The Completely Randomized Experiment and the Fisher Randomization Test

Peng Ding, Email: pengdingpku@berkeley.edu

1. Completely Randomized Experiment

An experiment with n units, in which n_1 receiving the treatment and n_0 receiving 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

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

Rubin (1980) called it the sharp null hypothesis in the sense that it can determine the whole Science Table based on the observed data. Other researchers called it the strong null hypothesis.

Conceptually, FRT works for any test statistic

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

where $\boldsymbol{Y} = (Y_1, \dots, Y_n)$ is the vector of the observed outcomes, and $\boldsymbol{Y}(z)$ is the vector of potential outcomes under treatment z ($z = 0, 1$). 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 completely known due to the design of the experiment. This distribution of T is sometimes called 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 ($n = 100, n_1 = 50, M > 10^2 9$), 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, FRTs are sometimes called the permutation tests.

3. Canonical choices of the test statistic

difference in means

Wilcoxon rank sum statistic

Kolmogorov–Smirnov statistic

References

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

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.