RANDOMIZATION INFERENCE

In the previous lesson, we discussed some randomized experiments:

1. Bernoulli, 2. completely randomized, 3. block randomized, 4. paired randomized

We now discuss randomization based inference for randomized experiments

Key idea: potential outcomes are fixed constants — Not Random Variables — we will write these using lower case y.

Probability enters through assignment rules in previous lesson; assignments Z are random variables.

We shall take up the sharp null hypothesis that there is absolutely no effect of treatment, for any unit — note this is much stronger than an average 0 effect.

RANDOMIZATION INFERENCE: 0 EFFECT

$y_i(\mathbf{z})$–the response of unit i under treatment assignment $\mathbf{z} \in \Omega$:

$\mathbf{y}(\mathbf{z}) = (y_1(\mathbf{z}), ..., y_n(\mathbf{z}))'$ the column vector of responses under treatment assignment $\mathbf{z}$.

H_0: For $i = 1, ..., n$, and for all assignments $\mathbf{z}$ and $\mathbf{z}' \in \Omega$, the response vector $\mathbf{y}(\mathbf{z}) = \mathbf{y}(\mathbf{z}') = (y_1, ..., y_n)$.

The idea behind testing H_0 or computing a p-value is quite ingenious and very simple, and exploits the fact that for every assignment of the n subjects, under H_0 the subject's data is the same, i.e, at least for the assignments in Ω, $y_i(0) = y_i(1)$.

RANDOMIZATION INFERENCE: 0 EFFECT

Under H_0, for any assignment of subjects, we know the subject's data — and we also know the probability of any assignment — so for a given test statistic, we can compare the value for the actual assignment with every other assignment and see if the value we got is in line with the other values or more extreme than we would expect if there was no effect

RANDOMIZATION INFERENCE: 0 EFFECT

1. Consider a test statistic $t(\mathbf{Z}, \mathbf{y})$.

2. Calculate $t = t(\mathbf{Z}, \mathbf{y})$ under the observed assignment $\mathbf{Z} = \mathbf{z}$.

3. Calculate $t(\mathbf{Z}, \mathbf{y})$ for all other assignments and use this to determine (for a one sided p-value)

$$p^* = \Pr_{H_0}(t(\mathbf{Z}, \mathbf{y}) \geq t) = \sum_{\mathbf{z}: t(\mathbf{Z}, \mathbf{y}) \geq T} \Pr(\mathbf{Z} = \mathbf{z}) \qquad (1)$$

the probability of obtaining, under H_0, a result as large or larger than that actually; obtained this is the p-value (significance level) for the test.

For all the randomized experiments we considered above, all assignment vectors in Ω are equally likely so

$$\sum_{\mathbf{z}:t(\mathbf{Z},\mathbf{y})\geq T} \Pr(\mathbf{Z}=\mathbf{z}) = |\Omega|^{-1} \sum_{z\in\Omega} 1_{(t(\mathbf{Z},\mathbf{y})\geq T)}(z) \qquad (2)$$

where $|\Omega|$ is the cardinality of Ω

RANDOMIZATION INFERENCE: 0 EFFECT

The famous British statistician R.A. Fisher had a colleague, Dr. Bristol, working in his lab who claimed she could tell whether the milk in a cup of tea with milk was poured into the cup before or after the tea. An experiment was conducted to test her claim. are 8 cups in total, 4 received milk first, 4 received tea first. Bristol then took a drink from each cup and responded tea first (denoted 0 here) or milk first (denoted 1). milk added first (i.e.,$Z_i = 1$) is $m = 4$. There are $8!/4!4! = 70$ such vectors. response vector $(y_{Z_1}, ..., y_{Z_8})$ should also contain 4 0's and 4 1's.

RANDOMIZATION INFERENCE: 0 EFFECT

Under H_0, Bristol gives the same sequence of responses no matter what the actual truth is.

For step 1, we choose as test statistic the number of correct responses.

For step 2, we see that Bristol gets 6 correct.

For step 3, we count how many Bristol would have gotten, given her fixed responses, under the other 69 possible assignment vectors. Then we find out that there are 17 ways to get 6 or more correct, so the p-value is .243.

Can you see how there are 17 ways to get 6 or more correct?

RANDOMIZATON INFERENCE: 0 EFFECT

A toy example in a completely randomized experiment: $S = 1$, $n = 4$, $m = 2$. There are 6 possible assignments, each with probability $1/6$. The assignment chosen is $\mathbf{z} = (0, 0, 1, 1)'$, and we observe $\mathbf{y} = (1, 3, 4, 6)$.

Step 1: As a test statistic, choose $\bar{Y}_t - \bar{Y}_c$

Step 2: We get $t = 3$ for the actual assignment.

Step 3. For the other 5 possible assignments, we get $t(\mathbf{Z}, \mathbf{y}) =$ 2,0,0,-2, -3. Thus the probability, under the null hypothesis, of getting a result of equal or greater magnitude is $1/6$, i.e, $\Pr_{H_0}(t(\mathbf{Z}, \mathbf{y}) \geq 3) = 1/6$.

Two sided p values can be dealt with.

RANDOMIZATION INFERENCE: 0 EFFECT

Many other test statistics have been used in practice.

The median test statistic computes the median q of the n responses, then counts the number of responses in the treatment group that exceed q under the assignment vectors $\mathbf{z} \in \Omega$:
$t(\mathbf{Z}, \mathbf{y}) = \sum_{i=1}^{n} Z_i 1_{y_i > q}(y_i)$

Of course, a different quantile could have been chosen.

RANDOMIZATION INFERENCE: 0 EFFECT

A famous statistic that is often used is the Wilcoxon rank sum test statistic (or Mann-Whitney test stastistic).

Here the responses are transformed to their ranks r_i, arranged in ascending order. We will ignore the fact that some values may have ties here; that can be dealt with. The test statistic is $t(\mathbf{Z}, \mathbf{Y}) = \sum_{i=1}^{n} Z_i r_i$, the sum of the ranks in the treatment group under the different assignment vectors.

This yields a one sided p-value p^*.

RANDOMIZATION INFERENCE: 0 EFFECT

For the case where $S > 1$, $n_s = 2$, $n_{s1} = 1$, i.e., the case of the paired randomized experiment, with binary data $y_{si'} = 0$ or 1, McNemar's test uses the number of 1's in the treated group, i.e. $t(\mathbf{Z}, \mathbf{y}) = \sum_{s=1}^{s} \sum_{i'=1}^{2} Z_{si\prime} y_{si'}$.

For the case $S > 1$, $n_s = 2$, $n_{s1} = 1$, with continuous data, the Wilcoxon signed rank test statistic is often used. For each pair $s = 1, ..., S$, without loss of generality, relabel the units so that is is the first unit ($i' = 1$) in each pair that receives treatment.

Then, for each pair, compute the absolute value of the difference between the treated and the control observation, and convert these absolute differences into ranks r_s. Then calculate the sum of the ranks for the subset of cases where the response of the treated unit exceeds that for the control unit: $\sum_{i=1}^{n} r_s 1_A(s)$, where $A = \{s : y_{s1} > y_{s2}\}$.

RANDOMIZATION INFERENCE

Tests for more than two treatments, ordinal data, partially ordered data, and multivariate outcome data may also be considered (see for example, Rosenbaum 2002 or Hollander and Wolfe (1999).

The same idea above can be used to assess a more general null hypothesis:$H_0 : y_i(1) - y_i(0) = \tau_i$, where τ_i is a specified constant. That is because for each observation i, we know either $y_i(0)$ or $y_i(1)$ and thus under H_0 we know the value of the potential outcome we don't see. Therefore we can write down the values for all responses under any assignment vector.

Typically, we would not know what value to use for each case, so we specify a constant effect: $\tau_i = \tau$ for all i. This is the null hypothesis of constant effect; this null hypothesis can be tested for different values of τ and a (1 - α)$\times$ 100 % confidence interval obtained by taking the set of values of τ for which H_0 is not rejected at level α.

RANDOMIZATION INFERENCE; CONSTANT EFFECT

For our toy example, suppose we want to test $H_0 : \tau = 1$ vs. $H_1 : \tau > 1$. Under H_0 $(y_1(0) = 1, y_2(0) = 3, y_3(0) = 3, y_4(0) = 5)$ and $(y_1(1) = 2, y_2(1) = 4, y_3(1) = 4, y_4(1) = 6)$. Thus, if $\mathbf{Z} = (1, 1, 0, 0)$, $\bar{Y}_1 - \bar{Y}_0 = 1$; continuing, we get under this H_0 $\Pr_{H_0}(t(\mathbf{Z}, \mathbf{y}) \geq 3) = 2/6$.

RANDOMIZATON INFERENCE

What test statistic to use? Here are three considerations:

1. sensitivity to departures of interest from H_0
2. robustness to outliers
3. power

RANDOMIZATION INFERENCE

In a completely randomized experiment, there aren $n!/n_1!n_0!$ assignments. Even with modern computers, the set of assignments can be too large to handle easily. For example, if n = 500, $n_1 = 250$, there are more than 10^{49} assignments.

Two approaches:

1. sample the assignments

2. derive mean and variance under H_0 and use a normal approximation for a normal theory test

RANDOMIZATION INFERENCE

Pros

1. makes minimal assumptions

2. easy to understand

CONS

1. sometimes the sample is of interest, but often want to extrapolate beyond sample—though this requires more assumptions

2. doesn't handle heterogeneity well—but see Rosenbaum (2010).

RANDOMIZATION INFERENCE

More generally, however, an investigator is likely to want a measure that summarizes the magnitude of the possibly heterogenous treatment effects, for example an average of the treatment effects. We turn to estimation now, focusing primarily on randomization based inference for the SATE.