Problem Set 1

Moonsoo Kim

01/21/2025

Contents

1 Potential Outcomes Notation

1.1 Explain the notation $Y_i(1)$.

Answer: $Y_i(1)$ refers to the potential outcome for individual i if they receive the treatment (denoted by 1).

1.2 Explain the notation $Y_1(1)$.

Answer: $Y_1(1)$ is similar to $Y_i(1)$ but focuses on individual 1 specifically. It denotes the potential outcome that would occur for the first individual if they receive the treatment

1.3 Explain the notation $E[Y_i(1)|d_i = 0]$.

Answer: represents the expected value of the potential outcome under treatment ($Y_1(1)$) for individuals who did not receive the treatment (d_i=0). This arises in cases where researchers use observational data and want to estimate the treatment effect for untreated individuals.

1.4 Explain the difference between the notation $E[Y_i(1)]$ and $E[Y_i(1)|d_i = 1]$

Answer: $E[Y_i(1)]$ represents the expected value of the potential outcome under treatment across the entire population whereas $E[Y_i(1)|d_i = 1]$ represents the expected value of the potential outcome under treatment, but only for individuals who actually received the treatment (d_i=1)

2 Potential Outcomes and Treatment Effects

2.1 Illustration

Use the values in the table below to illustrate that $E[Y_i(1)] - E[Y_i(0)] = E[Y_i(1) - Y_i(0)]$.

```
# Compute means
E_Y1 <- mean(table$y_1) # Mean of Y_i(1)
E_Y0 <- mean(table$y_0) # Mean of Y_i(0)
E_tau <- mean(table$tau) # Mean of tau

# Print results
list(
  "E[Y_i(1)]" = E_Y1,
  "E[Y_i(0)]" = E_Y0,
  "E[Y_i(1)] - E[Y_i(0)]" = E_Y1 - E_Y0,
  "E[Y_i(1) - Y_i(0)]" = E_tau
)
```

```
## $`E[Y_i(1)]`
## [1] 15
##
## $`E[Y_i(0)]`
## [1] 13
##
## $`E[Y_i(1)] - E[Y_i(0)]`
## [1] 2
##
## $`E[Y_i(1) - Y_i(0)]`
## [1] 2
```

Answer: Thus the equality $E[Y_i(1)] - E[Y_i(0)] = E[Y_i(1) - Y_i(0)]$ holds true, with both sides equal to 2.

2.2 Data Possibilities

Is it possible to collect all necessary values and construct a table like the one provided in real life? Explain why or why not?

Answer: No, because of the fundamental problem of causal inference, which states that we cannot observe both potential outcomes $Y_i(0)$ and $Y_i(1)$ for the same individual at the same time.

3 Visual Acuity

3.1 Treatment effect

Compute the individual treatment effect for each of the ten children.

```
# Add a column for the individual treatment effect
d[, tau := y_1 - y_0]
print(d)
```

```
##     child   y_0   y_1   tau
##     <int> <num> <num> <num>
##  1:     1   1.2   1.2   0.0
##  2:     2   0.1   0.7   0.6
##  3:     3   0.5   0.5   0.0
##  4:     4   0.8   0.8   0.0
##  5:     5   1.5   0.6  -0.9
##  6:     6   2.0   2.0   0.0
##  7:     7   1.3   1.3   0.0
##  8:     8   0.7   0.7   0.0
##  9:     9   1.1   1.1   0.0
## 10:    10   1.4   1.4   0.0
```

Answer: The individual treatment effects (tau) are calculated as y_1 - y_2.

3.2 Story time

Tell a "story" that could explain this distribution of treatment effects. In particular, discuss what might cause some children to have different treatment effects than others.

```
# Use this code chunk to show your code work (if needed)
```

Answer: ...

3.3 True ATE

For this population, what is the true average treatment effect (ATE) of playing outside.

```
# Calculate the true average treatment effect (ATE)
true_ATE <- mean(d$tau)
print(true_ATE)
```

```
## [1] -0.03
```

Answer: -0.03

3.4 Even-Odd split

Suppose we are able to do an experiment in which we can control the amount of time these children play outside for three years. We happen to randomly assign the odd-numbered children to treatment and the even-numbered children to control. What is the estimate of the ATE you would reach under this assignment? (Please describe your work.)

```
# Split the children into treatment (odd) and control (even)
d[, assignment := ifelse(child %% 2 == 1, "treatment", "control")]

# Calculate the observed ATE
observed_ATE <- mean(d[assignment == "treatment", y_1]) - mean(d[assignment == "control", y_0])
print(observed_ATE)
```

```
## [1] -0.06
```

Answer: -0.06. The observed ATE under this split is calculated as the difference in means between the treatment group (y_1 for odd-numbered children) and the control group (y_0 for even-numbered children)

3.5 Biased or Unbiased?

How different is the estimate from the truth? In your own words, why is there a difference? Does this mean that the estimator is a biased or an unbiased estimator? Does this mean that the estimate is biased or unbiased?

```
# True ATE
true_ATE <- mean(d$y_1 - d$y_0)

# Even-odd split assignment
treated <- d[child %% 2 == 1] # Odd-numbered children assigned to treatment
control <- d[child %% 2 == 0] # Even-numbered children assigned to control

# Estimated ATE under even-odd split
estimated_ATE <- mean(treated$y_1) - mean(control$y_0)

# Difference between estimate and truth
bias <- estimated_ATE - true_ATE

# Output results
list(
  true_ATE = true_ATE,
  estimated_ATE = estimated_ATE,
  bias = bias
)
```

```
## $true_ATE
## [1] -0.03
##
## $estimated_ATE
## [1] -0.06
##
## $bias
## [1] -0.03
```

Answer: The estimate might differ from the true ATE due to random variation in the assignment. This does not mean the estimator is biased—it is still an unbiased estimator because the expected value of the estimate equals the true ATE.

3.6 How many splits are possible?

We just considered one way (odd-even) an experiment might split the children. How many different ways (every possible way) are there to split the children into a treatment versus a control group (assuming at least one person is always in the treatment group and at least one person is always in the control group)?

```
# Compute the total number of possible splits
n <- 10  # Total number of children
total_splits <- sum(sapply(1:(n-1), function(k) choose(n, k)))
print(total_splits)
```

```
## [1] 1022
```

Answer: 1022

3.7 Thinking about your assignment strategy

Given there are as many ways to assign as you answered in the last sub-question, can you provide a rationale for why you might prefer one assignment strategy over another?

For concreteness, suppose that either (a) you can have a treatment assignment where one and only one of the kids is randomly assigned to treatment; or (b) you can have a treatment assignment where exactly five of the kids are randomly assigned to treatment.

As a small hint, you might note that $\left\{ \left[\sum_{i=1}^{n} Y_i(1) | d_i = 1 \right] - \left[\sum_{j=1}^{n} Y_j(0) | d_j = 0 \right] \right\} \equiv ATE$ is an estimator and there are some properties of estimators that we care about.

To make the question tractable, suppose that if you were to increase the size of the population procedure (a) would keep a single kid in treatment, while procedure (b) would keep 50% of the sample in treatment and 50% of the sample in control.

Answer:

A single individual in treatment leads to high variance and less reliable estimates. Assigning exactly half the sample to treatment (n/2) is generally preferable because it: - Minimizes the variance of the estimator. - Provides a more balanced comparison, reducing the influence of outliers. - Ensures greater statistical power for detecting differences.

3.8 Compute the MSE of these two designs

Because you have the entire population of kids, their entire scheduled of potential outcomes, and two proposed sampling procedures: conduct a simulation study. First, calculate all of the possible treatment effects that you might observe under each design. Then, compute the mean-squared error of each design. Which design – the one where you have a single kid in treatment, or the one where you have five kids in treatment – produces a lower MSE? (Hint the `combn` function might help you with your subsetting.)

```
# Define functions for MSE calculation
compute_mse <- function(d, group_size) {
  combs <- combn(10, group_size)
  mse <- mean(apply(combs, 2, function(treated_idx) {
    d[, assignment := ifelse(child %in% treated_idx, "treatment", "control")]
    obs_ate <- mean(d[assignment == "treatment", y_1]) - mean(d[assignment == "control", y_0])
    (obs_ate - true_ATE)^2
  }))
  return(mse)
}

# MSE for 1 treated and 9 control
mse_1 <- compute_mse(d, 1)

# MSE for 5 treated and 5 control
mse_5 <- compute_mse(d, 5)

print(list(mse_1 = mse_1, mse_5 = mse_5))
```

```
## $mse_1
## [1] 0.2339272
##
## $mse_5
```

```
## [1] 0.08987778
```

Answer: 0.2339272, 0.08987778

3.9 Observational study

Suppose that we decide it is too hard to control the behavior of the children, so we do an observational study instead. Children 1-5 choose to play an average of more than 10 hours per week from age 3 to age 6, while Children 6-10 play less than 10 hours per week. Compute the difference in means from the resulting observational data.

```
# Observational ATE calculation
obs_ate <- mean(d[child <= 5, y_1]) - mean(d[child > 5, y_0])
print(obs_ate)
```

```
## [1] -0.54
```

Answer: -0.54

3.10 Observational ATE

Compare your answer in `Observational study` to the true ATE. In your own words what causes the difference? Does this mean that the estimator is a biased or an unbiased estimator? Does this mean that the estimate is biased or unbiased?

```
# Observational ATE calculation
# Children 1-5 choose to play more than 10 hours, while 6-10 play less
obs_ate <- mean(d[child <= 5, y_1]) - mean(d[child > 5, y_0])
print(obs_ate)
```

```
## [1] -0.54
```

Answer:

Selection bias: Children 1-5 (who chose to play more) may systematically differ from children 6-10 in unobserved ways (e.g., initial visual acuity, genetic predispositions, environment). And the estimator is biased in this case because it does not consistently produce the true ATE due to systematic differences in the treated and untreated groups. And lastly, the specific estimate we computed here reflects that bias. It is a single realization of the biased estimator.

4 Randomization and Experiments

The following questions can be a little bit challenging. This is because the argument that you are being asked to make is based on the rote application of a definition. To begin with, it is useful for you to define what you mean when you write about either *an experiment* or *an observational study*. Then, with these definitions on hand, use the definitions to answer the following questions.

4.1 Define your terms

- **An experiment is:** a study in which the researcher controls the assignment of treatment to participants using a randomization process. This random assignment ensures that the treatment and control groups are comparable in terms of observed and unobserved characteristics.
- **An experiment provides the following statistical guarantees:**

1. Unbiased estimates of treatment effects if conducted properly.
2. Eliminates confounding due to random assignment.
3. Allows causal inference by ensuring comparability between treatment and control groups.

- **An observational study is:** a study in which the researcher observes and measures variables of interest without controlling or randomizing the assignment of treatments. Assignment to treatment or control is determined by natural conditions, choices, or external factors.
- **An observational study provides the following statistical guarantees:**

1. No guarantees of unbiased treatment effect estimates.
2. Susceptible to confounding from unobserved variables.
3. Causal inference is limited unless strong assumptions (e.g., no unobserved confounding) are justified.

4.2 Does a random, iid sample produce an unbiased treatment effect estimate?

Assume that a researcher takes a random sample of elementary school children and compares the grades of those who were previously enrolled in an early childhood education program with the grades of those who were not enrolled in such a program. Is this an experiment, an observational study, or something in between?

Answer: This is an observational study because the researcher does not control or randomize the assignment of treatment (early childhood education). Instead, the treatment (enrollment in the program) was determined by external factors (e.g., parental choice or availability).

4.3 What if an official agency produces the idd sample?

Assume that the researcher works together with an organization that provides early childhood education and offer free programs to certain children. However, which children that received this offer was not randomly selected by the researcher but rather chosen by the local government. (Assume that the government did not use random assignment but instead gives the offer to students who are deemed to need it the most) The research follows up a couple of years later by comparing the elementary school grades of students offered free early childhood education to those who were not. Is this an experiment, an observational study, or something in between? Explain!

Answer: This is an observational study, not an experiment because although the researcher collaborates with an official agency, the assignment to treatment is still not random; it is determined by the government's criteria (e.g., children deemed most in need). This introduces potential confounding variables that may bias the treatment effect estimate.

4.4 What if someone else randomly assigns

If the government assigned students to treatment and control by "coin toss", rather than simply sampling the population, would you say that the study is experimental or observational? Why? What, if any guarantees does this process provide?

Answer: This is an experiment because the treatment assignment is random. Random assignment ensures that treatment and control groups are comparable in expectation, eliminating confounding and allowing for causal inference.

5 Moral Panic

5.1 Explain the statements

Explain the statement $E[Y_i(0)|D_i = 0] = E[Y_i(0)|D_i = 1]$ in words. First, state the rote English language translation. Second, tell us the *meaning* of this statement. A full points solution will use the term "potential outcomes" twice.

Answer: 1. The expected test score when students listen to death metal at least once per week ($Y_i(0)$), the potential outcome under control) is the same whether the students were actually in the group that listens to death metal at least once per week ($D_i = 0$) or in the group that listens less than once per week ($D_i = 1$). 2. The potential outcome for test scores under the "control" condition ($Y_i(0)$) is independent of treatment assignment (D_i). Students' test scores in the absence of listening to death metal (their potential outcomes) would be the same, on average, regardless of whether they were in the "treatment" or "control" group. This independence is critical for unbiased estimation of causal effects because it ensures that treatment and control groups are comparable with respect to their potential outcomes.

5.2 Can you believe it

Do you expect that this circumstance actually matches with the meaning that you've just written down? Why or why not?

Answer: No, it is unlikely that this circumstance actually matches the meaning written above. The independence condition $E[Y_i(0)|D_i = 0] = E[Y_i(0)|D_i = 1]$. assumes there are no systematic differences between students who choose to listen to death metal frequently and those who do not. However, in reality, these groups likely differ in unobserved characteristics (e.g., personality, study habits, socioeconomic background) that influence both their music preferences and their test performance. Such differences would violate the independence assumption, leading to confounding and biased estimates of the treatment effect.