RANDOMIZATION BASED INFERENCE

In the previous module, we examined how randomization based inference allows for the estimation and hypothesis testing of the SATE. We assumed:

1. The i^{th} of n experimental units had two potential outcomes. One outcome, denoted $y_i(0)$, if assigned treatment 0, and one outcome, denoted $y_i(1)$, if assigned treatment 1.
2. We only observed one of the two potential outcomes for each unit, depending on the treatment assigned to that unit, Z_i. As a result, we could say that "half the data are missing".
3. We regarded the potential outcomes to be fixed constants, and the only source of randomness was that induced through the (random) assignment of treatments to subjects.

EXTENDING RANDOMIZATION BASED INFERENCE

In this lesson, we extend randomization based inference to the FATE, where the n units are a random sample from a population of N units.

We still regard $y_i(0)$ and $y_i(1)$ as fixed constants. However, in addition to missing half the potential outcomes for n of the units, we do not observe either potential outcome for the remaining $N - n$ units. Thus, randomness now has two sources: the (random) assignment of treatments to subjects, and the (random) selection of subjects from the population.

The estimator, $\bar{Y}(1) - \bar{Y}(0)$, stated in the previous module remains unbiased for the FATE, but the stated hypothesis test is no longer valid.

ESTIMATION OF THE FATE

Suppose the n units are a simple random sample from the population of N units. i.e. each possible sample of size n, taken without replacement from the N units is equally likely.

Let T_i denote the random variable that takes value 1 if unit i appears in the sample and 0 otherwise.

$$E(T_i) = n/N, i = 1, \ldots, N$$

Due to random sampling,

$$\bar{Y}(1) - \bar{Y}(0) := n^{-1} \sum_{i=1}^{n} (y_i(1) - y_i(0)) = n^{-1} \sum_{i=1}^{N} T_i (y_i(1) - y_i(0))$$

is now a random variable.

ESTIMATION OF THE FATE

Observe that

$$\bar{Y}(1) - \bar{Y}(0) := n^{-1} \sum_{i=1}^{N} T_i(y_i(1) - y_i(0))$$

is unbiased for the FATE

$$N^{-1} \sum_{i=1}^{N} (y_i(1) - y_i(0))$$

since

$$\begin{aligned} E(n^{-1} \sum_{i=1}^{N} T_i(y_i(1) - y_i(0)) &= n^{-1} \sum_{i=1}^{N} E(T_i(y_i(1) - y_i(0)) \\ &= n^{-1} \sum_{i=1}^{N} (n/N)(y_i(1) - y_i(0)) \qquad (1) \end{aligned}$$

ESTIMATION OF THE FATE

So for every sample,

$\bar{Y}_1 - \bar{Y}_0$ is unbiased for $\bar{Y}(1) - \bar{Y}(0)$ (over the distribution induced by randomization), and

$\bar{Y}(1) - \bar{Y}(0)$ is unbiased for $\mathrm{E}(y(1) - y(0))$ (over the distribution induced by sampling)

We write:

$$E_S E_R(\bar{Y}_1 - \bar{Y}_0) = N^{-1} \sum_{i=1}^{N} (y_i(1) - y_i(0)), \tag{2}$$

where E_R denotes the expectation over the randomization distribution and E_S denotes the expectation over the sampling distribution.

ESTIMATION OF THE FATE

As in the case of the SATE, the derivation of the variance of $\bar{Y}_1 - \bar{Y}_0$ is tedious and depends on the unknown unit effects. The result given earlier for the SATE is a special case.

$$V(\bar{Y}_1 - \bar{Y}_0) = \frac{1}{n_0(N-1)} \sum_{i=1}^{n} (y_i(0) - \bar{Y}(0))^2 + \frac{1}{n_1(N-1)} \sum_{i=1}^{n} (y_i(1) - \bar{Y}(1))^2 - \frac{1}{n(N-1)} \sum_{i=1}^{n} (y_i(1) - y_i(0) - (\bar{Y}(1) - \bar{Y}(0)))^2 \tag{3}$$

As in the case of the SATE, if the treatment effects are constant, the last term vanishes. Otherwise, a conservative estimate of the variance is obtained when the last term is ignored.

HYPOTHESIS TESTS FOR FATE

Hypothesis tests and confidence intervals for the the population average $\mathrm{E}(y(1) - y(0))$ may be obtained using a normal approximation for the distribution of $\bar{Y}(1) - Y(\bar{0})$.

Using this approach, as above, the ratio

$$\frac{\bar{Y}_1 - \bar{Y}_0}{(s_0^2/(n_0) + s^2(1)/n_1)^{1/2}}$$

is approximately $N(0, 1)$.

MODEL BASED INFERENCE

No doubt you are more familiar with the model based approach to inference, in which the potential outcomes are regarded as random variables. This is the approach that we will take throughout much of the remainder of the course.

We typically treat the n units as a sample from an infinite population, but we can also apply the model based approach to the case where the population has finite size N. We have already seen the basic ideas for how this approach works in module 1, but it is worthwhile to reiterate and formalize this here in the explicit context of randomized experiments and to contrast this with the previous approach in which outcomes are treated as fixed constants.

MODEL BASED INFERENCE

Let $(Y_i(0), Y_i(1))$ denote unit i's potential outcomes. Upper case notation indicates the potential outcomes are random variables.

Assume $E(Y_i(z)) = E(Y(z))$ for all i and $z = 0, 1$, as would be the case when the potential outcomes are independent and identically distributed.

For the completely randomized experiment, this means

$$\Pr(Z_i = z \mid Y_i(0), Y_i(1)) = \Pr(Z_i = z)$$

which we write more compactly as

$$Z_i \perp\!\!\!\perp (Y_i(0), Y_i(1))$$

.

MODEL BASED INFERENCE

Consider $\bar{Y}_1 - \bar{Y}_0$ as an estimator for the ATE. It is unbiased:

$$
\begin{aligned}
E(\bar{Y}_1 - \bar{Y}_0) &= E\left(\frac{\sum_{i=1}^n Z_i Y_i}{\sum_{i=1}^n Z_i} - \frac{\sum_{i=1}^n (1 - Z_i) Y_i}{\sum_{i=1}^n (1 - Z_i)}\right) \\
&= \sum_{i=1}^n E\left(\frac{Z_i Y_i(1)}{n_1} - \frac{(1 - Z_i) Y_i(0)}{n_0}\right) \\
&= \sum_{i=1}^n \frac{(n_1/n) E(Y_i(1)}{n_1} - \frac{(n_0/n) E(Y_i(0)}{n_0} \\
&= n^{-1} \sum_{i=1}^n (E(Y_i(1) - Y_i(0)) = E(Y(1) - Y(0) \qquad (4)
\end{aligned}
$$

where $E(Z_i Y_i(1)) = E(Z_i) E(Y_i(1)) = (n_1/n) E(Y_1)$ and
$E((1 - Z_i) Y_i(0)) = E(1 - Z_i) E(Y_i(0)) = (n_0/n) E(Y(0))$
since $Z_i \perp\!\!\!\perp (Y_i(0), Y_i(1))$

MODEL BASED INFERENCE

For a randomized block experiment,

$$\Pr(Z_i = z \mid Y_i(0), Y_i(1), \mathbf{X}_i) = \Pr(Z_i = z \mid \mathbf{X}_i)$$

which we write more compactly as

$$Z_i \perp\!\!\!\perp (Y_i(0), Y_i(1)) \mid \mathbf{X}_i$$

Thus, within covariate strata,

$$E(\bar{Y}_{s1} - \bar{Y}_{s0}) = E(Y(1) - Y(0) \mid S = s), \tag{5}$$

We can average using the stratum probabilities as weights, $\Pr(S = s) = N_s/N$ (where N_s is the number of units in the population from stratum s), to get the ATE:

$$E(\sum_{s=1}^{S}(\bar{Y}_{s1} - \bar{Y}_{s0})(N_s/N)) = E(Y(1) - Y(0)) \tag{6}$$