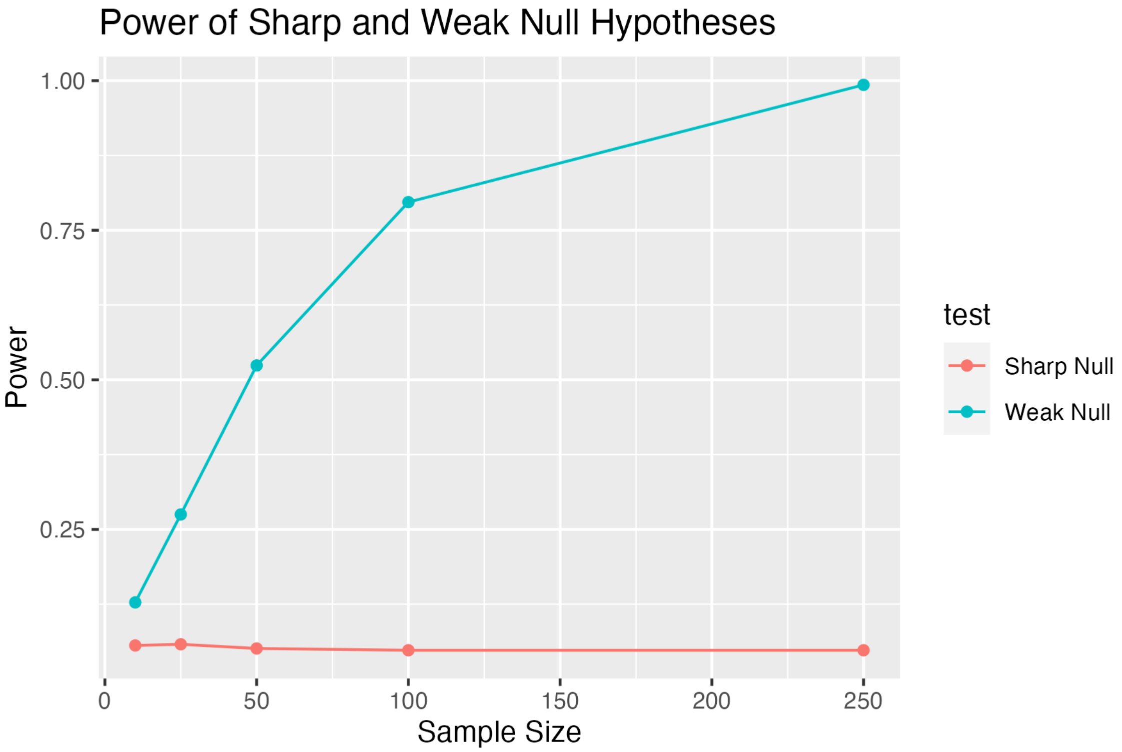
1. Set $n_1 = n_0 = \{10, 25, 50, 100, 250\}$.
2. Independently sample $Y_i(1) \sim N(\mu = 1/10, \sigma^2 = 1/16)$ and $Y_i(0) \sim N(\mu = 0, \sigma^2 = 1/16)$ for $i = 1, \dots, n = n_1 + n_0$. Note that the treatment effect is $\gamma_{ATE} = \mathbb{E}[Y(1) - Y(0)] = 1/10$.
3. Conduct $n_{sim} = 1000$ completely randomized treatment assignments, i.e., sample $A \sim Bern(p = 0.5)$, using the treatment assignment to reveal the potential outcomes.

4. Using the difference-in-means test statistic, evaluate evidence of there being a treatment effect, based on the weak and sharp null hypotheses, at significance level $\alpha = 0.05$. Test the sharp null, conduct at least $B = 10000$ repetitions in generating the test statistic's null distribution.
5. Compute the power for both tests at each of the three sample sizes $n = \{20, 50, 100, 200, 500\}$ and display your results in a figure.
6. Comment on your findings.

> ⚠ Warning
>
> This question requires that you write code to conduct a simulation study. As part of this process, make sure to write code that is modular and reusable, making use of functions or classes as necessary. Make sure to use variable names that are descriptive (e.g., not x or this_var) but concise. Write brief documentation for any functions or classes and any unit tests necessary to ensure that your code is working as expected. Finally, make sure to set a seed for your simulation experiment.

Answers to Question 6

(See code appendix at the end for the code.)



The visualization rendered shows how our power to test the weak null rapidly increases with our sample size, while the power to test the sharp null remains at basically zero. While this is not damning evidence against the sharp-null hypothesis per-se, it does represent a very real practical challenge in using the sharp-null hypothesis. While vigorous debates about whether the sharp null or weak null hypothesis is more appropriate may ensue, certainly most practically minded collaborators will appreciate that at any sample size we will have more power to test the weak null and if that suffices to answer key study questions, we have good reason to argue for testing a weak null as a satisfactory study endpoint.

Question 7

Consider a completely randomized experiment (CRE) having enrolled $i = 1, \ldots, n$ study units, m of which receive the treatment condition. Let A_i be the indicator of the i^{th} unit having received the treatment, and, further, define $Y^1 = \frac{1}{m}\sum_{i=1}^n A_i Y_i$ be the average outcome of the treated units, and similarly define Y^0. You have been tasked with estimating the average treatment effect (ATE), for which it suffices to solve the following least squares program:

$$\min_{\alpha,\beta} \frac{1}{2n}\sum_{i=1}^n (Y_i - \alpha - \beta A_i)^2$$

a) Solve the linear program in (α, β) to obtain solutions for each of the two parameters, denoting these $(\hat{\alpha}, \hat{\beta})$.

b) Is $\hat{\beta}$ a valid estimator of the ATE? Explain your answer.

Answers to Question 7

a) The least squares program can be solved by taking the partial derivatives of the objective function with respect to α and β, and setting them equal to zero. This yields the following system of equations:

$$\begin{aligned} \frac{\partial}{\partial \alpha}\frac{1}{2n}\sum_{i=1}^n (Y_i - \alpha - \beta A_i)^2 = 0 \\ \frac{\partial}{\partial \beta}\frac{1}{2n}\sum_{i=1}^n (Y_i - \alpha - \beta A_i)^2 = 0. \end{aligned}$$

Solving these equations yields the following solutions:

$$\begin{aligned} \hat{\alpha} &= \bar{Y} - \hat{\beta}\bar{A} \\ \hat{\beta} &= \frac{\sum A_i Y_i - \hat{\alpha}\sum A_i}{\sum A_i^2}. \end{aligned}$$

Now note that $A_i^2 = A_i$ since A_i is binary, and specifically $\sum A_i = m$, so we have that $\hat{\beta} = \frac{mY^1}{m} - \hat{\alpha} = Y^1 - \hat{\alpha}$.

b) As long as we have that $m > 0$, then yes, $\hat{\beta}$ is a valid estimator for the ATE. Since we are in the completely randomized experimental setting, we have that the necessary assumptions for $\hat{\beta}$ to be a valid estimator hold: namely, randomization, consistency, and positivity.

Since the treatment is binary, $\hat{\beta}$ captures the average difference between the treated and untreated group, and since we have randomized the treatment, we know that there is no potential for confounding.

Threats to the validity of $\hat{\beta}$ include if there were treatment non-compliance, interference, or spillover effects.

Code Appendix

Code for Question 6

```
set.seed(1234)
library(magrittr)
library(dplyr)
library(ggplot2)

n1 <- c(10, 25, 50, 100, 250)
n0 <- n1
n_sim <- 1000L
B <- 10000L
alpha <- 0.05

# Simulate data and test the sharp and weak nulls
simulate_and_test_null_hypotheses <- function(B, n1, n0) {

  sharp_p_values <- numeric(length=B)
  weak_p_values <- numeric(length=B)

  for (i in 1:B) {
    # simulate potential outcomes
    Y1 <- rnorm(n1+n0, mean = 1/10, sd = 1/4)
    Y0 <- rnorm(n1+n0, mean = 0, sd = 1/4)
    A <- rbinom(n1+n0, size = 1, prob = 0.5)

    test_stat <- (1/sum(A)) * sum(A * Y1) -
      (1/sum(1 - A)) * sum((1 - A) * Y0)

    # simulate the null distribution
    null_dist <- replicate(B, {
      Y1_perm <- Y1
      Y0_perm <- Y0
      A_perm <- sample(A)
      # test statistic calculated on the null distribution:
      (1/sum(A_perm)) * sum(A_perm * Y1_perm) -
        (1/sum(1 - A_perm)) * sum((1 - A_perm) * Y0_perm)
    })

    # calculate and store p-values
```

```r
    sharp_p_values[i] <- mean(null_dist > test_stat)
    weak_p_values[i] <- tryCatch({
      t.test(Y1[A == 1], Y0[A == 0])$p.value },
      error = function(e) { NA }) # handles when A is all 0 or all 1
  }

  return(data.frame(
    sharp_p_values_w_power = mean(sharp_p_values < alpha),
    weak_p_values_w_power = mean(weak_p_values < alpha)))
}

# Run simulations with different sample sizes
results <- list()
for (i in 1:length(n1)) {
  results[[length(results)+1]] <-
    bind_cols(n = n1[i], simulate_and_test_null_hypotheses(B, n1[i], n0[i]))
}

# Construct data frame of results
results_df <- bind_rows(results)

# Visualize the power curve
results_df |>
  tidyr::pivot_longer(
    cols = c(sharp_p_values_w_power, weak_p_values_w_power),
    names_to = "test",
    values_to = "power"
  ) |>
  mutate(test = ifelse(
    test == "sharp_p_values_w_power", "Sharp Null", "Weak Null")) |>
  ggplot(aes(x = n, y = power, color = test, group = test)) +
  geom_line() +
  geom_point() +
  xlab("Sample Size") +
  ylab("Power") +
  ggtitle("Power of Sharp and Weak Null Hypotheses")
```