A Primer on Power and Sample Size Calculations for Randomization Inference with Experimental Data[†]

Brandon Hauser
Department of Statistics
LMU Munich
hauserbd@alumni.wfu.edu

Mauricio Olivares
Department of Statistics
LMU Munich
m.olivares@lmu.de

July 13, 2025

Abstract

This paper revisits the problem of power analysis and sample size calculations in randomized experiments, with a focus on settings where inference on average treatment effects is conducted using randomization tests. While standard formulas based on the two-sample *t*-test are widely used in practice, we show that these calculations may yield misleading results when directly applied to randomization-based inference—unless certain assumptions are met.

We demonstrate that differences in potential outcome variances or unequal group sizes can distort the behavior of the randomization test, leading to incorrect power and flawed sample size calculations. However, a simple adjustment—studentizing the test statistic—restores the validity of the randomization test in large samples. This adjustment allows researchers to safely apply standard power and sample size formulas, even when using randomization inference.

We extend these results to a range of experimental designs commonly used in applied economics, including stratified randomization, matched pairs, and cluster-randomized trials. Throughout, we provide practical guidance to help researchers ensure that their design-stage calculations remain valid under the inferential methods they plan to use.

Keywords: power analysis, sample size calculations, randomization test, randomized controlled trial.

[†] The authors gratefully acknowledge financial support from the European Research Council (Starting Grant No. 852332).

1 Introduction

Power analysis and sample size calculations are essential components of the design of randomized controlled trials (RCTs). These calculations help ensure that a study is capable of detecting meaningful treatment effects with high probability, if such effects exist. In applied economics and related fields, it is common practice to base these calculations on standard formulas derived assuming some canonical version of the model, usually under normality (e.g. Lachin, 1981).

Although this approach is widely used, it is less clear whether these formulas are applicable when inference is conducted via randomization tests. In particular, practitioners may follow conventional power formulas while planning to use randomization inference, presuming the two approaches are interchangeable. This paper shows that they are not—at least not without additional care.

This paper aims to address the challenges of power analysis and sample size calculations in randomized experiments, where the goal is to make inference on the average treatment effect (ATE) using randomization tests. In this context, the randomization test constructs a reference distribution by shuffling the labels of the experimental units and then recalculating the difference-in-means across permutations of the data. Then, we use the quantiles of the reference distribution—the so-called randomization distribution—to perform statistical inference and power analysis.

Randomization tests have gained popularity in economics, in part because they provide exact control of the Type I error rate for any test statistic whenever experimental units are exchangeable under the null hypothesis (e.g., Lehmann and Romano, 2022, Chapter 17). For instance, Chung and Romano (2016) and Young (2019) document their widespread use in experimental economics; see also Ritzwoller, Romano, and Shaikh (2025) for a review.

Despite their popularity, a major obstacle for power analyses and sample size calculations for randomization-based inference is that the reference distribution itself—and thus its quantiles—is random. Due to its stochastic nature, the exact behavior of the randomization distribution and its quantiles might be intractable in practice, so we resort to approximations as the sample—drawn from a hypothetical superpopulation—grows large. While the superpopulation paradigm may appear restrictive at first glance, this is also the common approach when analyzing the power function of the two-sample t-test without imposing a parametric model (e.g. Van der Vaart, 2000).[1]

[1] One could also study the power of the randomization tests from a non-asymptotic perspective, e.g., Albert (2019), Kim, Balakrishnan, and Wasserman (2022). These developments, however, are expressed in terms of constant factors

The first result in this paper argues that when inference is conducted via a randomization test—rather than the usual two-sample t-test—, traditional power formulas may no longer hold, even asymptotically. In such cases, reliance on standard calculations derived for the two-sample t-test can lead to incorrect conclusions unless we impose additional assumptions, namely equality of variances between experimental groups, or that the experimental groups are of the same size.

This paper provides a systematic examination of these issues. Beginning with the analysis of completely randomized experiments under the potential outcomes framework, we illustrate where and how standard approaches to power analysis break down when applied in the context of randomization inference. This discrepancy arises because the randomization distribution based on the difference-in-means statistic does not mimic the sampling distribution under the null hypothesis. The resulting mismatch can lead to tests that fail to control size and exhibit distorted power, which in turn casts doubt on sample size calculations and the interpretation of minimum detectable effects.

We recommend a simple remedy: to studentize the test statistic by an appropriate estimate of the (asymptotic) variance. We show that after proper studentization, the randomization distribution of the modified statistic asymptotically aligns with the sampling distribution of the studentized statistic under the null. As a result, standard formulas for power and sample size can then be used safely, even in conjunction with randomization tests.

These insights extend beyond completely randomized designs. We show that similar concerns—and solutions—apply under covariate adaptive-randomization, matched pairs, and cluster-randomized experiments. In each case, we examine the asymptotic behavior of both the test statistic and its randomization distribution, showing how appropriate studentization restores the validity of power analyses.

Randomization tests are often conceived and motivated from a finite-sample perspective, where the random assignment is the "reasoned basis for inference" (e.g., Fisher, 1935). Thus, we discuss the distinction between superpopulation and finite-population frameworks, and their implications for power analysis using randomization inference. Finally, we highlight extensions to other target parameters and designs where theory remains underdeveloped, offering guidance for applied researchers interested in randomization-based inference.

that are often difficult to pinpoint without more assumptions, thus limiting their applicability for sample size calculations.

1.1 Overview of the Methods

In this paper, we briefly outline the main formulas and methods used to conduct power analyses based on randomization inference. As we argue, analyzing the statistical power of randomization tests in finite samples is inherently challenging. Accordingly, the formulas we present are based on large-sample approximations of the true power functions.

We distinguish between two types of analyses: *ex ante* and *ex post*. An ex ante analysis takes place at the design stage, prior to the implementation of the experiment and the collection of data. In this setting, the desired power level is typically specified by the researcher, and the objective is to determine the sample size needed to achieve that level of power.

By contrast, an ex post analysis is conducted after data collection—once the experiment has been implemented and the sample is fixed. The aim in this case is to assess the power of the different inference procedures. For instance, quantifying the power of the two-sample t-test and the randomization test based on the sample mean difference. While ex post analyses occur at the analysis stage, we emphasize that the researcher must specify such comparisons beforehand to avoid concerns about selective inference.

The methods we cover in this paper are useful for both ex ante and ex post analyses. As we will see, the formulas we present depend on user-chosen constants—such as the desired power, size of the test, or the effect we seek to detect—as well as unknown parameters, such as the variances of experimental groups, that must often be estimated using pilot or historical data.

NOTATION Throughout the paper, we maintain the following conventions. For a generic random variable indexed by i, W_i, $W^{(\mathrm{N})}$ stands for $(W_1, \ldots, W_{\mathrm{N}})$ and $\bar{W}$ is the sample average $\sum_{i=1}^{N} W_i/N$. The asymptotic results are understood as $N \to \infty$, unless otherwise specified. The indicator function is denoted $\mathbb{1}\{\cdot\}$. The expectation and variance are denoted $\mathbb{E}[\cdot]$ and $\mathbb{V}[\cdot]$, respectively.

2 Power Analysis in Completely Randomized Experiments

This section establishes the standard power analysis for a completely randomized experiment using a two-sample t-test. Our goal is to understand how the power of the test behaves in large samples and how it informs sample size decisions. For simplicity, we introduce the main ideas without covariates. We postpone the discussion on stratified randomization to Section 5.

2.1 Setup

Following the causal inference literature, we adopt the potential outcomes framework without interference. For each unit $i = 1, \ldots, N$, denote by $Y_i(1)$ the potential outcome under treatment, by $Y_i(0)$ the potential outcome under control, and by D_i the binary treatment status of the ith unit, where $D_i = 1$ indicates unit i receives treatment, $D_i = 0$ otherwise. For each unit, the observed data is $W_i = (Y_i, D_i)$, where the outcome Y_i obeys the relationship

$$Y_i = Y_i(1)D_i + Y_i(1 - D_i) \ ,$$

so that $Y_i(1) = Y_i$ among the treated, and $Y_i(0) = Y_i$ among the non-treated. Suppose that treatment status is randomly assigned so that the treatment is statistically independent of each potential outcome, denoted as $(Y^{\text{N}}(1), Y^{\text{N}}(0)) \perp D^{(\text{N})}$. In this section, we consider what is known as complete randomization. In simple terms, complete randomization states that $0 < m < N$ units receive treatment, and the remaining $n = N - m$ units receive no treatment. Formally, $D^{(\text{N})}$ is uniformly distributed over vectors $d^{(\text{N})} = (d_1, \ldots, d_{\text{N}})$ in which each $d_i \in \{0, 1\}$ and $\sum_{i=1}^{N} d_i = m$ for some known $0 < m < N$.

Throughout this paper the target parameter is the ATE, defined as $\Delta = \mathbb{E}[Y_i(1) - Y_i(0)]$. Notice that under random assignment, the ATE is identified by the mean difference between experimental groups:

$$\Delta = \mathbb{E}[Y|D = 1] - \mathbb{E}[Y|D = 0]$$

so that we can learn it from our data.

2.2 Statistical Inference

Our goal is to make inference about the ATE. Suppose we seek to test

$$H_0 \ : \ \mathbb{E}[Y_i(1) - Y_i(0)] = 0 \quad \text{vs.} \quad H_1 \ : \ \mathbb{E}[Y_i(1) - Y_i(0)] > 0 \tag{1}$$

at level $\alpha \in (0, 1)$ based on random sample $W^{(\text{N})} = \{(Y_i, D_i) \ : i = 1 \ldots, N\}$ obtained from an RCT under complete randomization. To test (1), consider a test statistic T_{N} such that "large" values provide evidence against the null hypothesis. Typically, such a test statistic is based on the sample-mean difference. Let $T_{\mathrm{N}} := T_{\mathrm{N}}(W^{(\mathrm{N})})$ be given by

$$T_{\mathrm{N}} = \sqrt{m}\left(\frac{1}{m}\sum_{i=1}^{N} Y_i D_i - \frac{1}{n}\sum_{i=1}^{N} Y_i(1-D_i)\right) . \tag{2}$$

Under mild regularity conditions, we can approximate the sampling distribution of (2) under the null hypothesis. Specifically, we can show that under the null hypothesis, the distribution of T_{N} is approximately $\mathcal{N}(0,\sigma^2)$ as the sample size grows large, where

$$\sigma^2 = \mathbb{V}[Y(1)] + \lambda\,\mathbb{V}[Y(0)] \ ,$$

and $\lambda = \lim(m/n)$ as $\min\{m,n\} \to \infty$, for some $\lambda > 0$. To gain further intuition on λ, let us consider the proportion of units that receive treatment, m/N. Observe that $m/N = (1+n/m)^{-1}$, so as the sample size increases, m/N approaches $\lambda/(1+\lambda)$. Thus, we can think of $\lambda/(1+\lambda)$, loosely speaking, as the unconditional probability of being treated.

The previous approximation gives rise to the two-sample t-test

$$\phi_{\mathrm{N}}^{\text{t-test}}\left(W^{(\mathrm{N})}\right) = \mathbb{1}\{T_{\mathrm{N}} > \sigma z_{1-\alpha}\} \ , \tag{3}$$

where $z_{1-\alpha}$ denotes the $1-\alpha$ quantile of a standard normal random variable. A few remarks are in order. First, we note that (3) is an asymptotic test in the sense that it is based on the normal *approximation* to the sampling distribution of T_{N} under the null hypothesis. Thus, we can perform inference without making parametric assumptions about the distribution governing the data. Secondly, $\phi_{\mathrm{N}}^{\text{t-test}}$ is (asymptotically) level α: the two-sample t-test controls the probability of a type-I error in large samples. Lastly, it can be shown that the two-sample t-test has power tending to one as the sample size increases, i.e., it is a consistent test against any fixed alternative $H_1 : \Delta > 0$.

Remark 1. (*Two-sided alternatives*). We focus on so-called one-sided alternatives as in (1) because it simplifies sample size calculations as opposed to, say, two-sided alternatives,

$$H_0 \ : \ \mathbb{E}[Y_i(1) - Y_i(0)] = 0 \quad \text{vs.} \quad H_1 \ : \ \mathbb{E}[Y_i(1) - Y_i(0)] \neq 0 \ . \tag{4}$$

However, we can deduce the power of a α-level test for (4) by that of a $\alpha/2$-level one-sided test; see Remark 3 for more details. ■

2.3 Local Asymptotic Power Analysis

Because the power function of a consistent test converges to one under fixed alternatives $H_1 : \Delta > 0$, we follow the standard approach in asymptotic analysis by considering *local alternatives* that shrink to H_0 with the sample size. This gives us a non-trivial and interpretable approximation to power. Specifically, we instead consider *sequences* of alternative hypotheses Δ_{N} approaching $\Delta = 0$ of the form $\Delta_{\text{N}} = h/\sqrt{N}$, for some constant $h > 0$. Under this paradigm, we can show that the so-called *local asymptotic power* function of the two-sample t-test with local parameter h is given by

$$1 - \Phi\left(z_{1-\alpha} - \sqrt{\frac{\lambda}{1+\lambda}} \cdot \frac{h}{\sigma}\right) , \tag{5}$$

where $\Phi(\cdot)$ is the cumulative distribution function (CDF) of a standard normal. To ease exposition, define a rescaled version of σ^2, denoted $\tilde{\sigma}^2$, by

$$\tilde{\sigma}^2 = \frac{1+\lambda}{\lambda} \cdot \sigma^2 = \frac{1+\lambda}{\lambda} \cdot \mathbb{V}[Y(1)] + (1+\lambda) \cdot \mathbb{V}[Y(0)] \ .$$

Then, we can rewrite the above power function, Eq. (5), as

$$1 - \Phi\left(z_{1-\alpha} - \frac{h}{\tilde{\sigma}}\right) \ . \tag{6}$$

In practice, we are only interested in the power function at a single Δ, not a sequence of alternatives Δ_{N}. To circumvent this, we usually fix N and a Δ and solve for the local parameter $h = \sqrt{N}\,\Delta$. Then, by plugging such a value in (6) we can approximate the power of the two-sample t-test at Δ of a completely randomized experiment. The following expression captures the approximate power under a local alternative

$$1 - \Phi\left(z_{1-\alpha} - \frac{\sqrt{N}\,\Delta}{\tilde{\sigma}}\right) \ . \tag{7}$$

Equation (7) reveals how power increases with the sample size N, the effect size Δ, and decreases with potential outcome variances $\mathbb{V}[Y(1)]$ and $\mathbb{V}[Y(0)]$. In other words, detecting smaller effects requires larger samples or less noisy potential outcomes.

Operationally, the unknown variances $\mathbb{V}[Y(1)]$ and $\mathbb{V}[Y(0)]$ are replaced by their corresponding estimators. For instance, $\mathbb{V}[Y(1)]$ and $\mathbb{V}[Y(0)]$ could be estimated, respectively, by

$$\frac{1}{m}\sum_{i=1}^{N} D_i(Y_i - \bar{Y}_1)^2 \quad \text{and} \quad \frac{1}{n}\sum_{i=1}^{N}(1-D_i)(Y_i - \bar{Y}_0)^2 \,, \tag{8}$$

where $\bar{Y}_1$ and $\bar{Y}_0$ represent the corresponding sample means of the treatment and control group. Therefore, in practice, all the unknown quantities are either set by the researcher *ex ante* or estimated *ex post*, so the power function (7) can be computed.

Remark 2. (*Minimum Detectable Effect*) As a byproduct of (7), we may also obtain an expression for the smallest ATE, Δ, that can be detected with pre-determined power, say $1-\beta$ for some $\beta \in (0,1)$, at significance level $\alpha \in (0,1)$:

$$\Delta = (z_{1-\alpha} + z_{1-\beta})\,\frac{\tilde{\sigma}}{\sqrt{N}} \,.$$

This is often referred to as the minimum detectable effect (MDE) in the literature. ■

Remark 3. (*Two-sided alternatives*, continued). Recall that T_{N} is approximately normally distributed in large samples. Then, symmetry of the normal distribution allows us to write the (local asymptotic) power of the two-sided test that rejects when $|T_{\mathrm{N}}| > \sigma z_{1-\alpha/2}$ as

$$1 - \Phi\left(z_{1-\alpha/2} - \frac{\Delta}{\tilde{\sigma}}\right) + \Phi\left(z_{\alpha/2} - \frac{\Delta}{\tilde{\sigma}}\right) \,.$$

■

Remark 4. Conventional power analyses recommend that the sizes of the treated and control groups should be proportional to their respective variances (Neyman, 1992). To see why, recall that $\lambda \approx m/n$. Thus, minimizing $\tilde{\sigma}^2$ with respect to λ gives the optimal choice, say λ^*:

$$\lambda^* = \sqrt{\frac{\mathbb{V}[Y(1)]}{\mathbb{V}[Y(0)]}} \,. \tag{9}$$

Then, it follows from (6) that the power is negatively affected whenever λ deviates from (9). ■

2.4 Sample Size Calculations

One of the key challenges in planning (*ex ante*) an RCT is to determine the sample size N required to tackle the testing problem (1) at level $\alpha \in (0,1)$ and desired power, say $1-\beta$ for some $\beta \in (0,1)$

(Lachin, 1981). Building on the power analysis from previous section, e.g., Eq. (7), we now establish the relationship between the sample size N and a pre-determined power $1-\beta$ for some $\beta \in (0,1)$ when we use the two-sample t-test to conduct inference for the ATE in a completely randomized experiment. Specifically, we find the unique integer N that solves the equation (7). This gives us the required sample size to achieve a given power

$$N = (z_{1-\alpha} + z_{1-\beta})^2 \, \frac{\tilde{\sigma}^2}{\Delta^2} \ , \tag{10}$$

where the only unknown is N after α, β, and λ are chosen by the researcher, and $\mathbb{V}[Y(1)]$ and $\mathbb{V}[Y(0)]$ are replaced by their corresponding estimators, say (8). Observe that the sample size in the previous display is (a) increasing in power $(1-\beta)$, (b) decreasing in the ATE (Δ), and (c) increasing in the potential outcome variances $\mathbb{V}[Y(1)]$ and $\mathbb{V}[Y(0)]$.

At first glance, the preceding calculations based on Eq. (10) may appear counterintuitive. The reason is twofold. First, it assumes that the variances $\mathbb{V}[Y(1)]$ and $\mathbb{V}[Y(0)]$ are known or can be readily estimated. However, sample size calculations are typically conducted *ex ante*, that is, prior to data collection and analysis. As a result, we must rely on auxiliary sources of information—such as historical data from the same or a comparable population, or preliminary results from a pilot study; see Glennerster and Takavarasha (2013, Chp. 6) or Duflo, Glennerster, and Kremer (2007, Sec. 4.6).

Second, the formulas derived in the preceding sections are based on asymptotic approximations as the sample size tends to *infinity*, which complicates the task of determining a *finite* total sample size N. However, these approximations are used not because we suppose we have an infinite amount of data, but because they become, roughly speaking, more precise as the sample size increases. Therefore, in a large-sample environment like ours, it is more appropriate to think of sample size calculations as answering a question of the form: *what is the sample size N such that a test for* (1) *based on the normal approximation gives us, say,* $(1-\beta) \times 100\%$ *power at* $\alpha \times 100\%$ *significance level?*

While the formulas in this section are widely used in practice, it is less clear whether they are applicable when inference is conducted using randomization tests. In the next section, we examine when we can safely apply these formulas in conjunction with randomization tests.

3 Randomization Test for the ATE under Complete Randomization

In this section, we shift the focus to randomization inference. Randomization inference has several key advantages. First, it is a nonparametric testing procedure, so we do not need to assume a parametric model for the data, such as normality. Second, while it is often motivated due to its finite-sample guarantees, randomization inference also offers a way to carry on robust inference in general settings when the sample size grows large. Lastly, in many interesting problems like the ones we consider here, the randomization test is as powerful as the test based on the normal approximation.

We begin with the basic construction under complete randomization. For the sake of exposition, we begin by casting our testing problem (1) as a two-sample testing problem of equality of means. To this end, we note that $Y_i(1) = Y_i$ among the treated, and $Y_i(0) = Y_i$ among the non-treated. Thus, we may think of the m potential outcomes under treatment $(Y_1(1), \ldots, Y_m(1))$ as a random sample from the distribution of $Y(1)$ with CDF $F_1(\cdot)$, so that we are implicitly assuming that we are sampling from a hypothetical superpopulation; see Section 6.1 below for further discussion. Analogously, $(Y_1(0), \ldots, Y_n(1))$ is a random sample from the distribution of $Y(0)$ with CDF $F_0(\cdot)$. Notice that the two samples are independent by virtue of random assignment. Then, (1) reduces to testing

$$H_0 : \mathbb{E}_{F_1}[Y(1)] = \mathbb{E}_{F_0}[Y(0)] \quad \text{vs.} \quad H_1 : \mathbb{E}_{F_1}[Y(1)] > \mathbb{E}_{F_0}[Y(0)]\ .$$

For the most part, our discussion and theoretical results in this section follow the exposition in Chapter 17 in Lehmann and Romano (2022). Ritzwoller, Romano, and Shaikh (2025) provide an excellent review of randomization tests and more recent developments. We do not include the proofs and instead focus on the intuition behind the main results. We also recommend the interested reader to consult Chung and Romano (2013) for a more in-depth exposition. Implementation of the randomization tests in this Section can be done using the `R` package `RATest`, available on CRAN.

3.1 Construction of a Randomization Test

Intuitively, the randomization test seeks to test (1) by constructing an auxiliary resampling distribution that approximates the sampling distribution of the test statistic T_{N}, so that we can use the quantiles of the auxiliary distribution as data-driven critical values. In our context, this resampling scheme is done by permuting the labels of the observations in both experimental groups, and then recomputing T_{N}

for each reshuffling of the data. Under some conditions, this auxiliary resampling distribution serves as a reference from which we obtain critical values to make valid inference and power calculations, at least in large samples.

We formalize the ongoing discussion. We begin by combining the data from the two samples as

$$Z^{(\mathrm{N})} = (Z_1, \ldots, Z_{\mathrm{N}}) = (Y_1(1), \ldots, Y_m(1), Y_1(0), \ldots, Y_n(0)) \ .$$

Let $\pi = (\pi(1), \ldots, \pi(N))$ be a permutation of indices $\{1, \ldots, N\}$, and $\boldsymbol{G}_{\mathrm{N}}$ denote the collection of all $N!$ permutations π of indices $\{1, \ldots, N\}$. For any $\pi \in \boldsymbol{G}_{\mathrm{N}}$, denote the permuted data as $Z_\pi^{(\mathrm{N})} = (Z_{\pi(1)}, \ldots, Z_{\pi(\mathrm{N})})$, and the permuted test statistic by

$$T_{\mathrm{N}}^{\pi} := T_{\mathrm{N}}\left(Z_\pi^{(\mathrm{N})}\right) = \sqrt{m}\left(\frac{1}{m}\sum_{i=1}^{m} Z_{\pi(i)} - \frac{1}{n}\sum_{j=1}^{n} Z_{\pi(m+j)}\right) \ .$$

The next steps illustrate the construction of a randomization test for (1) at fixed nominal level α:

Step 1 Given $Z^{(\mathrm{N})} = z^{(\mathrm{N})}$, recalculate the test statistic for all permuted samples. This process yields $N!$ recomputed test statistics as π varies in $\boldsymbol{G}_{\mathrm{N}}$. Collect them into $\{T_{\mathrm{N}}^{\pi} : \pi \in \boldsymbol{G}_{\mathrm{N}}\}$.

Step 2 Order the $\{T_{\mathrm{N}}^{\pi} : \pi \in \boldsymbol{G}_{\mathrm{N}}\}$ values obtained in Step 1 from smallest to largest, say

$$T_{\mathrm{N}}^{(1)}(z^{(\mathrm{N})}) \leq T_{\mathrm{N}}^{(2)}(z^{(\mathrm{N})}) \leq \cdots \leq T_{\mathrm{N}}^{(N!)}(z^{(\mathrm{N})}) \ .$$

Step 3 Let $k = N! - \lfloor \alpha N! \rfloor$, where $\lfloor \alpha N! \rfloor$ is the largest integer less than or equal to $\alpha N!$. Locate the corresponding k-th value among the ordered statistics:

$$T_{\mathrm{N}}^{(1)}(z^{(\mathrm{N})}) \leq \cdots \leq T_{\mathrm{N}}^{(k)}(z^{(\mathrm{N})}) \leq \cdots \leq T_{\mathrm{N}}^{(N!)}(z^{(\mathrm{N})}) \ ,$$

and from it, define

$$\begin{aligned}
N^{+}(z^{(\mathrm{N})}) &= \text{the number of } T_{\mathrm{N}}^{(j)}(z^{(\mathrm{N})}) \text{ that are greater than } T_{\mathrm{N}}^{(k)}(z^{(\mathrm{N})}) \ , \\
N^{0}(z^{(\mathrm{N})}) &= \text{the number of } T_{\mathrm{N}}^{(j)}(z^{(\mathrm{N})}) \text{ that are equal to } T_{\mathrm{N}}^{(k)}(z^{(\mathrm{N})}) \ , \\
a(z^{(\mathrm{N})}) &= \frac{\alpha N! - N^{+}(z^{(\mathrm{N})})}{N^{0}(z^{(\mathrm{N})})} \ .
\end{aligned}$$

Step 4 The randomization test is given by

$$\phi_{\mathrm{N}}^{\mathrm{rand}}(Z^{(\mathrm{N})}) = \mathbb{1}\{T_{\mathrm{N}} > T_{\mathrm{N}}^{(k)}\} + a(Z^{(\mathrm{N})})\, \mathbb{1}\{T_{\mathrm{N}} = T_{\mathrm{N}}^{(k)}\} \ . \tag{11}$$

Notice that $T_{\mathrm{N}}^{(k)}$—the kth-largest test statistic among the $\{T_{\mathrm{N}}^{j} \, : \, j = 1, \ldots, N!\}$—acts as a critical value using the quantiles of an auxiliary sampling distribution, namely the empirical distribution of the re-computed test statistics. More specifically, let

$$\hat{R}_{\mathrm{N}}^{T}(t) = \frac{1}{N!} \sum_{\pi \in \boldsymbol{G}_{\mathrm{N}}} \mathbb{1}\left\{T_{\mathrm{N}}^{\pi} \leq t\right\} \ . \tag{12}$$

Then $T_{\mathrm{N}}^{(k)}$ is the upper-α quantile of (12), denoted $\hat{r}_{\mathrm{N}}(1-\alpha)$ and given by

$$\hat{r}_{\mathrm{N}}(1-\alpha) = \inf\{t \, : \, \hat{R}_{\mathrm{N}}^{T}(t) \geq 1-\alpha\} \ . \tag{13}$$

Eq. (12) is the so-called randomization distribution of T_{N}, and $\hat{r}_{\mathrm{N}}(1-\alpha)$ is the data-driven critical value using the quantiles of the randomization distribution (12). From here, we see that the randomization test (11) rejects H_0 if the test statistic is bigger than $\hat{r}_{\mathrm{N}}(1-\alpha)$, fails to reject if $T_{\mathrm{N}} < \hat{r}_{\mathrm{N}}(1-\alpha)$, and otherwise randomizes the decision with success probability $a(Z^{(\mathrm{N})})$ whenever $T_{\mathrm{N}} = \hat{r}_{\mathrm{N}}(1-\alpha)$, a situation that may arise due to possible ties when permuting the data.

Remark 5. In most cases of empirical relevance, the sample size N is such that it is computationally prohibitive to re-calculate the test statistic for all $N!$ permutations of indices $\{1, \ldots, N\}$. For this reason, in practice, we resort to a stochastic approximation: rather than considering all permutations, set π_1 =identity permutation, and draw $\pi_2, \ldots, \pi_B$ permutations uniformly at random from $\boldsymbol{G}_{\mathrm{N}}$. For example, randomly permute the data 999 times and recalculate the test statistics each time, i.e., $B = 1000$ times in total, as we also calculate the test statistic using the original data (corresponding to π_1). Indeed, $\hat{R}_{\mathrm{N}}^{T}(t)$ can be approximated to any desired degree of accuracy when B is large without compromising the statistical properties of the randomization test (e.g. Lehmann and Romano, 2022, Chp. 17). Thus, for the rest of this paper, we write things in terms of the $N!$ permutations, though it is understood we calculate $\hat{R}_{\mathrm{N}}^{T}(t)$ and its critical values based on $\pi_1, \ldots, \pi_B$ permutations (e.g., B=1000). See Ramdas et al. (2023) for alternative approaches based on subsets of $\boldsymbol{G}_{\mathrm{N}}$. ■

Remark 6. (*Finite-Sample Exactness*). When the so-called "randomization hypothesis" holds (e.g., Lehmann and Romano, 2022, Def. 17.2.1), the randomization test yields an exact level α test for a fixed sample size. For instance, suppose that instead of (1), we were interested in testing the null hypothesis of "no treatment effect" $H_0 : Y(1) \overset{\mathrm{d}}{=} Y(0)$, where $\overset{\mathrm{d}}{=}$ denotes equality of distribution. In this particular situation, the randomization test attains exact finite-sample Type I error control. While this finite-sample exactness is certainly appealing, the null hypothesis $H_0 : Y(1) \overset{\mathrm{d}}{=} Y(0)$ is stronger than the actual hypothesis of interest (1). That is, equality in distribution implies equality of means, but not the other way around; see Chung (2017) for further discussion. ■

3.2 Asymptotic Behavior of the Randomization Distribution

The key challenge to understanding the asymptotic properties of the randomization test is that the decision to reject a null hypothesis depends on $\hat{r}_{\mathrm{N}}(1-\alpha)$, which is a random variable. Consider the power analysis of the previous sections. To show that a test has high power asymptotically, we typically need to show that the distribution of the test statistic is far from its critical value—for instance, 1.96—under the alternative hypothesis as the sample size increases. For the randomization test, this critical value is not a fixed number but rather a random quantity.

Despite these challenges, we can show that when the sample size grows large, the randomization distribution settles down to some nonrandom distribution, in probability (e.g., Chung and Romano, 2013, Theorem 2.1). Indeed, in our setting, we have that the randomization distribution of T_{N} behaves like the distribution of a normal random variable with mean zero and variance τ^2 given by

$$\tau^2 = \lambda\, \mathbb{V}[Y(1)] + \mathbb{V}[Y(0)] \ ,$$

and the random $\hat{r}_{\mathrm{N}}(1-\alpha)$ converges in probability to $\tau\, z_{1-\alpha}$, that is, the (random) quantile of the randomization distribution concentrates around the $(1-\alpha)$ quantile of the normal distribution with mean zero and variance τ^2, in probability.

However, this is *not* the asymptotic behavior of the test statistic T_{N} under H_0, in general. To see why, recall that we concluded in Section 2.2 that the sampling distribution of T_{N} behaves, under the null hypothesis, like the distribution of a normally distributed random variable with mean zero and variance

$$\sigma^2 = \mathbb{V}[Y(1)] + \lambda\, \mathbb{V}[Y(0)] \ ,$$

giving rise to the two-sample t-test that rejects H_0 if $T_{\mathrm{N}} > \sigma z_{1-\alpha}$. Therefore, unless $\mathbb{V}[Y(1)] = \mathbb{V}[Y(0)]$ or $m/n \to 1$ (equally-sized experimental groups), $\sigma^2 \neq \tau^2$. Then, the two distributions—and respective critical values—will not be the same, even asymptotically.

Figure 1 illustrates a scenario in which the randomization test yields conservative inference. It compares the large-sample behavior of the randomization distribution of T_{N} (yellow curve) with the corresponding sampling distribution of T_{N} under the null hypothesis (in blue) for suitable choices of τ^2 and σ^2. The vertical line marks $\hat{r}_{\mathrm{N}}(0.95)$, the 95th percentile of $\hat{R}^T_{\mathrm{N}}$. As shown, using this threshold leads to a rejection probability under the null of 0.017—substantially below the nominal level α—highlighting the test's conservativeness in this setting. However, just as this particular configuration results in underrejection, other parameter choices can lead to overrejection of the null, which is more problematic. This lack of robustness implies that randomization-based inference is not reliable, even asymptotically.

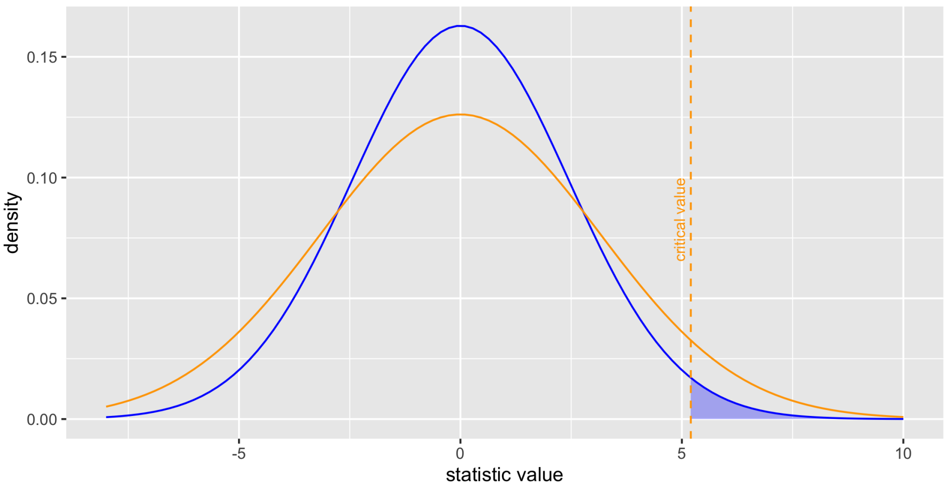

Figure 1: Asymptotic behavior of the randomization distribution of T_{N} (yellow) and the true unconditional distribution of T_{N} under the null (blue), where T_{N} is given by (2). The vertical line corresponds to $\hat{r}_{\mathrm{N}}(0.95)$, the 95th percentile of the randomization distribution of T_{N}.

3.3 Implication for Power Analysis in Completely Randomized Experiments

Unfortunately, traditional power analysis and sample size calculations may also be compromised if we blindly follow the calculations designed for the two-sample t-test but wish to perform randomization inference. As we argued in Section 3.2, this is so because the randomization distribution of T_{N} does not accurately approximate the true sampling distribution of the test statistic, leading to potentially misleading conclusions. That is, distorted conclusions about significance and potentially underpowered designs ensue if standard formulas are used.

To understand the detrimental effects on power analysis, we follow Section 2 and derive the (local asymptotic) power of the randomization test based on T_{N} in a completely randomized experiment. As before, consider sequences of alternatives of the form $\Delta_{\text{N}} = h/\sqrt{N}$ approaching $\Delta = 0$ as $N \to \infty$. We can show (e.g., Lehmann and Romano, 2022, Chapter 17.2.2) that the (local asymptotic) power function of the randomization test based on T_{N} in a completely randomized experiment is given by

$$1 - \Phi\left(\frac{\tau}{\sigma} z_{1-\alpha} - \sqrt{\frac{\lambda}{1+\lambda}} \cdot \frac{h}{\sigma}\right) . \tag{14}$$

Arguing as in Section 2.3, we approximate the power of the randomization test at Δ by

$$1 - \Phi\left(\frac{\tau}{\sigma} z_{1-\alpha} - \sqrt{\frac{\lambda}{1+\lambda}} \cdot \frac{\sqrt{N}\Delta}{\sigma}\right) = 1 - \Phi\left(\frac{\tau}{\sigma} z_{1-\alpha} - \frac{\sqrt{N}\Delta}{\tilde{\sigma}}\right) . \tag{15}$$

Observe that unless $\sigma^2 = \tau^2$, the power functions in (14) and (15) will not coincide with those of the two-sample t-test, eqs. (5) and (7), respectively. In general, $\sigma^2 \neq \tau^2$. However, there are two situations in which the power functions coincide. First, when the variances of the potential outcomes are the same across experimental groups, i.e., $\mathbb{V}[Y(1)] = \mathbb{V}[Y(0)]$. The second instance is when the experimental groups are of the same size, i.e., when m/N converges to $1/2$ as the sample size grows large, so that $\lambda = 1$. Though the latter condition is under the researcher's control, it is harder to claim $\mathbb{V}[Y(1)] = \mathbb{V}[Y(0)]$ without further assumptions; see Remark 7 below.

To illustrate the practical relevance of the ongoing discussion, suppose that $\lambda > 1$ and $\mathbb{V}[Y(1)] > \mathbb{V}[Y(0)]$ so that $\tau^2 > \sigma^2$. A direct comparison of the (local asymptotic) power functions of the two-sample t-test and randomization test based on T_{N} at $\Delta > 0$ yields

$$\underbrace{1 - \Phi\left(\frac{\tau}{\sigma} z_{1-\alpha} - \frac{\sqrt{N}\Delta}{\tilde{\sigma}}\right)}_{\text{randomization test}} < \underbrace{1 - \Phi\left(z_{1-\alpha} - \frac{\sqrt{N}\Delta}{\tilde{\sigma}}\right)}_{\text{two-sample } t\text{-test}} .$$

That is, for given sample size N, $\Delta > 0$, and $\alpha \in (0, 1)$, the randomization test exhibits a lower power relative to the two-sample t-test based on the asymptotic approximation. The detrimental effects of ignoring this phenomenon are immediate. First, the two-sample t-test will detect smaller deviations from the null hypothesis at a pre-specified power, say $1 - \beta$, than the randomization test based on T_{N}, since

$$\mathrm{MDE}^{\text{t-test}} = \frac{\tilde{\sigma}}{\sqrt{N}}\left(z_{1-\alpha} + z_{1-\beta}\right) < \frac{\tilde{\sigma}}{\sqrt{N}}\left(\frac{\tau}{\sigma}\, z_{1-\alpha} + z_{1-\beta}\right) = \mathrm{MDE}^{\text{rand}}$$

Second, sample size calculations in a completely randomized experiment are distorted too:

$$\underbrace{\left(z_{1-\alpha} + z_{1-\beta}\right)^2 \frac{\tilde{\sigma}^2}{\Delta^2}}_{N\text{: two-sample t-test}} < \underbrace{\left(\frac{\tau}{\sigma} z_{1-\alpha} + z_{1-\beta}\right)^2 \frac{\tilde{\sigma}^2}{\Delta^2}}_{N\text{: randomization test}} .$$

Therefore, naively using the formula for the two-sample t-test when inference is to be carried out by a randomization test based on T_{N} delivers incorrect sample size estimates. The effect of such a miscalculation will depend on the variances of the potential outcomes, the proportion of treated units, and the effect we seek to detect. For instance, suppose that $\lambda = 1/2$, $\mathbb{V}[Y(1)] = 0.7$, $\mathbb{V}[Y(0)] = 1.1$, and $\Delta = 0.5$. Then, at conventional values $\alpha = 0.05$ and $1 - \beta = 0.8$, eq. (10) dictates a sample size $N = 93$. Meanwhile, the sample size using the appropriate formula would be $N = 103$, approximately a 10% increase relative to the naive formula designed for the two-sample t-test.

Remark 7. (*Heterogeneous Treatment Effects*) Although it is generally difficult to justify the assumption $\mathbb{V}[Y(1)] = \mathbb{V}[Y(0)]$ without additional structure, one setting where this equality holds is under the absence of treatment effect heterogeneity. Formally, we say the treatment effect is homogeneous if $Y_i(1) - Y_i(0) = \delta$ for some (unknown) constant δ; otherwise, the treatment effect is heterogeneous, meaning it varies across individuals. Under treatment effect homogeneity, it follows that $F_1(y + \delta) = F_0(y)$, i.e., the CDF of treatment and control groups differ only by a constant shift δ, where δ coincides with the ATE. In this case, the treatment affects only the *location* of the distributions of $Y(1)$ and $Y(0)$—not their shape—, implying $\mathbb{V}[Y(1)] = \mathbb{V}[Y(0)]$. While the constant treatment effect assumption is inherently untestable—we never observe $Y_i(1)$ and $Y_i(0)$ for the same individual—it is possible to test the implication $F_1(y + \delta) = F_0(y)$; see for example Ding, Feller, and Miratrix (2016) or Chung and Olivares (2021). ■

Remark 8. It is often the case in randomized experiments with binary treatments to set $\mathrm{P}(D_i = 1) = 1/2$ so as to ensure treatment and control groups are of the same size. In this case, we would have in the limit that $\lambda = 1$ and so $\tau^2 = \sigma^2$. Thus, the power analysis based on the formulas for the two-sample t-test is applicable for randomization tests, though perhaps inadvertently by practitioners. Nevertheless, it is not unusual to find situations where the proportion of treated units differs from 1/2, e.g. Karlan and List (2007), Bloom et al. (2013), Banerjee et al. (2015), Bloom et al. (2020). ■

4 Solutions in Completely Randomized Experiments

The reason why traditional power analysis breaks when making inference for the ATE in a completely randomized experiment is straightforward: the randomization distribution and the sampling distribution of the test statistic under H_0 do not agree due to the mismatch in variances. This suggests a solution. Since σ^2 and τ^2 do not agree unless we impose additional restrictions, we could studentize the test statistic T_{N} to ensure that its asymptotic variance does not depend on the variances of the potential outcomes or the proportion of treated units at all. It turns out, when this studentization is done properly, the randomization distribution of the studentized test statistic settles around the true unconditional distribution of the test statistic. This way, we can restore valid inference and power analysis, at least asymptotically.

To set the stage, consider the studentized test statistic, $S_{\mathrm{N}} := S_{\mathrm{N}}(Z^{(\mathrm{N})})$, given by

$$S_{\mathrm{N}} = \frac{T_{\mathrm{N}}}{\sqrt{\hat{\sigma}_1^2 + \frac{m}{n}\hat{\sigma}_0^2}} \ , \tag{16}$$

where $\hat{\sigma}_1^2 := \hat{\sigma}_1^2(Z_1, \ldots, Z_m)$ and $\hat{\sigma}_0^2 := \hat{\sigma}_0^2(Z_{m+1}, \ldots, Z_{\mathrm{N}})$ are the sample variances

$$\hat{\sigma}_1^2 = \frac{1}{m}\sum_{i=1}^{m}\left(Z_i - \bar{Z}_m\right)^2 \quad \text{and} \quad \hat{\sigma}_0^2 = \frac{1}{n}\sum_{j=1}^{n}\left(Z_{m+i} - \bar{Z}_n\right)^2 \ ,$$

and $\bar{Z}_m$ and $\bar{Z}_n$ are the sample means of treatment and control group, respectively. For completely randomized experiments, we can show that S_{N} converges in distribution under the null hypothesis (1) to the distribution of a standard normal random variable. That is, its asymptotic distribution does not depend on additional parameters, such as the variances or λ.

The previous asymptotic result gives rise to the *studentized* two-sample t-test based on the normal approximation. Thus, in the same spirit as in $\phi_{\mathrm{N}}^{\text{t-test}}$ in (3), the *studentized* two-sample t-test rejects H_0 if $S_{\mathrm{N}} > z_{1-\alpha}$, and fails to reject otherwise. Indeed, we can view the *studentized* two-sample t-test as the usual t-test on the slope parameter associated with the treatment indicator in a linear regression of outcome Y on a constant and D, with heteroskedasticity-robust standard errors.

As before, we will define the randomization test using the quantiles of a reference distribution, the so-called randomization distribution of S_{N}, denoted $\hat{R}_{\mathrm{N}}^{S}(\cdot)$. To this end, we proceed in the same manner as in Section 3.1, except we now replace T_{N} by S_{N}, i.e., we re-calculate S_{N}, say S_{N}^{π}, for each permutation of labels $\pi \in \boldsymbol{G}_{\text{N}}$. However, unlike the randomization distribution of T_{N}, we now must re-calculate $\hat{\sigma}_1^2$ and $\hat{\sigma}_1^2$, besides T_{N}, for each permutation of data. Therefore, for a fixed $\alpha \in (0,1)$, the randomization test based on S_{N} rejects the null hypothesis if $S_{\text{N}} > \hat{r}_{S,\text{N}}(1-\alpha)$, fails to reject if $S_{\text{N}} < \hat{r}_{S,\text{N}}(1-\alpha)$, and randomizes the decision when $S_{\text{N}} = \hat{r}_{S,\text{N}}(1-\alpha)$, where $\hat{r}_{S,\text{N}}(1-\alpha)$ is the $(1-\alpha)$ quantile of the randomization distribution of S_{N}, given by

$$\hat{R}^S_{\text{N}}(t) = \frac{1}{N!} \sum_{\pi \in \boldsymbol{G}_{\text{N}}} \mathbb{1}\left\{S^{\pi}_{\text{N}} \leq t\right\} \ . \tag{17}$$

Importantly, we can then show that the randomization distribution of S_{N} is uniformly asymptotically equivalent to the distribution of a standard normal distribution (e.g., Chung and Romano, 2013, Theorem 2.2). Therefore, the randomization test based on S_{N} controls the probability of a type-I error in large samples, even if the underlying variances of the potential outcomes are not the same—possibly due to treatment effect heterogeneity—or the experimental groups' sizes differ.

4.1 Power Analysis in Completely Randomized Experiments: Revisited

Since the randomization distribution of S_{N} settles around the true unconditional distribution of the test statistic S_{N} regardless of whether the null hypothesis holds or not, we can study its power properties along the same lines as in Section 2.3.

First, we observe that the (local asymptotic) power function of S_{N} against alternative hypotheses of the form $\Delta = h/\sqrt{N}$ in fully randomized experiments is still given by (5) since $\hat{\sigma}_1^2$ and $\hat{\sigma}_0^2$ are consistent estimators for $\mathbb{V}[Y(1)]$ and $\mathbb{V}[Y(0)]$, respectively.

Since the randomization distribution based on S_{N} behaves like the distribution standard normal random variable, we have that its upper-α quantile, $\hat{r}_{S,\text{N}}(1-\alpha)$, converges in probability to the upper-α quantile of the standard normal distribution, $z_{1-\alpha}$. Therefore, we can show by a contiguity argument (e.g., Lehmann and Romano, 2022, Chapter 17.2.2) that the (local asymptotic) power function of the randomization test based on S_{N} is given by (5), that is,

$$1 - \Phi\left(z_{1-\alpha} - \frac{h}{\tilde{\sigma}}\right) \ .$$

Thus, there is no loss in power when using the data-driven critical values from the randomization distribution of S_{N}. By the same token as in Section 2.3, the power of the randomization test based on

S_{N} at $\Delta > 0$ is given, with high probability, by

$$1 - \Phi\left(z_{1-\alpha} - \frac{\sqrt{N}\Delta}{\tilde{\sigma}}\right)$$

as the sample size grows large. Importantly, sample size calculation formulas based on the studentized two-sample t-test apply when one seeks to perform inference on the ATE using randomization inference based on S_{N}.

5 Broader Implications

One of the important lessons from previous Sections is that the randomization test may lead to erroneous decisions and misleading power analyses. The reason is that the randomization distribution does not always mimic the true unconditional distribution of the test statistic. When making inference on the ATE, we show that this "mismatch" occurs because the randomization distribution of T_{N} and the sampling distribution of T_{N} have different asymptotic variances in general.

In this section, we show that this phenomenon is not unique to completely randomized experiments. In fact, a similar issue appears under covariate-adaptive randomization, matched-pair designs, and experiments with cluster randomization. Thus, the same caveats as in completely randomized designs are warranted—inference and power analyses based on randomization tests might lead to erroneous conclusions.

This drawback is not insurmountable. Indeed, we will present modern developments in randomization inference that overcome these issues. The solution echoes Section 4: proper studentization of the test statistic restores the validity of the randomization test. For the sake of concreteness, we omit all the technical details, but encourage the readers to consult Bugni, Canay, and Shaikh (2018), Bai, Romano, and Shaikh (2022), and Bugni et al. (2025) for an in-depth discussion on the theoretical properties of inference methods under covariate-adaptive randomization, matched-pair designs, and cluster randomized experiments, respectively.

As before, we focus on testing problems about the ATE. However, since the aforementioned randomization schemes incorporate pre-determined characteristics to inform the treatment, we now make the following assumption to ensure identification of the ATE. Denote X_i the vector of baseline characteristics for individual i. We will assume that the joint distribution of the treatment status only depends on pre-determined characteristics:

$$\left(Y^{(\mathrm{N})}(1),\, Y^{(\mathrm{N})}(0)\right) \perp D^{(\mathrm{N})} \,|\, X^{(\mathrm{N})}\ .$$

In practice, researchers use predetermined covariates not only as features of the stratification rule but also as controls to improve the precision of their estimates in stratified experiments. Although adjusting for covariates can improve statistical power and estimation precision, our focus here is on their role in the design stage—specifically, how they are used for stratification. This is because, as we will argue, stratification may have unintended negative consequences for randomization inference.

5.1 Covariate-Adaptive Randomization

While complete randomization takes care of selection bias, it does not guard researchers against imbalances over these baseline covariates after randomization. This situation may result in loss of statistical efficiency or low estimation precision, even if these imbalances occur purely by chance (Imbens and Rubin, 2015, Chp. 9).

In such circumstances, covariate-adaptive randomization (CAR) is a popular randomization technique that exploits observable characteristics to inform the treatment and achieve balance over baseline covariates. In plain terms, CAR proceeds in two steps. First, it groups individuals into strata based on the baseline covariate levels. Then, it treats individuals within each stratum using a randomization technique to achieve balance. For instance, we could assign treatment in the second stage by using complete, permuted-block, or biased-coin randomization; see Rosenberger and Lachin (2015).

Although CAR is widely used in practice (e.g., Bruhn and McKenzie, 2009), it has been established that the two-sample t-test can be conservative under this design, with a limiting rejection probability under the null hypothesis that falls below the nominal significance level (Bugni, Canay, and Shaikh, 2018). The source of the problem lies in the dependence among treatment assignments induced by CAR—both across individuals and between treatment status and baseline covariates.

This conservativeness directly leads to a reduction in power when making inference on the ATE. Consequently, standard power analyses and sample size calculations that ignore the impact of stratification on the asymptotic distribution of T_{N} may be misleading, for reasons analogous to those discussed earlier.

However, if we studentize T_{N} correctly, we can show that the two-sample t-test based on the adjusted T_{N} leads to a valid test under CAR in large samples; see Bugni, Canay, and Shaikh (2018) Eq. (24) and Theorem 4.2. Building on this result, it follows by standard arguments that the properly studentized test statistic has limiting power given by (5). Therefore, we could carry on valid inference for the ATE under CAR using the two-sample (adjusted) t-test based on the modified statistic proposed by Bugni, Canay, and Shaikh (2018, Eq. (24)), and Eq. (7) for sample calculations.

Remark 9. (*t-test with Strata Fixed Effects*) It is also possible to construct an alternative test for the ATE under CAR, namely, the usual t-test on the slope coefficient in a regression of Y on D and strata fixed effects with heteroskedasticity-robust standard errors (Bruhn and McKenzie, 2009). As with the two-sample t-test, the t-test with strata fixed effects is conservative in general, though proper studentization yields an exact test in large samples (Bugni, Canay, and Shaikh, 2018). ■

5.1.1 Randomization Test under CAR

As previously discussed, the randomization test based on T_{N} can lead to incorrect sample size calculations under complete randomization without additional assumptions. Since CAR introduces additional dependencies across treatment assignments, there is little reason to expect improved sample size calculations from randomization-based inference when applied in this setting using the same construction as in Section 3.

One might hope that permuting treatment assignments *within strata* would preserve the dependency structure induced by CAR, potentially enabling valid power analysis via randomization tests based on T_{N}. Indeed, such a test does control the type I error rate asymptotically, but only under specific conditions: the allocation proportion must satisfy $\lambda = 1$, and the randomization scheme in the second step of CAR must also achieve the so-called "strong balance" (see Bugni, Canay, and Shaikh, 2018, Sec.2). While certain randomization mechanisms—such as permuted block randomization and Efron's biased-coin design—satisfy this condition, it excludes commonly used schemes, including simple randomization.

More generally, the randomization test will fail to control the probability of a Type I error—and hence compromise sample size calculations—if $\lambda \neq 1$ unless we impose stronger assumptions about the variance and conditional expectations of potential outcomes. As before, a simple fix is possible: calculate the randomization test that permutes treatment status within strata based on the modified test statistic proposed by Bugni, Canay, and Shaikh (2018, Eq. (24)). This test is asymptotically valid for the ATE under CAR schemes with strong balance even if $\lambda \neq 1$.

5.2 Matched Pairs

Another popular stratified randomization scheme in RCT is the matched-pairs design. As in CAR, the idea is to form strata according to covariates X_i, and then assign treatment within each stratum according to a certain treatment assignment mechanism. The key difference is that in matched-pair designs, strata have exactly two units, and only one receives treatment with probability $1/2$. Therefore, assuming N is an even number, we will have that $N = 2m$, which in turn implies that the relative group size is always equal to one.

As before, this form of stratified randomization has an effect on the asymptotic behavior of the test statistic T_{N}. Specifically, Bai, Romano, and Shaikh (2022) show that the sampling distribution of the test statistic behaves, as the sample size increases, like the normal distribution with mean zero and variance

$$\sigma^2_{\text{m-pair}} := \mathbb{V}[Y(1)] + \mathbb{V}[Y(0)] - \frac{1}{2}\,\mathbb{E}\left[\left\{(\mathbb{E}[Y(1)|X] - \mathbb{E}[Y(1)]) + (\mathbb{E}[Y(0)|X] - \mathbb{E}[Y(0)])\right\}^2\right]$$

Therefore, ignoring this effect may yield flawed power analysis and misleading sample size calculations. Indeed, Bai, Romano, and Shaikh (2022) show that the two-sample t-test is generally conservative. However, it is possible to studentize the test statistic and thereby obtain a modified test statistic whose limit distribution does not depend on the way the pairs are formed. The main challenge is that we seek to estimate the conditional variances of potential outcomes within pairs, but only have one treated unit. Bai, Romano, and Shaikh (2022, Eq. (20)) develop a consistent estimator for $\sigma^2_{\text{m-pair}}$, say $\hat{\sigma}^2_{\text{m-pair}}$ by utilizing adjacent pairs that are "similar" in terms of X.

5.2.1 Randomization Test under Matched-Pairs Design

To construct a randomization test under a matched-pairs design, we need to take into account that treatment status is assigned at the pair level. Therefore, we adapt the construction presented in Section 3 to reflect this feature and permute treatment assignment *within pairs*. That is, construct the randomization distribution in match-pair designs by recalculating the test statistic for all permutations of treatment status within each pair.

Denote $\hat{r}_{\text{m-pair,N}}(1-\alpha)$ the upper-α quantile of the randomization distribution based on $T_{\text{N}}/\hat{\sigma}_{\text{m-pair}}$. Consider the randomization test that rejects H_0 if $T_{\text{N}}/\hat{\sigma}_{\text{m-pair}} > \hat{r}_{\text{m-pair,N}}(1-\alpha)$, fails to reject if $T_{\text{N}}/\hat{\sigma}_{\text{m-pair}} < \hat{r}_{\text{m-pair,N}}(1-\alpha)$, and randomizes the decision otherwise.

As before, we seek to approximate the power function of the randomization test against a sequence of alternatives that shrink to the null hypothesis as N grows large. In view of Bai, Romano, and Shaikh (2022, Theorem 3,5) and a standard contiguity argument, we could show that, in RCTs under a matched-pairs design, the randomization test based on the properly studentized test statistic has (local) asymptotic power against alternatives of the form $\Delta_{\text{N}} = h/\sqrt{N}$

$$1 - \Phi\left(z_{1-\alpha} - \frac{h}{\sqrt{2} \cdot \sigma_{\text{m-pair}}}\right) ,$$

where the $\sqrt{2}$ factor comes from the fact that $N = 2m$ in matched-pair designs by construction. Thus, the power of the randomization test based on $T_{\text{N}}/\hat{\sigma}_{\text{m-pair}}$ at $\Delta > 0$ is given by

$$1 - \Phi\left(z_{1-\alpha} - \frac{\sqrt{N}\Delta}{\sqrt{2} \cdot \sigma_{\text{m-pair}}}\right)$$

as the sample size grows large, leading to sample size formulas that resemble those from previous sections, albeit with the appropriate asymptotic variance.

Remark 10. Given the fact that $N = 2m$ in matched-pair designs, we could have considered the power of the randomization test against alternatives of the form $h/\sqrt{m}$. Observe that power analyses are invariant to this choice since we rescaled the asymptotic variance. For instance, MDE formula satisfies

$$\Delta = (z_{1-\alpha} + z_{1-\beta}) \frac{\sqrt{2} \cdot \sigma_{\text{m-pair}}}{\sqrt{N}} = (z_{1-\alpha} + z_{1-\beta}) \frac{\sigma_{\text{m-pair}}}{\sqrt{m}} .$$

■

5.3 Cluster Randomized Experiments

In some cases, we might be interested in cluster randomized experiments (CRE)— experiments where the treatment is assigned at the cluster level, not at the individual level. For instance, clusters could be schools or villages, so if a school is treated, then all the members of said school receive treatment;

see Section 8 in Athey and Imbens (2017) for a more detailed exposition.

Arguing as in Bugni et al. (2025), we can think of the sampling process in CREs as a two-stage process, where we first draw a random sample of G clusters from a superpopulation of clusters—so the clusters are random variables—, and then we sample a subset of individual units within each cluster. Two remarks are in order. First, the size of each cluster, denoted N_g, $g = 1, \ldots, G$, and the cluster characteristics, X_g, are random variables themselves. Thus, the cluster sizes and characteristics are heterogeneous once realized, even though they are drawn from the same distribution. Second, the researcher might not sample all units within a cluster. We will denote the subset of sample units from cluster g by $\mathcal{M}_g \subset \{1, \ldots, N_g\}$, and by $|\mathcal{M}_g|$ the number of elements in $\mathcal{M}_g$ for each $g = 1, \ldots, G$.

We adapt the notation from previous sections to reflect the fact that this is a CRE. Let $Y_{i,g}(1)$ denote the potential outcomes of individual i in cluster g if treated, and similarly $Y_{i,g}(0)$ as the potential outcomes of individual i in cluster g if not treated. Once the observations in a cluster are realized, the researcher assigns treatment at the cluster level. We assume that the entire cluster is either treated or not treated. Denote by D_g the cluster g treatment indicator taking value 1 if cluster g is treated, 0 otherwise. As is standard in the potential outcomes framework, observed and potential outcomes are linked via $Y_{i,g} = Y_{i,g}(1)D_g + Y_{i,g}(0)(1 - D_g)$.

We discuss three different parameters of interest. First, consider the ATE where the clusters are the units of interest. This parameter of interest is referred to as the *equally-weighted cluster-level average treatment effect*, given by

$$\theta_1 := \mathbb{E}\left[\frac{1}{N_g}\sum_{i=1}^{N_g}(Y_{i,g}(1) - Y_{i,g}(0))\right] . \tag{18}$$

The second parameter of interest is the ATE where the individuals are the units of interest. This is the so-called *size-weighted cluster-level average treatment effect*, given by

$$\theta_2 := \mathbb{E}\left[\frac{1}{\mathbb{E}[N_g]}\sum_{i=1}^{N_g}(Y_{i,g}(1) - Y_{i,g}(0))\right] . \tag{19}$$

Lastly, we will consider the *sample-weighted cluster-level average treatment effect*,

$$\theta_3 := \mathbb{E}\left[\frac{1}{\mathbb{E}[|\mathcal{M}_g|]}\sum_{i\in\mathcal{M}_g}(Y_{i,g}(1) - Y_{i,g}(0))\right] . \tag{20}$$

Remark 11. In general, θ_1, θ_2, and θ_3 are different. However, they coincide in some special cases. For instance, if all clusters are of the same size, say k, or the treatment effects are constant, then $\theta_1 = \theta_2$. On the other hand, θ_3 equals θ_1 if we sample the same number of individuals across clusters ($\mathrm{P}[|\mathcal{M}_g| = k] = 1$ for all $g = 1, \ldots, G$). Lastly, θ_3 equals θ_2 if we sample the same fraction of units across clusters, i.e., $\mathrm{P}[|\mathcal{M}_g| = \gamma N_g] = 1$ for some $\gamma \in (0, 1]$, for all $g = 1, \ldots, G$. ■

Following Bugni et al. (2025), we estimate (18)–(20) using an appropriate linear regression in each case. Starting with (20), the estimator is simply the difference-in-means estimator, given by the estimate of the slope parameter of a linear regression of $Y_{i,g}$ on the cluster treatment indicator D_g and a constant. Similarly, we may estimate (19) with $\hat{\theta}_2$, given by the estimate of the slope parameter in a weighted linear regression of $Y_{i,g}$ on the cluster treatment indicator D_g and a constant, using weights $N_g/|\mathcal{M}_g|$. Lastly, consider (18). Since the θ_1 is an ATE where the clusters are the units of interest, we begin by aggregating sampled individuals by cluster as

$$\bar{Y}_g = \frac{1}{|\mathcal{M}_g|} \sum_{i \in \mathcal{M}_g} Y_{i,g} \ .$$

Then, we estimate θ_1 using $\hat{\theta}_1$ given by the estimate of the slope parameter in the linear regression of $\bar{Y}_g$ on D_g and a constant.

5.3.1 Power Analysis for Randomization Tests in CRE

The inference theory for CRE with non-ignorable cluster sizes is being developed contemporaneously, with current efforts primarily centered on tests based on asymptotic normal approximations. This framework enables power analyses and sample size calculations for asymptotic tests in the same spirit as in Section 2.3. In contrast, the theoretical foundations for randomization tests under non-ignorable clustering remain less developed. In particular, we lack a general characterization of the asymptotic behavior of the randomization distribution in this setting—at least one that would support power analysis and sample size calculations with comparable scope and practical applicability.

For this reason, we focus in this section on a simplified setup where we can derive the asymptotic properties of the randomization test and provide sample size calculations for applied researchers. In this section, we will only consider CRE under complete randomization, where $1 \leq m < G$ clusters receive treatment, and $n = G - m$ clusters receive no treatment. Therefore, we omit covariates and cluster characteristics for the sake of simplicity; see Bugni et al. (2025) for asymptotic inference based on the normal approximation in CRE under CAR, and Bai et al. (2024) for matched-pair designs.

Under complete randomization, we can readily provide power analyses for randomization inference for θ_1. To see why, recall that θ_1 can be seen as the ATE when the units of interest are the clusters themselves. Thus, this case boils down to the two-sample problem discussed in Sections 2 and 3, where the treated sample is now given by the collection of $\bar{Y}_g$ for all treated clusters g, and the control group is the corresponding $\bar{Y}_g$ for all non-treated clusters $g = 1, \ldots, G$. Similarly, the effective sample size is now G as opposed to N, so the approximations are now conceived as G grows large.

Building on the lessons from Section 4, we want to establish the asymptptoc behavior of the randomization distribution based on the studentized test statistics $S_{\mathrm{N}}^{\text{cluster}} = T_{\mathrm{N}}/\hat{\sigma}_{\text{cluster}}$, where T_{N} is given by the difference-in-means estimator

$$T_{\mathrm{N}}^{\text{cluster}} = \sqrt{m}\left(\frac{\sum_{g=1}^{G} \bar{Y}_g D_g}{\sum_{g=1}^{G} D_g} - \frac{\sum_{g=1}^{G} \bar{Y}_g (1 - D_g)}{\sum_{g=1}^{G} (1 - D_g)}\right) \tag{21}$$

and $\hat{\sigma}_{\text{cluster}}$ is a consistent estimator of the asymptotic variance of $T_{\mathrm{N}}^{\text{cluster}}$ under complete randomization (e.g., Theorem 3.4 of Bugni et al. (2025) and related discussion therein). Then, it follows by standard arguments that the randomization distribution of the studentized statistic $T_{\mathrm{N}}^{\text{cluster}}$ in (21) behaves like a distribution of a standard normal random variable as the number of clusters G grows large (e.g, by adapting the arguments in Lehmann and Romano (2022, Chapter 17.3)). Since also $T_{\mathrm{N}}^{\text{cluster}}$ converges in distribution to a standard normal random variable by Theorems 3.2 and 3.4 in Bugni et al. (2025), it follows that the randomization distribution mimics the large sample behavior of the sampling distribution of $T_{\mathrm{N}}^{\text{cluster}}$.

An immediate corollary of the ongoing discussion is that we can approximate the (local) asymptotic power in completely randomized CREs when the goal is to make inferences on the equally weighted cluster-level ATE. Specifically, the power of the randomization test based on $T_{\mathrm{N}}^{\text{cluster}}$ against alternatives of the form $\Delta_G = h/\sqrt{G}$ is given by

$$1 - \Phi\left(z_{1-\alpha} - \sqrt{\frac{\lambda}{1+\lambda}} \cdot \frac{h}{\sigma_{\text{cluster}}}\right) \ , \ \text{where} \ \ \sigma_{\text{cluster}}^2 = \mathbb{V}[\bar{Y}_g(1)] + \lambda \cdot \mathbb{V}[\bar{Y}_g(0)] \ .$$

Thus, the power of the randomization test based on (21) at $\Delta > 0$, as the number of clusters increases, is approximated by

$$1 - \Phi\left(z_{1-\alpha} - \sqrt{\frac{\lambda}{1+\lambda}} \cdot \frac{\sqrt{G}\Delta}{\sigma_{\text{cluster}}}\right) ,$$

so we obtain sample size formulas as in previous sections. Moreover, these formulas reveal that, since the target parameter is θ_1 so the units of interest are the clusters themselves, the *sample size* now refers to the number of clusters (the sampled units within clusters are aggregated into $\bar{Y}_g(d)$, $d \in \{0,1\}$).

Building on the previous power analysis, we may also obtain sample size calculations for θ_2 and θ_3 under additional assumptions. For instance, suppose that the target parameter is θ_3 instead. First, we note that θ_3 equals θ_1 if we sample the same number of individuals across clusters. Moreover, under the same assumption, we can show that $\hat{\theta}_3$ reduces to $\hat{\theta}_1$. Therefore, under this additional assumption but otherwise under the same set of assumptions as before, the randomization-based power analysis for θ_1 carries over to θ_3. See Remark 11 for more conditions under which θ_1, θ_2, and θ_3 are equal.

6 Practical Recommendations for Applied Researchers

6.1 Sampling from a Finite- vs Superpopulation

We have performed power analyses under a paradigm that presumes our sample $\{(Y_i(1), Y_i(0), X_i) : 1 \leq i \leq N\}$ is drawn independently and identically distributed according to some probability distribution. In particular, when analyzing randomized experiments, we view the potential outcomes as random variables. This perspective on sampling is often referred to as sampling from a hypothetical infinite "superpopulation," e.g., Van der Vaart (2000).

However, one could also adopt a sampling perspective that treats the data as a sample from a finite population instead (Neyman, 1923, 1935). In this framework, we typically assume that the potential outcomes are nonrandom, and the only source of randomness comes entirely from the treatment assignment. While this paradigm is somewhat mainstream in statistics, it has also become popular in econometrics, often referred to as "design-based" uncertainty, e.g., Abadie et al. (2020). See also Imbens and Rubin (2015), Athey and Imbens (2017), and Ding (2024) for textbook expositions.

The distinction between these two approaches is not innocuous, and it may result in qualitatively different analyses.[2] In particular, this distinction is relevant for randomization inference. While a

[2] We could see the superpopulation paradigm as an approximation to the finite-population paradigm; see Ding, Li, and Miratrix (2017) for more details.

formal comparison between these two approaches is beyond the scope of this paper, let us focus on one aspect that is relevant for power analyses and sample size calculations based on randomization inference when the object of interest is the ATE.

We have seen that randomization-based power analyses rely on the fact that we may correctly mimic the sampling behavior of the test statistic in large samples. Moreover, we showed that to achieve this result, we relied on the proper studentization of the test statistic by an estimator of the asymptotic variance. Specifically, for a suitable consistent estimator of the variance, we may ensure that the limiting distribution of the test statistic does not depend on unknown parameters.

However, while it is possible to construct such a consistent estimator in a superpopulation paradigm, we cannot always achieve the same in a finite-population paradigm (Bai, Shaikh, and Tabord-Meehan, 2025).[3] The reason is that, from a finite-population perspective, the variance of the difference-in-means estimator of the ATE (e.g., T_{N} in Section 2.2) depends on the unit-level treatment effects, $Y_i(1) - Y_i(0)$, which are fundamentally unobservable for the same unit. This term—which does not show in the super-population paradigm—renders randomization inference conservative (Wu and Ding, 2021), thus affecting sample calculations from a super-population perspective.

Given these concerns, we emphasize that applied researchers should be mindful of these differences, both conceptually and mathematically. This is especially relevant for randomization inference, as randomization tests are often conceived and motivated in the context of design-based uncertainty, e.g., Fisher (1935). However, if we seek to use the sample size calculations based on the asymptotic approximations in the previous sections, then it is important to clarify that we adhere to a framework that supposes sampling from an infinite population. Whether this modeling decision is appropriate or not depends on a given application; see Abadie et al. (2020) for more discussion.

6.2 Considerations for Different Designs or Target Parameters

In practice, we might care about different target parameters beyond the ATE. For instance, we may consider quantile treatment effects or hypotheses about the entire distributions of the potential outcomes. Indeed, it is possible to show that the randomization test has the same limiting power as the asymptotic case against so-called “contiguous alternatives” in some of these scenarios, so there is no loss in power when using the critical values from the randomization test when performing inference

[3] The large-sample asymptotic analysis from a finite-population perspective can be found in Li and Ding (2017); Wu and Ding (2021). See also Example 12.2.2 in Lehmann and Romano (2022).

and power analyses. This is the case, for instance, when testing for heterogeneity in the treatment effect in completely randomized experiments (e.g., Chung and Olivares, 2021, 2025).

However, it might not be a trivial task to justify and generalize the asymptotic approach from Sections 3 in more complex settings without further assumptions. For instance, it is difficult to analyze the behavior of the randomization distribution and get analytically tractable power functions that enable us to carry on sample calculations in high-dimensional settings, where the large-sample behavior of the test statistic is non-normal; see Kim, Balakrishnan, and Wasserman (2022) and Dobriban (2022) for more details.

Similarly, we might be interested in different randomization schemes beyond complete randomization, CAR, or matched-pairs designs. For instance, we could use rerandomization (Morgan and Rubin, 2012) to achieve covariate balance. Though sample size calculations have been derived for the two-sample t-test from a finite-sample perspective (Branson, Li, and Ding, 2024), the power of the randomization test remains largely unknown to the best of our knowledge.

7 Concluding Remarks

This paper examines the challenges and remedies associated with power and sample size calculations for randomization tests in experimental research when the goal is to make inference on the ATE.

There are two takeaway messages. First, we show that naive applications of classical formulas—developed for the two-sample t-test—can yield misleading power analyses when used in conjunction with randomization-based methods. In particular, when the limiting behavior of the test statistic depends on unknown parameters under the null hypothesis, the randomization distribution may, in general, fail to mirror the behavior of the sampling distribution, thereby distorting Type I error control, power, and sample size calculations regardless of whether we are at the design stage (*ex ante*) or analysis stage (*ex post*).

The second message is a positive one: to address the issues above, a simple solution is to properly studentize the test statistic so that the modified statistic's large-sample behavior is free of unknown parameters. In doing so, we show that the randomization distribution of the newly modified statistic behaves like the true sampling distribution of the test statistic, restoring the validity of classical power analysis without sacrificing the nonparametric appeal of randomization inference.

These insights hold not only in completely randomized designs but also extend to more complex settings, including covariate-adaptive randomization, matched-pairs designs, and cluster-randomized experiments—though more research is needed to cover recent developments that are largely underdeveloped in the context of randomization inference. And while our focus has been on the ATE, similar principles may apply to other target parameters—including quantile treatment effects—though the solutions could be less automatic in some cases due to the high-dimensionality.

Future research could extend this framework to richer designs and target parameters where the behavior of randomization distributions remains analytically elusive, such as in high-dimensional inference, nonlinear test statistics, or in designs involving rerandomization. Strengthening the connection between design-based uncertainty and practical power analysis in these contexts remains an important avenue for both theoretical and applied work.

References

Abadie, A., Athey, S., Imbens, G. W., and Wooldridge, J. M. (2020). Sampling-based versus design-based uncertainty in regression analysis. *Econometrica*, 88(1):265–296.

Albert, M. (2019). Concentration inequalities for randomly permuted sums. In *High Dimensional Probability VIII: The Oaxaca Volume*, pages 341–383. Springer.

Athey, S. and Imbens, G. W. (2017). The econometrics of randomized experiments. In *Handbook of economic field experiments*, volume 1, pages 73–140. Elsevier.

Bai, Y., Liu, J., Shaikh, A. M., and Tabord-Meehan, M. (2024). Inference in cluster randomized trials with matched pairs. *Journal of Econometrics*, 245(1-2):105873.

Bai, Y., Romano, J. P., and Shaikh, A. M. (2022). Inference in experiments with matched pairs. *Journal of the American Statistical Association*, 117(540):1726–1737.

Bai, Y., Shaikh, A. M., and Tabord-Meehan, M. (2025). A primer on the analysis of randomized experiments and a survey of some recent advances. *forthcoming in the Journal of Political Economy: Microeconomics*.

Banerjee, A., Duflo, E., Glennerster, R., and Kinnan, C. (2015). The miracle of microfinance? evidence from a randomized evaluation. *American economic journal: Applied economics*, 7(1):22–53.

Bloom, N., Eifert, B., Mahajan, A., McKenzie, D., and Roberts, J. (2013). Does management matter? evidence from india. *The Quarterly journal of economics*, 128(1):1–51.

Bloom, N., Mahajan, A., McKenzie, D., and Roberts, J. (2020). Do management interventions last? evidence from india. *American Economic Journal: Applied Economics*, 12(2):198–219.

Branson, Z., Li, X., and Ding, P. (2024). Power and sample size calculations for rerandomization. *Biometrika*, 111(1):355–363.

Bruhn, M. and McKenzie, D. (2009). In pursuit of balance: Randomization in practice in development field experiments. *American economic journal: applied economics*, 1(4):200–232.

Bugni, F., Canay, I. A., Shaikh, A. M., and Tabord-Meehan, M. (2025). Inference for cluster randomized experiments with nonignorable cluster sizes. *Journal of Political Economy Microeconomics*, 3(2):255–288.

Bugni, F. A., Canay, I. A., and Shaikh, A. M. (2018). Inference under covariate-adaptive randomization. *Journal of the American Statistical Association*, 113(524):1784–1796.

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

Chung, E. and Olivares, M. (2021). Permutation test for heterogeneous treatment effects with a nuisance parameter. *Journal of Econometrics*, 225(2):148–174.

Chung, E. and Olivares, M. (2025). Quantile-based test for heterogeneous treatment effects. *Journal of Applied Econometrics*, 40(1):3–17.

Chung, E. and Romano, J. P. (2013). Exact and asymptotically robust permutation tests. *The Annals of Statistics*, 41(2):484–507.

Chung, E. and Romano, J. P. (2016). Asymptotically valid and exact permutation tests based on two-sample u-statistics. *Journal of Statistical Planning and Inference*, 168:97–105.

Ding, P. (2024). *A first course in causal inference*. CRC Press.

Ding, P., Feller, A., and Miratrix, L. (2016). Randomization inference for treatment effect variation. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 78(3):655–671.

Ding, P., Li, X., and Miratrix, L. W. (2017). Bridging finite and super population causal inference. *Journal of Causal Inference*, 5(2):20160027.

Dobriban, E. (2022). Consistency of invariance-based randomization tests. *The Annals of Statistics*, 50(4):2443–2466.

Duflo, E., Glennerster, R., and Kremer, M. (2007). Using randomization in development economics research: A toolkit. *Handbook of development economics*, 4:3895–3962.

Fisher, R. A. (1935). *The Design of Experiments*. Oliver & Boyd.

Glennerster, R. and Takavarasha, K. (2013). Running randomized evaluations: A practical guide. In *Running randomized evaluations*. Princeton University Press.

Imbens, G. W. and Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge university press.

Karlan, D. and List, J. A. (2007). Does price matter in charitable giving? evidence from a large-scale natural field experiment. *American Economic Review*, 97(5):1774–1793.

Kim, I., Balakrishnan, S., and Wasserman, L. (2022). Minimax optimality of permutation tests. *The Annals of Statistics*, 50(1):225–251.

Lachin, J. M. (1981). Introduction to sample size determination and power analysis for clinical trials. *Controlled clinical trials*, 2(2):93–113.

Lehmann, E. L. and Romano, J. P. (2022). *Testing Statistical Hypotheses*. Springer.

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

Morgan, K. L. and Rubin, D. B. (2012). Rerandomization to improve covariate balance in experiments. *The Annals of Statistics*, 40:1263–1282.

Neyman, J. (1923). On the application of probability theory to agricultural experiments. essay on principles. section 9 (translated). *Reprinted Statistical Science*, pages 465–472.

Neyman, J. (1935). Statistical problems in agricultural experimentation. *Journal of the Royal Statistical Society*, 2(2):107–180.

Neyman, J. (1992). On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. In *Breakthroughs in statistics: Methodology and distribution*, pages 123–150. Springer.

Ramdas, A., Barber, R. F., Candès, E. J., and Tibshirani, R. J. (2023). Permutation tests using arbitrary permutation distributions. *Sankhya A*, 85(2):1156–1177.

Ritzwoller, D. M., Romano, J. P., and Shaikh, A. M. (2025). Randomization inference: Theory and applications. *forthcoming in the Journal of Political Economy Microeconomics*.

Rosenberger, W. F. and Lachin, J. M. (2015). *Randomization in clinical trials: theory and practice*. John Wiley & Sons.

Van der Vaart, A. W. (2000). *Asymptotic statistics*, volume 3. Cambridge university press.

Wu, J. and Ding, P. (2021). Randomization tests for weak null hypotheses in randomized experiments. *Journal of the American Statistical Association*, 116(536):1898–1913.

Young, A. (2019). Channeling fisher: Randomization tests and the statistical insignificance of seemingly significant experimental results. *The quarterly journal of economics*, 134(2):557–598.