STAT-511 **2018 Spring**

Lecture 3: Randomization Tests

Instructor: Naveen N. Narisetty   Scribe: Yutong Dai   Last Modified: 2018-05-05

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications, and they may contain factual and/or typographic errors.*

General Design:

1. Collect N subjects in total.

2. Randomly assign subjects to the treatment group and the control group. For example, assign n out of N subjects to a medicine and assign $m = N - n$ subjects to a placebo.

The task is to make inferences based on such a design.

3.1 Randomization Model

Principle:

1. Regard the set of subjects as the population rather than samples from population(s).
2. The statistical randomness comes from the random assignment of each subject rather than the sampling. Therefore, the distribution of a test statistic depends on how subjects are randomly assigned to treatment/placebo rather on the underlying distribution of the population(s) where samples come from.

Example 3.1 Detailed descriptions refer to Section 17.1 from John Marden's book.

Remarks:

1. Under the null hypothesis, we literally know the "complete" subjects, i.e., each pair of (x_i, y_i).

2. The randomization comes from which particular permutation matrix is associated with the observed $\mathbf{Z}_N$.

3. Under the null hypothesis, $\mathbf{U}_N \overset{d}{=} P\mathbf{Z}_n \sim F(\mathbf{U}_N)$ where $P \sim \text{Uniform}\{S_N\}$. So we can traverse through all possible P and then know the exact quantile of the distribution $F(\mathbf{U}_N)$ and also for the test statistics $T(P\mathbf{Z}_N)$.

4. From 3, we know that $P\mathbf{Z}_N$ follows a discrete distribution.

3.1.1 Asymptotical Distribution of the Test Statistics

When it's is computaionally imposible to traverse through all the possible P from Uniform$\{S_N\}$, that is N is fairly large, we consider the asymptotic distribution of the $T(P\mathbf{Z}_N)$.

The randomization distribution of test statistic (linear transformation only) is given by

$$T(P\mathbf{Z}_N) = a_N^T P\mathbf{Z}_N, \quad P \sim \text{Uniform}\{S_N\}.$$

Here, $\mathbf{Z}_N$, under the null, represents the population (all are known).

Lemma 3.2 Denote $\mathbf{U_N} = P\mathbf{Z}_n = (U_1, ..., U_N)$, then we have

$$E(U_i) = \bar{\mathbf{Z}}_N, Var(U_i) = s^2_{\mathbf{Z}_N}, Cov(U_i, U_j) = \frac{-1}{N-1} s^2_{\mathbf{Z}_N}$$

, where $\bar{\mathbf{Z}}_N = \frac{1}{N} 1_N^T \mathbf{Z}_N$ and $S^2_{\mathbf{Z}_N} = \frac{1}{N} \sum_i (z_i - \bar{\mathbf{Z}}_N)^2$.

Theorem 3.3 Denote

$$V_n = \frac{T(P\mathbf{Z}_n) - E[T(P\mathbf{Z}_n)]}{\sqrt{Var(T(P\mathbf{Z}_n))}},$$

under certain conditions, $V_n \xrightarrow{D} N(0, 1)$.

3.2 Randomization tests for sampling models

P-value dervied from sampling model is the same as the P-value derived from the randomzation model.