The Completely Randomized Experiment and the Fisher Randomization Test

Peng Ding, Email: pengdingpku@berkeley.edu

1. Completely Randomized Experiment

An experiment with n units, n_1 receive the treatment and n_0 receive the control.

Treatment assignment mechanism:

$$\text{pr}(\boldsymbol{Z} = \boldsymbol{z}) = 1 \Big/ \binom{n}{n_1},$$

where $\boldsymbol{z} = (z_1, \ldots, z_n)$ satisfies $\sum_{i=1}^n z_i = n_1$ and $\sum_{i=1}^n (1 - z_i) = n_0$. Here we treat the Science Table as fixed, or equivalently, the potential outcome vector under treatment $\boldsymbol{Y}(1) = (Y_1(1), \ldots, Y_n(1))$ and the potential outcome vector under control $\boldsymbol{Y}(0) = (Y_1(0), \ldots, Y_n(0))$ as fixed. If not, we can condition on them and the treatment assignment mechanism becomes

$$\text{pr}\{\boldsymbol{Z} = \boldsymbol{z} \mid \boldsymbol{Y}(1), \boldsymbol{Y}(0)\} = 1 \Big/ \binom{n}{n_1}.$$

Fisher (1935) pointed out the following advantages of randomization:

(1) It creates comparable treatment and control groups on average.

(2) It serves as a "reasoned basis" for statistical inference.

Point (1) is very intuitive, and most people understand it well. Point (2) is more subtle: What Fisher meant in his book is that randomization justifies a statistical test. This statistical test is call the Fisher Randomization Test (FRT).

2. Fisher Randomization Test

Fisher (1935)'s null hypothesis is

$$H_{0\mathrm{F}} : Y_i(1) = Y_i(0) \text{ for all units } i = 1, \ldots, n.$$

Rubin (1980) called it the sharp null hypothesis in the sense that it can determine the whole Science Table based on the observed data: $\boldsymbol{Y}(1) = \boldsymbol{Y}(0) = \boldsymbol{Y} = (Y_1, \ldots, Y_n)$, the vector of the observed outcomes. Other researchers called it the strong null hypothesis.

Conceptually, FRT works for any test statistic

$$T = T(\boldsymbol{Z}, \boldsymbol{Y}) = T(\boldsymbol{Z}, \boldsymbol{Y}(1), \boldsymbol{Y}(0)). \tag{1}$$

The first identity in (1) states that the test statistic is a function of the observed data, and the second identity states that the test statistic is a function of the treatment vector $\boldsymbol{Z}$ and the fixed Science Table. So the only random component in the test statistic T is the treatment vector $\boldsymbol{Z}$.

In a completely randomized experiment, $\boldsymbol{Z}$ is uniform over the set

$$\{\boldsymbol{z}^1, \cdots, \boldsymbol{z}^M\}$$

where $M = \binom{n}{n_1}$, and the $\boldsymbol{z}^m$'s are all possible vectors with n_1 1's and n_0 0's. As a consequence, T is uniform over the set (with possible duplications)

$$\{T(\boldsymbol{z}^1, \boldsymbol{Y}), \cdots, T(\boldsymbol{z}^M, \boldsymbol{Y})\}.$$

That is, the distribution of T is known due to the design of the completely randomized experiment. We will call this distribution of T the randomized distribution.

If larger values are more extreme for T, we can use the following tail probability to measure the extremeness of the test statistic with respect to its randomization distribution:

$$p = M^{-1} \sum_{m=1}^{M} I\{T(\boldsymbol{z}^m, \boldsymbol{Y}) \geq T(\boldsymbol{Z}, \boldsymbol{Y})\}, \tag{2}$$

which is called the p-value by Fisher. In practice, M is often to large ($n = 100, n_1 = 50, M > 10^29$), and it is computationally infeasible to enumerate all possible values of the treatment vector. We often approximate p by Monte Carlo. To be more specific, we take random draws from the possible values of the treatment vector, or, equivalently, we randomly permute $\boldsymbol{Z}$, and approximate p by

$$p \approx R^{-1} \sum_r I\{T(\boldsymbol{z}^r, \boldsymbol{Y}) \geq T(\boldsymbol{Z}, \boldsymbol{Y})\}, \tag{3}$$

where the $\boldsymbol{z}^r$'s the R random permutations of $\boldsymbol{Z}$.

Note that the p-value in (2) is finite-sample exact for any choice of test statistic, and the p-value in (3) has Monte Carlo error decreasing fast with R. Because the calculation of the p-value in (3) involves permutations, the FRT is sometimes called the permutation test.

3. Canonical choices of the test statistic

From the above discussion, the FRT generates finite-sample exact p-value for any choice of test statistic. This is a feature of the FRT. However, this feature should not encourage arbitrary choice of the test statistic. Below I will review some canonical choices.

Example 1 (difference-in-means)**.** The difference-in-means statistic is

$$\hat{\tau} = n_1^{-1} \sum_{Z_i=1} Y_i - n_0^{-1} \sum_{Z_i=0} Y_i = n_1^{-1} \sum_{i=1}^n Z_i Y_i - n_0^{-1} \sum_{i=1}^n (1 - Z_i) Y_i \equiv \hat{\bar{Y}}(1) - \hat{\bar{Y}}(0).$$

Under $H_{0\mathrm{F}}$, it has mean

$$E(\hat{\tau}) = n_1^{-1} \sum_{i=1}^n E(Z_i) Y_i - n_0^{-1} \sum_{i=1}^n E(1 - Z_i) Y_i = 0$$

and variance

$$\begin{aligned}
\text{var}(\hat{\tau}) &= \text{var}\left\{ n_1^{-1} \sum_{i=1}^{n} Z_i Y_i - n_0^{-1} \sum_{i=1}^{n} (1 - Z_i) Y_i \right\} \\
&= \text{var}\left(\frac{n}{n_0} \frac{1}{n_1} \sum_{i=1}^{n} Z_i Y_i \right) \\
&= \frac{n^2}{n_0^2} \left(1 - \frac{n_1}{n} \right) \frac{s^2}{n_1} \\
&= \frac{n}{n_1 n_0} s^2,
\end{aligned}$$

where the above calculations follow from a Lemma of survey sampling with the sample mean and variance of all the observed outcomes defined as

$$\bar{Y} = n^{-1} \sum_{i=1}^{n} Y_i, \quad s^2 = (n-1)^{-1} \sum_{i=1}^{n} (Y_i - \bar{Y})^2.$$

Furthermore, the randomization distribution of $\hat{\tau}$ is approximately normal due to the finite population central limit theorem:

$$\frac{\hat{\tau}}{\sqrt{\frac{n}{n_1 n_0} s^2}} = \frac{n_1^{-1} \sum_{Z_i=1} Y_i - n_0^{-1} \sum_{Z_i=0} Y_i}{\sqrt{\frac{n}{n_1 n_0 (n-1)} \sum_{i=1}^{n} (Y_i - \bar{Y})^2}} \xrightarrow{\text{d}} \text{N}(0, 1)$$

Then we can calculate an approximate p-value which is close to the p-value from t.test() with "var.equal = TRUE".

The observed data are $\{Y_i : Z_i = 1\}$ and $\{Y_i : Z_i = 0\}$, so the problem is essentially a two-sample problem. The FRT with $\hat{\tau}$ effectively uses a pooled variance ignoring the heteroskedasticity between these two groups. In classical statistics, the two-sample problem with heteroskedastic Normal outcomes is called the Behrens–Fisher problem. In the Behrens–Fisher problem, a standard choice of the test statistic is the studentized statistic below.

Example 2 (studentized statistic)**.** The studentized statistic is

$$t = \frac{\hat{\tau}}{\sqrt{\frac{\hat{S}^2(1)}{n_1} + \frac{\hat{S}^2(0)}{n_0}}},$$

where

$$\hat{S}^2(1) = (n_1 - 1)^{-1} \sum_{Z_i=1} \{Y_i - \hat{\bar{Y}}(1)\}^2, \quad \hat{S}^2(0) = (n_0 - 1)^{-1} \sum_{Z_i=0} \{Y_i - \hat{\bar{Y}}(0)\}^2.$$

We will motivate this statistic from another perspective in the next lecture, and show that it is asymptotically Normal under $H_{0\mathrm{F}}$:

$$t \xrightarrow{\mathrm{d}} \mathrm{N}(0, 1).$$

Then we can calculate an approximate p-value which is close to the p-value from t.test() with "var.equal = FALSE".

Example 3 (Wilcoxon rank sum)**.** The difference-in-means statistic uses the original outcomes, and its sampling distribution depends on the second moments of the outcomes. This makes it sensitive to outliers. Another popular test statistic is based on the ranks of the pooled observed outcomes. Let R_i denote the rank of Y_i in the pooled samples $\boldsymbol{Y}$: $R_i = \#\{j : Y_j \leq Y_i\}$. The Wilcoxon rank sum statistic is the sum of the ranks under treatment:

$$W = \sum_{i=1}^{n} Z_i R_i.$$

For algebraic simplicity, we assume that there are no ties in the outcomes, although the Fisher randomization test can be applied regardless of the existence of ties. For the case with ties, see Lehmann (1975, Chapter 1 Section 4). Because the sum of the ranks of the pooled samples are fixed at $1 + 2 + \cdots + n = n(n+1)/2$, the Wilcoxon statistic is equivalent to the difference in the means of the ranks under treatment and control. We can use the Lemma for survey sampling to show that under $H_{0\mathrm{F}}$, W has mean

$$E(W) = \sum_{i=1}^{n} E(Z_i) R_i = \frac{n_1}{n} \sum_{i=1}^{n} i = \frac{n_1}{n} \times \frac{n(n+1)}{2} = \frac{n_1(n+1)}{2}$$

and variance

$$\begin{aligned}
\text{var}(W) &= \text{var}\left(\sum_{i=1}^{n} Z_i R_i\right) \\
&= \frac{n_1 n_0}{n(n-1)} \sum_{i=1}^{n} \left(R_i - \frac{n+1}{2}\right)^2 \\
&= \frac{n_1 n_0}{n(n-1)} \sum_{i=1}^{n} \left(i - \frac{n+1}{2}\right)^2 \\
&= \frac{n_1 n_0}{n(n-1)} \left\{\sum_{i=1}^{n} i^2 - n\left(\frac{n+1}{2}\right)^2\right\} \\
&= \frac{n_1 n_0}{n(n-1)} \left\{\frac{n(n+1)(2n+1)}{6} - n\left(\frac{n+1}{2}\right)^2\right\} \\
&= \frac{n_1 n_0 (n+1)}{12}.
\end{aligned}$$

Furthermore, under $H_{0\text{F}}$, the randomization distribution of $\widehat{\tau}$ is approximately normal:

$$\frac{\sum_{i=1}^{n} Z_i R_i - \frac{n_1(n+1)}{2}}{\sqrt{\frac{n_1 n_0 (n+1)}{12}}} \xrightarrow{\text{d}} \text{N}(0,1). \tag{4}$$

Based on (4), we can conduct an asymptotic test. In R, wilcox.test() can compute both exact and asymptotic p-values based on the statistic $W - n_1(n_1+1)/2$. Based on some asymptotic analyses, Lehmann (1975) showed that the FRT using W has reasonable powers over a wide range of data generating processes.

Example 4. The treatment may affect the outcome in different ways. It seems natural to measure the treatment outcomes and control outcomes based on the empirical distributions:

$$\hat{F}_1(y) = n_1^{-1} \sum_{i=1}^{n} Z_i I(Y_i \le y), \quad \hat{F}_0(y) = n_0^{-1} \sum_{i=1}^{n} (1 - Z_i) I(Y_i \le y).$$

This yields the famous Kolmogorov–Smirnov statistic

$$D = \max_y \left|\hat{F}_1(y) - \hat{F}_0(y)\right|.$$

It is a challenging mathematics problem to derive the distribution of D (Van der Vaart 2000):

$$\text{pr}\left(\frac{n_1 n_0}{n} D \leq x\right) \to \frac{\sqrt{2\pi}}{x} \sum_{j=1}^{\infty} e^{-(2j-1)^2 \pi^2/(8x^2)},$$

based on which we calculate an asymptotic p-value. In R, ks.test() can compute both exact and asymptotic p-values.

4. A case study of the LaLonde experimental data

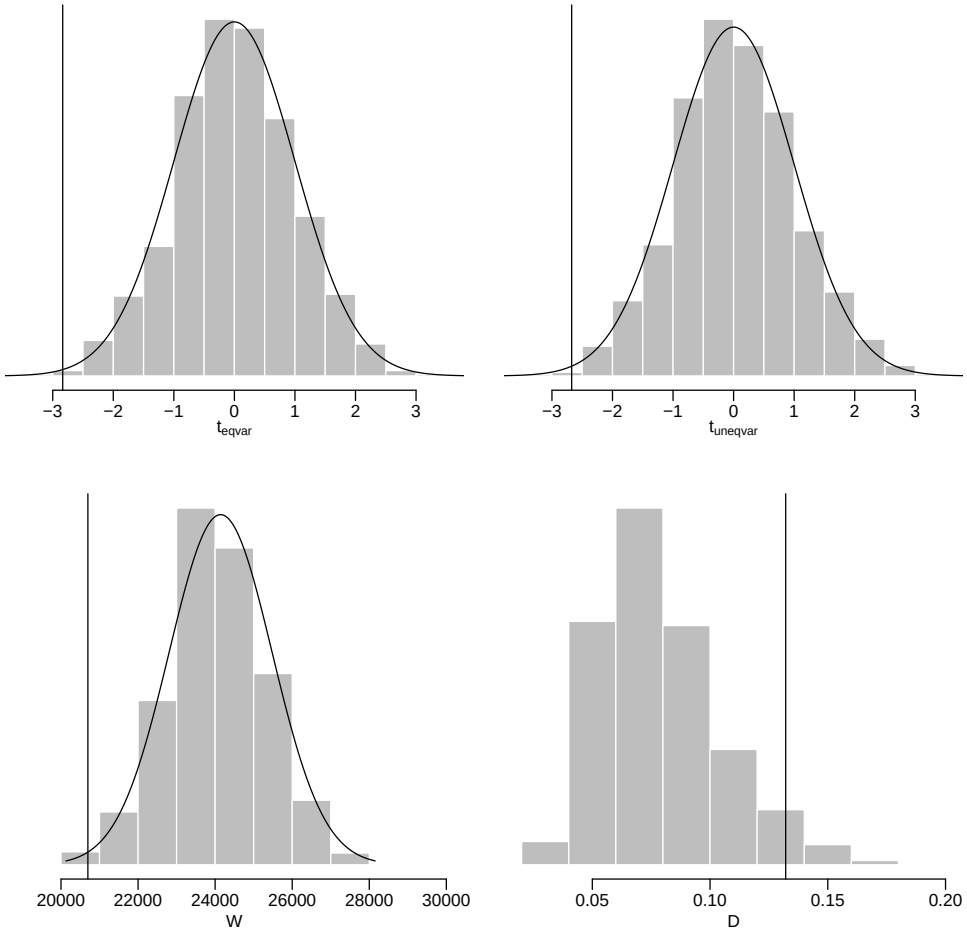

Figure 1: The permutation distribution of four test statistics based on the LaLonde experimental data

References

Fisher, R. A. (1935). *The Design of Experiments, 1st Edition.* Edinburgh, London: Oliver and Boyd.

Lehmann, E. L. (1975). *Nonparametrics: Statistical Methods Based on Ranks.* California: Holden-Day, Inc.

Rubin, D. B. (1980). Comment on "Randomization analysis of experimental data: the Fisher randomization test" by D. Basu. *Journal of American Statistical Association*, 75:591–593.

Van der Vaart, A. W. (2000). *Asymptotic Statistics.* Cambridge: Cambridge University Press.