

Practical: Randomization inference

Qingyuan Zhao (Statistical Laboratory)*

July 21, 2023

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. Skim through the abstract and read the section called "Design" in 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?

Answer: This study is using a "stepped-wedge cluster-randomized" design. In the first trial, one of the 6 wards in each hospital is randomly chosen in each calendar month to stop weekend allied health services.

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("pmed.1002412.s002.xlsx")
data <- subset(data, hospital == "Dandenong" & study1 == 1)
```

Answer: The second line selects all patients involved in the first trial in the Dandenong hospital.

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 exposure and outcome of interest ("los" is short for length of stay).

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

Answer: `cross_over_realized = c(3, 5, 7, 6, 2, 4)`.

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

Answers: It appears that the confidence intervals obtained by the two linear models do not overlap.

*qyzhao@statslab.cam.ac.uk

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 exposure is not the same as the treatment being randomized in the experiment (crossover order). What assumption do you incur while defining your null hypothesis? Give an example in which this assumption is not satisfied.

Answer: Let $Y_i(0)$ denote the potential outcome of patient i when there is weekend health services, and $Y_i(1)$ denote the potential outcome when there is no weekend health services. The null hypothesis is that $Y_i(0) = Y_i(1)$ for all i . When using these potential outcomes, we implicitly assumed that there is no time-varying effect of terminating weekend health services (this would be violated if, for example, the hospital staff need some time to adjust to the new system.)

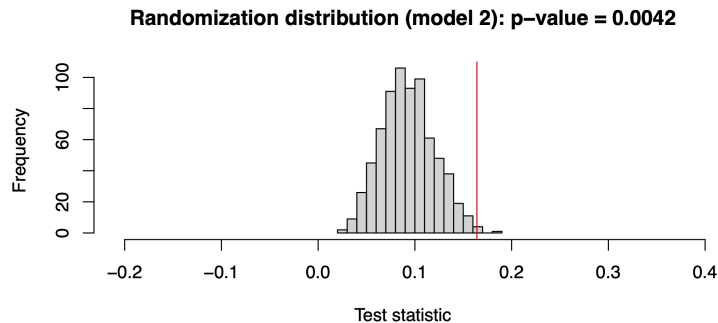
7. Read the following code, then execute it in your R session (you may need to install the package `combinat` which contains a function `permn` that generates all the permutations of a vector). For your reference, the expected output is included.

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

Answer (to some of the questions): The randomization distribution of the second test statistic is much more concentrated, and the second test gives a much smaller p-value. The normal linear model makes a strong assumption on the relationship of the variables, and the inference is based on the assumed model. On the other hand, the randomization test is based on the randomization introduced by the experimenter.