1 The randomization model

Most of this course will be devoted to the study of treatment effects in the absence of random assignment of subjects to treatments. As we will see, performing causal inference in the absence of random treatment assignment requires that we make fairly strong assumptions. In contrast, when treatment is assigned randomly, treatment effects can be estimated with very mild assumptions and, very importantly, the hypothesis of no treatment effect can be tested without assumptions of any kind. In this section, we will study the basics of randomization inference so that in the remaining of the course we can think of observational studies as departures from this benchmark.

Let the term “experimental unit” refer to the opportunity to apply or withhold the treatment. In general, experimental units will be persons who will either receive or not receive the treatment. But units could also be families, classrooms, or, as we have seen in lecture, cups of tea.

The randomization model has several distinctive features:
1. Experimental units are not a random sample from a population of units. This is, experimental units are fixed and are not assumed to be drawn from a common population.

2. Experimental units are not assumed to be homogeneous.

3. Chance enters the experiment solely through the assignment of the units to treatment and control, a process which is controlled by the experimenter.

Before we discuss the randomization model in more detail, let’s define what we mean by the “hypothesis of no treatment effect”.

1.1 The hypothesis of no treatment effect

In experimental designs such as the one described by the randomization model, we are often interested in testing the hypothesis that the treatment has no effect. The reason is that, under randomization, this hypothesis can be tested with no assumptions whatsoever. If the randomization is successful, randomization tests of the hypothesis of no effect can be done with no assumptions at all.

The definition of no effect of treatment is very simple. Each unit exhibits a certain outcome after treatment, and hence we say that the treatment has no effect if the unit would exhibit the same value of the outcome whether assigned to treatment or control. As we will see in more detail below, when the treatment is without effect, the outcome of the unit is fixed in the sense that its outcome would not change if a different assignment were selected.

1.2 A discussion of the randomization model

Under the randomization model, the way we think about potential outcomes changes radically. In the Rubin Causal Model, the potential outcomes are random variables, and that’s
why we talk about their expectations (the ATT, the ATE, etc.). In contrast, in the randomization model, *under the assumption that the treatment is without effect*, the potential outcomes or responses are determined before the assignment of units to treatment in the sense that the responses of the units would not change if a different treatment assignment was selected. Since the potential outcomes are fixed and do not depend on whether a unit receives treatment, we can think of each unit’s potential outcome as being *attached* to the unit even before the assignments to treatment and control are made. Thus, it is as if potential outcomes are randomly assigned together with the units. In other words, under the null hypothesis responses are fixed quantities and not random variables.

Moreover, the fact that under the null hypothesis the only random variable is the assignment of treatment implies that the distribution of the chosen test statistic under the null is completely determined by the randomization distribution of the treatment assignment. This is, the randomization of treatment is directly incorporated in the estimation.

Finally, randomization inference can be used not only to test hypothesis but also to compute non-parametric confidence intervals (by inverting the tests) and point estimates (e.g. Hodges-Lehman).

In what follows, we will study Fischer’s Lady Tasting Tea example, which is the canonical example of how randomization can be used as a “reasoned basis for inference” (in the words of Fischer). This is, when the conditions of the randomization model are met, valid inferences about the effects of a given treatment on a given outcome exhibited by the experimental units require only that the treatment be allocated at random to the units. No further assumptions are needed.
2 The Lady Tasting Tea

As we saw in lecture, Fisher’s famous example goes as follows. A lady declares that when she tests a cup of tea made with milk, she can discriminate whether the milk was poured before the tea or the tea was poured before the milk. So Fisher proposes the following experiment. Mix 8 cups of tea with milk, four with milk poured first and four with tea poured first, and present these cups to the lady in random order. The lady knows in advance how the experiment is designed, so she knows that there are exactly four cups with poured milk first and four cups with tea poured first. The 8 cups are presented to the lady randomly, and she must divide the cups into two groups of 4 cups according to the treatment received (i.e., according to whether milk or tea was poured first).

Imagine that we suspect that the lady is lying, and we want to test the hypothesis that she cannot discriminate between the two different cups of tea. Our null hypothesis is then that the lady has no faculty of discrimination. How can we test this hypothesis?

If she really has no faculty to discriminate both types of cups of tea, then she will only classify the cups correctly by chance. Now, there are \( \binom{8}{4} = \frac{8!}{4!(8-4)!} = \frac{8!}{4!4!} = \frac{8 \times 7 \times 6 \times 5}{4 \times 3 \times 2} = 1680 \) possible divisions of 8 cups into two groups of four and only one of these is the correct arrangement. This means that the lady has probability \( p = \frac{1}{1680} \approx 0.0006 \) of correctly discriminating the cups by chance. This probability is the significance level for testing the null hypothesis that the lady has no ability to discriminate. In other words, if the lady were lying and she had no ability to discriminate whether milk or tea was poured first into the cups, she would group the cups correctly with a probability of 0.014. Since this is probability is small, if she does classify the cups correctly we can make a justified inference that she has the ability to discriminate, i.e. we can reject the null hypothesis. Note that the only probability used in this experiment is the probability created by the experimenter.
2.1 Test statistics

Let’s introduce some notation. Suppose that there are $N$ units available for experimentation. Every unit will be assigned to either treatment or control. These $N$ units are subdivided into $S$ strata in the basis of some characteristics measured prior to the assignment of treatment (i.e., on the basis of some covariates). There are $n_s$ units in stratum $s$ for $s = 1, 2, \ldots, S$. Hence $N = \sum_{s=1}^{S} n_s$. Let $Z_{is}$ be an indicator variable equal to one if unit $i$ in stratum $s$ receives treatment. Let $m_s$ denote the number of units in stratum $s$, so that $m_s = \sum_{i=1}^{n_s} Z_{is}$ and $0 \leq m_s \leq n_s$. Let $\mathbf{Z}$ be an $N$-dimensional column vector whose elements are the $Z_{si}$ for all units. So,

$$\mathbf{Z} = \begin{bmatrix} Z_{11} \\ Z_{12} \\ \vdots \\ Z_{1n_1} \\ \vdots \\ Z_{S1} \\ \vdots \\ Z_{Sn_s} \end{bmatrix} = \begin{bmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_S \end{bmatrix}$$

where $\mathbf{Z}_s = \begin{bmatrix} Z_{s1} \\ Z_{s2} \\ \vdots \\ Z_{sn_s} \end{bmatrix}$.

If no covariates are used to subdivide the units, then $S = 1$. When there is only a single stratum, we write $Z_i$ instead of $Z_{1i}$.

In a randomized experiment, the experimenter randomly determines the assignment of units to treatment and control. In other words, the experimenter determines the value of $\mathbf{Z}$ using a random mechanism. To say that the randomization mechanism is known is to say that the distribution of the random variable $\mathbf{Z}$ is known because it was created by the experimenter. One requirement that is placed on this random mechanism is that prior to the assignment of treatments all units have a positive probability of being assigned both the
treatment and control. Formally, \(0 < \Pr(Z_{si} = 1) < 1\) for \(s = 1, \ldots, S\) and \(i = 1, \ldots, N\). Let \(\Omega\) be the set of all possible values of \(Z\), i.e. all values of \(Z\) that are given positive probability by the random mechanism. Of course, many different random mechanisms can be used in practice. The one used in the Lady Tasting Tea example is just one of those.

Let’s go back to testing the hypothesis of no treatment effect. As we said, to say that the treatment had no effect is to say that the outcome under treatment is equal to the outcome under control. Let \(r_{si}\) be the outcome of unit \(i\) in stratum \(s\) and the \(N\)-tuple of outcomes for the \(N\) units be \(r\). Note a very important point: under the null hypothesis of no treatment effect, \(r\) is a fixed quantity and not a random variable (the only random variable is \(Z\)).

A test-statistic \(t(Z, r)\) is a quantity that is computed from the treatment assignment \(Z\) and the outcome \(r\). (For example, the treatment-minus-control difference in sample means is the test-statistic \(t(Z, r) = \frac{Z^T r}{Z^T 1} - \frac{(1-Z^T) r}{(1-Z^T) 1}\) where \(\mathbf{1}\) is an \(N\)-dimension vector of ones).

Given any test-statistic \(t(Z, r)\), the goal is to compute a significance level for a test that rejects the null hypothesis of no treatment effect when \(t(Z, r)\) is large. This can stated more precisely:

1. The null hypothesis of no effect is assumed to hold, so \(r\) is assumed to be fixed.
2. A treatment assignment \(Z\) has been selected from \(\Omega\) using a known random mechanism.
3. The observed value, \(T\), of the test-statistic \(t(Z, r)\) has been calculated.
4. We seek the probability of a value of the test-statistic as large or larger than the one observed if the null hypothesis were true.

The significance level is the sum of the randomization probabilities of assignments \(z \in \Omega\) that lead to values of \(t(Z, r)\) greater than or equal to the observed value \(T\). Formally,

\[
\Pr(t(Z, r) \geq T) = \sum_{z \in \Omega} \mathbb{I}\{t(z, r) \geq T\} \Pr(Z = z)
\]
where $\mathbb{I}\{\cdot\}$ is the indicator function and $\Pr (Z = z)$ is determined by the known random mechanism that assigned treatments.

In the case of a uniform randomized experiment where each possible assignment is given the same probability, we have $\Pr (Z = z) = \frac{1}{|\Omega|}$. So the significance level is just the proportion of treatment assignments $z \in \Omega$ that yield values of the test statistic $t(Z, r)$ greater than or equal to $T$. If we let $|\Omega| = K$ we can write

$$
\Pr (t(Z, r) \geq T) = \frac{|\{z \in \Omega : t(z, r) \geq T\}|}{K}
$$

### 2.2 Back to the Lady Tasting Tea

We can now analyze the Lady Tasting Tea example using this set-up. The Lady tastes 8 cups, $N = 8$, and there is only one stratum, $S = 1$. A treatment assignment is an 8-tuple containing four 1s and four 0s. For example $Z = (1, 1, 0, 0, 1, 0, 1, 0)$ represents the situation where the first, second, fifth and seventh cups had milk added first and the rest of the cups had tea added first. The set of treatment assignments $\Omega$ contains all possible arrangements of four 1s and four 0s, so $|\Omega| = K = \binom{8}{4} = 70$. The fact that treatment was assigned at random means that $\Pr (Z = z) = \frac{1}{K} = \frac{1}{70}$ for all $z \in \Omega$.

The Lady’s outcome for cup $i$ is $r_i$, where $r_i = 1$ if the lady classifies cup $i$ as milk first and $r_i = 0$ if the lady classifies cup $i$ as tea first. This means that $r = (r_1, \ldots, r_8)^T$. Since the Lady must classify exactly 4 cups as milk first and 4 cups as tea first, we have $1^T r = 4$. The test-statistic is the number of cups correctly identified:

$$
t(Z, r) = Z^T r + (1 - Z)^T (1 - r) =
\begin{align*}
&= Z^T r + 1^T 1 - 1^T r - Z^T 1 + Z^T r \\
&= Z^T r + 8 - 4 - 4 + Z^T r \\
&= 2Z^T r
\end{align*}
$$

7
Suppose that the lady classifies the first, third, fifth and seventh cup as milk first, so \( \mathbf{r} = (1, 0, 1, 0, 1, 0, 1, 0)^T \). There is only one treatment assignment \( \mathbf{z} \in \Omega \) that leads to perfect agreement with the lady’s choice, namely \( \mathbf{z} = (1, 0, 1, 0, 1, 0, 1, 0)^T \). Under perfect agreement between the lady’s choices and the way milk and tea are added to the cups, the observed value of the test-statistic is

\[
T = 2\mathbf{z}^T \mathbf{r} = 2 \cdot 4 = 8
\]

Therefore, the significance level is

\[
\Pr (t(\mathbf{Z}, \mathbf{r}) \geq 8) = \frac{|\{\mathbf{z} \in \Omega : t(\mathbf{z}, \mathbf{r}) > 8\}|}{70} = \frac{1}{70} = 0.014
\]

This means that the chance of perfect agreement by accident is 0.014, a very small chance. So if we observe that the lady classifies all eight cups correctly, we can make a justified inference and say that she does recognize the order in which milk and tea are added to the cups.

What if we allow the lady to make mistakes? In this example, the lady cannot misclassify only one cup, since for any cup that she misclassifies she must have misclassified another cup. However, we can calculate the significance level of the test if we allow the lady to make at most two mistakes. We saw that there was only one assignment \( \mathbf{z} \in \Omega \) that led to perfect agreement. How many \( \mathbf{z} \in \Omega \) assignments lead to six agreements? Take any \( \mathbf{z} \), say \( \mathbf{z} = (1, 1, 1, 0, 0, 0, 0) \) and consider a lady’s response with perfect agreement \( \mathbf{r} = (1, 1, 1, 1, 0, 0, 0, 0) \). To get exactly six agreements, we must change a 0 by 1 and a 1 by a 0 in \( \mathbf{r} \). Since there are four 1s and four 0s, there are \( \binom{4}{1} \times \binom{4}{1} = \frac{4!}{(4-1)!1!} \times \frac{4!}{(4-1)!1!} = 16 \) ways in which we can do this. So there are 16 assignments with exactly \( t(\mathbf{Z}, \mathbf{r}) = 6 \) agreements. Since there is one assignment with \( t(\mathbf{Z}, \mathbf{r}) = 8 > 6 \), there are 17 assignments leading to six or more agreements. The significance level in this case is

\[
\Pr (t(\mathbf{Z}, \mathbf{r}) \geq 6) = \frac{|\{\mathbf{z} \in \Omega : t(\mathbf{z}, \mathbf{r}) > 6\}|}{70} = \frac{17}{70} = 0.24
\]
Note that this is no longer a small probability. In other words, if the lady has no ability to discriminate the order in which milk and tea are poured into the cups, she will correctly classify six cups by chance with probability 0.24. From observing that she classifies six cups correctly we are not justified in inferring that she does have the ability to discriminate between the cups.

Obviously, this experimental design is not very sensitive, since we cannot draw a justified inference unless we observe that the lady correctly classifies the eight cups. What if we do a binomial randomization where each cup has probability \( p = \frac{1}{2} \) of having milk first? We know that for an experiment that consists of \( N \) trials, the probability that the number of successes, \( Y \), is equal to a particular number, \( y \), is:

\[
P(Y = y) = \binom{N}{y} p^y (1 - p)^{N-y}, \quad y = 0, 1, 2, \ldots, N
\]

So the chance that the lady correctly identifies 8 cups under this experimental design is

\[
P(Y = 8) = \binom{8}{8} \left(\frac{1}{2}\right)^8 \left(\frac{1}{2}\right)^0 = \frac{1}{256}
\]

Note that this is significantly lower than the \( \frac{1}{70} \) chance that we had under the previous experimental design. The chance of incorrectly classifying one cup is

\[
P(Y = 7) = \binom{8}{7} \left(\frac{1}{2}\right)^7 \left(\frac{1}{2}\right)^1 = \frac{8}{256} = 0.031
\]

Unlike the previous case, the binomial design is sensitive enough to reject the null hypothesis if the lady makes one mistake.