Chapter 41
Causality and Design-Based Inference

Jake Bowers and Thomas Leavitt

Abstract

Counterfactual causal quantities cannot be observed, but researchers can use statistical procedures – namely, estimators and hypothesis tests – to draw inferences from data that can be observed. In this chapter, we present a unified account of estimation and testing for causal inference, showing how a study's research design can provide a foundation for both estimators and tests. We show how certain characteristics of research designs can justify claims that a given estimator or test has 'good' properties (e.g., unbiasedness, consistency, controlled error rates). We first develop ideas in the context of a randomized controlled experiment. In that context, we juxtapose estimations of and tests about causal effects and then provide an explicit comparison of Fisherian and Neymanian hypothesis tests. We then extend our analysis to research designs that are either partially controlled (e.g., experiments with noncompliance and/or attrition) or uncontrolled (e.g., observational studies). We show the ways in which knowledge and assumptions about research design – as well as assessments of how inferences would change should these assumptions be false – constitute a reliable basis for causal inference. We conclude by discussing the value of design-based causal inference in light of recent debates on its role in social scientific inquiry more broadly.

Design-Based Causal Inference

No one knows the true causal effect of an intervention. In an experiment, a researcher can assign some units to treatment and others to control; yet, one cannot see how treated units would have acted were they assigned to control nor how the control units would have acted were they assigned to treatment.[1] In the face of this fundamental ignorance, statisticians have developed two prominent approaches to inferring unobservable causal effects using data that can be observed. An analyst can either (1) generate a guess about (usually average) treatment effects or (2) posit a hypothesis about the effects of a treatment (such as the hypothesis that a treatment had no effects) and then assess the consistency of observable data with that null hypothesis, relative to a class of alternative hypotheses (such as the hypothesis that a treatment had a positive effect).

In what follows, we will define criteria by which a procedure qualifies as 'good' in the context of both estimation and testing and subsequently explain the role that research design plays in whether estimators and tests satisfy these criteria. We consider estimators and tests about causal effects first in the context of a randomized study design under full control of the researcher and second in cases in which the researcher does not fully control the study design. We show the ways in which either complete knowledge or assumptions (and the ways in which they could be violated) about the study design constitute what Fisher (1935) referred to as a 'reasoned basis for inference'.

Causality and Research Design

Defining Causal Effects

Consider a study in which there is a finite population of $1, \ldots, N$ units and the index $i \in \{1, \ldots, N\}$ runs over these units. Each individual, i, can be in either the treatment condition, $z_i = 1$, or the control condition, $z_i = 0$. Under the Stable Unit Treatment Value Assumption (SUTVA) (Cox, 1958; Rubin, 1980, 1986), each individual has a treated potential outcome, $y_{t,i}$ (unit i's outcome if given the intervention), and a control potential outcome, $y_{c,i}$ (unit i's outcome if not given the intervention).[2] An individual causal effect, τ_i, for each of the $i \in \{1, \ldots, N\}$ units is a function

[1] Holland (1986, 947) refers to the inability to observe both potential outcomes for a single unit at the same moment in time as the 'fundamental problem of causal inference.'

[2] For simplicity, we consider studies in which there are two conditions—treatment and control—although the same general principles apply to studies with multiple conditions (see Dasgupta et al., 2015).

of each unit's two potential outcomes, $\tau_i \equiv f\left(y_{c,i}, y_{t,i}\right)$, such as $\tau_i \equiv \frac{y_{t,i}}{y_{c,i}}$ or $\tau_i \equiv y_{t,i} - y_{c,i}$. For this chapter, we focus specifically on the additive, individual causal effect defined as $\tau_i \equiv y_{t,i} - y_{c,i}$. Researchers, however, can never observe both potential outcomes for each unit; instead, one can observe only y_i, which can be equal to either $y_{c,i}$ or $y_{t,i}$, depending on whether unit i is assigned to treatment ($z_i = 1$) or control ($z_i = 0$). We therefore represent observed outcomes by the function $y_i = z_i y_{t,i} + (1 - z_i)\, y_{c,i}$. For example, researchers may want to make an inference about the $1, \ldots, N$ individual causal effects, which we collect into the vector $\boldsymbol{\tau}' = \begin{bmatrix} \tau_1 & \tau_2 & \cdots & \tau_N \end{bmatrix}$, or about a function of these individual causal effects, such as the average causal effect, $\bar{\tau} = \left(\frac{1}{N}\right) \sum_{i=1}^{N} \tau_i$. We say that researchers want to 'make an inference' because neither $\boldsymbol{\tau}$ nor $\bar{\tau}$ can be directly observed. Researchers, of course, don't simply want to 'make an inference': they want to make inferences that can reliably track the true causal quantity of interest. This chapter shows how inferential procedures based on the research design can have such a reliable relationship with true causal quantities and explains what it means for a procedure to be 'based on research design'.

Defining a Research Design

Although a research design can certainly be more than this, for the purposes of this chapter a research design refers to the process by which units come to be in one study condition instead of another, i.e., each z_i comes to equal 1 or 0. More formally, we denote the collection of the values of z_i for all $i \in \{1, \ldots, N\}$ units by the vector $\mathbf{z}' = \begin{bmatrix} z_1 & \cdots & z_N \end{bmatrix}$ and define a research design as (1) a set of possible ways (events) in which the whole vector $\mathbf{z}$ could occur and (2) a probability distribution on this set of possible events. In a controlled study design (i.e., an experiment), we think of a researcher as 'assigning' conditions to all units in the study. When a researcher does not control how a unit i takes on a value of z_i, we think of that unit as 'selecting' into its own condition. As we lay the groundwork of concepts and notation, we write 'assignment' and assume control by the researcher, but we will apply the general framework later to uncontrolled research designs in which units 'select' into study conditions.

The set of possible ways in which $\mathbf{z}$ can occur depends on the process by which units are assigned to study conditions. If individuals can be in either the treatment or control condition irrespective of any other individual in the population, then we call this process individual assignment. As

a simple example, consider the 'coin flip' assignment process: in this case, the proportion of N individuals in either the treatment or control condition can vary across different assignments. We refer to this process as *simple* individual assignment. A researcher can implement simple individual assignment via an actual physical, stochastic process, such as N flips of a (potentially biased) coin – although in practice researchers will typically use random number generators (RNGs).

Under completely unconstrained simple individual assignment, the number of units in the treatment condition can range from 0 to N and the number of units in the control condition can likewise range from $N-0$ to $N-N$. More formally, we write the set, Ω, of possible ways that a researcher can assign all individuals to study conditions as follows:

$$\Omega = \{0,1\}^N = \left\{ \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ \vdots \\ 0 \\ 0 \end{bmatrix}, \cdots, \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ \vdots \\ 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \\ 1 \end{bmatrix} \right\}. \tag{1}$$

We can write the number of possible assignments in the set Ω by $|\Omega|$ (the 'cardinality of Omega'), under simple assignment as follows:

$$\begin{aligned} |\Omega| &= \binom{N}{0} + \binom{N}{1} + \cdots + \binom{N}{N-1} + \binom{N}{N} \\ &= \sum_{n_t=0}^{N} \binom{N}{n_t}, \end{aligned}$$

where $n_t = \sum_{i=1}^{N} z_i$ is the number of units in the treatment condition, which can range from 0 to N, and $\binom{N}{n_t} = \frac{N!}{n_c! n_t!}$ is the number of ways to choose n_t units from a total of N units. Conversely, $n_c = \sum_{i=1}^{N} (1 - z_i)$ is the number of units in the control condition, which can range from $N-0$ to $N-N$. In practice, researchers who control the assignment process will typically forbid assignments in which all units are in either condition (the first and last assignments in Equation (1)), in which case $|\Omega| = \sum_{n_t=1}^{N-1} \binom{N}{n_t}$.

The 'coin flipping' design helps us introduce the formal elements of a research design. In practice, individual coin flips can lead to lopsided designs in which many units are in one condition

or another. An alternative design that enables the researcher to control the numbers of units in each condition is *complete* assignment.

Complete individual assignment differs from simple, individual assignment only in that the value of n_t is fixed across all possible assignments. We have described simple individual assignment via the example of coin flips. Complete individual assignment can be thought of as draws from an urn. Imagine, for example, that an urn contains N balls, of which n_t are blue balls and $N - n_t = n_c$ are red balls. The researcher could draw the first ball from the urn and assign the first unit in the study to the treatment condition if the ball is blue and to the control condition if the ball is red. The second draw could follow the same rule for the assignment of the second unit, and so on and so forth until no more balls remain in the urn. This form of assignment ensures that exactly n_t units are in the treatment condition and $N - n_t = n_c$ units are in the control condition. More formally, complete individual assignment excludes any assignment, $\mathbf{z}$, with more or less treatment units, n_t, than that which is predetermined by the researcher. Therefore, under complete assignment, the number of possible assignments is simply $|\Omega| = \binom{N}{n_t}$.

Simple and complete assignment can also happen at the *cluster* (as opposed to the *individual*) level. In this setup, we not only have a set of $1, \ldots, N$ individuals, but also a set of $1, \ldots, K$ clusters, where each cluster, $k \in \{1, \ldots, K\}$, contains $N_k \geq 1$ individual units and $N = \sum_{k=1}^{K} N_k$. In cluster assignment designs, all of the $i \in \{1, \ldots, N_k\}$ units in the kth cluster are assigned to either the treatment condition, $z_{i,k} = 1$, or the control condition, $z_{i,k} = 0$. In simple cluster assignment, the number of possible assignments is given by $|\Omega| = \sum_{k_t=0}^{K} \binom{K}{k_t}$, where k_t denotes the number of treatment clusters, although (just as in simple individual assignment) researchers will typically ensure that $k_t \notin \{0, K\}$. Under complete cluster assignment, the number of treatment clusters is fixed, such that the number of assignments is $|\Omega| = \binom{K}{k_t}$.

Lastly, *blocked assignment* is when individuals or clusters are assigned (either simply or completely) to study conditions within blocks, which we index from $b \in \{1, \ldots, B\}$. Blocks are typically constructed on the basis of individuals' or clusters' values of baseline covariates. Baseline covariates are measured prior to assignment and hence their values are fixed regardless of the condition to which a unit or cluster is assigned. Under simple individual block assignment, the number of

possible assignments is $|\Omega| = \prod_{b=1}^{B} \left(\sum_{n_{t,b}=1}^{N_b-1} \binom{N_b}{n_{t,b}} \right)$, where N_b is the number of units in block b, $n_{t,b}$ is the number of units in the treatment condition in block b and $n_{t,b} \notin \{0, N_b\}$ for all b. Under complete individual block assignment $|\Omega| = \prod_{b=1}^{B} \binom{N_b}{n_{t,b}}$, one can analogously deduce the number of possible assignments under either simple or complete cluster block assignment. As we will explain in subsequent sections, block assignment carries important implications for properties of both estimators and hypothesis tests.

Given a set of possible assignments, Ω, arising from an assignment mechanism, the remaining component of a research design is a probability distribution on this set of assignments. In a *uniform randomized experiment*, the probability of each assignment is simply $\frac{1}{|\Omega|}$ for all $\mathbf{z}$, whereby each assignment has an identical probability of realization. Yet the probability distribution on Ω need not be uniform, even in a randomized experiment. Design-based inference means only that the stochastic properties of estimators and tests be based on this probability distribution on Ω, regardless of whether that distribution is uniform or not. As we now move to discussions of both estimation and testing, notice throughout that whenever we refer to random quantities, the randomness of those quantities stems solely from the probability distribution on Ω.

An Illustrative Example

In the sections to follow, we demonstrate our arguments via a simple hypothetical example that consists of $N = 6$ units and an individual assignment process (complete individual assignment) in which three units are assigned to treatment ($n_t = 3$) and to control ($n_c = 3$). Let's further imagine that (unbeknownst to the researcher) the six units' potential outcomes and individual causal effects are as follows in Table 41.1:

$\mathbf{y_c}$	$\mathbf{y_t}$	$\boldsymbol{\tau}$
20	22	2
8	12	4
11	11	0
10	15	5
14	18	4
1	4	3

Table 41.1: True values of $\mathbf{y_c}$, $\mathbf{y_t}$ and $\boldsymbol{\tau}$, where $\tau_i = y_{t,i} - y_{c,i}$, for the study population

Based on the complete individual assignment process in which there are $N = 6$ units and of

which $n_t = 3$ are assigned to treatment, the set of $|\Omega| = \binom{6}{3} = 20$ possible assignments is given by Equation (2):

$$\Omega = \left\{ \begin{bmatrix} 1 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \cdots, \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 1 \end{bmatrix} \right\}. \tag{2}$$

The assignment that one draws from the set, Ω, determines which potential outcomes one observes. One can observe treatment potential outcomes only for units assigned to treatment, and control potential outcomes only for units assigned to control (recall the function $y_i = z_i y_{t,i} + (1 - z_i)\, y_{c,i}$, which determines the potential outcome that one observes for each individual i). As Table 41.2 shows, for each of the $\binom{6}{3} = 20$ possible assignments, there are $\binom{6}{3} = 20$ corresponding possible realizations of observed data, where '?' throughout this chapter denotes an unobserved and hence unknown potential outcome.

$\mathbf{z}_1$	$\mathbf{y_c}$	$\mathbf{y_t}$	$\mathbf{y}_1$
1	?	22	22
1	?	12	12
1	?	11	11
0	10	?	10
0	14	?	14
0	1	?	1

$\mathbf{z}_2$	$\mathbf{y_c}$	$\mathbf{y_t}$	$\mathbf{y}_2$
1	?	22	22
1	?	12	12
0	11	?	11
1	?	15	15
0	14	?	14
0	1	?	1

$\cdots$

$\mathbf{z}_{19}$	$\mathbf{y_c}$	$\mathbf{y_t}$	$\mathbf{y}_{19}$
0	20	?	20
0	8	?	8
1	?	11	11
0	10	?	10
1	?	18	18
1	?	4	4

$\mathbf{z}_{20}$	$\mathbf{y_c}$	$\mathbf{y_t}$	$\mathbf{y}_{20}$
0	20	?	20
0	8	?	8
0	11	?	11
1	?	15	15
1	?	18	18
1	?	4	4

Table 41.2: All possible realizations of experimental data from a completely randomized study with 6 units and 3 treated units.

Only one such possible realization of data in Table 41.2 can be observed; but, knowing that there are 20 possible realizations allows the researcher to use procedures – e.g., estimators or hypothesis tests – to make inferences about unobservable causal quantities based on the single observed realization. We want procedures for drawing causal inferences to have properties that are 'good' (a notion that we will define more precisely in later sections). These properties describe or measure a procedure's performance in two contexts: (1) studies with a fixed, finite population and (2) a hypothetical scenario in which the size of a given study increases towards ∞ while all other relevant factors remain constant. We refer to the latter context as one of *asymptotic growth*, which we conceptualize below.

If we have an experimental pool of six units, that is not a sample using a known procedure from a well-defined population, what does 'asymptotic growth' mean? We follow Brewer (1979) and Middleton and Aronow (2015) in using the idea of 'copies' as a way to talk about how estimators and tests behave as study sizes increase. In short, this conception of asymptotic growth states that (1) the original population of N units is copied $h-1$ times, (2) within each of the h copies, exactly n_t units are assigned to the treatment condition and the remaining $n_c = N - n_t$ units are assigned to the control condition and (3) the h copies are then collected into a single population with hN total units, hn_t treated units and hn_c control units.

In the context of our working example, this conception of growth stipulates that the study population of $N = 6$ units is embedded in a sequence of populations of increasing sizes in which the initial population is simply copied $h-1$ times.

$\mathbf{y_c}$	$\mathbf{y_t}$	$\boldsymbol{\tau}$
20	22	2
8	12	4
11	11	0
10	15	5
14	18	4
1	4	3

$\mathbf{y_c}$	$\mathbf{y_t}$	$\boldsymbol{\tau}$
20	22	2
8	12	4
11	11	0
10	15	5
14	18	4
1	4	3
20	22	2
8	12	4
11	11	0
10	15	5
14	18	4
1	4	3

$\mathbf{y_c}$	$\mathbf{y_t}$	$\boldsymbol{\tau}$
20	22	2
8	12	4
11	11	0
10	15	5
14	18	4
1	4	3
20	22	2
8	12	4
11	11	0
10	15	5
14	18	4
1	4	3
20	22	2
8	12	4
11	11	0
10	15	5
14	18	4
1	4	3

$\mathbf{y_c}$	$\mathbf{y_t}$	$\boldsymbol{\tau}$
20	22	2
8	12	4
11	11	0
10	15	5
14	18	4
1	4	3
20	22	2
8	12	4
11	11	0
10	15	5
14	18	4
1	4	3
20	22	2
8	12	4
11	11	0
10	15	5
14	18	4
1	4	3
20	22	2
8	12	4
11	11	0
10	15	5
14	18	4
1	4	3

$\ldots$

Table 41.3: Finite populations under asymptotic growth in which $h \in \{1, 2, 3, 4, \ldots\}$

Notice that over this sequence of increasing finite populations shown in Table 41.3, all relevant factors other than N remain constant: the proportions of treatment and control units remain fixed and the mean of control and treatment potential outcomes remain fixed, as do their variances and their covariance. Notice, however, that the number of possible assignments increases over this sequence of increasing finite populations from $\binom{6}{3} = 20$ to $\binom{12}{6} = 924$ and from $\binom{18}{9} = 48620$ to

$\binom{24}{12} = 2704156$ and so forth.

We will show that, in either the finite context – given in Table 41.1 – or in the asymptotic context – given in Table 41.3 – whether a procedure is 'good' depends on whether it maintains fidelity to the research design – i.e., the probability distribution on the set Ω. In other words, we will show that in a randomized experiment, a 'good' procedure is one that heeds the dictum of Senn (2004, 3729) who, in the voice of R. A. Fisher, states that '[a]s ye randomise so shall ye analyse'.

Estimation

As we have mentioned above, no one can observe both potential outcomes for any given unit in a given study population. One can, however, generate a guess about some function of the study population's individual causal effects (e.g., the mean causal effect) using observed outcomes. We call this unobservable causal quantity the *estimand*. The *estimator*, by contrast, refers to the procedure that generates a guess about the estimand. An *estimate* is the actual output of the estimator once it is applied to a given data set.

One estimand is the mean causal effect, $\bar{\tau} = \left(\frac{1}{N}\right) \sum_{i=1}^{N} \tau_i$, which, to return to the example from Table 41.1, is $\bar{\tau} = \frac{2+4+0+5+4+3}{6} = 3$. A procedure for generating a guess about $\bar{\tau}$ is the Difference-in-Means estimator, which we can define in terms of observable quantities as follows:

$$\begin{aligned} \hat{\bar{\tau}}\left(\mathbf{Z}, \mathbf{Y}\right) &= \frac{\mathbf{Z}'\mathbf{Y}}{\mathbf{Z}'\mathbf{Z}} - \frac{(\mathbf{1}-\mathbf{Z})'\mathbf{Y}}{(\mathbf{1}-\mathbf{Z})'(\mathbf{1}-\mathbf{Z})} \\ &= \left(\frac{1}{\sum_{i=1}^{N} Z_i}\right) \sum_{i=1}^{N} Z_i Y_i - \left(\frac{1}{\sum_{i=1}^{N} (1-Z_i)}\right) \sum_{i=1}^{N} (1-Z_i)\, Y_i. \end{aligned} \tag{3}$$

In the example from Table 41.1, the random vectors[3] of $\mathbf{Z}$ and $\mathbf{Y}$ can take on any of the possible values, $(\mathbf{z}_1, \mathbf{y}_1), \cdots, (\mathbf{z}_{20}, \mathbf{y}_{20})$, given in Table 41.2. If we apply the estimator in Equation (3) to the possible realizations of data in Table 41.2, then there are 20 possible estimates that correspond

[3]To distinguish between fixed quantities and quantities that can take on different values with certain probabilities (i.e., random quantities), we now use uppercase letters for random quantities and lowercase letters for fixed or realized quantities.

to each of the 20 possible realizations of data:

$$\hat{\bar{\tau}}\left(\mathbf{z}_1, \mathbf{y}_1\right) = 6.6667,\ \hat{\bar{\tau}}\left(\mathbf{z}_2, \mathbf{y}_2\right) = 7.6667,\ \cdots,\ \hat{\bar{\tau}}\left(\mathbf{z}_{19}, \mathbf{y}_{19}\right) = -1.6667,\ \hat{\bar{\tau}}\left(\mathbf{z}_{20}, \mathbf{y}_{20}\right) = -0.6667.$$

The researcher can observe only one of these 20 possible estimates and this single estimate should be generated by an estimator that is 'good'. More specifically, three 'good' properties of an estimator are *unbiasedness*, *consistency* and *precision*. An unbiased estimator is one in which, although any single estimate may be close to or far from the true value of the estimand, the *expected value of the estimator* – i.e., the probability-weighted mean of all possible estimates – is equal to the value of the estimand. Consistency states that as the number of units in the study increases asymptotically, holding all other factors constant, the probability distribution of an estimator concentrates increasingly around the truth. (For any fixed $\varepsilon > 0$, the probability that the estimate and its target differ by no more than ε tends to 1.) Lastly, a precise estimator is one in which the expected distance of an estimate from the true causal effect is small.

In the following discussion, we show the role that research design plays in whether the Difference-in-Means estimator is unbiased, consistent and/or precise with respect to the estimand $\bar{\tau} = \left(\frac{1}{N}\right)\sum_{i=1}^{N}\tau_i$. We also show how designs can yield estimators that are more or less precise. Researchers may want to estimate quantities other than $\bar{\tau}$. Although we do not discuss such cases, the general principles for determining whether an estimator is unbiased, consistent and/or precise with respect to the causal quantity of interest is the same: researchers can define an estimand that they seek to infer, define an estimator by which they would estimate this quantity under all possible realizations of data and subsequently assess whether this estimator is unbiased, consistent and/or precise based only on the probabilities with which possible data are realized.

Unbiasedness

We now show that a 'good' estimator of the unknown estimand, $\bar{\tau}$, is the estimator given in Equation (3), $\hat{\bar{\tau}}\left(\mathbf{Z}, \mathbf{Y}\right)$. In particular, we will show that this estimator satisfies the criterion of

[4]When the numbers of treatment and control units are *not* fixed, such as in simple, individual assignment, the Difference-in-Means estimator remains unbiased in a uniform randomized experiment so long as at least one unit is always in the treatment and control conditions, respectively. In general, when the numbers of treatment and control are *not* fixed, the Difference-in-Means estimator is not necessarily unbiased, such as in cluster uniform random assignment when clusters are of unequal sizes (see Middleton and Aronow, 2015).

no systematic error – i.e., unbiasedness – in a *uniform randomized experiment*, when the numbers of treatment and control units are both fixed.[4] Whether the Difference-in-Means estimator is unbiased with respect to $\bar{\tau}$ depends solely on the known research design, i.e., whether or not there is a uniform probability distribution on the set of assignments.

Returning to the example from Table 41.1, recall that if we apply the estimator in Equation (3) to the possible realizations of data in Table 41.2, then there are 20 possible estimates that correspond to each of the 20 possible realizations of data:

$$\hat{\bar{\tau}}\left(\mathbf{z}_1, \mathbf{y}_1\right) = 6.6667,\ \hat{\bar{\tau}}\left(\mathbf{z}_2, \mathbf{y}_2\right) = 7.6667,\ \ldots,\ \hat{\bar{\tau}}\left(\mathbf{z}_{19}, \mathbf{y}_{19}\right) = -1.6667,\ \hat{\bar{\tau}}\left(\mathbf{z}_{20}, \mathbf{y}_{20}\right) = -0.6667.$$

Informally, an unbiased estimator produces a guess about the estimand with no systematic error. Slightly more formally, an estimator is unbiased if the average of all possible estimates is equal to the true value of the estimand. This average of estimates, however, must be weighted by the probabilities of observing each possible estimate; we call this average the 'expected value' and denote the expected value of the Difference-in-Means estimator by $\mathbb{E}\left[\hat{\bar{\tau}}\left(\mathbf{Z}, \mathbf{Y}\right)\right]$.

To calculate the expected value of the Difference-in-Means estimator to assess properties like bias and consistency, we need to know the probability associated with each of these 20 possible estimates. We know that the estimator is a function of two random quantities, $\mathbf{Z}$ and $\mathbf{Y}$, but $\mathbf{Y}$ inherits randomness only from $\mathbf{Z}$, since $Y_i = Z_i y_{t,i} + (1 - Z_i)\, y_{c,i}$ for all $i \in \{1, \ldots, N\}$ units. Therefore, each probability associated with its corresponding estimate depends only on $\mathbf{Z}$. So, we calculate the expected value of the estimator in general as follows:

$$\mathbb{E}\left[\hat{\bar{\tau}}\left(\mathbf{Z}, \mathbf{Y}\right)\right] = \hat{\bar{\tau}}\left(\mathbf{z}_1, \mathbf{y}_1\right) \Pr\left(\mathbf{Z} = \mathbf{z}_1\right) + \cdots + \hat{\bar{\tau}}\left(\mathbf{z}_{|\Omega|}, \mathbf{y}_{|\Omega|}\right) \Pr\left(\mathbf{Z} = \mathbf{z}_{|\Omega|}\right).$$

In the context of the running example, there are 20 possible estimates corresponding to each of the $\mathbf{z}_1, \ldots, \mathbf{z}_{20}$ possible assignments, and the probability of each of those possible assignments is

$\frac{1}{20}$. Therefore, the expected value of the Difference-in-Means estimator is

$$\begin{aligned}
\mathbb{E}\left[\hat{\bar{\tau}}\left(\mathbf{Z},\mathbf{Y}\right)\right] &= \hat{\bar{\tau}}\left(\mathbf{z}_1,\mathbf{y}_1\right)\Pr\left(\mathbf{Z}=\mathbf{z}_1\right)+\cdots+\hat{\bar{\tau}}\left(\mathbf{z}_{20},\mathbf{y}_{20}\right)\Pr\left(\mathbf{Z}=\mathbf{z}_{20}\right) \\
&= 6.6667\left(\frac{1}{20}\right)+\cdots+-0.6667\left(\frac{1}{20}\right) \\
&= 3.
\end{aligned}$$

In this example, the expected value of the estimator, $\mathbb{E}\left[\hat{\bar{\tau}}\left(\mathbf{Z},\mathbf{Y}\right)\right]$, is exactly equal to the true mean causal effect, $\bar{\tau}$. The estimator is unbiased given the design. If $\Pr\left(\mathbf{Z}=\mathbf{z}\right)$ did not equal $\frac{1}{20}$ for all $\mathbf{z}$ – i.e., if some assignments were more or less probable than others – then the Difference-in-Means estimator might not be unbiased. In general, the equality between $\mathbb{E}\left[\hat{\bar{\tau}}\left(\mathbf{Z},\mathbf{Y}\right)\right]$ and $\bar{\tau}$ holds when (1) units are assigned to study conditions as individuals (not as groups, i.e., *clusters*), (2) there is always at least one unit in the treatment condition and one in the control condition and (3) each possible assignment has an identical probability of realization. In other words, the Difference-in-Means estimator is unbiased with respect to the mean additive causal effect in a uniform randomized experiment under either complete individual assignment or simple individual assignment so long as there is always at least one unit in each of the study conditions.

Notice that the Difference-in-Means estimator did not require large numbers of units or assumptions about the distributions of potential outcomes to be unbiased. The potential outcomes could have been any set of values and the property of unbiasedness would still have held. However, the unbiasedness property did require that there be a uniform probability distribution on the set of possible assignments. But in a controlled research design, like a randomized experiment, the researcher knows whether or not this condition is true.

Consistency

'No systematic error' is not the same as 'close to the truth'. Achen (1982, 36) explains the need for another conception of a 'good' estimator when he writes '[u]nbiasedness is too weak a property, since it says nothing about approximating the truth'. While we know that an unbiased estimator yields estimates that are, on average, equal to the true value of the estimand, any single estimate might be far from the truth. In our running example, not one of the 20 possible estimates is actually

equal to the true mean causal effect of $\bar{\tau} = 3$, even though the probability-weighted average of those 20 estimates is equal to 3. Another 'good' characteristic of an estimator is to produce values close to the truth, especially as the size of the experiment grows larger while other factors remain constant, and more information is supplied to the estimator from the design.

In contrast to unbiasedness, consistency states that as the number of units in the study grows asymptotically, holding all other factors constant, the probability of an estimate within an arbitrarily small distance, ε, from the truth is equal to 1. More formally, we can define consistency as follows:

$$\lim_{h \to \infty} \Pr\left(\left| \hat{\bar{\tau}} \left(\mathbf{Z}, \mathbf{Y} \right) - \bar{\tau} \right| < \varepsilon \right) = 1 \text{ for all } \varepsilon > 0 \tag{4}$$

or equivalently as

$$\lim_{h \to \infty} \Pr\left(\hat{\bar{\tau}} \left(\mathbf{Z}, \mathbf{Y} \right) \in \left(\bar{\tau} - \varepsilon, \bar{\tau} + \varepsilon \right) \right) = 1 \text{ for all } \varepsilon > 0, \tag{5}$$

where, referring back to the conception of asymptotic growth in Table 41.3, h is the number of copies of the original finite population from Table 41.1.

To unpack Equations (4) and (5), Figure 41.1 shows what happens to the distribution of the Difference-in-Means estimator under complete random assignment as $h \to \infty$.

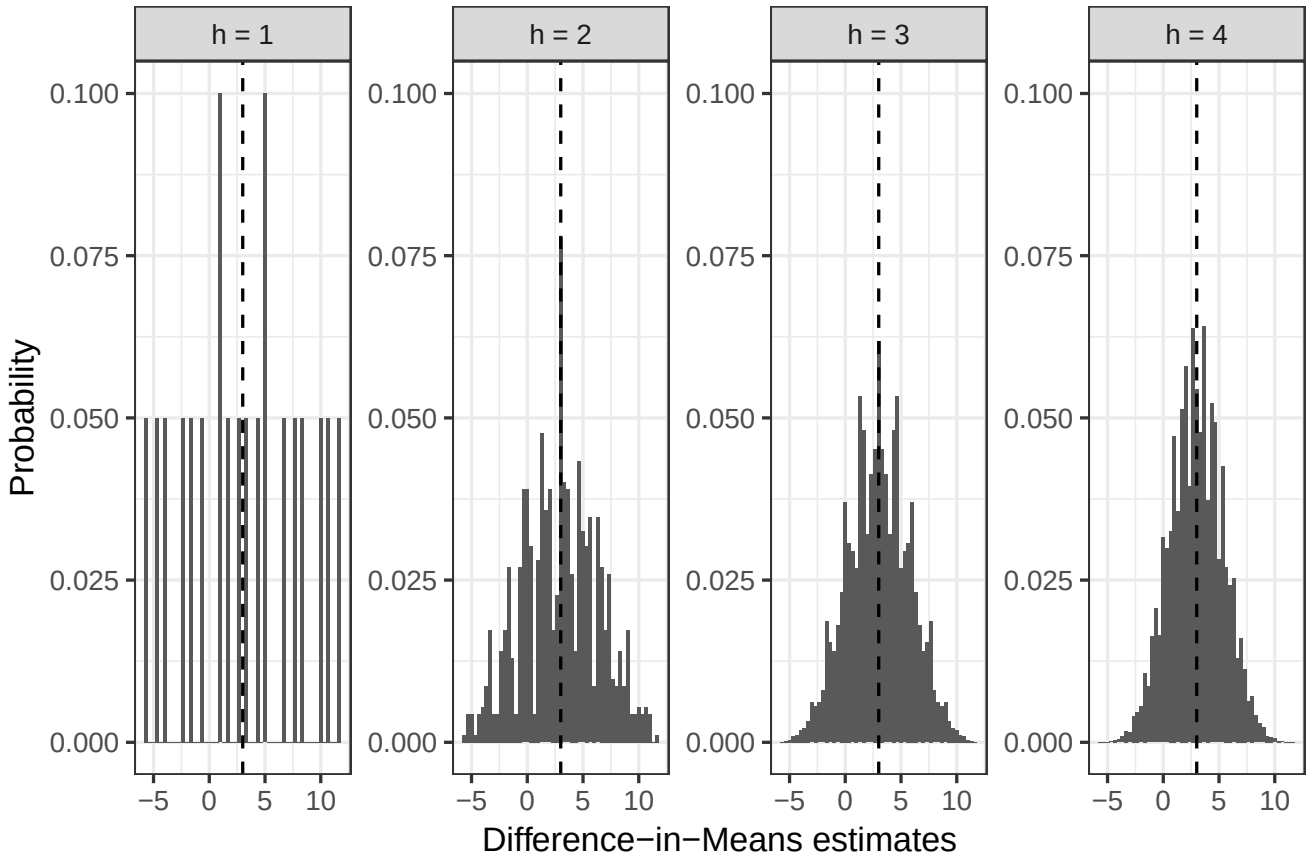

Figure 41.1: Distribution of Difference-in-Means estimator as $h \to \infty$

The general trend is that the probability of estimates close to the true mean causal effect, $\bar{\tau} = 3$, grows larger and larger and ultimately converges in probability (over the sequence of increasing finite populations) to 1. For example, following Equation (5), consider the probability that an estimate lies on the interval $(3 - \varepsilon, 3 + \varepsilon)$ and let $\varepsilon = 1$. The respective probabilities of an estimate on this interval for $h \in \{1, 2, 3, 4\}$ are 0.1, 0.2359, 0.2559 and 0.2994, and as $h \to \infty$, that probability tends to 1. This property holds for $\varepsilon = 1$ as well as for any positive value of ε that one could choose. For example, we could have let $\varepsilon = 0.5$, in which case the corresponding probabilities for $h \in \{1, 2, 3, 4\}$ are 0.1, 0.1234, 0.1522 and 0.1643, and the limiting probability as $h \to \infty$ is also 1. In general, it is not necessary that each probability be greater than its predecessor for $h \in \{1, 2, 3, 4, \ldots\}$. Consistency states only that there exists some number in which, for any h greater than that number, the estimator will lie within an interval of ε from the true mean causal effect.

In this particular case, consistency follows (in part) from unbiasedness. As the size of N in-

creases towards ∞ while all other factors remain constant, the probability of an estimate arbitrarily close to the estimator's expected value is equal to 1. Unbiasedness ensures that the expected value of the estimator is equal to the true value of the estimand. Therefore, as the size of the study population grows towards ∞, the estimator produces a value arbitrarily close to the truth (not just to the estimator's expected value) with a probability of 1. Both the unbiasedness and consistency of the Difference-in-Means estimator, moreover, arise solely from the research design. Even though the distribution of the estimator starts to look more and more normal as the size of the study population increases to ∞, we made no such distributional assumptions to show the estimator's unbiasedness and consistency.

Precision

While unbiased and consistent estimators are desirable, such estimators may yield estimates far from the truth, with high probability in actual experiments with fixed study populations. One estimator is more precise that another estimator for a given study design if it produces estimates that are closer to the truth on average. In other words, a 'good' estimator also has a low variance; in our case, a more precise estimator than the Difference-in-Means estimator would make the 20 possible estimates in Table 41.2 closer to the true mean causal effect, on average. We now consider first the factors that make the Difference-in-Means estimator produce guesses with lower expected distance from the true mean causal effect and second the procedure one can use to conservatively estimate the variance of the Difference-in-Means estimator.

Neyman (1923) derived an exact analytic expression for the variance of the Difference-in-Means estimator based solely on the research design of a randomized experiment, as follows:

$$\sigma^2_{\hat{\tau}} = \frac{1}{N-1}\left(\frac{n_t\sigma^2_{y_c}}{n_c} + \frac{n_c\sigma^2_{y_t}}{n_t} + 2\sigma_{y_c,y_t}\right), \tag{6}$$

where $\sigma^2_{y_c}$ is the variance of control potential outcomes, $\sigma^2_{y_t}$ is the variance of treated potential outcomes and σ_{y_c,y_t} is the covariance of control and treated potential outcomes.

Equation (6) suggests that one can increase the precision of the Difference-in-Means estimator (i.e., reduce the estimator's variance) by increasing the number of treatment units and/or the

number of control units. Precision can also be increased by decreasing the variances of treatment, $\sigma^2_{y_t}$, and control, $\sigma^2_{y_c}$, potential outcomes. For a simple and clear account of the factors that increase precision, see Gerber and Green (2012, section 3.2).

A standard design choice that researchers can make to increase precision is blocking. That is, a researcher can first construct blocks that are similar in terms of covariates related to potential outcomes and second assign units to study conditions within blocks. Blocked assignment works by excluding assignments that yield estimates far from the true mean effect, on average. To see this point, we return to the example in Table 41.1 and introduce $\mathbf{x}$, which is a vector of baseline covariates, x_i, for all $i \in \{1, \ldots, N\}$ units. The vector $\mathbf{x}$ is a fixed quantity that is measured for all units prior to assignment; hence, $\mathbf{x}$ cannot change as a function of whichever assignment is realized.

$\mathbf{y_c}$	$\mathbf{y_t}$	$\boldsymbol{\tau}$	$\mathbf{x}$
20	22	2	1
8	12	4	1
11	11	0	0
10	15	5	1
14	18	4	1
1	4	3	0

Table 41.4: True values of $\mathbf{y_c}$, $\mathbf{y_t}$, $\boldsymbol{\tau}$ and the baseline covariate $\mathbf{x}$

From Table 41.4, we can see that $\mathbf{x}$ is related to both $\mathbf{y_c}$ and $\mathbf{y_t}$. The treatment potential outcomes are greater, on average, among units whose baseline covariate values are equal to 1 compared to units whose baseline covariate values are equal to 0. The same is true for control potential outcomes. If the researcher puts units whose covariate values are equal to 1 in one block and units whose covariate values are equal to 0 in another, and then assigns half of the units to treatment and control within blocks, the set of possible assignments, Ω_b, would be as follows:

$$\Omega_b = \left\{ \begin{bmatrix} 1 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}, \ldots, \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 1 \end{bmatrix} \right\}. \tag{7}$$

The set Ω_b above has only 12 possible assignments as opposed to the 20 possible assignments

in Equation (2) under complete random assignment without blocks. In particular, the assignments of $\mathbf{z}_2, \mathbf{z}_3, \mathbf{z}_7, \mathbf{z}_8, \mathbf{z}_{13}, \mathbf{z}_{14}, \mathbf{z}_{18}, \mathbf{z}_{19} \in \Omega$ are excluded from Ω_b.

Figure 41.2 shows the eight estimates corresponding to the eight assignments that were included in unblocked assignment but excluded in blocked assignment. On average, these eight estimates are farther from the true mean effect than are the other 12 estimates. More concretely, the average squared distance of the eight excluded estimates from the truth is 36.91667 and the same average squared distance of the 12 included assignments is 18.12963. Hence, this blocked design increases precision by reducing the probability (to 0) of estimates that are, on average, far from the truth and by increasing the probability of estimates that are, on average, closer to the truth.[5]

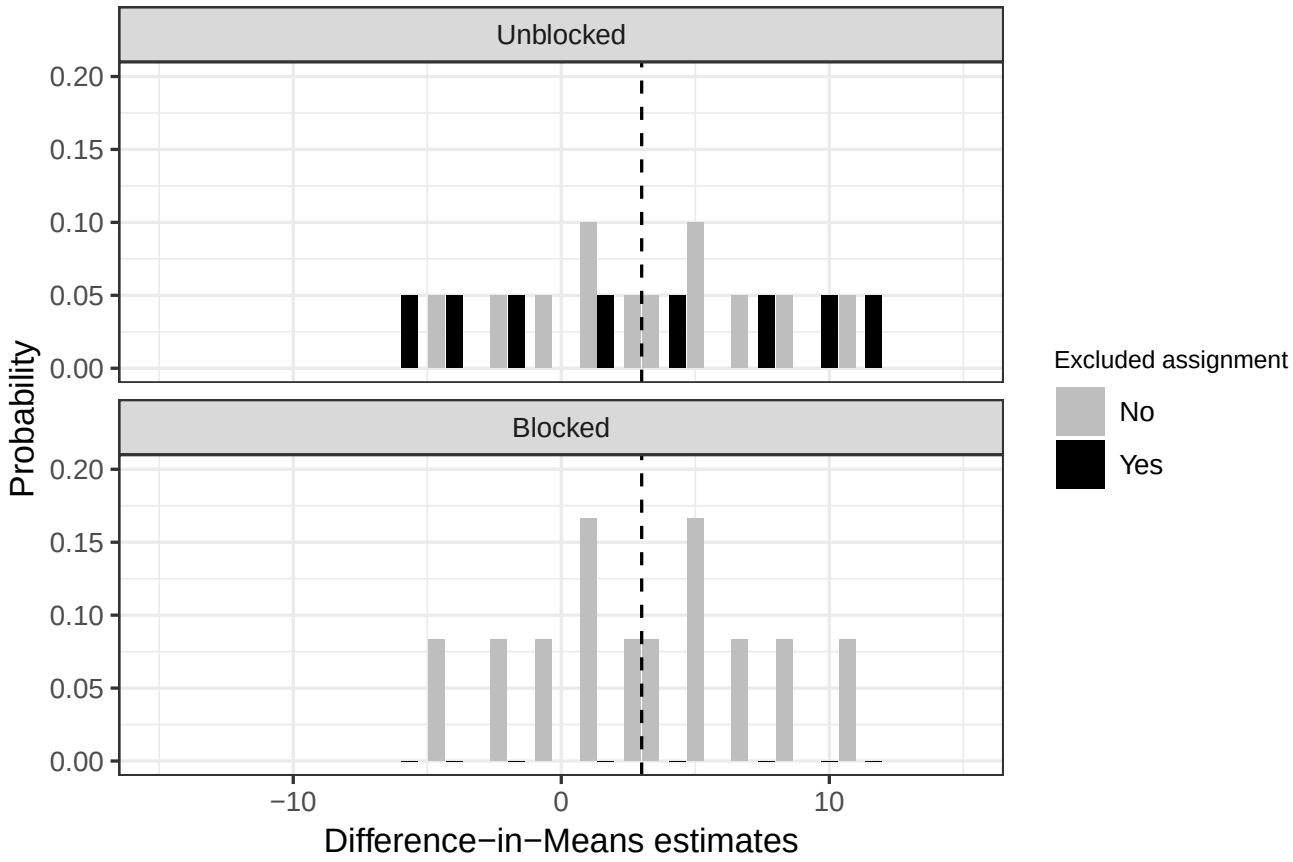

Figure 41.2: Distribution of Difference-in-Means estimates under (1) unblocked and (b) blocked assignment

Thus far, we have focused on design choices that can decrease the variance of the Difference-in-

[5]Note that, if treatment assignment probabilities differ *across* blocks (but are uniform *within* blocks), then the standard Difference-in-Means estimator may be biased. In such cases, an unbiased estimator would be the Difference-in-Means estimator that generates an estimate within each block and subsequently weights each block-specific estimate by the proportion of units in that block.

Means estimator. But, note that the variance of the Difference-in-Means estimator is, like the true mean causal effect, a fixed, unobservable quantity. As we can see from Equation (6), the variance of the Difference-in-Means estimator depends on the variance of treatment and control potential outcomes as well as their covariance, none of which can be directly observed. We need a 'good' procedure by which we can estimate the variance of the Difference-in-Means estimator if we are to reasonably infer its precision. We will use such a variance estimator in the context not only of evaluating estimators but also hypothesis testing.

One can unbiasedly estimate two of the three unknown quantities in Equation (6). Following Cochran (1977), unbiased estimators of $\sigma^2_{y_c}$ and $\sigma^2_{y_t}$, respectively, are: $\hat{\sigma}^2_{y_c} = \left(\frac{N-1}{N(n_c-1)}\right) \sum_{i:Z_i=0}^{N} \left(y_{c,i} - \hat{\mu}_{y_c}\right)^2$ and $\hat{\sigma}^2_{y_t} = \left(\frac{N-1}{N(n_t-1)}\right) \sum_{i:Z_i=1}^{N} \left(y_{t,i} - \hat{\mu}_{y_t}\right)^2$, where $\hat{\mu}_{y_c} = \left(\frac{1}{n_c}\right) \sum_{i=1}^{N} (1 - Z_i)\, y_{c,i}$ and $\hat{\mu}_{y_t} = \left(\frac{1}{n_t}\right) Z_i y_{t,i}$. We cannot write an unbiased estimator for σ_{y_c,y_t} since no two potential outcomes for any unit can be jointly observed. Neyman (1923) noted, however, that one could use a conservative procedure for estimating the quantity in Equation (6) by assuming the largest possible value of $2\sigma_{y_c,y_t}$, which, by the Cauchy–Schwarz inequality and the AM–GM inequality (i.e., inequality of arithmetic and geometric means), is $\sigma^2_{y_c} + \sigma^2_{y_t}$.

After substituting $\sigma^2_{y_c} + \sigma^2_{y_t}$ for $2\sigma_{y_c,y_t}$, the analytic expression for the variance of the Difference-in-Means estimator (assuming $2\sigma_{y_c,y_t} = \sigma^2_{y_c} + \sigma^2_{y_t}$) is

$$\frac{1}{N-1}\left(\frac{n_t\sigma^2_{y_c}}{n_c} + \frac{n_c\sigma^2_{y_t}}{n_t} + \sigma^2_{y_c} + \sigma^2_{y_t}\right),$$

which can be simplified to

$$\frac{N}{N-1}\left(\frac{\sigma^2_{y_c}}{n_c} + \frac{\sigma^2_{y_t}}{n_t}\right). \tag{8}$$

Now there are only two unknown quantities in Equation (8), each of which can be unbiasedly estimated. Hence, one can now unbiasedly estimate the quantity in (8) via the conservative

variance estimator of

$$\hat{\sigma}^2_{\hat{\bar{\tau}}} = \frac{N}{N-1}\left(\frac{\hat{\sigma}^2_{y_c}}{n_c} + \frac{\hat{\sigma}^2_{y_t}}{n_t}\right). \tag{9}$$

This estimator is conservative because, since it unbiasedly estimates the quantity in (8), its expected value is equal to or greater than the true variance of the estimator given in (6).[6]

Thus far, we have explained the role that research design – i.e., the probability distribution on the set of assignments, Ω – plays in determining whether estimators are unbiased, consistent and precise. We have also explained how one can infer the variance of an estimator via a conservative procedure. We have used a simple example of complete uniform assignment to illustrate these points; yet an estimator that is unbiased, consistent and/or relatively precise in this design may not be so in another design. For example, the Difference-in-Means estimator is not necessarily unbiased when there is a non-uniform probability distribution on Ω; however, the Horvitz–Thompson (i.e., inverse probability weighted) estimator (Horvitz and Thompson, 1952) is unbiased in such a design (see Aronow and Middleton, 2013). Such design-based inference differs from model-based inferences in that the former remains reliable without the need to impose a probability model on potential outcomes or to model the functional form (e.g., a linear model) that links the treatment variable to potential outcomes. The only probability model in design-based inference is the assignment process itself, which, in the case of a randomized experiment, is known to be the true model of the data-generating process.

Hypothesis Testing

Focusing on $\bar{\tau}$ engages with the fundamental problem of causal inference by aggregation across units in the study and $\hat{\bar{\tau}}$ can be shown to be an unbiased, consistent and, depending on the study design, relatively precise estimator. An alternative approach begins with claims about causal effects rather than guesses about them, stating hypotheses about $\bar{\tau}$ or even about individual effects, τ_i. This approach then tests those claims, and so the procedures that we will assess in this section

[6] Recent work has derived a consistent estimator for an upper bound on the term $2\sigma_{y_c,y_t}$ that is always less than or equal to $\sigma^2_{y_c} + \sigma^2_{y_t}$ (Aronow et al., 2014). Such an estimator enables researchers to more precisely estimate the the variance of the Difference-in-Means estimator and, as we will discuss later, increase the power of hypothesis tests about the mean causal effect.

are properties of tests rather than of estimators. A researcher who performs a hypothesis test for causal inference first states a null hypothesis about a relationship between unobserved potential outcomes and a (usually composite) alternative hypothesis, which we will define more precisely below. The researcher can then assess the probability that the research design generates data more extreme than the observed data under the null hypothesis, where the data are summarized via a test statistic that maps the data – the observed outcome, perhaps adjusted to reflect the implications of the hypothesis, and treatment assignment – to a single number. For example, $t(\mathbf{Z}, \mathbf{Y})$ could be $\left(\frac{1}{\sum_{i=1}^{N} Z_i}\right) \sum_{i=1}^{N} Z_i Y_i - \left(\frac{1}{\sum_{i=1}^{N} (1-Z_i)}\right) \sum_{i=1}^{N} (1 - Z_i) Y_i$, which is the same formula that we labeled $\hat{\bar{\tau}}$ in Equation (3) but, in the context of hypothesis testing, is not an estimator but a data summary. We call the probability of a test statistic more extreme than the observed test statistic a probability-value or p-value. Typically, when the p-value is lower than a pre-specified 'significance level' of the test, which we denote by $\alpha \in (0, 1)$, a researcher *rejects* the null hypothesis, meaning that the researcher declares that the observed data are not consistent with the hypothesis. When the p-value is greater than or equal to the significance level of the test, the researcher *fails to reject* the null hypothesis – meaning that the researcher declares that there is not enough information to state that the observed data are inconsistent with the hypothesized state of the world. Sometimes, we talk about hypothesis testing as an attempt to distinguish signal from noise; a high p-value tells us that we cannot distinguish signal from noise, and a low p-values tells us that we can do so.

Hypothesis tests are subject to at least two types of errors: first, one could reject the null hypothesis when it is true (a type I error) or, second, fail to reject the null hypothesis when it is false (a type II error). Two features of hypothesis tests related to these two potential errors are the α size of the test and the *power* of the test. We now define the α size (distinct from the α level) and power-of-hypothesis tests.

A test's α *level* is, in the words of Rosenbaum (2010), that test's 'promise' that the probability of a Type I error (i.e., the probability of a p-value that is less than α when the null hypothesis is true) is less than or equal to the α level. The test's α *size*, on the other hand, is the test's true probability of a Type I error, which, in general, can be greater than, equal to or less than the α level 'promised' by the test. In contrast to the α level and size of a test, a test's power is the

probability of a p-value that is less than the α level when the null hypothesis is false. In other words, power is 1 minus the Type II error probability; hence, as the power of a test increases, the Type II error probability decreases.

In the subsections to follow, we first define 'good' properties of hypothesis tests. We then describe tests of causal hypotheses in two distinct traditions traceable to Fisher (1935) (subsequently developed most extensively by Rosenbaum, 2002, 2010), and Neyman and Pearson (1933). We then explain the role that research design plays in justifying whether tests in either of these two traditions have good properties.

What Makes a Hypothesis Test a 'Good' Test?

We have already discussed three properties of good estimators – namely, unbiasedness, consistency and precision – but what makes a test of a null hypothesis relative to an alternative hypothesis a 'good' test? Just as we did for estimation, we describe 'good' features of hypothesis tests in the context of a fixed, finite population and in the context of a hypothetical scenario in which a study's size increases towards ∞ by increasing the number of copies of the study. The first two 'good' properties (a Type I error probability less than the α level and an unbiased test) refer to the former context, and the third property (a consistent test) refers to the latter context. Informally, a good hypothesis test should rarely mislead us: it should rarely encourage us to declare that we have discovered a signal in the noise when no signal exists and it should often find signals when they do exist.

Regardless of the size of a given study population, a hypothesis test first ought to control its α size (true Type I error probability) such that it is less than or equal to the test's α level. Second, a hypothesis test ought to be an *unbiased test* (not to be confused with an *unbiased estimator*), i.e., the probability of rejecting the null hypothesis when it is false and the alternative hypothesis is true should be at least as great as the probability of rejecting the null hypothesis when it is true and the alternative hypothesis is false (see Lehmann and Romano, 2005, chapter 4). Intuitively, we want to *reject* something that is *false* and we want to *not reject* something that is *true*. A test that leads us to reject true nulls with greater probability than we reject false nulls does not yield inferences that track the true causal effect. A good hypothesis test ought to be an unbiased test

in this sense.

Turning now to the asymptotic context described in the section 'An Illustrative Example', a hypothesis test ought to be a *consistent test* (not the same as a *consistent estimator*); that is, as the size of the study population increases asymptotically while all other relevant factors remain constant, the probability of rejecting the null hypothesis when it is false and the alternative is true should tend to 1 (see Lehmann and Romano, 2005, chapter 11). We now show the role that research design plays in enabling 'good' hypothesis tests in two different design-based traditions: Fisherian (Fisher, 1935) and Neymanian (Neyman and Pearson, 1933).

Fisherian Hypothesis Testing

Hypothesis testing in the tradition of Fisher (1935), later developed most extensively by Rosenbaum (2002, 2010), assesses the consistency of the observed data with a null hypothesis vis-a-vis an alternative hypothesis. A strong null hypothesis,[7] which we denote by $\boldsymbol{\tau}_0$, postulates an individual treatment effect for all $i \in \{1, \ldots, N\}$ units in a given study population. For example, one strong null hypothesis is $\boldsymbol{\tau}_0' = \begin{bmatrix} 5 & 5 & \cdots & 5 & 5 \end{bmatrix}$ and another might be $\boldsymbol{\tau}_0' = \begin{bmatrix} 0.5 & -10 & \cdots & 200 & -74.25 \end{bmatrix}$. The *strong null hypothesis of no effect* (which we henceforth refer to as 'the strong null of no effect') specifically postulates that $\tau_i = 0$ for all $i \in \{1, \ldots, N\}$ units – i.e., that $\boldsymbol{\tau}_0' = \begin{bmatrix} 0 & 0 & \cdots & 0 & 0 \end{bmatrix}$.

The consistency of the observed data with a strong null hypothesis vis-a-vis an alternative is typically assessed via p-values. To reiterate, a p-value is the probability of a test statistic at least as extreme as the observed test statistic from the perspective of the null hypothesis: as we will show below, the hypothetical world of the null generates, along with the known research design, the probability distribution that we compare against our single observed test statistic. In the context of Fisherian hypothesis tests, we can formally represent upper (p_u), lower (p_l) and two-sided (p_t)

[7] We use the term 'strong' instead of 'sharp,' as used by Fisher (1935), to contrast *strong* null hypotheses with *weak* null hypotheses which we discuss below.

p-values as follows:

$$
\begin{aligned}
p_u &= \sum_{m=1}^{|\Omega|} \mathbb{1}\left[t\left(\mathbf{z}_m, \mathbf{y}_{0_m}\right) \geq T\right] \Pr\left(\mathbf{Z} = \mathbf{z}_m\right) \\
p_l &= \sum_{m=1}^{|\Omega|} \mathbb{1}\left[t\left(\mathbf{z}_m, \mathbf{y}_{0_m}\right) \leq T\right] \Pr\left(\mathbf{Z} = \mathbf{z}_m\right) \\
p_t &= \min\left\{1, 2\min\left\{p_u, p_l\right\}\right\},
\end{aligned} \tag{10}
$$

where the index $m \in \left\{1, \ldots, |\Omega|\right\}$ runs over all possible assignments in the set of assignments Ω, $\mathbb{1}$ is an indicator function that is 1 if the argument $[\cdot]$ is true and 0 if false, $t\left(\mathbf{z}_m, \mathbf{y}_{0,m}\right)$ is the null test statistic (using $\mathbf{y}_{0,m}$ to refer to the vector of observed outcomes for the mth assignment implied by the null hypothesis, H_0) and T is the observed test statistic.[8]

To provide an illustration of Fisherian p-values, we return to the example in Table 41.1 and imagine that the assignment $\mathbf{z}_8' = \begin{bmatrix} 1 & 0 & 0 & 1 & 1 & 0 \end{bmatrix}$ happened to be the one randomly selected. In this case, the realization of data would be as follows.

$\mathbf{z}_8$	$\mathbf{y_c}$	$\mathbf{y_t}$	$\mathbf{y}_8$
1	?	22	22
0	8	?	8
0	11	?	11
1	?	15	15
1	?	18	18
0	1	?	1

Table 41.5: Realization of Data if $\mathbf{z}_8$ were the randomly drawn assignment

If the researcher uses the Difference-in-Means test statistic to provide a single, numerical summary of the observed data in Table 41.5, then the observed test statistic would be $t(\mathbf{z}, \mathbf{y}) = 11.6667$. Let's assume that the researcher wants to assess the consistency of this observed test statistic with the strong null of no effect – i.e., that $\tau_i = 0$ for all i – relative to the alternative hypothesis of a positive effect – i.e., that τ_i is nonnegative for all i and positive for at least one i.

Potential outcomes are only partially observed, but the researcher can 'fill in' the missing

[8] The expression for a two-sided p-value, $p_t = \min\left\{1, 2\min\left\{p_u, p_l\right\}\right\}$, comes from Rosenbaum (2010, 33) who states that '[i]n general, if you want a two-sided P-value, compute both one-sided P-values, double the smaller one, and take the minimum of this value and 1.' The rationale is that doubling a one-sided p-value compensates for, in essence, testing twice.

potential outcomes following the strong null hypothesis $H_0 : y_{t,i} = y_{c,i}$ for all i. Below, we can also show what this hypothesis implies for the observed outcomes y_i, recalling that $y_i = Z_i y_{t,i} + (1 - Z_i) y_{c,i}$ and writing $y_{c0,i}$ to mean 'value of $y_{c,i}$ under H_0':

$$
\begin{aligned}
y_{c0,i} &= y_i - z_i \tau_{0_i} \\
y_{t0,i} &= y_i + (1 - z_i)\, \tau_{0_i},
\end{aligned}
\tag{11}
$$

which in the case of the strong null of no effect implies that units' null potential outcomes and observed outcomes are as they appear in Table 41.6.

$\mathbf{z}_8$	$\mathbf{y_{c0}}_8$	$\mathbf{y_{t0}}_8$	$\mathbf{y}_8$
1	22	22	22
0	8	8	8
0	11	11	11
1	15	15	15
1	18	18	18
0	1	1	1

Table 41.6: Null potential outcomes if $\mathbf{z}_8$ were the realized assignment and under the hypothesis that $y_{t,i} = y_{c,i}$ for all i

Considering the strong null of no effect as a model of unobserved causal relationships for the sake of argument, the researcher knows exactly what all other possible realizations of data would look like under each possible assignment in Ω. Hence, the researcher can summarize all other possible realizations of data under the null with the same Difference-in-Means test statistic – generating a probability distribution of those null test statistics – and then calculate the probability of a null test statistic greater than or equal to the observed test statistic of 11.6667 (see Figure 41.3).

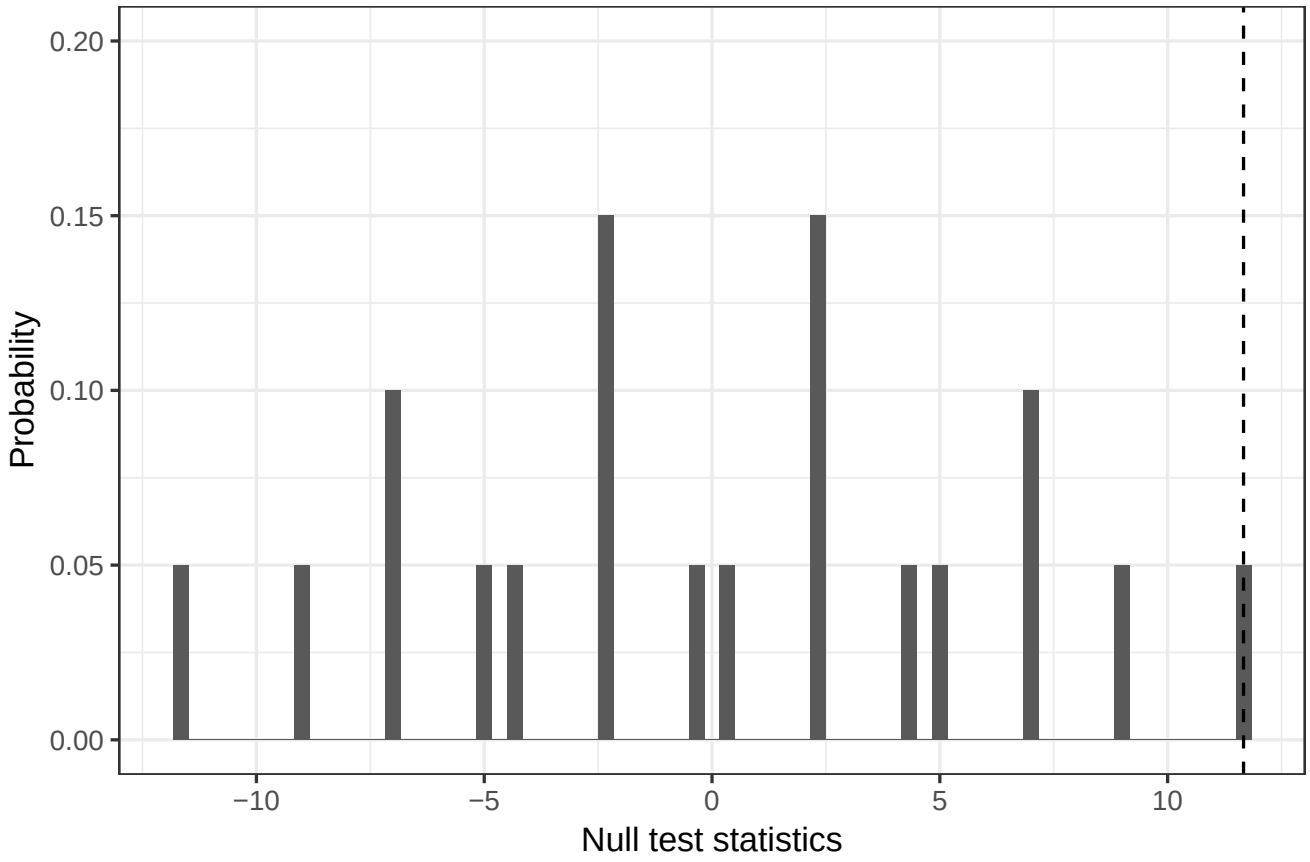

Figure 41.3: Distribution of Difference-in-Means test statistic under strong null of no effect when $\mathbf{z}_8$ is the realized assignment

In this case, the upper p-value – p_u from Equation (10) – is 0.05. If the value of α had been pre-set to a level greater than 0.05, then the researcher would reject the null. Table 41.1 shows that the strong null of no effect is false ($y_{t,i} \neq y_{c,i}$ for all i) and the alternative of a positive effect is true; hence, this particular choice to reject the strong null of no effect relative to the alternative of a positive effect happened to be a good one.

To assess whether the hypothesis-testing procedure is a good one overall, we want to establish that the test (1) has a true Type I error probability less than the α level, (2) is unbiased and (3) is consistent. We will now illustrate these three properties in turn.

To see this point, we return to the example given in Table 41.1, where the true vector of individual causal effects is $\boldsymbol{\tau}' = \begin{bmatrix} 2 & 4 & 0 & 5 & 4 & 3 \end{bmatrix}$, and Table 41.2 describes the 20 possible realizations of data. When the null hypothesis, $\boldsymbol{\tau}_0$, is false, the potential outcomes implied by the null hypothesis vary depending on which data are realized. However, when the null hypothesis, $\boldsymbol{\tau}_0$,

is true, i.e., when $\boldsymbol{\tau}_0 = \boldsymbol{\tau}$, the potential outcomes implied by the null hypothesis are fixed across all possible realizations of data, as shown by Table 41.7.

$\mathbf{z}_1$	$\mathbf{y_{c0}}$	$\mathbf{y_{t0}}$	$\mathbf{y}_1$
1	20	22	22
1	8	12	12
1	11	11	11
0	10	15	10
0	14	18	14
0	1	4	1

$\mathbf{z}_2$	$\mathbf{y_{c0}}$	$\mathbf{y_{t0}}$	$\mathbf{y}_2$
1	20	22	22
1	8	12	12
0	11	11	11
1	10	15	15
0	14	18	14
0	1	4	1

...

$\mathbf{z}_{19}$	$\mathbf{y_{c0}}$	$\mathbf{y_{t0}}$	$\mathbf{y}_{19}$
0	20	22	20
0	8	12	8
1	11	11	11
0	10	15	10
1	14	18	18
1	1	4	4

$\mathbf{z}_{20}$	$\mathbf{y_{c0}}$	$\mathbf{y_{t0}}$	$\mathbf{y}_{20}$
0	20	22	20
0	8	12	8
0	11	11	11
1	10	15	15
1	14	18	18
1	1	4	4

Table 41.7: Null Potential Outcomes for all Possible Realizations of Data when the Null Hypothesis is True. The observed outcomes column, $\mathbf{y}_j = \mathbf{y}_{c0} + \mathbf{z_j}\boldsymbol{\tau}'$.

If we set the significance level of the test to $\alpha = 0.10$, then it is true by definition that the probability that an observed test statistic lies in the lower tail of its distribution is less than or equal to 0.10, and the same is true for the probability that an observed test statistic lies in the upper tail of its distribution. Figure 41.4 shows the distribution of this test statistic, in which both the lower and upper tails according to $\alpha = 0.10$ are shaded in red.

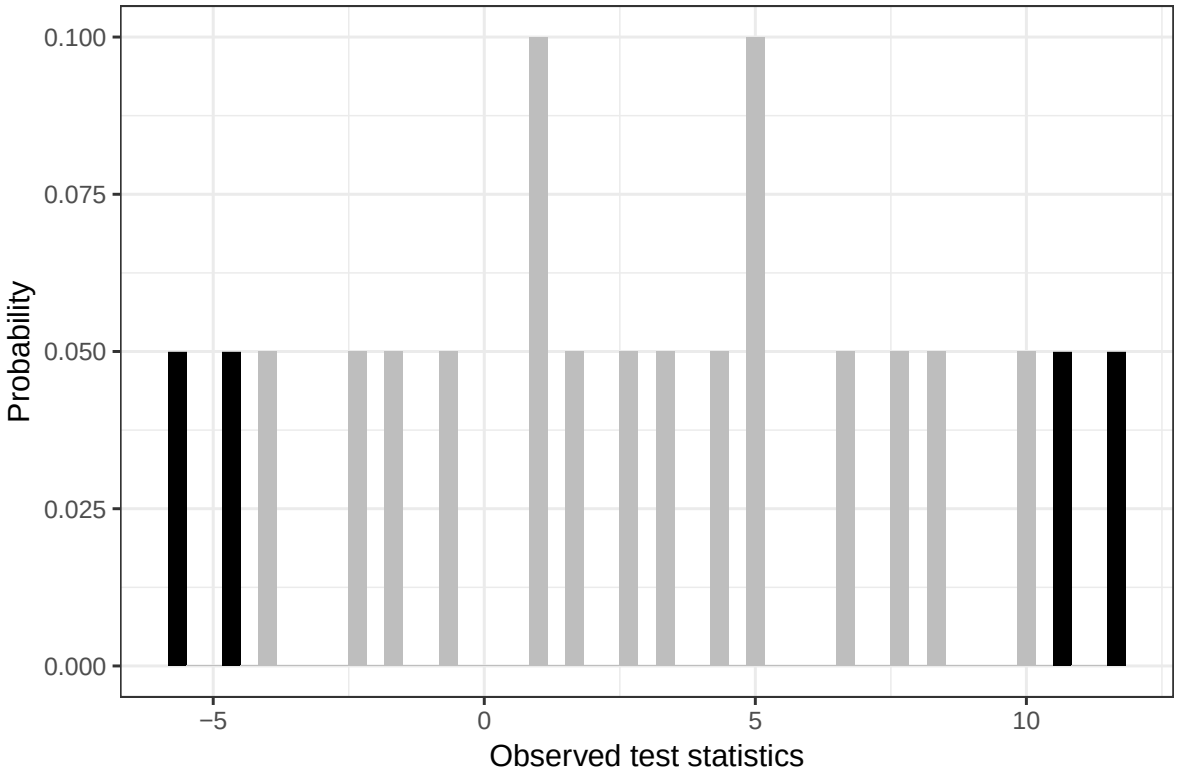

Figure 41.4: Distribution of observed test statistic. The upper and lower tails containing 10% of the area under the curve are shaded in red.

Notice that no matter which of the 20 possible realizations of data is actually realized, the null potential outcomes are identical to the true potential outcomes in Table 41.1, and hence each of the 20 possible null distributions of the test statistic are identical to the distribution of the

observed test statistic. Since all 20 possible null distributions are identical to the distribution of the observed test statistic, the probability that an observed test statistic lies in one of the tails of the null distribution must also be less than or equal to 0.10. This means that the probability of a null test statistic that is more extreme than the observed test statistic must be less than or equal to α.

Fisherian hypothesis tests possess the property that the Type I error probability is less than or equal to the test's α level. Yet such a property does *not* imply that the power of the test is greater than the test's Type I error probability (i.e., that the test is unbiased). To show that a test is unbiased relative to a specific class of alternative hypotheses, we now need to more precisely define the alternative to the null hypothesis.

In the previous sections, we referred to causal effects by the vector $\boldsymbol{\tau}$, but now we define the size of causal effects using the vectors of control and treatment potential outcomes: $\mathbf{y_c}$ and $\mathbf{y_t}$, respectively. More specifically, following Rosenbaum (2002), we define a treatment effect that is 'larger' than another treatment effect as follows:

Definition 1. *One treatment effect* $(\mathbf{y_c}^*, \mathbf{y_t}^*)$ *has a larger effect than another treatment effect* $(\mathbf{y_c}, \mathbf{y_t})$ *if and only if* $y^*_{t,i} \geq y_{t,i}$ *and* $y^*_{c,i} \leq y_{c,i}$ *for all* $i \in \{1, \ldots, N\}$ *units, where* $\mathbf{y_t}^* \neq \mathbf{y_t}$ *or* $\mathbf{y_c}^* \neq \mathbf{y_c}$.

Such an ordering of causal effects is consistent with many models of treatment effects, such as additive, multiplicative, tobit and dilated effects (Rosenbaum, 1999, 2002, 2010), not solely with the model of a constant, additive effect that we use here.

We now use this formal definition of 'larger effect' in terms of potential outcomes to define a desirable property of a test statistic: 'larger' effects yield test statistic values greater than those produced by 'smaller' effects. Rosenbaum (2002, chapter 2) shows that an 'effect-increasing' test statistic with respect to two possible realizations of data, $(\mathbf{z}, \mathbf{y})$ and $(\mathbf{z}, \mathbf{y}^*)$, satisfies this property. Following Rosenbaum (2002), we define an effect-increasing test statistic as follows:

Definition 2. *A test statistic,* $t(\cdot, \cdot)$*, is effect increasing when* $t(\mathbf{z}, \mathbf{y}) \leq t(\mathbf{z}, \mathbf{y}^*)$ *whenever* $y_i \leq y^*_i$ *for all* $i \in \{1, \ldots, N\} : z_i = 1$ *and* $y^*_i \leq y_i$ *for all* $i \in \{1, \ldots, N\} : z_i = 0$.

An effect-increasing test statistic ensures that, when the null hypothesis is false and the alternative of a larger effect is true, each possible realization of data yields a test statistic value that is greater than or equal to the corresponding test statistic value when the null hypothesis is true and the alternative of a larger effect is false. To understand this property, consider the following example in Table 41.8 of two possible causal effects, $(\mathbf{y_c}, \mathbf{y_t})$ and $(\mathbf{y_c}^*, \mathbf{y_t}^*)$, in which the former is a null effect and the latter is a larger, positive causal effect (see definition 1).

$\mathbf{y_c}$	$\mathbf{y_t}$
20	20
8	8
11	11
10	10
14	14
1	1

$\mathbf{y_c}^*$	$\mathbf{y_t}^*$
20	22
3	12
10	11
10	15
9	19
1	4

Table 41.8: No-Effects, $(\mathbf{y_c}, \mathbf{y_t})$, and Positive-Effects, $(\mathbf{y_c}^*, \mathbf{y_t}^*)$

Regardless of whichever three out of the six units are assigned to treatment in $\mathbf{z}$, the larger causal effect, $(\mathbf{y_c}^*, \mathbf{y_t}^*)$, will always yield a value of the observed outcome for all treated units that is greater than or equal to the outcome we would see with the same $\mathbf{z}$ in the no-causal-effect state, and we would also see a value of for all control units that is less than or equal to what we would see in the no-causal-effect state. Table 41.9 shows that an effect-increasing test statistic ensures that the observed test statistic of a larger effect is always greater than or equal to the observed test statistic of a smaller effect.

$\mathbf{Z}$	No-Effects $t(\mathbf{Z}, \mathbf{Y})$	Positive-Effects $t(\mathbf{Z}, \mathbf{Y}^*)$
$\mathbf{z}_1$	4.67	8.33
$\mathbf{z}_2$	4.00	9.67
$\mathbf{z}_3$	6.67	10.67
$\mathbf{z}_4$	-2.00	3.00
$\mathbf{z}_5$	6.00	11.67
$\mathbf{z}_6$	8.67	12.67
$\mathbf{z}_7$	0.00	5.00
$\mathbf{z}_8$	8.00	14.00
$\mathbf{z}_9$	-0.67	6.33
$\mathbf{z}_{10}$	2.00	7.33
$\mathbf{z}_{11}$	-2.00	2.67
$\mathbf{z}_{12}$	0.67	3.67
$\mathbf{z}_{13}$	-8.00	-4.00
$\mathbf{z}_{14}$	0.00	5.00
$\mathbf{z}_{15}$	-8.67	-2.67
$\mathbf{z}_{16}$	-6.00	-1.67
$\mathbf{z}_{17}$	2.00	7.00
$\mathbf{z}_{18}$	-6.67	-0.67
$\mathbf{z}_{19}$	-4.00	0.33
$\mathbf{z}_{20}$	-4.67	1.67

Table 41.9: Observed mean-difference test statistics under all possible assignments for both a no-effects and positive-effects true causal effect.

In addition to the property that increasing causal effects map monotonically onto increasing test statistics, we also want the p-values for tests of the null hypothesis when the null is false and the alternative is true to be smaller compared to the p-values when the null is true and the alternative is false. An effect-increasing test statistic also suffices for this property (for a formal proof, see Rosenbaum, 2002, chapter 2).

Table 41.10 shows that for all $\mathbf{z}_1, \ldots, \mathbf{z}_{20}$, the p-value of the strong null of no effect when the positive effect, $(\mathbf{y_c}^*, \mathbf{y_t}^*)$, is true is less than or equal to the strong null's p-value when no-effect, $(\mathbf{y_c}, \mathbf{y_t})$, is true.

$\mathbf{Z}$	No-Effects P-Value for $(\mathbf{y_c}, \mathbf{y_t})$	Positive-Effects P-Value for $(\mathbf{y_c}^*, \mathbf{y_t}^*)$
$\mathbf{z}_1$	0.25	0.20
$\mathbf{z}_2$	0.30	0.10
$\mathbf{z}_3$	0.15	0.05
$\mathbf{z}_4$	0.70	0.40
$\mathbf{z}_5$	0.20	0.10
$\mathbf{z}_6$	0.05	0.05
$\mathbf{z}_7$	0.55	0.40
$\mathbf{z}_8$	0.10	0.05
$\mathbf{z}_9$	0.60	0.35
$\mathbf{z}_{10}$	0.40	0.20
$\mathbf{z}_{11}$	0.70	0.45
$\mathbf{z}_{12}$	0.45	0.30
$\mathbf{z}_{13}$	0.95	0.90
$\mathbf{z}_{14}$	0.55	0.30
$\mathbf{z}_{15}$	1.00	0.80
$\mathbf{z}_{16}$	0.85	0.65
$\mathbf{z}_{17}$	0.40	0.25
$\mathbf{z}_{18}$	0.90	0.80
$\mathbf{z}_{19}$	0.75	0.65
$\mathbf{z}_{20}$	0.80	0.60

Table 41.10: Comparing p-values for tests of the strong null of no effect when no effects are true ($(\mathbf{y_c}, \mathbf{y_t})$) and false ($(\mathbf{y_c}^*, \mathbf{y_t}^*)$).

We have just shown that Fisherian tests using effect-increasing test statistics are unbiased: they provide more evidence against false claims than against true claims. In addition to being unbiased, we would also like our tests to be consistent, i.e., as the size of the study population grows towards ∞, while other factors remain constant, the power of the test tends to 1. Returning to the example in Table 41.1, notice that the probability of a p-value less than α increases along the sequence of finite populations of increasing size, given in Table 41.3. For example, with an α level of $\alpha = 0.10$, Figure 41.5 shows that the probability of a p-value less than $\alpha = 0.10$ grows greater and greater.

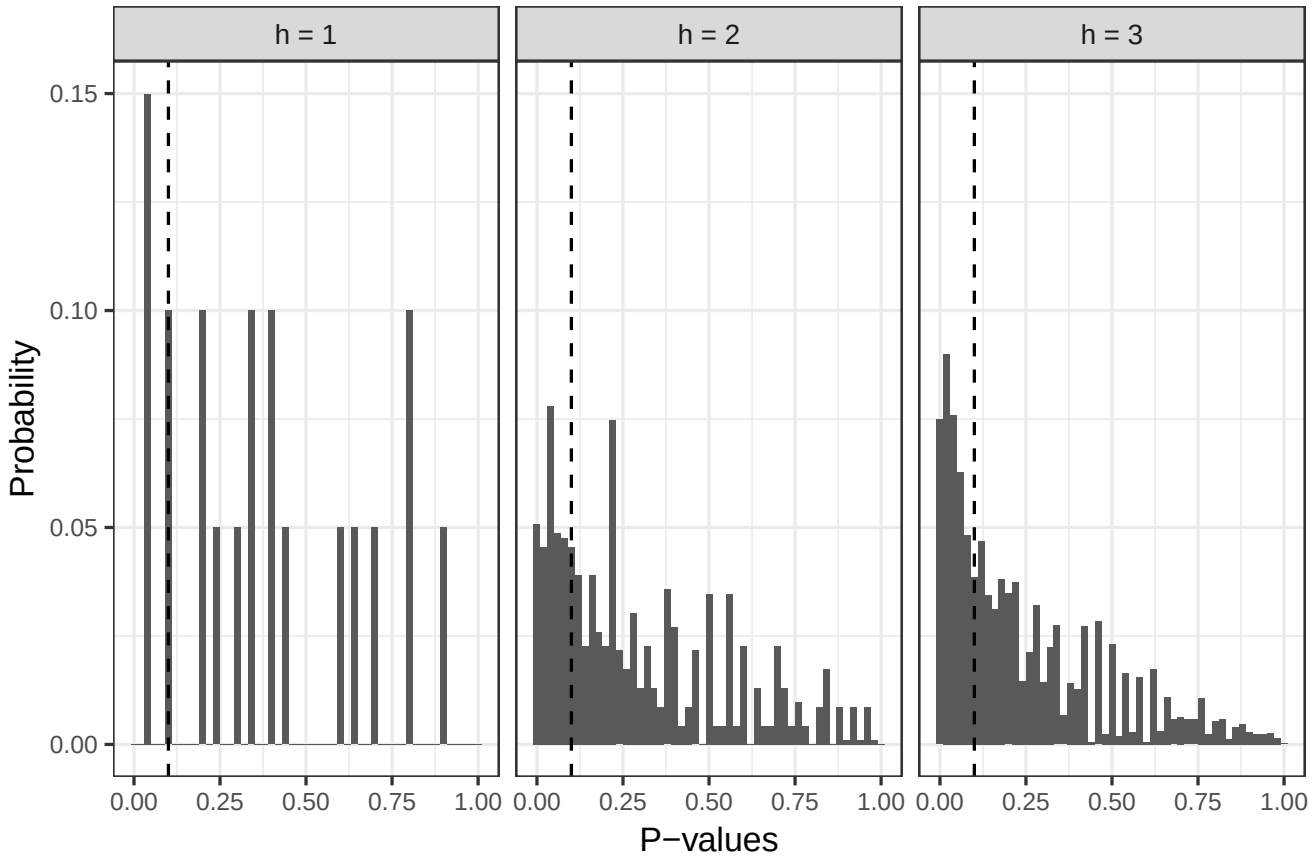

Figure 41.5: Distribution of Fisherian p-values for test of strong null of no effect under all realizations of data as the size of the experimental pool grows from one copy of the study, $h = 1$ ($n = 6$)); to two copies of the study, $h = 2$ ($n = 12$); to three copies of the study, $h = 3$ ($n = 18$).

Figure 41.5 shows that when the study population size grows from 6 to 12 and 18 units while all other factors are held constant, the power increases from 0.15 to roughly 0.2933 and 0.3741, respectively. As the population size increases along the sequence given in Table 41.3, the power of the test of the strong null of no effect will tend to 1. In other words, as we draw upon more and more data, we *reject* a *false* null with greater and greater probability.

We used the Difference-in-Means estimator as a test statistic for our example data in order to demonstrate the properties of Fisherian hypothesis tests in randomized experiments. However, we could have also used rank-based test statistics, standardized mean differences or many other functions of treatment and outcomes. All of those test statistics, like the Difference-in-Means test statistic, are effect-increasing. In fact, one of the challenges of Fisherian testing is its flexibility in terms of test statistics. We do not engage with those decisions here but encourage interested readers to see (Rosenbaum, 2010, chapter 2) for some discussion on using this flexibility to assess

substantively interesting hypotheses about pareto optimal causal effects, as well as Caughey et al. (2018) and Bowers et al. (2013, 2016) for examples on the propagation of causal effects on networks.

Hypothesis Testing in the Neymanian Tradition

While Fisherian tests allow the use of many different kinds of test statistics, Neymanian hypothesis tests are much more closely related to the estimation of mean causal effects (see the third section). In any given study, one can observe only a single *estimate*. However, a researcher can postulate (provisionally, for the sake of argument) a weak null hypothesis, $H_0 : \bar{\tau} = \bar{\tau}_0$, relative to some alternative hypothesis, such as $H_a : \bar{\tau} > \bar{\tau}_0$, and subsequently assess the probability of an estimate more extreme than the estimate the researcher actually observed if the weak null hypothesis were true. Such Neymanian hypothesis tests differ from Fisherian tests in several fundamental ways. Neymanian hypothesis tests (1) hypothesize about the mean causal effect, not the individual causal effect for each unit in the study, (2) require that the Difference-in-Means estimator be unbiased such that a hypothesis about the mean causal effect implies the same value for the mean (i.e., expected value) of the estimator, (3) require that the researcher estimate the variance of the Difference-in-Means estimator (recall that the distribution of the test statistic in the Fisherian test is known under random assignment and a strong null hypothesis) and (4) draws upon the finite population central limit theorem (Erdös and Rényi, 1959; Hájek, 1960; Li and Ding, 2017), which implies that the product of $\sqrt{N}$ multiplied by the difference between the estimator and its expected value converges to a normal distribution with mean equal to 0 and variance equal to $\sigma^2_{\hat{\bar{\tau}}}$, which, due to Slutsky's theorem, can be equivalently stated as $\frac{\hat{\bar{\tau}} - \mathbb{E}\left[\hat{\bar{\tau}}\right]}{\sqrt{\sigma^2_{\hat{\bar{\tau}}}}}$ converges in distribution to a standard normal (i.e., normal distribution with mean 0 and variance equal to 1).

We can see points (1)–(4) by looking at the common expressions for Neymanian p-values:

$$
\begin{aligned}
p_u &= 1 - \Phi\left(\frac{\hat{\bar{\tau}} - \bar{\tau}_0}{\sqrt{\hat{\sigma}^2_{\hat{\bar{\tau}}}}}\right) \\
p_l &= \Phi\left(\frac{\hat{\bar{\tau}} - \bar{\tau}_0}{\sqrt{\hat{\sigma}^2_{\hat{\bar{\tau}}}}}\right) \\
p_t &= 2\left(1 - \Phi\left(\frac{|\hat{\bar{\tau}} - \bar{\tau}_0|}{\sqrt{\hat{\sigma}^2_{\hat{\bar{\tau}}}}}\right)\right),
\end{aligned}
\tag{12}
$$

where $\hat{\bar{\tau}}$ is the familiar Difference-in-Means estimator (from the third section) now interpreted as a test statistic, not an estimator, $\hat{\sigma}^2_{\hat{\bar{\tau}}}$ is the conservative variance estimator (also from the third section) now used to describe the reference distribution for a null hypothesis rather than the precision of an estimator, $\bar{\tau}_0$ is a weak null hypothesis and $\Phi(\cdot)$ is the standard normal cumulative distribution function (CDF).

The expressions in Equation (12) return the probability of an estimate at least as extreme as the observed estimate if the weak null hypothesis, $\bar{\tau}_0$, were true. Notice, though, that the weak null hypothesis, $\bar{\tau}_0$, is technically a claim about the mean of the Difference-in-Means test statistic, which we can denote by $\mathbb{E}_0\left[\hat{\bar{\tau}}\right]$. However, because the Difference-in-Means estimator is unbiased, its expected value is always equal to the mean causal effect; this is why we use $\bar{\tau}_0$ in Equation (12) rather than $\mathbb{E}_0\left[\hat{\bar{\tau}}\right]$. Finally, note that the true variance of the Difference-in-Means test statistic is unknown, but the normal CDF requires two arguments – a value for the mean and a value for the variance – to assign a probability to estimates as least as extreme as the observed estimate. Rather than postulate a hypothesis about the variance of the estimator, like one does for the mean of the estimator, Neymanian tests use an estimate from the conservative variance estimator to calculate a p-value via the standard normal CDF.

In many situations, we can easily justify the assumption that $\frac{\hat{\bar{\tau}} - \bar{\tau}_0}{\sqrt{\hat{\sigma}^2_{\hat{\bar{\tau}}}}}$ is well approximated by a normal distribution by appealing to the aforementioned finite population Central Limit Theorem (CLT) and associated theory (see, e.g., Höglund, 1978). The Difference-in-Means test statistic

scaled by $\sqrt{N}$ is indeed asymptotically normal and the p-values in Equation (12) are all asymptotically valid – i.e., as N grows towards ∞, the probability of a Type I error is less than or equal to the α level of the test.

However, in small experiments in which the normal approximation is poor, tests of a weak null hypothesis relative to an alternative may have either a Type I error probability greater than the test's α level (when the null is true) or low power (when the null is false). Table 41.5 and Figure 41.3 demonstrate Neymanian p-values, in which the assignment $\mathbf{z}_8$ happened to be the one randomly drawn by the researcher. In this example, the observed Difference-in-Means test statistic is 11.6667 and the conservatively estimated variance is 12.8889. If we were to test the weak null hypothesis of no effect, i.e., that $\bar{\tau}_0 = 0$, against the alternative hypothesis that $\bar{\tau}_a > 0$, then the upper one-tailed p-value would be as follows:

$$\left(1 - \Phi\left(\frac{11.6667 - 0}{\sqrt{12.8889}}\right)\right) \approx 0.0006, \tag{13}$$

which yields a smaller p-value than the upper p-value we calculated via the Fisherian test of the strong null of no effects (which was $p = 0.05$). Figure 41.6 illustrates all upper p-values for a test of the weak null over all 20 possible realizations of data.

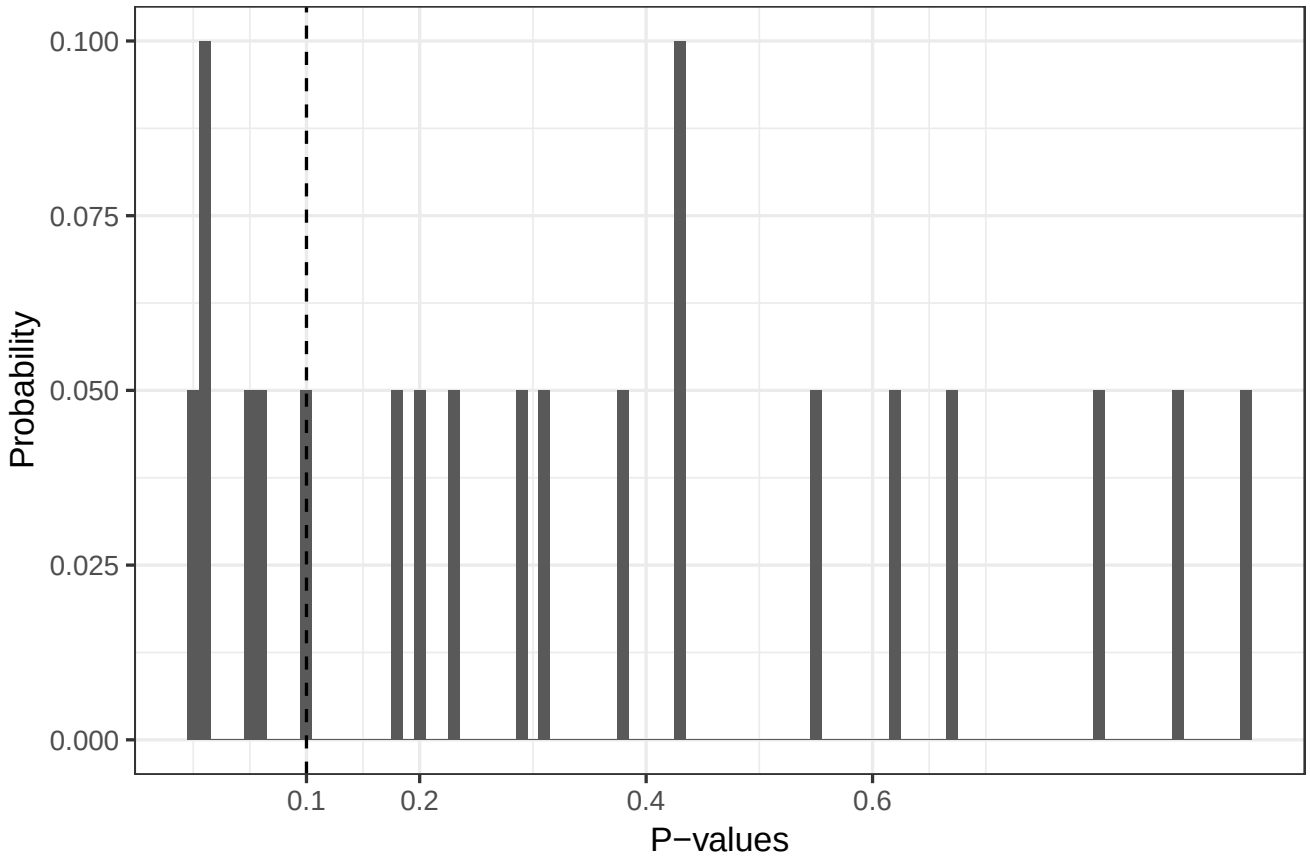

Figure 41.6: Distribution of Neymanian p-values under all realizations of data

In this particular case in which the weak null hypothesis is false, we can see that the Neymanian hypothesis test has high power. But in general, Neymanian tests can have bad properties in small experiments, such as a Type I error probability greater than α. For example, imagine that the weak null hypothesis of no effect were true as is depicted in Table 41.11 below:

$\mathbf{y_c}$	$\mathbf{y_t}$	$\boldsymbol{\tau}$
22	22	0
8	8	0
11	11	0
15	15	0
18	18	0
1	1	0

Table 41.11: Values of $\mathbf{y_c}$, $\mathbf{y_t}$ and $\boldsymbol{\tau}$ when weak null hypothesis is true

In this hypothetical experiment in which the weak null is true (and the strong null also happens to be true but need not be), Figure 41.7 shows that for some α levels like $\alpha = 0.05$, the type I error probability is greater than the test's α level (the black line is above the red line). For other

α levels, the type I error probability is less than the α level (the black line is below the red line), which makes the test at that level a conservative test. In short, although Neymanian hypothesis tests are asymptotically valid, such tests (particularly in small experiments) may yield type I error probabilities that do not fulfill the 'promise' made by a given α level.

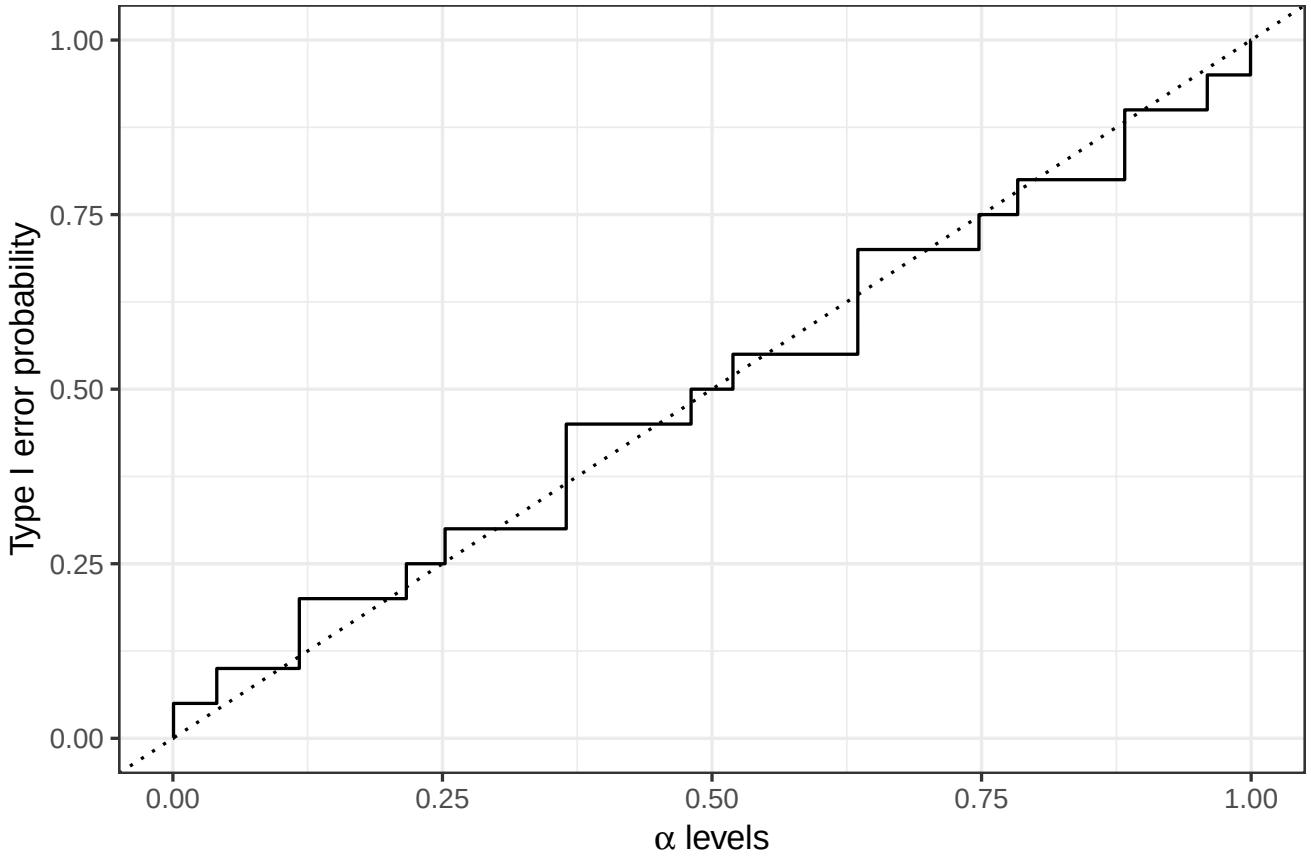

Figure 41.7: Distribution of Type I error probabilities for different α levels

An additional implication of the differences between Neymanian and Fisherian hypothesis tests is that, although the former is consistent[9], it is not necessarily unbiased in finite contexts, even when the Difference-in-Means test statistic is well approximated by a normal distribution. For intuition on this point, note that when the alternative of a larger effect is true and the null of no effects is false, the Difference-in-Means test statistic will yield values that are systematically larger than the values it would produce if the null of no mean effect were true – larger causal effects lead to larger test statistic values. Yet the alternative of a larger effect could be such that when it is true, the *variance estimates* are systematically larger than what the same variance estimates would

[9]We can see this property indirectly from Wu and Ding (2018) and Lin (2013).

be if the null were true – e.g., if a positive mean causal effect is caused by a few outliers that react strongly to treatment. Since the z-score scales the Difference-in-Means estimates by the variance estimates, the systematically greater variance estimates when the alternative is true could yield *smaller* z-scores (and hence larger p-values) compared to when the null is true. Hence, Neymanian hypothesis tests of weak causal hypotheses are not necessarily unbiased in finite contexts, even when the assumption of normality holds. For more on Fisherian versus Neymanian hypothesis tests, see Ding (2017), as well as the discussion from Aronow and Offer-Westort (2017), Chung (2017), Bailey (2017) and Loh et al. (2017).

Up to this point, we have discussed the role that research design plays in the quality of procedures for causal inference: estimation and testing. To simplify the exposition, we have been referring to situations where the researcher completely controls and thus knows the research design. We now consider situations in which the researcher does not control or only partially controls the research design.

Partially Controlled Research Designs: Noncompliance and Attrition

We refer to designs with imperfect compliance and/or attrition as partially controlled designs, in that the researcher *does* control the probability distribution on the set of possible assignments but does *not* control whether units actually comply with the assigned treatment or report their outcomes. Causal inferences in such partially controlled designs often require more assumptions to make causal inferences, which are typically defined no longer on the whole study population but a specific stratum of units in the study.

Noncompliance

Noncompliance occurs when units who are assigned to receive treatment or control do not actually receive it, where the random variable $D_i \in \{0, 1\}$ is an indicator variable for whether unit i has received or not received the assigned treatment. We are often substantively interested in the causal effect of actual receipt of treatment and not its mere assignment. Under complete uniform random assignment, we have shown above that the probability $Z_i = 1$ is identical for all $i = 1, \ldots, N$ units. Yet, we're interested in the causal effect of D_i, and since D_i is an outcome of Z_i (i.e., measured after Z_i), the probability that $D_i = 1$ is no longer identical for all units. Hence, a naive estimator

of the difference in observed outcomes between those who did and did not receive the treatment (i.e., the per-protocol estimator) is not necessarily unbiased.

Since whether or not units actually receive (or comply with) the treatment is an outcome variable measured after assignment, we can define units' compliance status in terms of their unobservable potential outcomes (Table 41.12).

$z_i = 0$	$z_i = 1$	Stratum
$d_{c,i} = 1$	$d_{t,i} = 1$	*Always-Taker*
$d_{c,i} = 0$	$d_{t,i} = 1$	*Complier*
$d_{c,i} = 1$	$d_{t,i} = 0$	*Defier*
$d_{c,i} = 0$	$d_{t,i} = 0$	*Never-Taker*

Table 41.12: Compliance Strata

Notice that the probability that $D_i = 1$ for an *Always Taker* is 1 and the probability that $D_i = 1$ for a *Never Taker* is 0. The only types of units for which the probability that $D_i = 1$ remain identical and on the interval $(0, 1)$ are *Compliers* and *Defiers*.

Angrist et al. (1996) show that scholars can *consistently estimate* the mean causal effect among Compliers under three further assumptions in addition to SUTVA and a uniform probability distribution on the set of assignments, Ω, which is sometimes referred to as the exogeneity of the instrument, $\mathbf{Z}$. These three assumptions, in addition to SUTVA and uniform random assignment, are:

1. Exclusion restriction – i.e., the instrument affects the outcome only through the receipt of treatment.
2. No Defiers – i.e., that there are no units for which $d_{c,i} = 1$ and $d_{t,i} = 0$.
3. At least one Complier – i.e., there exists at least one unit in the study for which $d_{c,i} = 0$ and $d_{t,i} = 1$.

To see this point, note that one can write the mean causal effect as a sum of the causal effects among the four strata (Always-Takers, Compliers, Defiers and Never-Takers), weighted by the proportion of units in each stratum:

$$\bar{\tau} = \delta_{AT}\pi_{AT} + \delta_C\pi_C + \delta_D\pi_D + \delta_{NT}\pi_{NT},$$

where $\delta_s, \pi_s : s \in \{AT, C, D, NT\}$ represent the mean causal effect in each stratum and the proportion of units in that stratum, respectively.

By the exclusion restriction assumption, the causal effect of Z on Y must be 0 for Always Takers and Never Takers (i.e., $\delta_{AT} = 0$ and $\delta_{NT} = 0$). By the assumption of no Defiers, the proportion of Defiers, π_D, is 0. Hence, the mean causal effect among Compliers is

$$\begin{aligned} \bar{\tau} &= \delta_C \pi_C \\ \frac{\bar{\tau}}{\pi_C} &= \delta_C, \end{aligned}$$

i.e., the mean causal effect on all units scaled by the proportion of Compliers is equal to the mean causal effect among Compliers.

We showed in the section on unbiasedness that under complete uniform random assignment, the Difference-in-Means estimator is unbiased regardless of the distributions of potential outcomes. Both $\mathbf{Y}$ and $\mathbf{D}$ are outcomes of $\mathbf{Z}$; hence, the Difference-in-Means estimators of $\hat{\bar{\tau}}(\mathbf{Z}, \mathbf{Y})$ and $\hat{\bar{\tau}}(\mathbf{Z}, \mathbf{D})$ unbiasedly estimate $\bar{\tau}$ and π_C, respectively. The Wald (or IV) estimator (Wald, 1940), which is also sometimes referred to as the Bloom estimator (Bloom, 1984) or the CACE (Complier Average Causal Effect) estimator (Gerber and Green, 2012), is defined as $\frac{\hat{\bar{\tau}}(\mathbf{Z},\mathbf{Y})}{\hat{\bar{\tau}}(\mathbf{Z},\mathbf{D})}$. In short, it is the ratio of these two unbiased estimators. This ratio estimator consistently, though not necessarily unbiasedly, estimates $\frac{\bar{\tau}}{\pi_C}$ (see Angrist and Pischke, 2008, chapters 4.6 and 4.7).

A design-based test of the hypothesis of no mean causal effect among *Compliers* is difficult due to the absence of an analytic expression for the variance of the CACE ratio estimator (for more on this topic, see Imbens and Rosenbaum, 2005; Kang et al., 2018). On the other hand, tests of the null hypothesis that the causal effect is 0 for all Compliers is relatively straightforward.

The null hypothesis of no causal effect among Compliers implies the strong null of no effect under the assumptions of the exclusion restriction and no Defiers. Since these two assumptions jointly imply that the individual causal effect of assignment is 0 for all Always Takers and Never Takers, and that there are no Defiers, then no causal effect among Compliers implies no causal effect among all units. For example, let's imagine that (like in Table 41.5) $\mathbf{z}_8$ was the assignment that the researcher happened to randomly draw, except that now we observe the following imperfect

compliance, as shown in Table 41.13:

$\mathbf{z}_8$	$\mathbf{y_c}$	$\mathbf{y_t}$	$\mathbf{y}_8$	$\mathbf{d}_8$	$\mathbf{d_c}$	$\mathbf{d_t}$
1	?	22	22	0	?	0
0	8	?	8	0	0	?
0	11	?	11	0	0	?
1	?	15	15	1	?	1
1	?	18	18	0	?	0
0	1	?	1	1	1	?

Table 41.13: Realization of Data if $\mathbf{z}_8$ were the randomly drawn assignment

We know that if $d_{t,i} = 0$, then, by the no-Defiers assumption, that unit must be a Never Taker, and if $d_{c,i} = 1$, then, by the same assumption, that unit must be an Always Taker. We don't know, however, whether treated units for which $d_t = 1$ are Compliers or Always Takers, and we don't know whether control units for which $d_c = 0$ are Compliers or Never Takers. Yet under the null hypothesis of no causal effect among Compliers, the individual causal effect is 0 regardless of whether a given unit is a Complier, Always Taker or Never Taker. Hence, we can fill in missing potential outcomes without knowing those units' compliance strata, as follows in Table 41.14.

$\mathbf{z}_8$	$\mathbf{y_c}$	$\mathbf{y_t}$	$\mathbf{y}_8$	$\mathbf{d}_8$	$\mathbf{d_c}$	$\mathbf{d_t}$
1	22	22	22	0	0	0
0	8	8	8	0	0	?
0	11	11	11	0	0	?
1	15	15	15	1	?	1
1	18	18	18	0	0	0
0	1	1	1	1	1	1

Table 41.14: Potential outcome under strong null of no effect if $\mathbf{z}_8$ were realized assignment

One can now use an effect-increasing test statistic, such as the Difference-in-Means test statistic, to test the hypothesis of no Complier causal effect against either the alternative of a positive Complier causal effect or a negative Complier causal effect. Hansen and Bowers (2009) present an application of this idea in the context of the one-sided compliance in a cluster-randomized get-out-the-vote campaign with a binary outcome, replacing what would be a complex two-stage logistic multilevel model with a relatively simple Fisherian hypothesis test. Imbens and Rosenbaum (2005) show how the Fisherian approach produces valid confidence intervals, where the Neyman-style approach using two-stage least squares fails to control the false positive rate when the instrument is weak (i.e., there are few Compliers).

Under the assumptions of the exclusion restriction and no Defiers, we do not need to know which units are Compliers in order to assert the hypothesis of no Complier causal effect. However,

if one were to posit a hypothesis other than no Complier causal effect, one would also need to posit a hypothesis about which units are Compliers and which are not. Hypothesis testing with imperfect compliance is therefore more complicated when testing hypotheses other than that of no Complier causal effect (for more on Fisherian approaches to instrumental variable analysis, see Kang et al., 2018; Rosenbaum, 1996, among others).

Attrition or Missing Outcomes

Our second step away from the ideal case of complete control over the research design is to allow for the possibility of missing outcomes. Let $r_{t,i}$ be an indicator, i.e., $r_{t,i} \in \{0, 1\}$, for whether subject i would respond to an attempt to measure an outcome, and let $r_{c,i} \in \{0, 1\}$ be an indicator for whether subject i would respond if assigned to control. We can represent whether an individual's outcomes are missing or not, based on Equation (14):

$$Y_i = \begin{cases} y_{c,i} + [y_{t,i} - y_{c,i}]Z_i & \text{if } R_i = 1 \\ \text{NA} & \text{if } R_i = 0, \end{cases} \tag{14}$$

where $R_i = Z_i r_{t,i} + (1 - Z_i)\, r_{c,i}$.

From Equation (14), we can see that if $R_i = 1$, then the researcher will observe $y_{c,i}$ for unit i if $Z_i = 0$ and $y_{t,i}$ for unit i if $Z_i = 1$. By contrast, if $R_i = 0$, then Y_i will be unobserved – i.e., NA.

We can define four distinct strata of subjects (see Table 41.15) with regard to attrition in order to help us understand how attrition can affect the properties of estimators and hypothesis tests. Just as we focused on a subgroup of units in the case of uncontrolled compliance, we can only infer the causal effects on certain subgroups when outcomes are missing (even when assignments are randomized).

$z_i = 0$	$z_i = 1$	Stratum
$r_{ci} = 1$	$r_{ti} = 1$	*Always Reporter*
$r_{ci} = 0$	$r_{ti} = 1$	*If Treated Reporter*
$r_{ci} = 1$	$r_{ti} = 0$	*If Untreated Reporter*
$r_{ci} = 0$	$r_{ti} = 0$	*Never Reporter*

Table 41.15: Attrition Strata

In the context of attrition, can we define a set of assumptions, as we did under imperfect compliance, in which estimators and tests satisfy the properties they should on some subset of the experimental data? It turns out that we can only do so if the question of whether a unit attrits or not is independent of treatment assignment. Without further assumptions, we know only that the random variable R_i is independent of the random variable Z_i among only *Always Reporters* and *Never Reporters*.[10] Therefore, if we assume that all experimental units belong to one of these two types, then our estimators and tests maintain the properties they should among the set of *Always Reporters*.[11] (We cannot observe outcomes for *Never Reporters* and hence cannot estimate or test hypotheses about causal effects on units that are *Never Reporters*.) On this set of *Always Reporters*, one can estimate causal effects and test strong or weak causal hypotheses, as we did above.

Uncontrolled Research Designs: Observational Studies

In this section, we finally relax the assumption that the researcher has control over how the treatment variable is assigned. The key distinction between experimental and observational studies is that in a randomized experiment, the researcher knows the probabilities with which units are *assigned* to treatment and control conditions; however, in an observational study, the researcher observes units only after they have *selected* into study conditions with unknown probabilities. How, then, is one to generate statistical inferences using estimators and tests focusing on causal effects when the probability distribution on the set of assignments, Ω, is unknown? A common design-based approach to this problem is to define units' treatment assignment probabilities as an unknown function of a set of baseline covariates. In the ideal (and often unattainable) case, by appropriate conditioning on these baseline covariates, the researcher can estimate and test hypotheses about causal effects via procedures that meet the desirable properties described at the outset of this chapter. Much of the work on observational studies emphasizes appropriate conditioning

[10] Note that R_i is independent of Z_i if and only if the probability that R_i takes on any value in its sample space does not vary conditional on any value that Z_i takes on in its sample space—i.e., that $\Pr\left(R_i = 1 \mid Z_i = 1\right) = \Pr\left(R_i = 1 \mid Z_i = 0\right)$ and $\Pr\left(R_i = 0 \mid Z_i = 1\right) = \Pr\left(R_i = 0 \mid Z_i = 0\right)$. The only two types of subjects who satisfy such independence are Always Reporters and Never Reporters.

[11] Other approaches to estimation and testing relax the assumption that outcome missingness is independent of treatment assignment and devise procedures that bound inferences about treatment effects under best- or worst-case scenarios, e.g., 'trimming bounds' (Lee, 2009) and 'extreme value bounds' (Manski, 1990).

on baseline covariates, as well as methods to diagnose the success of such conditioning strategies (e.g., Hansen and Bowers, 2008; Hartman and Hidalgo, 2018, among others). Realistically, the design-based approach in observational studies might be called an 'as-if-randomized' approach, e.g., a researcher might make choices about comparison groups such that within a group, treatment selection *appears* random.

The model of an observational study states that units are individually assigned to treatment or control by N *independent* (but not necessarily *identically distributed*) coin tosses. More specifically, for all $i \in \{1, \ldots, N\}$ units, we let $\Pr(Z_i = 1)$ be equal to $\lambda(\mathbf{x}_i)$, where $\lambda(\cdot)$ is an unknown function and $\mathbf{x}_i$ is unit i's fixed vector of baseline covariate values. Even if we don't know $\lambda(\cdot)$, if $\mathbf{x}_i = \mathbf{x}_j$ for any two units i and $j \neq i$, then it follows that $\Pr(Z_i = 1) = \Pr\big(Z_j = 1\big)$. Of course, we still don't know the function $\lambda(\cdot)$ and hence don't know the actual values of $\Pr(Z_i = 1)$ and $\Pr\big(Z_j = 1\big)$; we know only that these two values are equal. Therefore, if we construct a block, b, that consists of units i and $j \neq i$, then each possible assignment within that block has an equal probability of realization. If each possible assignment has an equal probability, then an observational study can be analyzed as if it is a uniform, block randomized experiment (for more on the analysis of block, randomized experiments, see Gerber and Green, 2012, chapter 4). This approach allows us to avoid directly estimating $\lambda(\cdot)$, although there are alternative approaches that do so and subsequently use these estimated values (known as estimated propensity scores) as a basis for inference (see Robins et al., 2000).

Scholars can therefore attempt to make an observational study as experiment-like as possible by creating matched blocks on the basis of observed covariates that determine units' treatment assignment probabilities. A range of matching algorithms exist to improve covariate balance and hence make observational studies like experiments as much as possible (at least on the basis of observed covariates). For more on this topic, see Hansen (2004), Diamond and Sekhon (2013), Sävje et al. (2017) and Zubizarreta (2012), among others. After matching (or perhaps weighting or non-parametric modeling) and favorable comparisons with actually randomized designs, researchers then must confront the fact that their observational studies are not randomized studies. This leads directly to the topic of sensitivity analyses.

Sensitivity to Assumptions in Uncontrolled Research Designs

Thus far, we have considered cases in which either the probability distribution on Ω is known by random assignment or units' assignment probabilities are a function of only *observed* covariates. But in an observational study, we rarely know – let alone observe – all relevant covariates. We now consider deviations from the assumption that units' assignment probabilities are determined by only observed covariates and subsequently assess how one's inferences would change under violations of this assumption.

A powerful, design-based framework for such a sensitivity analysis is given by Rosenbaum (2002). Before explaining this framework, we need to define a few additional terms. First, the *treatment odds* for unit $i \in \{1, \ldots, N\}$ is $\frac{\pi_i}{(1-\pi_i)}$, which is simply the ith unit's probability of assignment to treatment divided by that unit's probability of assignment to control. The *treatment odds ratio* for any two units i and $j \neq i$ is simply the ratio of the ith unit's treatment odds and the jth unit's treatment odds. If units' treatment odds are a function of only observed covariates *and* the researcher is able to obtain balance on all of these observed covariates, then the treatment odds for units $i, j \neq i : \mathbf{x}_i = \mathbf{x}_j$ is identical and their treatment odds ratio is 1.

Rosenbaum (2002) considers what happens when units' treatment odds are a function not only of observed covariates, $\mathbf{x}$, but also an unobserved covariate, u. Under the assumption of a logistic functional form between all units' treatment odds and baseline covariates, as well as the constraint that $0 \leq u \leq 1$, one can write the treatment odds of the ith unit as follows:

$$\begin{aligned} \frac{\pi_i}{(1-\pi_i)} &= \exp\left\{\kappa\left(\mathbf{x}_i\right) + \gamma u_i\right\} \\ \log\left(\frac{\pi_i}{(1-\pi_i)}\right) &= \kappa\left(\mathbf{x}_i\right) + \gamma u_i, \end{aligned}$$

where $\kappa\left(\cdot\right)$ is an unknown function and γ is an unknown parameter, and the treatment odds ratio

for units i and j is:

$$\begin{aligned}\frac{\left(\frac{\pi_i}{1-\pi_i}\right)}{\left(\frac{\pi_j}{1-\pi_j}\right)} &= \frac{\exp\left\{\kappa\left(\mathbf{x}_i\right)+\gamma u_i\right\}}{\exp\left\{\kappa\left(\mathbf{x}_j\right)+\gamma u_j\right\}} \\ &= \exp\left\{\left(\kappa\left(\mathbf{x}_i\right)+\gamma u_i\right)-\left(\kappa\left(\mathbf{x}_j\right)+\gamma u_j\right)\right\}.\end{aligned}$$

If $\mathbf{x}_i = \mathbf{x}_j$, then $\kappa\left(\mathbf{x}_i\right) = \kappa\left(\mathbf{x}_j\right)$ and, hence, the treatment odds ratio is simply:

$$\exp\left\{\gamma\left(u_i - u_j\right)\right\}.$$

Since $u_i, u_j \in [0,1]$, the minimum and maximum possible values of $\left(u_i - u_j\right)$ are -1 and 1. Therefore, the minimum and maximum possible values of the treatment odds ratio are $\exp\{-\gamma\}$ and $\exp\{\gamma\}$. After noting that $\exp\{-\gamma\} = \frac{1}{\exp\{\gamma\}}$, we can bound the treatment odds ratio between i and j as follows:

$$\frac{1}{\exp\{\gamma\}} \leq \frac{\left(\frac{\pi_i}{1-\pi_i}\right)}{\left(\frac{\pi_j}{1-\pi_j}\right)} \leq \exp\{\gamma\}. \tag{15}$$

We can denote $\exp\{\gamma\}$ by Γ and subsequently consider how one's inferences would change for various values of Γ.

For example, let's say that a researcher obtains balance via stratification on all observed covariates – such that the design closely resembles a uniform, block randomized experiment – and subsequently tests a strong null hypothesis under the assumption that all units' treatment odds are identical. Now the researcher considers deviations from this assumption. Different assumptions about u and γ imply differing probabilities of possible assignments, which, as Rosenbaum (2002, chapter 4) shows, can be represented by

$$\Pr\left(\mathbf{Z} = \mathbf{z}\right) = \frac{\exp\left\{\gamma \mathbf{z}'\mathbf{u}\right\}}{\sum_{\mathbf{z}\in\Omega}\exp\left\{\gamma \mathbf{z}'\mathbf{u}\right\}}. \tag{16}$$

To return to the working example from Table 41.1, let's imagine that the realized assignment

was $\mathbf{z}_8$, which yielded an observed test statistic of 11.6667 and a p-value of 0.05 (see Figure 41.3). We calculated that p-value under the assumption that all units had identical probabilities of assignment. We could relax that assumption and assume that units do *not* have identical assignment probabilities. For example, we might first assume that $\Gamma = 2$, which implies that $\gamma = \log(2) \approx 0.6931$. The quantity γ is the coefficient of a unit's unobserved baseline covariate u_i. If all units have the same value of this unobserved covariate — i.e., $u_i = u_j$ for all $i \neq j$ — then all units' assignment probabilities will remain identical. A conservative approach therefore might instead find a vector $\mathbf{u}$ that, given a value of γ, will maximize the p-value of a test of the strong null hypothesis of no effect. In this particular example, it is straightforward to see that $\mathbf{u}' = \begin{bmatrix} 1 & 0 & 0 & 1 & 1 & 0 \end{bmatrix}$ maximizes the p-value for an upper one-sided test of the strong null hypothesis of no effect. In general, different procedures exist for finding the vector $\mathbf{u}$ that maximizes (or minimizes) the p-value for a given value of Γ (see Gastwirth et al., 2000; Rosenbaum, 2018; Rosenbaum and Krieger, 1990). Figure 41.8 illustrates how the respective null distributions would differ under $\Gamma = 1$ and $\Gamma = 2$ in which $\mathbf{u}' = \begin{bmatrix} 1 & 0 & 0 & 1 & 1 & 0 \end{bmatrix}$.

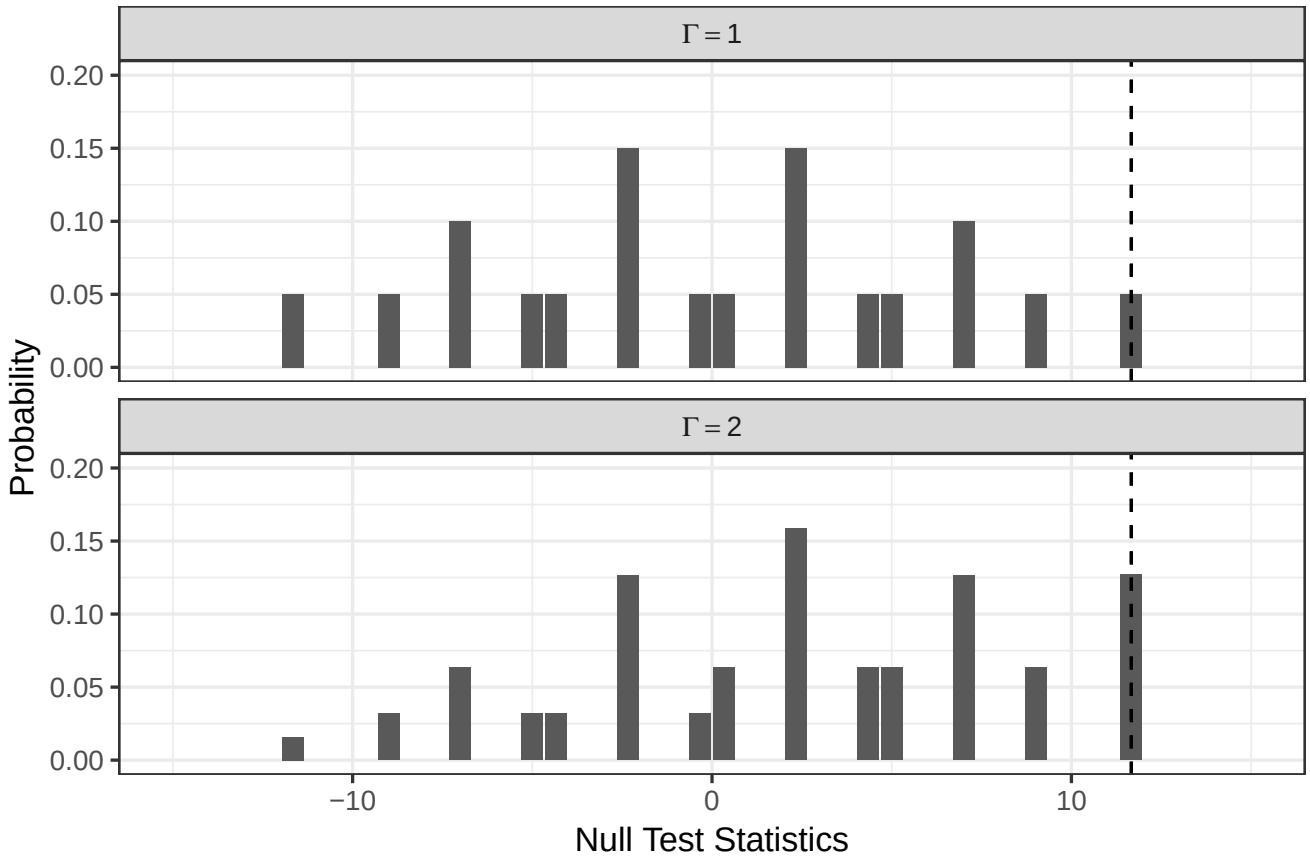

Figure 41.8: Distributions of Difference-in-Means test statistic under strong null of no effect when $\mathbf{z}_8$ is the realized assignment under (a) a model where an unobserved covariate has no effect on odds of treatment, $\Gamma = 1$, and (b) a model where an observed covariate doubles the odds of treatment, $\Gamma = 2$.

Notice that under $\Gamma = 1$ and $\Gamma = 2$, the observed test statistic remains fixed at 11.6667. The set of 20 null test statistic also remains fixed for $\Gamma = 1$ and $\Gamma = 2$. The value of Γ changes only the probability associated with each of the null test statistics. As we can see, when $\Gamma = 1$, the p-value is 0.05, but when $\Gamma = 2$, the p-value increases to approximately 0.1270. The researcher is no longer able to reject the strong null of no effect at a level of $\alpha = 0.10$ when $\Gamma = 2$ compared to when $\Gamma = 1$, and this is one way in which researchers can assess the sensitivity of their inferences to assumptions about the research design.

This approach is not the only way to formalize the impact of the assumptions underlying the 'as-if randomized' research designs used for causal inference when researchers have little to no control over the selection process of the main explanatory variable. See also Hosman et al. (2010) for an approach focusing on regression coefficients and regression-based adjustment, as well as an

application by Chaudoin et al. (2018) of ideas like these to problems in international relations.

Conclusion

Statistics and research design help us learn about general and abstract social science theory using concrete and specific observations. Observation helps us learn about theory, but observation occurs using the tools of research design and they are summarized, described and interpreted using the tools of statistics. In fact, we have shown that certain counterfactual causal quantities can never be directly observed, and that our ability to report with confidence about such unobserved causal effects depends heavily on statistical tools, which, in turn, depend on research design for their operation. The persuasiveness of claims about links from an estimate or p-value to an unobserved causal effect and general theory depends on the clarity of each step. At the most nitty-gritty level, we want our tools to work well – our estimators and tests should err in known and controlled manners and should err rarely.[12] We have asked, 'How can we know that tools do the work that we want them to do?' Additionally, we have shown how to use the facts of research design to answer that question and to justify use of these common tools. This means that when we want to persuade an audience that our findings support a given theory (or urge modification of it), we do not need our audiences to believe that (1) we have a random sample from a well-defined population, (2) that the outcome arises from some known probability process (like a normal or zero-inflated Poisson distribution) or (3) that the treatment or selection process relates to background covariates in some known (often linear and additive) fashion. Instead, in a randomized experiment, we ask that a reader believe that our research design correctly describes the physical processes that occurred in the research (a request that is not hard to verify and assess). In an observational study, we ask readers to agree that the as-if randomized approach is reasonable, and we present direct evidence comparing observational designs to randomized experiments in order to make the provisional as-if randomized approach easier to grant.

In explaining reliable procedures, we have used a very simple set of examples. Our simplifications include the use of only two experimental conditions (treatment and control), yet the general modes of inference described in this chapter can be straightforwardly applied to factorial

[12] We did not assess tools like randomizers or sampling schemes in this paper, but those tools are equally important in the effort to advance theory through observation.

experiments and other contexts with multiple treatments (see Dasgupta et al., 2015). A second such simplification has been the Stable Unit Treatment Value Assumption (SUTVA) (Cox, 1958; Rubin, 1980, 1986), which, in the case of a binary treatment variable, implies that all units have only two potential outcomes. Yet both estimation (see Aronow and Samii, 2017; Hudgens and Halloran, 2008) and testing (see Bowers et al., 2013, 2016; Rosenbaum, 2007) are possible when units have more than two potential outcomes due to interference between units, for example, in the context of experiments on networks. Furthermore, we have focused on randomized experiments and only briefly pointed to the strategy of 'as-if randomized' approaches to estimation and testing in observational studies. We explained that such approaches, when paired with sensitivity analyses, can enable persuasive statistical inferences about causal effects when the researcher lacks control over the design. In other words, once an observational research design compares favorably to the standard of an equivalent experiment (the way that a matched design can be compared to a block-randomized experiment), statistical inference about causal effects can use the same procedures, provisionally justified in the same way, as in a randomized experiment if also followed by a sensitivity analysis.

This focus on the basics – on ensuring that our statistical tools do what they should – leaves larger questions unexplored. For example, some researchers would prefer to make inferences about not only counterfactual causal effects among units in a given study but also about future units in data contexts that differ from the one under study. One might desire unbiased and consistent estimators, as well as valid and powerful tests, not only based on the research design generating the data collected here and now, but also for unknown future research designs guiding data collection elsewhere and at other times. Forecasting causal effects is an active research area (for only a few recent works on the topic, see Bisbee et al., 2017; Coppock et al., 2018; Dehejia et al., 2019; Pearl, 2015; Stuart et al., 2015). Whether or not a researcher or policy-maker would like a formal forecast of the causal effects of an intervention from one study to a new context (in time, space and/or units), information provided by a single study to the motivating theoretical question still depends on the reliability of the tools used to conduct and analyze the study. We have focused on showing how the reliability of such procedures is based on the research design itself and leave questions

about the properties of procedures for forecasting causal effects as a separate, though important, topic.

Design-based causal inference emphasizes inferring a counterfactual quantity, not a quantity in a population or of an outcome-generating model. Such an emphasis arises naturally from a wide range of social science research contexts, such as (1) when units are not a random sample, e.g., when the units are administrative units like schools or countries or convenience samples arising without a known chance process, or (2) when probability models of outcomes – and their structural relationship to explanatory variables – are difficult to write or justify, e.g., when an outcome can be plausibly modeled by multiple different likelihood functions. In such contexts, simple comparisons based on the research design can advance social science theory and avoid debates about data models. When strong theory generates clear probability and structural models, a model-based justification for statistical inference might be preferred, although we would want such modes of inferences to satisfy the same properties discussed in this chapter: tests should not mislead and estimators should produce estimates close to the truth. The model-based approach to showing whether these characteristics hold is well described in most statistics textbooks, and we recommend Cox (2006) for an overview.

A general question nevertheless remains: are design-based procedures better or worse than model-based procedures for advancing social science theory or policy learning? Design-based inference is simple, easily interpretable and can ensure that estimators and tests are reliable based on few assumptions, where the assumptions tend to be easily defended in terms of the known features of the research design. But does design-based inference possess reliable properties only for narrowly defined research questions? To be sure, there is nothing intrinsic to design-based inference that requires scholars to use only specific estimators that focus only on specific estimands, such as mean causal effects, or to test only certain hypotheses about no causal effects instead of others. And, although we did not show it here, it is straightforward to assess the properties of non-standard estimators and tests by representing the research design and simulating from it (see Blair et al., 2019, for an example of a framework for simulation based assessment of estimators and tests). For only one example of the flexibility of design-based approaches, imagine that we wondered whether

a certain non-linear structural model described well a relationship between a causal driver (like an experimental treatment) and an outcome. In this case, Bowers et al. (2013, 2016) show how the evidence against structural models of unobserved potential outcomes can be generated and hypothesis tests created, i.e., there is nothing about design-based approaches that precludes the use of structural models. That said, a clear difference of means can often teach enough about a complicated structural theory such that there is no need to complicate the research design or statistical inference tasks. In the end, all else equal, reliable procedures advance scientific knowledge more than unreliable procedures do. For this reason, one of the benefits of engaging with design-based inference is that it brings clarity to the task of judging and choosing our statistical tools and provokes us to directly confront and grapple with the conditions under which evidence can be reliably interpreted as evidence for or against causal claims.

References

Achen, C. H. (1982). *Interpreting and Using Regression*. Number 07-029 in Quantitative Applications in the Social Sciences. Newbury Park, CA: Sage Publications. 11

Angrist, J. D., G. W. Imbens, and D. B. Rubin (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association 91*(434), 444–455. 37

Angrist, J. D. and J.-S. Pischke (2008). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton, NJ: Princeton University Press. 38

Aronow, P. M., D. P. Green, D. K. Lee, et al. (2014). Sharp bounds on the variance in randomized experiments. *The Annals of Statistics 42*(3), 850–871. 18

Aronow, P. M. and J. A. Middleton (2013). A class of unbiased estimators of the average treatment effect in randomized experiments. *Journal of Causal Inference 1*(1), 135–154. 18

Aronow, P. M. and M. R. Offer-Westort (2017). Understanding ding's apparent paradox. *Statistical Science 32*(3), 346–348. 36

Aronow, P. M. and C. Samii (2017). Estimating average causal effects under general interference, with application to a social network experiment. *Annals of Applied Statistics 11*(4), 1912–1947. 48

Bailey, R. A. (2017). Inference from randomized (factorial) experiments. *Statistical Science 32*(3), 352–355. 36

Bisbee, J., R. Dehejia, C. Pop-Eleches, and C. Samii (2017). Local instruments, global extrapolation: External validity of the labor supply–fertility local average treatment effect. *Journal of Labor Economics 35*(S1), S99–S146. 48

Blair, G., J. Cooper, A. Coppock, and M. Humphreys (2019). Declaring and diagnosing research designs. *The American Political Science Review 113*(3), 838–859. 49

Bloom, H. S. (1984). Accounting for no-shows in experimental evaluation designs. *Evaluation Review 8*(2), 225–246. 38

Bowers, J., M. Fredrickson, and C. Panagopoulos (2013). Reasoning about interference between units: A general framework. *Political Analysis 21*(1), 97–124. 31, 48, 50

Bowers, J., M. M. Fredrickson, and P. M. Aronow (2016). Research note: A more powerful test statistic for reasoning about interference between units. *Political Analysis 24*(3), 395–403. 31, 48, 50

Brewer, K. (1979). A class of robust sampling designs for large-scale surveys. *Journal of the American Statistical Association 74*(368), 911–915. 7

Caughey, D., A. Dafoe, and L. Miratrix (2018, June). Beyond the sharp null: Randomization inference, bounded null hypotheses, and confidence intervals for maximum effects. Working Paper. 31

Chaudoin, S., J. Hays, and R. Hicks (2018). Do we really know the wto cures cancer? *British Journal of Political Science 48*(4), 903–928. 47

Chung, E. (2017). Randomization-based tests for "no treatment effects". *Statistical Science 32*(3), 349–351. 36

Cochran, W. G. (1977). *Sampling Techniques* (3rd ed.). Hoboken, NJ: John Wiley & Sons. 17

Coppock, A., T. J. Leeper, and K. J. Mullinix (2018). Generalizability of heterogeneous treatment effect estimates across samples. *Proceedings of the National Academy of Sciences of the United States of America 115*(49), 12441–12446. 48

Cox, D. R. (1958). *Planning of Experiments*. New York, NY: Wiley. 1, 48

Cox, D. R. (2006). *Principles of Statistical Inference*. New York, NY: Cambridge University Press. 49

Dasgupta, T., N. S. Pillai, and D. B. Rubin (2015). Causal inference from 2k factorial designs by using potential outcomes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 77*(4), 727–753. 1, 48

Dehejia, R., C. Pop-Eleches, and C. Samii (2019). From local to global: External validity an a fertility natural experiment. *Journal of Business & Economic Statistics* (just-accepted), 1–48. 48

Diamond, A. and J. S. Sekhon (2013). Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies. *Review of Economics and Statistics 95*(3), 932–945. 42

Ding, P. (2017). A paradox from randomization-based causal inference. *Statistical Science 32*(3), 331–345. 36

Erdös, P. and A. Rényi (1959). On the central limit theorem for samples from a finite population. *Publications of the Mathematics Institute of the Hungarian Academy of Science 4*, 49–61. 31

Fisher, R. A. (1935). *The Design of Experiments*. Edinburgh, SCT: Oliver and Boyd. 1, 20, 21

Gastwirth, J. L., A. M. Krieger, and P. R. Rosenbaum (2000). Asymptotic separability in sensitivity analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 63*(3), 545–555. 45

Gerber, A. S. and D. P. Green (2012). *Field Experiments: Design, Analysis, and Interpretation*. New York, NY: W.W. Norton. 15, 38, 42

Hájek, J. (1960). Limiting distributions in simple random sampling from a finite population. *Publications of the Mathematics Institute of the Hungarian Academy of Science 5*, 361–374. 31

Hansen, B. B. (2004). Full matching in an observational study of coaching for the sat. *Journal of the American Statistical Association 99*(467), 609–618. 42

Hansen, B. B. and J. Bowers (2008). Covariate balance in simple, stratified and clustered comparative studies. *Statistical Science 23*(2), 219–236. 42

Hansen, B. B. and J. Bowers (2009). Attributing effects to a cluster-randomized get-out-the-vote campaign. *Journal of the American Statistical Association 104*(487), 873–885. 39

Hartman, E. and F. D. Hidalgo (2018). An equivalence approach to balance and placebo tests. *American Journal of Political Science 62*(4), 1000–1013. 42

Höglund, T. (1978). Sampling from a finite population. a remainder term estimate. *Scandinavian Journal of Statistics 5*(1), 69–71. 32

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association 81*(396), 945–960. 1

Horvitz, D. G. and D. J. Thompson (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association 47*(260), 663–685. 18

Hosman, C. A., B. B. Hansen, and P. W. Holland (2010). The sensitivity of linear regression coefficients' confidence limits to the omission of a confounder. *The Annals of Applied Statistics 4*(2), 849–870. 46

Hudgens, M. G. and M. E. Halloran (2008). Toward causal inference with interference. *Journal of the American Statistical Association 103*(482), 832–842. 48

Imbens, G. W. and P. R. Rosenbaum (2005). Robust, accurate confidence intervals with a weak instrument: Quarter of birth and education. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 168*(1), 109–126. 38, 39

Kang, H., L. Peck, and L. Keele (2018). Inference for instrumental variables: A randomization inference approach. *Journal of the Royal Statistical Society. Series A: Statistics in Society 181*(4), 1231–1254. 38, 40

Lee, D. S. (2009). Training, wages, and sample selection: Estimating sharp bounds on treatment effects. *The Review of Economic Studies 76*(3), 1071–1102. 41

Lehmann, E. L. and J. P. Romano (2005). *Testing Statistical Hypotheses* (Third ed.). Springer Texts in Statistics. New York, NY: Springer-Verlag. 20, 21

Li, X. and P. Ding (2017). General forms of finite population central limit theorems with applications to causal inference. *Journal of the American Statistical Association 112*(520), 1759–1769. 31

Lin, W. (2013). Agnostic notes on regression adjustments to experimental data: Reexamining freedman's critique. *The Annals of Applied Statistics 7*(1), 295–318. 35

Loh, W. W., T. S. Richardson, and J. M. Robins (2017). An apparent paradox explained. *Statistical Science 32*(3), 356–361. 36

Manski, C. F. (1990). Nonparametric bounds on treatment effects. *The American Economic Review 80*(2), 319–323. 41

Middleton, J. A. and P. M. Aronow (2015). Unbiased estimation of the average treatment effect in cluster-randomized experiments. *Statistics, Politics and Policy 6*(1-2), 39–75. 7, 9

Neyman, J. (1923). Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes. *Roczniki Nauk Rolniczych 10*, 1–51. 14, 17

Neyman, J. and E. S. Pearson (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London 231*(694–706), 289–337. 20, 21

Pearl, J. (2015). Generalizing experimental findings. *Journal of Causal Inference 3*(2), 259–266. 48

Robins, J. M., M. A. Hernán, and B. Brumback (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology 11*(5), 550–560. 42

Rosenbaum, P. R. (1996). Identification of causal effects using instrumental variables: Comment. *Journal of the American Statistical Association 91*(434), 465–468. 40

Rosenbaum, P. R. (1999). Reduced sensitivity to hidden bias at upper quantiles in observational studies with dilated treatment effects. *Biometrics 55*(2), 560–564. 26

Rosenbaum, P. R. (2002). *Observational Studies* (Second ed.). New York, NY: Springer. 20, 21, 26, 28, 43, 44

Rosenbaum, P. R. (2007). Interference between units in randomized experiments. *Journal of the American Statistical Association 102*(477), 191–200. 48

Rosenbaum, P. R. (2010). *Design of Observational Studies*. New York, NY: Springer. 19, 20, 21, 22, 26, 30

Rosenbaum, P. R. (2018). Sensitivity analysis for stratified comparisons in an observational study of the effect of smoking on homocysteine levels. *Annals of Applied Statistics 12*(4), 2312–2334. 45

Rosenbaum, P. R. and A. M. Krieger (1990). Sensitivity of two-sample permutation inferences in observational studies. *Journal of the American Statistical Association 85*(410), 493–498. 45

Rubin, D. B. (1980). Comment on 'randomization analysis of experimental data in the fisher randomization test' by basu, d. *Journal of the American Statistical Association 75*(371), 591–593. 1, 48

Rubin, D. B. (1986). Which ifs have causal answers? (comment on 'statistics and causal inference' by paul w. holland). *Journal of the American Statistical Association 81*, 961–962. 1, 48

Sävje, F., M. J. Higgins, and J. S. Sekhon (2017). Generalized full matching. Working Paper. 42

Senn, S. (2004). Controversies concerning randomization and additivity in clinical trials. *Statistics in Medicine 23*(24), 3729–3753. 8

Stuart, E. A., C. P. Bradshaw, and P. J. Leaf (2015). Assessing the generalizability of randomized trial results to target populations. *Prevention Science 16*(3), 475–485. 48

Wald, A. (1940). The fitting of straight lines if both variables are subject to error. *The Annals of Mathematical Statistics 11*(3), 284–300. 38

Wu, J. and P. Ding (2018, October). Randomization tests for weak null hypotheses. In *eprint arXiv:1809.07419v2*. 35

Zubizarreta, J. R. (2012). Using mixed integer programming for matching in an observational study of kidney failure after surgery. *Journal of the American Statistical Association 107*(500), 1360–1371. 42