Chapter 1

Completely Randomized Experiments

1.1 Introduction

In the first part of this book we discuss the analysis and design of randomized experiments with binary treatments. In this first chapter the focus is on the simplest setting, with a completely randomized experiment. In this setting we discuss two specific methods for analyzing such experiments: first the calculation of exact p-values for sharp null hypotheses (primarily the null hypothesis of no effects of the treatment whatsoever), following the work by R. A. Fisher from the 1920s (Fisher [1925, 1935]), and second, estimation of, and inference for, average treatment effects, following the work by Jerzy Neyman (Neyman [1923/1990, 1935]). In both cases the uncertainty, captured by p-values in the Fisher analyses and by standard errors and confidence intervals in Neyman's approach, arisesarise from the *design* of the experiment, leading to *randomization-based* or *design-based* inference. This in contrast to much work in econometrics that focuses on *sampling-based* uncertainty, arising from the random sampling from a large population.

Starting an econometrics text with a discussion of randomized experiments is somewhat unusual, with "Mostly Harmless Econometrics" (Angrist and Pischke [2008b]) an exception. Although randomized experiments have become more widespread in economics in the last two decades, they are still fairly rare, and most of the time economists have no choice but to analyze data from observational studies. Nevertheless, in most, though not

all cases, causal inferences are what economists are ultimately interested in, and causal inferences are most easily obtained and understood in randomized experiments. As Freedman wrote: "Experiments offer more reliable evidence on causation than observational studies" (Freedman [2006], abstract). Understanding that simplest of settings well is crucial before it is possible to fully appreciate and understand the complications arising in observational studies, where in many cases we analyze the data as if they arose from a (possibly more complicated) randomized experiment.

Within the setting of a completely randomized experiment, the analyses delivering exact p-values, and the methods for estimation of, and inference for, average treatment effects, are the most fundamental in the sense of requiring the fewest assumptions. Although in practice the p-value analyses are ultimately of limited value for analyzing even experimental data, understanding them is of great importance for understanding more sophisticated analyses for experimental data, as well as for understanding observational studies, which all involve additional, typically controversial, assumptions, and/or large sample, asymptotic approximations. After methods for calculating exact p-values, methods for estimating, and conducting inference for, average treatment effects, are the most basic tools.

The limitation to experiments with binary treatments is largely for convenience because many of the results readily extend to cases with discrete multi-valued treatments. Extensions to continuous treatments, or ordered treatments with many values for the treatment are more complicated. We discuss those in Part IX of this text.

1.2 Randomized Experiments

Suppose the population of interest consists of N individuals eligible for a job training program. Let $W_i \in \{0, 1\}$ denote the binary treatment assignment for unit i, with $\mathbf{W}$ the N-vector of treatment assignments. In a completely randomized experiment, we fix the number of treated units, say $N_{\rm t}$, and randomly select which $N_{\rm t}$ units out of the population of N units are to be treated. The remaining $N_{\rm c} = N - N_{\rm t}$ units are assigned to the control

group. The implication is that the distribution of the assignment vector $\mathbf{W}$ is

$$\text{pr}(\mathbf{W}=\mathbf{w}) = \left(\begin{array}{c} N \\ N_{\rm t} \end{array} \right)^{-1}, \quad \text{for all } \mathbf{w} \in \{0,1\}^N \text{ s.t. } \sum_{i=1}^{N} w_i = N_{\rm t}.$$

For each unit in the population we observe an outcome, denoted by $Y_i^{\rm obs}$, with $\mathbf{Y}^{\rm obs}$ denoting the N-component vector with i-th element equal to $Y_i^{\rm obs}$. Later we drop the "obs" superscript when it is clear what is observed and what is not. We may also observe pre-treatment variables for these units, denoted by the K-component row vector X_i for unit i, with $\mathbf{X}$ denoting the $N \times K$ matrix with i-th row equal to X_i.

As a running example in this chapter we use the experimental evaluation of the 1990s California GAIN (Greater Avenues to INdependence) job training program. See Riccio et al. [1989] and Hotz et al. [2006] for more details on these data. In this chapter we use the data from one of the four sites where the program was evaluated using a randomized experiment, Riverside. The two outcomes we focus on are total earnings in the first and seventh year after the program. In Table 1.1 observations on four individuals from this data set are presented (with the names made up). The two covariates presented are sex and age.

Table 1.1: Four Observations from the GAIN Experiment in Riverside: Earnings in First Year Post-randomization in Thousands of Dollars

Individual	Treatment W_i	Earnings $Y_i^{\rm obs}$	Sex X_{i1}	Age X_{i2}
Donald	0	0.00	M	38
Josh	1	0.45	M	34
Susan	1	12.49	F	38
Gary	0	0.00	M	27

The framework used in this book for studying randomized experiments in particular, and causal effects in general, is associated with the work by Donald Rubin (Rubin [1974], Holland [1986], Imbens and Rubin [2015]), and referred to as the *potential outcome framework* or the *Rubin Causal Model*. It builds heavily on the earlier work by Neyman (Neyman

[1923/1990, 1935]). A key notion is that of *potential outcomes*, each corresponding to the various levels of a treatment. Each of these potential outcomes are a priori observable, if the unit were to receive the corresponding treatment level, but, ex post, once a treatment is actually applied, at most one of the potential outcomes can be observed.

Following this potential outcome framework, we postulate, for each possible value of the vector of assignments $\mathbf{W}$, the existence of a potential outcome for each unit, with $Y_i(\mathbf{W})$ denoting the potential outcome for unit i given treatment vector $\mathbf{W}$. In principle, each of the N elements of $\mathbf{W}$ can take on two values, so $\mathbf{W}$ can take on $2^N = 16$ different values. Let $\mathbb{W}$ denote the set of possible values for each element of $\mathbf{W}$, in this case $\mathbb{W} = \{0, 1\}$, and let $\mathbb{W}^N$ denote the set of values for the N-component vector $\mathbf{W}$. Note that at this stage we allow the potential outcomes $Y_i(\mathbf{W})$ to depend on the full vector of assignments $\mathbf{W}$, not solely on the assignment W_i for unit i. The assumption that the potential outcomes depend only on the treatment for unit i is worth considering, but in many cases it will be controversial. To make this more specific, consider the set of four individuals from the Riverside GAIN program displayed in Table 1.1. The actual treatment vector is $\mathbf{w} = (0, 1, 1, 0)$, and so we see $Y_i^{\text{obs}} = Y_i(\mathbf{w}) = Y_i(0, 1, 1, 0)$ for individual i, but not $Y_i(0, 0, 0, 0)$, or $Y_i(1, 0, 0, 0)$, or any other of the $2^4 - 1$ potential outcomes corresponding to treatment assignments not observed.

1.3 Randomized Experiments: Testing Sharp Null Hypotheses through Fisher Exact P-values

For the setup with a completely randomized experiment Fisher [1925, 1935] showed how we can calculate exact p-values for *sharp null hypotheses*. A sharp null hypothesis is a hypothesis such that there are no nuisance parameters. In the current context means that given the observed data we can infer all the missing potential outcomes. Here the missing potential outcomes are all the $Y_i(\mathbf{w})$ for values of $\mathbf{w}$ other than the realized assignment, in this case, $\mathbf{w} = (0, 1, 1, 0)$. The most common null hypothesis is that of no effect of the treatment whatsoever:

$$H_0: \quad Y_i(\mathbf{w}) = Y_i(\mathbf{w}') \quad \text{for all } \mathbf{w}, \mathbf{w}' \in \mathbb{W}^N, \text{ and all } i = 1, \ldots, N.$$

Fisher's approach does not require us to explicitly specify an alternative hypothesis: any set of potential outcomes that is not consistent with the null hypothesis is part of the alternative hypothesis, so that:

$$H_a: \quad Y_i(\mathbf{w}) \neq Y_i(\mathbf{w}') \text{ for some } \mathbf{w}, \mathbf{w}' \in \mathbb{W}^N, \text{ and some } i.$$

We can use more general null hypotheses, for example, one where there is a particular constant additive effect of the own treatment, but it is rare that such hypotheses are of substantial interest.

1.3.1 Basics

Because the null hypothesis is sharp one can infer the distribution of any test statistic $T : \mathbb{W}^N \times \mathbb{Y}^N \times \mathbb{X}^N \mapsto \mathbb{R}$, that is, a function of the assignment vector, $\mathbf{W}$, the realized outcomes, $\mathbf{Y}^{\text{obs}}$, and any observed pre-treatment variables $\mathbf{X}$. This distribution is generated by the randomization of the treatment $\mathbf{W}$, both through the dependence of the statistic $T(\cdot)$ on $\mathbf{W}$ directly, and indirectly through the dependence of the statistic on $\mathbf{Y}^{\text{obs}}$, which itself is a function of $\mathbf{W}$. We refer to the distribution determined this way as the *randomization distribution* of the test statistic. Using this distribution, we can compare the realized value of the test statistic, $T^{\text{obs}} = T(\mathbf{W}, \mathbf{Y}^{\text{obs}}, \mathbf{X})$, to its distribution under the null hypothesis. An observed value that is "very unlikely," given the null hypothesis and the implied distribution, will be taken as evidence against the null hypothesis. How "unlikely" is given by the probability that a value as extreme as, or more extreme than, the observed value T^{obs}, would be observed, referred to as the significance level or p-value. Hence given a sharp null hypothesis, Fisher's approach involves a single choice, the test statistic. The scientific nature of the problem should govern the choice of the statistic. The statistic should be chosen in order to be sensitive to the difference between the null hypothesis and one or more substantially interesting alternative hypotheses that the researcher wants to assess.

Let us go back to Table 1.1. The assignment was $\mathbf{w} = (0, 1, 1, 0)$, and so we observe $Y_i(0, 1, 1, 0)$, for $i \in \{\text{Donald}, \text{Josh}, \text{Susan}, \text{Gary}\}$, and we do not see $Y_i(\mathbf{w}')$ for any other value $\mathbf{w}' \neq (0, 1, 1, 0)$, for any i. However, under the sharp null hypothesis that the program

had absolutely no effect on earnings, the unobserved potential outcomes are equal to the observed outcomes for each individual. Thus we can fill in $Y_i(\mathbf{w})$ for all four units and for all elements of $\mathbb{W}^N$. Table 1.2 lists part of the expanded data set. This is the first key point of the Fisher exact p-value approach; under the sharp null hypothesis, all the missing potential outcome values can be inferred from the observed ones.

Table 1.2: Four Observations from the GAIN Experiment in Riverside: Average Quarterly Earnings Post-randomization

Individual	Potential Outcomes $Y_i(0,1,1,0)$	$Y_i(1,1,0,0)$	$Y_i(1,0,1,0)$	...	Treatment	Outcome	Rank	Sex	Age
Donald	0.00	(0.00)	(0.00)		0	0.00	-1.0	M	38
Josh	0.45	(0.45)	(0.45)		1	0.45	0.5	M	34
Susan	12.49	(12.49)	(12.49)		1	12.49	1.5	F	38
Gary	0.00	(0.00)	(0.00)		0	0.00	-1.0	M	27

Table 1.3: Randomization Distribution for four observations from Riverside GAIN data. Observed Values in Boldface (R_i is Rank(Y_i))

$\mathbf{W}$	T^{dif}	T^{rank}
$(0,0,1,1)$	6.02	0.5
$(\mathbf{0},\mathbf{1},\mathbf{1},\mathbf{0})$	**6.47**	**2.0**
$(0,1,0,1)$	-6.02	-0.5
$(1,0,1,0)$	6.02	0.5
$(1,0,0,1)$	-6.47	-2.0
$(1,1,0,0)$	-6.02	-0.5

1.3.2 The Choice of Statistic

The most common choice of statistic is the difference in average outcomes by treatment status:

$$T^{\text{dif}} = \overline{Y}_{\text{t}}^{\text{obs}} - \overline{Y}_{\text{c}}^{\text{obs}}, \qquad \text{where } \overline{Y}_{\text{t}}^{\text{obs}} = \frac{\sum W_i Y_i^{\text{obs}}}{\sum W_i}, \text{ and } \overline{Y}_{\text{c}}^{\text{obs}} = \frac{\sum (1-W_i) Y_i^{\text{obs}}}{\sum (1-W_i)}.$$

If we use this for the four individuals in the population in Table 1.1, we find that the value of the statistic is $T^{\rm dif} = 6.472$. How extreme is this value in the distribution generated by the randomization? Not very extreme, because two assignment vectors lead to a value at least as large as 6.47 in absolute value, so the p-value is $2/6 = 1/3$.

An alternative to the simple difference in average outcomes by treatment status is to transform the outcomes to ranks. The normalized rank of the outcome $Y_i^{\rm obs}$ for unit i is calculated as:

$$R_i^{\rm obs} = R(i:\mathbf{Y}^{\rm obs}) = \sum_{j=1}^{N} \mathbf{1}_{Y_j^{\rm obs}<Y_i^{\rm obs}} + \frac{1}{2}\left(1+\sum_{j=1}^{N}\mathbf{1}_{Y_j^{\rm obs}=Y_i^{\rm obs}}\right) - \frac{N+1}{2}.$$

In this definition we account for ties among the values, and we subtract the mean rank $(N+1)/2$ so that the rank has an average value of zero over the population. The difference in average ranks by treatment status is

$$T^{\rm rank} = \overline{R}_{\rm t}^{\rm obs} - \overline{R}_{\rm c}^{\rm obs} \qquad \text{where } \overline{R}_{\rm t}^{\rm obs} = \frac{\sum W_i R_i^{\rm obs}}{\sum W_i}, \text{ and } \overline{R}_{\rm c}^{\rm obs} = \frac{\sum (1-W_i) R_i^{\rm obs}}{\sum 1-W_i}.$$

If we re-assign the treatments, the value of the rank for individual i, $R_i^{\rm obs}$, would not actually change, but the rank statistic $T^{\rm rank}$ might change, generating a distribution for this statistic. For the population of four individuals we find that the difference in average ranks for the actual assignment is 2, with a p-value of 1/3. In general, using ranks is a useful statistic if there are outliers in the distribution of the original outcomes. In some cases, simply taking logarithms may suffice to deal with the outliers, but transforming to ranks works well even when taking logarithms is not sufficient because the outliers are very extreme, or when it is not feasible, because there are negative numbers or zeros among the outcomes. With a large fraction of zeros, however, using ranks may not work very well.

There are many other statistics one could consider, including some that depend on the covariates. Let us consider one more statistic that can be useful in settings where the outcome distribution has a large number of zeros and a thick-tailed distribution for the non-zero outcomes. Let $q_{\rm c}$ and $q_{\rm t}$ be the fraction of positive outcomes for the control and treated sample, and let $\overline{Y}_{\rm c}^{\rm obs,pos}$ and $\overline{Y}_{\rm t}^{\rm obs,pos}$ be the average outcome for control and treated units with a positive outcome. One statistic is based on the fraction of units with positive

outcomes in the two treatment groups:

$$T^{\text{pos}} = \frac{(q_{\text{t}} - q_{\text{c}})^2}{q_{\text{c}}(1-q_{\text{c}})/N_{\text{c}} + q_{\text{t}}(1-q_{\text{t}})/N_{\text{t}}}.$$

This statistic is sensitive to changes in the fraction of positive outcomes.

Another statistic we can use compares the average rank for the outcomes with a positive value. Let $N^{\text{pos}} = \sum_{i=1}^{N} \mathbf{1}_{Y_i^{\text{obs}}>0}$ be the number of units with a positive outcome. Define

$$R_i^{\text{obs,pos}} = R^{\text{pos}}(i : \mathbf{Y}^{\text{obs}}) = \sum_{j:Y_j^{\text{obs}}>0} \mathbf{1}_{Y_j^{\text{obs}}<Y_i^{\text{obs}}} + \frac{1}{2}\left(1 + \sum_{j:Y_j^{\text{obs}}>0} \mathbf{1}_{Y_j^{\text{obs}}=Y_i^{\text{obs}}}\right) - \frac{N^{\text{pos}}+1}{2},$$

and

$$T^{\text{rank,pos}} = \frac{\left(\overline{R}_{\text{t}}^{\text{obs,pos}} - \overline{R}_{\text{c}}^{\text{obs,pos}}\right)^2}{s^2_{R_{\text{t}}^{\text{obs,pos}}}/N_{\text{t}}^{\text{pos}} + s^2_{R_{\text{c}}^{\text{obs,pos}}}/N_{\text{c}}^{\text{pos}}}.$$

Then we can create a single combined statistic that is has power against both types of alternatives:

$$T^{\text{comb}} = T^{\text{pos}} + T^{\text{rank,pos}}.$$

Alternatively we could use a convex combination of T^{pos} and $T^{\text{rank,pos}}$ with weights depending on the variances.

1.3.3 Computation of p-values

In cases with N very small, it may be possible to calculate the exact p-values by simply calculating the value of the statistic for all values of the assignment vector, and then calculating the fraction of the values of the statistic that is as large as, or larger than, in absolute value, the observed value of the statistic. If N is large (and the fraction of treated units is not close to zero or one), this is not feasible. In that case a more practical choice may be to approximate the p-value by sampling from the randomization distribution. We would draw B values $\mathbf{w}_b$ from the distribution of $\mathbf{W}$, calculate the value of the statistic for each draw $\mathbf{w}_b$, and approximate the p-value as

$$\hat{p} = \frac{1}{B}\sum_{b=1}^{B} \mathbf{1}_{|T(\mathbf{w}_b,\mathbf{Y}(\mathbf{w}_b),\mathbf{X})| \geq |T^{\text{obs}}|}.$$

If B is chosen sufficiently large, this will approximate the true p-value accurately. Note that the researcher can control the accuracy of the approximation by choosing the value of B.

1.3.4 P-values for the GAIN Experiments

Let us illustrate these ideas for the Riverside GAIN data, with $N = 5,445$ individuals in the sample. We use two outcome measures, earnings in the first and earnings in the seventh year after the start of the program. The number of individuals in the treatment group is $N_{\rm t} = 4,405$, and the number of individuals in the control group is $N_{\rm c} = 1,040$. Summary statistics are presented in Table 1.4.

Table 1.4: Summary Statistics for the Riverside Gain Data, Earnings in 1,000's of dollars, for 5445 Individuals

	Year 1 Earnings	Year 7 Earnings
Mean	2.37	4.29
Standard Deviation	4.95	8.76
Fraction of Individuals with Zero Earnings	0.53	0.61

We focus on three statistics, the difference in average outcomes by treatment group $T^{\rm dif}$, the difference in average ranks, $T^{\rm rank}$, and the combination statistic $T^{\rm comb}$. Note that the fraction of individuals with zero earnings is approximately 53% in the first year post-treatment. The results are reported in Table 1.5. The p-values are based on $B = 1,000,000$ draws from the randomization distribution, implying that they are highly accurate: the standard error of the p-values is $\sqrt{p(1-p)/1,000,000} \leq 0.0005$. For the Year 1 earnings all p-values are less than 0.0005. For Year 7 earnings the p-values vary between 0.056 and 0.036, with the statistic $T^{\rm comb}$ that explicitly takes account of the mass point at zero earnings the smallest.

Figure 1.1 shows the distribution of the statistic $T^{\rm dif}$, for seventh year earnings, under

Table 1.5: Exact P-values for Fisher Randomization Tests (Based on $B = 1,000,000$ draws from the Randomization Distribution): The Riverside Gain Data

	Year 1 Earnings	Year 7 Earnings
T^{dif}	0.000	0.056
T^{rank}	0.000	0.055
T^{comb}	0.000	0.036

the null hypothesis. The vertical line shows the value of the statistic, 0.575. For comparison we include a normal approximation to the distribution of the statistic, centered at zero, and with a standard deviation equal to the standard deviation under the randomization distribution, 0.302. Figures 1.2 and 1.3 repeat this exercise for random samples of size 500, 100, and 20 from this population of size 5,445. This illustrates how the quality of the approximation by a Normal distribution descreases as the sample size goes down.

Figure 1.1: Distribution of the statistic T^{dif} for seventh year earnings, N = 5445

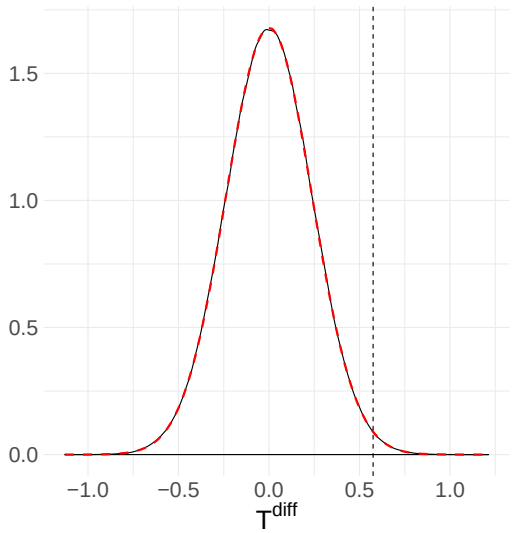


Figure 1.2: Distribution of the statistic T^{dif} for seventh year earnings, N = 500

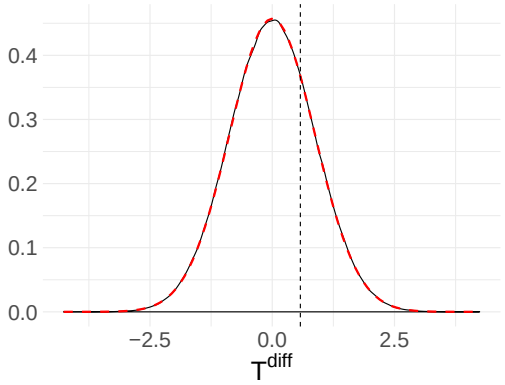


Figure 1.3: Distribution of the statistic T^{dif} for seventh year earnings, N = 100

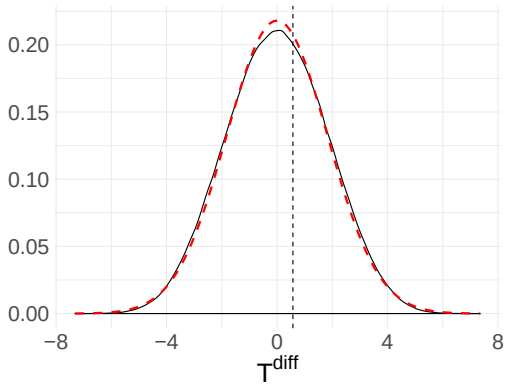
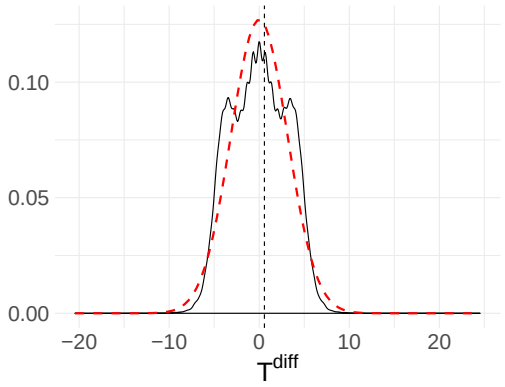


Figure 1.4: Distribution of the statistic T^{dif} for seventh year earnings, N = 20



1.4 Randomization Inference for Average Treatment Effects

In this section we focus on a different question in the context of completely randomized experiments. Instead of asking whether there is any evidence of causal effects, we are interested in estimating, and doing inference for, average causal effects. This requires assumptions beyond the random assignment of the treatment, and, for the inference part of the question, relies on large sample approximations. We will be precise about the nature of the large sample approximations here, and draw a clear distinction between estimands defined for the sample at hand and estimands defined in terms of a larger population that the sample is drawn from. Mostly we focus on the sample average treatment effect, the difference between the average outcome in the sample if every unit in the sample was treated and the average outcome in the sample if no unit in the sample was treated. A leading alternative is the population average treatment effect, the average causal effect in the super-population that the sample was drawn from. Although the difference between these two estimands is immaterial for estimation, it is important for inference as will be shown below.

1.4.1 The Stable Unit Treatment Value Assumption

In principle there are for each of the N units 2^N different potential outcomes, one for each possible value for $\mathbf{w}$. Any comparison of $Y_i(\mathbf{w}')$ and $Y_i(\mathbf{w})$ for pairs $\mathbf{w} \neq \mathbf{w}'$ is a causal effect.

A natural estimand is the average outcome if all units are treated ($\mathbf{w} = \mathbf{1}$) versus nobody is treated ($\mathbf{w} = \mathbf{0}$):

$$\tau^{\text{fs}} = \frac{1}{N} \sum_{i=1}^{N} \Big(Y_i(\mathbf{1}) - Y_i(\mathbf{0}) \Big).$$

However, we do not observe $Y_i(\mathbf{0})$ or $Y_i(\mathbf{1})$ for any unit, and so without more structure on the problem it is difficult to learn much about this, or any other average or individual causal effects. One common assumption is potential outcomes for unit i do not vary with

assignments for other units. This assumption is often referred to as the no-interference assumption or stable-unit-treatment-value assumption (sutva, Rubin [1980]):

Assumption 1.1 (Stable Unit Treatment Value Assumption)
If $\mathbf{w}'_i = \mathbf{w}_i$, *then* $Y_i(\mathbf{w}') = Y_i(\mathbf{w})$.

Under this assumption there are only two potential outcomes for each individual, which we denote by $Y_i(0)$ and $Y_i(1)$, depending only on the treatment assigned to unit i. The sample average effect of the treatment can now be written as

$$\tau^{\text{fs}} = \frac{1}{N}\sum_{i=1}^{N}\Big(Y_i(1) - Y_i(0)\Big).$$

The superscript "fs" (finite-sample) captures the notion that this is the average effect for the sample at hand, rather than the average effect for a possibly hypothetical population that the sample is drawn from.

1.4.2 Unbiased Estimation of the Average Treatment Effect

The assignment mechanism is the same as in the discussion of the calculation of exact p-values. Out of the sample of N units, N_{t} randomly selected units are assigned to treatment and the remaining N_{c} are assigned to control. Given randomization, the intuitive estimator for the average treatment effect is the difference in the average outcomes for those assigned to the treatment versus those assigned to the control:

$$\hat{\tau} = \overline{Y}_{\text{t}}^{\text{obs}} - \overline{Y}_{\text{c}}^{\text{obs}},$$

where

$$\overline{Y}_{\text{t}}^{\text{obs}} = \frac{1}{N_{\text{t}}}\sum_{i:W_i=1} Y_i^{\text{obs}}, \qquad \text{and} \quad \overline{Y}_{\text{c}}^{\text{obs}} = \frac{1}{N_{\text{c}}}\sum_{i:W_i=0} Y_i^{\text{obs}}.$$

To see that this estimator, $\hat{\tau} = \overline{Y}_{\text{t}}^{\text{obs}} - \overline{Y}_{\text{c}}^{\text{obs}}$, is unbiased for τ_{fs}, consider the statistic

$$T_i = \frac{W_i Y_i^{\text{obs}}}{N_{\text{t}}/N} - \frac{(1-W_i)Y_i^{\text{obs}}}{N_{\text{c}}/N}.$$

The average of this statistic, over the finite population, is equal to the estimator, $\hat{\tau} = \overline{Y}_{\mathrm{t}}^{\mathrm{obs}} - \overline{Y}_{\mathrm{c}}^{\mathrm{obs}}$:

$$\frac{1}{N}\sum_{i=1}^{N} T_i = \frac{1}{N}\sum_{i=1}^{N} \frac{W_i Y_i^{\mathrm{obs}}}{N_{\mathrm{t}}/N} - \frac{1}{N}\sum_{i=1}^{N} \frac{(1-W_i)Y_i^{\mathrm{obs}}}{(N_{\mathrm{c}})/N} = \overline{Y}_{\mathrm{t}}^{\mathrm{obs}} - \overline{Y}_{\mathrm{c}}^{\mathrm{obs}}.$$

Using the fact that Y_i^{obs} is equal to $Y_i(1)$ if $W_i = 1$ and $Y_i(0)$ if $W_i = 0$, we can rewrite T_i in terms of the potential outcomes as:

$$T_i = \frac{W_i Y_i(1)}{N_{\mathrm{t}}/N} - \frac{(1-W_i)Y_i(0)}{N_{\mathrm{c}}/N} = Y_i(1) - Y_i(0) + (W_i - N_{\mathrm{t}}/N)\frac{Y_i(1)(/nc/N) + Y_i(0)(N_{\mathrm{t}}/N)}{(N_{\mathrm{t}}/N)(N_{\mathrm{c}}/N)}. \tag{1.1}$$

Writing the statistic in this form is important because it separates it cleanly into stochastic and deterministic parts. Specifically, because we take the potential outcomes as fixed, the only component in this statistic that is stochastic is the treatment assignment, W_i. Given our completely randomized experiment (N units, with N_{t} randomly assigned to treatment), W_i has a Bernoulli distribution with parameter N_{t}/N. Hence the expectation of W_i—the probability of receiving the active treatment—is equal to N_{t}/N and the probability of receiving the control treatment, $\mathbb{E}[1 - W_i] = 1 - \mathbb{E}[W_i]$, is equal to N_{c}/N.

Using these results it follows that the expectation of T_i is equal to the unit-level causal effect, $Y_i(1) - Y_i(0)$:

$$\begin{aligned}\mathbb{E}[T_i|\mathbf{Y}(0), \mathbf{Y}(1)] &= Y_i(1) - Y_i(0) - \mathbb{E}[W_i - N_{\mathrm{t}}/N]\frac{Y_i(1)(N_{\mathrm{c}}/N) + Y_i(0)(N_{\mathrm{t}}/N)}{(N_{\mathrm{t}}/N)(N_{\mathrm{c}}/N)} \\ &= Y_i(1) - Y_i(0),\end{aligned}$$

where $\mathbf{Y}(0)$ and $\mathbf{Y}(1)$ are the N-vectors with typical element $Y_i(0)$ and $Y_i(1)$, respectively. Because each T_i is unbiased for the unit-level causal effect $Y_i(1) - Y_i(0)$, and the average of T_i is equal to our estimator, $\overline{Y}_{\mathrm{t}}^{\mathrm{obs}} - \overline{Y}_{\mathrm{c}}^{\mathrm{obs}}$, it follows directly that the expected value of our estimator is equal to the population average treatment effect:

$$\begin{aligned}\mathbb{E}\left[\overline{Y}_{\mathrm{t}}^{\mathrm{obs}} - \overline{Y}_{\mathrm{c}}^{\mathrm{obs}}\middle|\, \mathbf{Y}(0), \mathbf{Y}(1)\right] &= \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N} T_i \middle|\, \mathbf{Y}(0), \mathbf{Y}(1)\right] = \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}[T_i|\mathbf{Y}(0), \mathbf{Y}(1)] \\ &= \frac{1}{N}\sum_{i=1}^{N}\Big(Y_i(1) - Y_i(0)\Big) = \tau_{\mathrm{fs}}.\end{aligned}$$

Hence our estimator, $\overline{Y}_{\mathrm{t}}^{\mathrm{obs}} - \overline{Y}_{\mathrm{c}}^{\mathrm{obs}}$, is unbiased for the average treatment effect τ_{fs}, given data from a completely randomized experiment.

1.4.3 Inference for the Sample Average Treatment Effect

Neyman [1923/1990, 1935] was also interested in inference for the average treatment effect Deriving methods for conducting inference involved two steps: first, deriving the variance of the estimator for the average treatment effect; and second, developing estimators for this variance. In addition, Neyman sought to create confidence intervals for the population average treatment effect which also requires an appeal to the central limit theorem for large sample normality.

In this section we focus on the first step, calculating the variance of the estimator $\hat{\tau} = \overline{Y}_{\rm t}^{\rm obs} - \overline{Y}_{\rm c}^{\rm obs}$, given the assumption of a completed randomized experiment. This assumption complicates the derivation of the variance because the assignments for different units are not independent. With the number of treated units fixed at $N_{\rm t}$, the fact that unit 1 is assigned to the treatment lowers the probability that unit 2 will received the same treatment.

To calculate the variance of $\overline{Y}_{\rm t}^{\rm obs} - \overline{Y}_{\rm c}^{\rm obs}$, we need the second and cross moments of the random variable W_i, $\mathbb{E}[W_i^2]$ and $\mathbb{E}[W_i W_j]$. Because $W_i \in \{0, 1\}$, it follows that $W_i^2 = W_i$ and thus

$$\mathbb{E}\left[W_i^2\right] = \mathbb{E}[W_i] = \frac{N_{\rm t}}{N},$$

and

$$\mathbb{E}[W_i W_j] = {\rm pr}(W_i = 1){\rm pr}(W_j = 1|W_i = 1) = \frac{N_{\rm t}(N_{\rm t} - 1)}{N(N-1)},$$

for $i \neq j$ (and not $N_{\rm t}^2/N^2 = pr(W_i = 1){\rm pr}(W_j = 1)$), because conditional on $W_i = 1$ there are $N_{\rm t} - 1$ treated units remaining out of $N - 1$ total remaining.

A conceptually simple but long calculation, given in the Appendix, shows that the variance of $\overline{Y}_{\rm t}^{\rm obs} - \overline{Y}_{\rm c}^{\rm obs}$ is equal to:

$$\mathbb{V}(\overline{Y}_{\rm t}^{\rm obs} - \overline{Y}_{\rm c}^{\rm obs}) = \frac{S_{\rm c}^2}{N_{\rm c}} + \frac{S_{\rm t}^2}{N_{\rm t}} - \frac{S_{\rm ct}^2}{N}, \tag{1.2}$$

where $S_{\rm c}^2$ and $S_{\rm t}^2$ are the variance of $Y_i(0)$ and $Y_i(1)$ in the population, defined as:

$$S_{\rm c}^2 = \frac{1}{N-1}\sum_{i=1}^{N}\Big(Y_i(0) - \overline{Y}(0)\Big)^2, \qquad \text{and} \quad S_{\rm t}^2 = \frac{1}{N-1}\sum_{i=1}^{N}\Big(Y_i(1) - \overline{Y}(1)\Big)^2,$$

and $S^2_{\rm ct}$ is the population variance of the unit-level treatment effect:

$$S^2_{\rm ct} = \frac{1}{N-1}\sum_{i=1}^{N}\Big(Y_i(1) - Y_i(0) - \tau_{\rm fs}\Big)^2.$$

The first two components of the variance are fairly intuitive. Recall that the sample average treatment effect is the difference in average potential outcomes for the two treatment groups: $\tau_{\rm fs} = \overline{Y}(1) - \overline{Y}(0)$. To estimate $\tau_{\rm fs}$, we first estimate $\overline{Y}(1)$, the population average potential outcome under treatment, with the sample average outcome for the $N_{\rm t}$ treated units, $\overline{Y}^{\rm obs}_{\rm t}$. This estimator is unbiased for the average of the potential outcome given treatment, $\overline{Y}(1)$. The population variance of $Y_i(1)$ is $S^2_{\rm t} = \sum_i (Y_i(1) - \overline{Y}_1)^2/(N-1)$. Given this variance for $Y_i(1)$, the variance for an average from a random sample of size $N_{\rm t}$ from a large population is $S^2_{\rm t}/N_{\rm t} = \sum_i (Y_i(1) - \overline{Y}(1))^2/(N_{\rm t}(N-1))$. Similarly the average outcome for the $N_{\rm c}$ units assigned to control, $\overline{Y}^{\rm obs}_{\rm c}$, is unbiased for the population average outcome under the control, $\overline{Y}(0)$, and its variance is $S^2_{\rm c}/N_{\rm c}$. These results follow by direct calculation or by using standard results from the analysis of simple random samples. Given a completely randomized experiment, the $N_{\rm t}$ treated units provide a simple random sample of the N $Y_i(1)$ values and the $N_{\rm c}$ control units provide a simple random sample of the N $Y_i(0)$ values.

The third component of this variance, $S^2_{\rm ct}/N$, is trickier. It depends on the sample variance of the unit-level treatment effect, $Y_i(1) - Y_i(0)$. This element is difficult to estimate because we do not observe any unit-level treatment effects. If the treatment effect were constant within the population, this third term is equal to zero. If the treatment effect varies, $S^2_{\rm ct}$ is positive. Because it is subtracted from the sum of the first two elements in the expression for the variance, (1.2), a positive value for S^2_{01} reduces the variance of the estimator for the average treatment effect.

Given the variance of the estimator, $\overline{Y}^{\rm obs}_{\rm t} - \overline{Y}^{\rm obs}_{\rm c}$, the next step is to develop an estimator for this variance. To do this, consider separately each of the three components of the variance in equation (1.2).

The numerator of the first term, the population variance of the potential control out-

come vector, $\mathbf{Y}(0)$, is equal to

$$S_{\rm c}^2 = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i(0) - \overline{Y}(0))^2.$$

An unbiased estimator for this numerator is

$$s_{\rm c}^2 = \frac{1}{N_{\rm c}-1}\sum_{i:W_i=0}\left(Y_i(0) - \overline{Y}_{\rm c}^{\rm obs}\right)^2 = \frac{1}{N_{\rm c}-1}\sum_{i:W_i=0}\left(Y_i^{\rm obs} - \overline{Y}_{\rm c}^{\rm obs}\right)^2.$$

Similarly we can estimate $S_{\rm t}^2$, the population variance of $Y_i(1)$, by

$$s_{\rm t}^2 = \frac{1}{N_{\rm t}-1}\sum_{i:W_i=1}\left(Y_i(1) - \overline{Y}_{\rm t}^{\rm obs}\right)^2 = \frac{1}{N_{\rm t}-1}\sum_{i|W_i=1}\left(Y_i^{\rm obs} - \overline{Y}_{\rm t}^{\rm obs}\right)^2.$$

The third term, $S_{\rm ct}^2$ is difficult to estimate because we do not observe the unit-level treatment effect for any unit. We therefore have no direct observations on the variation in the treatment effect across the population and there is no unbiased estimator for $S_{\rm ct}^2$. If the treatment effect is additive (so that $Y_i(1) - Y_i(0) = \tau$ for all i), then this variance is equal to zero and the third term vanishes. Under this circumstance we can obtain an unbiased estimator for the variance as:

$$\hat{\mathbb{V}}_{\rm neyman} = \hat{\mathbb{V}}\left(\overline{Y}_{\rm t}^{\rm obs} - \overline{Y}_{\rm c}^{\rm obs}\right) = \frac{s_{\rm c}^2}{N_{\rm c}} + \frac{s_{\rm t}^2}{N_{\rm t}}. \tag{1.3}$$

This unbiased estimator for the variance under constant treatment effects is widely used, even when the assumption of an additive treatment effect is inappropriate. There are two reasons for this estimator's popularity. First, by ignoring the third element of this variance, the expected value of $\hat{\mathbb{V}}$ is at least as large as the true variance, irrespective of the heterogeneity in the treatment effect. Hence in large samples, confidence intervals generated using this estimator of the variance will be *conservative* with actual coverage at least as large, but not necessarily equal to, the nominal coverage. Note that this statement still needs to be qualified by the clause "in large samples," because we rely on the large sample approximations in order to apply a central limit theorem to construct confidence intervals.

Let $\hat{\mathbb{V}}$ be an estimate of the variance of $\hat{\tau}$, and suppose we wish to construct a 95% confidence interval for $\tau_{\rm fs}$. We use a normal approximation to the randomization distribution of $\hat{\tau}$. This is somewhat inconsistent with the stress in this chapter on finite sample

properties of the estimator and the variance, but it is necessary because of the lack of a parametric model for the distribution of the potential outcomes. Normality is often a good approximation to the randomization distribution of $\hat{\tau}$ even in fairly small samples. We then use the 2.5th and 97.5th percentile of the standard Normal distribution, -1.96 and 1.96, to calculate an implied 95% confidence interval equal to:

$$\text{CI}_{0.95}\left(\tau^{\text{fs}}\right) = \left(\hat{\tau} - 1.96\sqrt{\hat{\mathbb{V}}}, \hat{\tau} + 1.96\sqrt{\hat{\mathbb{V}}}\right).$$

More generally, if we wish to construct a confidence interval with confidence level $(1-\alpha)\times 100\%$, as usual we look up the $\alpha/2$ quantile of the standard Normal distribution, denoted by $c_{\alpha/2}$, and construct the confidence interval:

$$\text{CI}_{1-\alpha}(\tau) = \left(\hat{\tau} - c_{\alpha/2}\sqrt{\hat{\mathbb{V}}}, \hat{\tau} + c_{\alpha/2}\sqrt{\hat{\mathbb{V}}}\right).$$

This approximation applies to all estimates of the variance, and in large samples the resulting confidence intervals are valid under the same assumptions that make the corresponding sample variance an unbiased or upwardly biased estimate of the true variance.

1.4.4 Inference for Super-population Average Treatment Effects

Now consider the setting where the sample with N units in the completely randomized experiment is itself a randomly drawn sample from a larger population of size n. Mostof the time we will think of the case where n is large enough so we can think of it as infinitely large, relative to the sample. Considering our N units as a random sample of the population of interest, rather than as the population of interest itself, induces a distribution on the potential outcomes and pre-treatment variables. The potential outcome values of a specific unit i in the sample can be viewed as a draw from the distributions within the full population and are stochastic. The joint distribution of the two potential outcomes in turn induces a distribution on the unit-level treatment effect and thus on the average of the unit-level treatment effect within the experimental sample. To be explicit about this super population perspective, we will index the average treatment effect by "sp" to denote the super-population average treatment effect, to contrast with the superscript fs which denotes the average effect in the sample. Thus $\tau^{\text{sp}} = \mathbb{E}[Y(1) - Y(0)]$ is the expected value

(population average) of the unit-level treatment effect under the distribution induced by random sampling from the super population. Let $\mu_{\rm c}$, $\mu_{\rm t}$, $\sigma^2_{\rm c}$, and $\sigma^2_{\rm t}$ be the popularion mean and variance of the two potential outcomes, and let $\sigma^2_{\rm ct}$ be the variance of the unit-level treatment effect within this super-population.

Because the sample of size N is now assumed to be a simple random sample from a large super-population, the average treatment effect within the sample, $\tau^{\rm fs} = \overline{Y}(1) - \overline{Y}(0)$, can itself be viewed as a random variable with expectation equal to the average treatment effect in the super-population, $\tau^{\rm sp}$:

$$\mathbb{E}\left[\tau^{\rm fs}\right] = \mathbb{E}[\overline{Y}(1) - \overline{Y}(0)] = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[Y_i(1) - Y_i(0)] = \mathbb{E}\left[Y_i(1) - Y_i(0)\right] = \tau^{\rm sp}.$$

Thus, under this random sampling assumption, the average treatment effect in the sample is unbiased for the average treatment effect in the super-population. Given our second assumption that the variance of the unit-level treatment effect within the super-population is equal to $\sigma^2_{\rm ct}$, the variance of $\tau_{\rm fs}$ is equal to

$$\mathbb{V}(\tau_{\rm fs}) = \mathbb{V}(\overline{Y}(1) - \overline{Y}(0)) = \sigma^2_{\rm ct}/N. \tag{1.4}$$

The variance of the estimator $\hat{\tau} = \overline{Y}^{\rm obs}_{\rm t} - \overline{Y}^{\rm obs}_{\rm c}$, given this super-population interpretation can be written as:

$$\begin{aligned}\mathbb{V}\left(\hat{\tau}\right) &= \mathbb{E}\left[(\hat{\tau} - \tau^{\rm sp})^2\right] = \mathbb{E}\Big[\Big(\overline{Y}^{\rm obs}_{\rm t} - \overline{Y}^{\rm obs}_{\rm c} - \mathbb{E}\left[\overline{Y}^{\rm obs}_{\rm t} - \overline{Y}^{\rm obs}_{\rm c}\right]\Big)^2\Big] \\ &= \mathbb{E}\Big[\Big(\overline{Y}^{\rm obs}_{\rm t} - \overline{Y}^{\rm obs}_{\rm c} - \mathbb{E}\left[\overline{Y}(1) - \overline{Y}(0)\right]\Big)^2\Big],\end{aligned}$$

where the second equality holds because both $\mathbb{E}[\overline{Y}^{\rm obs}_{\rm t} - \overline{Y}^{\rm obs}_{\rm c}]$ and $\mathbb{E}[\overline{Y}(1) - \overline{Y}(0)]$ are equal to $\tau_{\rm sp}$, as shown above. Adding and subtracting $\tau_{\rm fs} = \overline{Y}(1) - \overline{Y}(0)$ within the expectation, this variance is equal to:

$$\begin{aligned}\mathbb{V}\left(\hat{\tau}\right) &= \mathbb{E}\Big[\Big(\overline{Y}^{\rm obs}_{\rm t} - \overline{Y}^{\rm obs}_{\rm c} - \big(\overline{Y}(1) - \overline{Y}(0)\big) + \big(\overline{Y}(1) - \overline{Y}(0)\big) - \mathbb{E}\left[\overline{Y}(1) - \overline{Y}(0)\right]\Big)^2\Big] \\ &= \mathbb{E}\Big[\Big(\overline{Y}^{\rm obs}_{\rm t} - \overline{Y}^{\rm obs}_{\rm c} - \big(\overline{Y}(1) - \overline{Y}(0)\big)\Big)^2\Big] \\ &+\mathbb{E}\Big[\Big(\big(\overline{Y}(1) - \overline{Y}(0)\big) - \mathbb{E}\left[\overline{Y}(1) - \overline{Y}(0)\right]\Big)^2\Big] \\ &+2\mathbb{E}\Big[\Big(\overline{Y}^{\rm obs}_{\rm t} - \overline{Y}^{\rm obs}_{\rm c} - \big(\overline{Y}(1) - \overline{Y}(0)\big)\Big)\Big(\big(\overline{Y}(1) - \overline{Y}(0)\big) - \mathbb{E}\left[\overline{Y}(1) - \overline{Y}(0)\right]\Big)\Big].\end{aligned}$$

The third term of this last equation is equal to zero because the expectation of the first factor, $\overline{Y}_{\mathrm{t}}^{\mathrm{obs}} - \overline{Y}_{\mathrm{c}}^{\mathrm{obs}} - \big(\overline{Y}(1) - \overline{Y}(0)\big)$, conditional on the N–vectors $\mathbf{Y}(0)$ and $\mathbf{Y}(1)$, is zero. Hence the variance reduces to:

$$\begin{aligned}\mathbb{V}\left(\hat{\tau}\right) = \mathbb{E}\Big[\Big\{\overline{Y}_{\mathrm{t}}^{\mathrm{obs}} - \overline{Y}_{\mathrm{c}}^{\mathrm{obs}} - \Big(\overline{Y}(1) - \overline{Y}(0)\Big)\Big\}^2\Big] \\ + \mathbb{E}\Big[\Big\{\overline{Y}(1) - \overline{Y}(0) - \mathbb{E}[Y(1) - Y(0)]\Big\}^2\Big]. \qquad (1.5)\end{aligned}$$

Earlier we showed that $\mathbb{E}\left[\overline{Y}_{\mathrm{t}}^{\mathrm{obs}} - \overline{Y}_{\mathrm{c}}^{\mathrm{obs}} \middle| \mathbf{Y}(0), \mathbf{Y}(1)\right] = \overline{Y}(1) - \overline{Y}(0)$, hence by iterated expectations the first term on the right side is equal to the expectation of the conditional variance of $\overline{Y}_1^{\mathrm{obs}} - \overline{Y}_0^{\mathrm{obs}}$ (conditional on the N-vector of potential outcomes $\mathbf{Y}(0)$ and $\mathbf{Y}(1)$). This is equal to $S_{\mathrm{c}}^2/N_{\mathrm{c}} + S_{\mathrm{t}}^2/N_{\mathrm{t}} - S_{\mathrm{ct}}^2/N$, as in equation (1.2). Recall that these earlier calculations were made when assuming that the sample N represented the full population, and thus were conditional on $\mathbf{Y}(0)$ and $\mathbf{Y}(1)$. The unconditional expectation of this value within the super-population is $\sigma_{\mathrm{c}}^2/N_{\mathrm{c}} + \sigma_{\mathrm{t}}^2/N_{\mathrm{t}} - \sigma_{\mathrm{ct}}^2/N$. The expectation of the second term on the right side of equation (1.5), the variance of the average treatment effect in the super-population, is equal to σ_{ct}^2/N, as we saw in equation (1.4). Thus the variance of $\overline{Y}_1 - \overline{Y}_0$ in the super-population is equal to:

$$\mathbb{V}_{\mathrm{sp}}(\overline{Y}_{\mathrm{t}}^{\mathrm{obs}} - \overline{Y}_{\mathrm{c}}^{\mathrm{obs}}) = \frac{\sigma_{\mathrm{c}}^2}{N_{\mathrm{c}}} + \frac{\sigma_{\mathrm{t}}^2}{N_{\mathrm{t}}},$$

which we can estimate by substituting s_{c}^2 and s_{t}^2 for σ_{c}^2 and σ_{t}^2, respectively:

$$\hat{\mathbb{V}}_{\mathrm{sp}} = \frac{s_{\mathrm{c}}^2}{N_{\mathrm{c}}} + \frac{s_{\mathrm{t}}^2}{N_{\mathrm{t}}}.$$

Notice that $\hat{\mathbb{V}}_{\mathrm{sp}}$ is equal to the previously introduced conservative estimator of the variance for the finite population average treatment effect estimator $\hat{\mathbb{V}}_{\mathrm{neyman}}$ presented in Equation (1.3).

1.4.5 Least Squares Regression and Estimating Average Treatment Effects

One can also view the estimator $\hat{\tau} = \overline{Y}_{\mathrm{t}}^{\mathrm{obs}} - \overline{Y}_{\mathrm{c}}^{\mathrm{obs}}$ as the result of minimizing the sum of squared residuals in a linear regression. Consider the regression function

$$Y_i^{\mathrm{obs}} = \alpha + \tau W_i + \varepsilon_i.$$

Suppose we estimate α and τ by least squares:

$$(\hat{\alpha}^{\rm ols},\hat{\tau}^{\rm ols}) = \arg\min_{\alpha,\tau} \sum_{i=1}^{N} \left(Y_i^{\rm obs} - \alpha - \tau W_i\right)^2.$$

This leads to

$$\hat{\alpha}^{\rm ols} = \overline{Y}_{\rm c}^{\rm obs}, \qquad \hat{\tau}^{\rm ols} = \hat{\tau} = \overline{Y}_{\rm t}^{\rm obs} - \overline{Y}_{\rm c}^{\rm obs}.$$

Now consider the standard variance estimators for $\hat{\tau}_{\rm ols}$, to compare to the Neyman variance we discussed before. Define the residual

$$\hat{\varepsilon}_i = Y_i^{\rm obs} - \hat{\alpha}^{\rm ols} - \hat{\tau}^{\rm ols} W_i = \begin{cases} Y_i(0) - \overline{Y}_{\rm c}^{\rm obs} & \text{if } W_i = 0, \\ Y_i(1) - \overline{Y}_{\rm t}^{\rm obs} & \text{if } W_i = 1. \end{cases}$$

The homoskedastic variance estimator is

$$\hat{\mathbb{V}}_{\rm homo} = \frac{\sum_{i=1}^N \hat{\varepsilon}_i^2/N}{\sum_{i=1}^N (W_i - \overline{W})^2}.$$

The robust, Eicker-Huber-White variance estimator is

$$\hat{\mathbb{V}}_{\rm ehw} = \frac{\sum_{i=1}^N \hat{\varepsilon}_i^2 (W_i - \overline{W})^2/N}{\left(\sum_{i=1}^N (W_i - \overline{W})^2\right)^2}.$$

The latter can be writte as

$$\hat{\mathbb{V}}_{\rm ehw} = \frac{(N_{\rm c}-1)S_{\rm c}^2}{N_{\rm c}^2} + \frac{(N_{\rm t}-1)S_{\rm t}^2}{N_{\rm t}^2},$$

very close to, but not identical to, the Neyman estimator $\hat{\mathbb{V}} = S_{\rm c}^2/N_{\rm c} + S_{\rm t}^2/N_{\rm t}$. The difference is a finite sample issue, the degrees of freedom adjustment in the estimator for $\sigma_{\rm c}^2$ and $\sigma_{\rm t}^2$.

1.4.6 Estimates of the Average Treatment Effects for the Riverside GAIN Experiment

For the full sample of the Riverside Gain data we present the point estimates the standard errors for the effect of the training on earnings in the first and seventh year after the training started.

Table 1.6: Point Estimates, Standard Errors, and Confidence Intervals for Average Treatment Effects In Thousands of Dollars: The Riverside Gain Data

	Year 1 Earnings	Year 7 Earnings
$\hat{\tau}$	1.14	0.58
s.e.	(0.13)	(0.29)
95% CI	$(0.88, 1.40)$	$(0.00, 1.15)$

Notes

Standard texts on randomized experiments in the biostatistics literature include Cochran and Cox [1950], Cox [1958], Kempthorne [1952], Wu and Hamada [2011] and Berry et al. [2010].

The Freedman quote given in the introduction notwithstanding there continue to be controversies regarding the merits of randomization. See Deaton [2010], Imbens [2010], Deaton and Cartwright [2018], Imbens [2018] for recent exchanges. My view is in line with Freedman's view that randomized experiments are unparalleled in the way they allow the researcher to draw credible causal inferences. The two main concerns are, first, that in many cases it is not feasible to carry out randomized experiments and we do rely in those cases on observational studies, and second, that because of informed consent there is more concern about external validity.

There are many studies analyzing the GAIN job training programs. Early studies from MDRC (the Manpower Demonstration Research Company) include Riccio et al. [1989, 1994], and another early academic study is Friedlander and Robins [1995]. A more recent study looking at long term impacts is Hotz et al. [2006].

For early references on randomization inference see Kempthorne [1955]. There is also a substantial amount of recent work on randomization-based p-values. See Rosenbaum [2002] for references, and Athey and Imbens [2017] for a recent discussion in econometrics.

Appendix

In the finite population case the central limit theorems are slightly non-standard. Here we discuss the formal results briefly, drawing on Abadie et al. [2017] and Li and Ding [2016].