Causal Inference Practical

Qingyuan Zhao (Statistical Laboratory)*

February 14, 2021

1 Randomization in design and analysis

Randomized controlled trials (RCTs) are widely regarded as the "gold standard" of establishing causality. The often forgotten component of the RCTs is that they can be objectively analyzed by randomization test. Haines and coauthors investigated the impact of disinvestment from weekend allied health services. We will use their dataset to explore the concept of randomization in the design and analysis of an experiment.

1. [Group] Skim through the abstract and read the section called "Design" of their article. Then answer the following questions: What is the name of the design of the experiment in this study? How was it carried out?
2. Download the patient-level data, then run the following code in R (you may need to install the `readxl` package first by `install.packages("readxl")`). What does the second line do?

   ```
   data <- readxl::read_excel("S2 Data.xlsx")
   data <- subset(data, hospital == "Dandenong" & study1 == 1)
   ```

3. Unfortunately, this dataset is not very well annotated. The columns `index_ward` and `sw_step` contain the identifiers for hospital ward and time step (in calendar month), respectively. In which order do you think the wards crossed over to no weekend health services? You may find the following `R` code useful.

   ```
   table(data[, c("index_ward", "sw_step", "no_we_exposure")])
   ```

4. Construct a vector called `cross_over_realized` that contains the calendar month in which the 6 hospital wards crossed over. Then use the following code to define the treatment and outcome of interest ("los" is short for length of stay).

   ```
   data$treatment_status <- as.numeric(data$sw_step >= cross_over_realized[data$index_ward])
   data$log_acute_los <- log(data$acute_los)
   ```

5. [Group] Execute the following code in your `R` session. Then comment on the two interval estimators of the treatment effect (of no weekend health services on log length of stay).

   ```
   confint(lm(log_acute_los ~ treatment_status, data))
   confint(lm(log_acute_los ~ treatment_status + as.factor(index_ward), data))
   ```

6. [Group] Next, we explore the randomization analysis of this dataset. First, use potential outcomes to define the null hypothesis that stopping weekend health services has no effect whatsoever. Notice that the treatment is not the same as the variable randomized in the experiment (crossover order). What assumption do you incurred while defining your null hypothesis? Give an example in which this assumption is not satisfied.

*qyzhao@statslab.cam.ac.uk

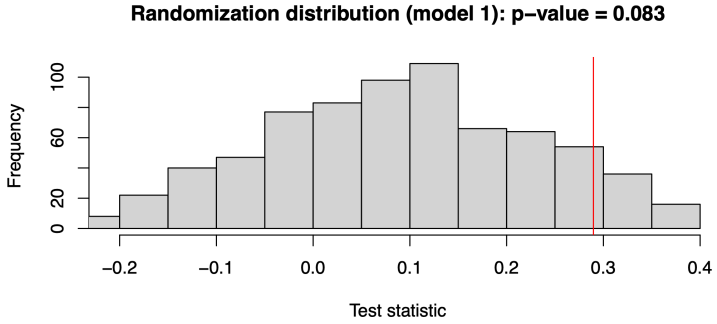
7. Read the following code, then execute it in your `R` session (you may need to install the package `combinat` which contains a function `permn` that generates all the permutations of a vector). For your reference, the expected output is included.

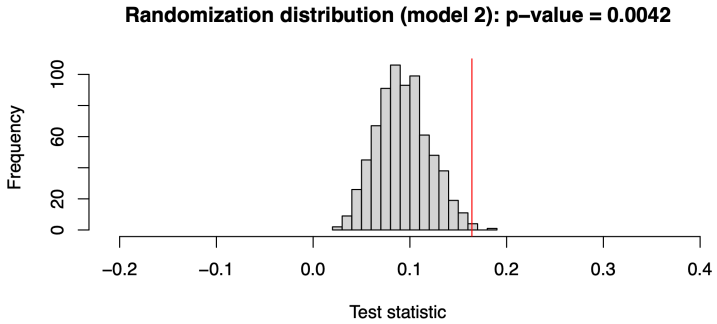
```
get_statistic <- function(index_ward,
                          sw_step,
                          log_acute_los,
                          cross_over) {
  treatment_status <- sw_step >= cross_over[index_ward]
  c(lm(log_acute_los ~ treatment_status)$coef[2],
    lm(log_acute_los ~ treatment_status + as.factor(index_ward))$coef[2])
}

T_obs <- get_statistic(data$index_ward, data$sw_step,
                       data$log_acute_los, cross_over_realized)

T_random <- sapply(combinat::permn(2:7),
                   get_statistic,
                   index_ward = trial1$index_ward,
                   sw_step = trial1$sw_step,
                   log_acute_los = trial1$log_acute_los)

par(mfrow = c(2, 1))
for (m in 1:2) {
  hist(T_random[m, ], 20,
       main = paste0("Randomization distribution (model ", m, "): ",
                     "p-value = ", signif(mean(T_random[m, ] >= T_obs[m]), 2)),
       xlab = "Test statistic", xlim = range(T_random))
  abline(v = T_obs[m], col = "red")
}
```





8. [Group] Explain what the code above does and discuss the results. Here are some points you may consider
    - How do the two randomization tests compare with each other?
    - How do the randomization tests compare with the normal linear model? How would you interpret their results?
    - The randomization distribution of the second test statistic is clearly not centered at 0. Why?
    - How can you "invert" the randomization tests to obtain an interval estimator of the treatment effect?

2 [Group] Causal diagrams and causal identification

In this group exercise, we will read the article titled "A Note on Posttreatment Selection in Studying Racial Discrimination in Policing".

1. Read the section "Review". Using Figure 1, explain the causal inference problem under investigation. Why do the authors say "the naive treatment effect Δ [in equation (1)] can be quite misleading when used to represent the causal effect of race on police violence"? *Hint:* M is a collider.
2. Use your own words to explain Assumption 1.

You may skip the section "Average treatment effects conditional on the mediator".

1. Read the first half of the section "A new estimator for the causal risk ratio", then use your own words to explain Equation 3. Can we use the police admin data to estimate the "bias factor" in this equation?
2. Read the first three paragraphs in "A reanalysis of the NYPD stop-and-frisk dataset", then use your own words to explain the results in Table 1. Use Equation 3 and Figure 2 to explain the large discrepancy between the naive and adjusted estimators in Table 1.