# Covariate-adjusted Fisher randomization tests for the average treatment effect

Anqi Zhao and Peng Ding *

**Abstract**

Fisher's randomization test (FRT) delivers exact $p$-values under the strong null hypothesis of no treatment effect on any units whatsoever and allows for flexible covariate adjustment to improve the power. Of interest is whether the resulting covariate-adjusted procedure could also be valid for testing the weak null hypothesis of zero average treatment effect. To this end, we evaluate two general strategies for conducting covariate adjustment in FRTs: the *pseudo-outcome* strategy that uses the residuals from an outcome model with only the covariates as the pseudo, covariate-adjusted outcomes to form the test statistic, and the *model-output* strategy that directly uses the output from an outcome model with both the treatment and covariates as the covariate-adjusted test statistic. Based on theory and simulation, we recommend using the ordinary least squares (OLS) fit of the observed outcome on the treatment, centered covariates, and their interactions for covariate adjustment, and conducting FRT with the robust $t$-value of the treatment as the test statistic. The resulting FRT is finite-sample exact for testing the strong null hypothesis, asymptotically valid for testing the weak null hypothesis, and more powerful than the unadjusted counterpart under alternatives, all irrespective of whether the linear model is correctly specified or not. We start with complete randomization, and then extend the theory to cluster randomization, stratified randomization, and rerandomization, respectively, giving a recommendation for the test procedure and test statistic under each design. Our theory is design-based, also known as randomization-based, in which we condition on the potential outcomes but average over the random treatment assignment.

**Keywords**: Finite-population inference; permutation test; randomization distribution; robust standard error; studentization; super-population inference

*Anqi Zhao, Department of Statistics and Applied Probability, National University of Singapore, 117546, Singapore (E-mail: staza@nus.edu.sg). Peng Ding, Department of Statistics, University of California, Berkeley, CA 94720 (E-mail: pengdingpku@berkeley.edu). Peng Ding was partially funded by the U.S. National Science Foundation (grant # 1945136). We thank the Associate Editor and three referees for their most insightful comments. We thank Jason Wu, Cheng Gao, Kevin Guo, Thomas Richardson, Avi Feller, Xiaokang Luo, Xinran Li, Zhichao Jiang, Mengsi Gao, Bin Yu, and Philip Stark for helpful suggestions.

# 1. Fisher's randomization test with covariate adjustment

Fisher (1935) viewed randomization as a "reasoned basis" for inference and proposed the randomization test as a universal way to generate finite-sample exact $p$-values without imposing modeling assumptions on the experimental outcomes. Fisher's randomization test (FRT) becomes increasingly important with the popularity of field experiments in social sciences in addition to the traditional biomedical experiments. Proschan and Dodd (2019) reviewed the use of FRTs in randomized clinical trials and highlighted its strength in analyzing complex data. There is an increasing interest in economics and related fields to use FRTs to analyze various types of empirical data (Freedman and Lane 1983; Kennedy 1995; Lee and Shaikh 2014; Cattaneo et al. 2015; Canay et al. 2017; Ganong and Jäger 2018; Athey et al. 2018; Bugni et al. 2018; Young 2019; Heckman and Karapakula 2021; MacKinnon and Webb 2020; Heckman et al. 2020).

The flexibility of FRT enables two natural strategies to incorporate covariate information via statistical modeling. First, we can fit a statistical model of the outcome on the covariates to obtain the residuals as the covariate-adjusted pseudo outcomes, and proceed with the usual FRT in a covariate-free fashion. Tukey (1993) used it with linear models, Gail et al. (1988) used it with generalized linear models, Raz (1990) used it with nonparametric regressions, Stephens et al. (2013) used it with the generalized estimating equation for clustered data, and Rosenbaum (2002) reviewed and extended it to not only randomized experiments but also matched observational studies. Second, we can directly fit an outcome model with both the treatment and covariates, and use the model output, such as the coefficient of the treatment or the corresponding $t$-value, as the test statistic. The canonical choice is a linear model on the treatment and covariates, often known as the analysis of covariance (Fisher 1935; Freedman 2008; Lin 2013; Young 2019). Brillinger et al. (1978) gave an early application of this strategy with more complex statistical models. This defines two model-assisted approaches to conducting covariate-adjusted FRTs.

The strong guarantees of FRT hold only under the strong null hypothesis of zero individual treatment effects, which is often criticized for being too restrictive for many practical applications. Adaptation to the weak null hypothesis of zero average treatment effect is one important direction for broadening its application. A natural class of test statistics for this purpose are the coefficients of the treatment from various outcome models, with or without covariate adjustment, along with their classic or robustly-studentized $t$-statistics (Eicker 1967; Huber 1967; White 1980). These coefficients are consistent estimators of the average treatment effect and are thus sensitive to deviations from both the strong and weak null hypotheses (Freedman 2008; Lin 2013). Of interest is whether these intuitive test statistics can preserve the correct type one error rates when only the weak null hypothesis holds (Romano 1990; Chung and Romano 2013; Wu and Ding 2020), and if covariate adjustment delivers additional power under alternatives.

To this end, we extend the discussion by Ding and Dasgupta (2018) and Wu and Ding (2020) on the utility of FRT for testing the weak null hypothesis to the presence of covariates under the finite-population framework, and examine the asymptotic operating characteristics of nine

covariate-adjusted test statistics as the coefficients from three common outcome models and their respective classic and robustly-studentized $t$-statistics. The results establish the permutational limiting theorems of the OLS coefficients and standard errors under the possibly misspecified linear models, and shed light on the utility of model-assisted covariate adjustment for testing the weak null hypothesis via FRT. Building upon previous work on using studentized statistics for permutation tests under the super-population (Janssen 1997; Chung and Romano 2013; Pauly et al. 2015; Bugni et al. 2018) and finite-population (Wu and Ding 2020) frameworks, respectively, we extend the discussion to the covariate-adjusted test statistics under the finite-population framework, and show the necessity of robust studentization for ensuring the asymptotic validity of the covariate-adjusted test statistics when only the weak null hypothesis holds. The robustly studentized $t$-statistic based on Lin (2013)'s estimator, as it turns out, guarantees both asymptotic validity and the highest power for testing the weak null hypothesis. The estimator equals the coefficient of the treatment in the OLS fit of the outcome on the treatment, centered covariates, and their interactions, but the aforementioned theoretical guarantees hold irrespective of whether the linear model is correctly specified or not. Together with its finite-sample exactness under the strong null hypothesis, it is thus our final recommendation for testing both the strong and weak null hypotheses under complete randomization.

We first focus on complete randomization and then generalize the theory to other types of design. The extension to cluster randomization and stratified randomization is direct whereas that to rerandomization (Morgan and Rubin 2012) has some distinct features. In particular, covariate adjustment becomes more crucial since studentization alone no longer ensures the appropriateness of FRT for the weak null hypothesis. In addition, it is common that the designer and analyzer do not communicate (Bruhn and McKenzie 2009; Heckman et al. 2020; Heckman and Karapakula 2021), and if this happens, we recommend using FRT pretending that the experiment was completely randomized. In this non-ideal case, the proposed FRT is no longer finite-sample exact under the strong null hypothesis unless the original experiment is indeed completely randomized, but at least it preserves the correct type one error rates under the weak null hypothesis. Based on extensive theoretical investigations, we make final recommendations for FRT and the test statistic in each experimental design.

We will use the following notation for permutations. Let $\Pi$ be the set of all $N!$ random permutations of $\{1, \ldots, N\}$, indexed by $\pi$. For an $N \times 1$ vector $a = (a_1, \ldots, a_N)^{\mathrm{T}}$, let $a_\pi = (a_{\pi(1)}, \ldots, a_{\pi(N)})^{\mathrm{T}}$ be a permutation of its elements. If $b = b(a)$ is a function of $a$, let $b^\pi = b(a_\pi)$ be its value evaluated at $a_\pi$. Without introducing new notation, use $\pi$ to also represent a random draw from $\Pi$, namely $\pi \sim \mathrm{Unif}(\Pi)$, with meaning clear from the context. With a slight abuse of notation, assume sets like $\{a_\pi : \pi \in \Pi\}$ to contain $|\Pi| = N!$ elements defined by $\pi \in \Pi$ throughout, such that $a_\pi$ and $a_{\pi'}$ are two distinct elements so long as $\pi \neq \pi'$, even if $a_\pi = a_{\pi'}$. For two random variables $A$ and $B$, let $A \leq_{\mathrm{st}} B$ denote $A$ is stochastically dominated by $B$.

# 2. Basic setup under complete randomization

## 2.1. Potential outcomes and Fisher's randomization test

Consider an intervention of two levels, $z = 0, 1$, and a finite population of $N$ units, $i = 1, \ldots, N$. Let $Y_i(z)$ be the potential outcome of unit $i$ under treatment $z$ (Neyman 1923/1990). The individual treatment effect is $\tau_i = Y_i(1) - Y_i(0)$, and the finite-population average treatment effect is $\tau = N^{-1} \sum_{i=1}^N \tau_i$. We focus on the finite-population inference in the main text, and will refer to $\tau$ as the *average treatment effect* when no confusion would arise. Let $x_i = (x_{i1}, \ldots, x_{iJ})^{\mathrm{T}}$ be the covariates for unit $i$, concatenated as an $N \times J$ matrix $X = (x_1, \ldots, x_N)^{\mathrm{T}}$. Center the covariates at $\bar{x} = N^{-1} \sum_{i=1}^N x_i = 0_J$ to simplify the presentation.

The designer assigns $N_z$ units to receive level $z$ with $N_1 + N_0 = N$ and $(p_1, p_0) = (N_1/N, N_0/N)$. Let $Z_i$ denote the treatment level received by unit $i$, with $Z_i = 1$ for treatment and $Z_i = 0$ for control, vectorized as $Z = (Z_1, \ldots, Z_N)^{\mathrm{T}}$. Complete randomization samples $Z$ uniformly from the set $\mathcal{Z}$ that contains all permutations of $N_1$ 1's and $N_0$ 0's. The observed outcome is $Y_i = Z_i Y_i(1) + (1 - Z_i) Y_i(0)$ for unit $i$, vectorized as $Y = (Y_1, \ldots, Y_N)^{\mathrm{T}}$. A test statistic is a function of the treatment vector, observed outcomes, and covariates, denoted by $T = T(Z, Y, X)$.

Write $Y = Y(Z)$ and $T = T(Z, Y(Z), X)$ to highlight the dependence of the observed outcomes and the test statistic on the treatment vector. Complete randomization induces a uniform distribution over $\{T(\boldsymbol{z}, Y(\boldsymbol{z}), X) : \boldsymbol{z} \in \mathcal{Z}\}$ as the *sampling distribution* of $T$. Fisher (1935) considered testing the strong null hypothesis

$$H_{0\mathrm{F}} : Y_i(1) = Y_i(0) \quad \text{for all } i = 1, \ldots, N$$

and proposed FRT to compute the $p$-value as

$$p_{\mathrm{FRT}} = |\Pi|^{-1} \sum_{\pi \in \Pi} 1\{T(Z_\pi, Y, X) \geq T(Z, Y, X)\}, \tag{1}$$

assuming a one-sided test. Each $Z_\pi$ is a permutation of $Z$, and by symmetry, all possible values of $Z_\pi$ over $\pi \in \Pi$ consist of $\mathcal{Z}$. FRT thus induces a uniform distribution over $\{T(\boldsymbol{z}, Y(Z), X) : \boldsymbol{z} \in \mathcal{Z}\}$ conditioning on the observed $Z$, known as the *randomization distribution* of $T$. Let $T^\pi = T(Z_\pi, Y(Z), X)$, where $\pi \sim \mathrm{Unif}(\Pi)$, be a random variable from this distribution conditioning on $Z$. The $p_{\mathrm{FRT}}$ in (1) gives the right-tail probability of the observed value of the test statistic with regard to its randomization distribution. Under $H_{0\mathrm{F}}$, the randomization distribution equals the sampling distribution, and thereby ensures the finite-sample exactness of FRT for arbitrary $T$.

In practice, we need to choose a test statistic that is sensitive to deviations from $H_{0\mathrm{F}}$. Computationally, FRT involves randomly permuting the treatment vector to generate $Z_\pi$. This justifies *permutation test* as its other name. If $|\Pi| = N!$ is too large, we can take a simple random sample from $\Pi$ to obtain a Monte Carlo approximation of $p_{\mathrm{FRT}}$.

**Remark 1.** Based on the definition in (1), FRT does not require any algebraic group structure of

the treatment assignment mechanism. Therefore, it is more general than the usual definition of permutation test which requires certain invariance under group transformations (Hoeffding 1952; Lehmann and Romano 2005). Nevertheless, we will focus on FRTs that can be implemented by permutations or restricted permutations in this paper. See Basse et al. (2019) for more discussion on the connections and distinctions.

The nice properties of FRT under $H_{0\text{F}}$ inspire endeavors to extend it to other types of hypotheses. Consider the weak null hypothesis of zero average treatment effect (Neyman 1935):

$$H_{0\text{N}} : \tau = 0.$$

We can proceed with FRT by permuting the treatment vector $Z$ and report $p_{\text{FRT}}$ by (1) as if we were testing $H_{0\text{F}}$. Under $H_{0\text{N}}$, $Y(\boldsymbol{z})$ varies with $\boldsymbol{z} \in \mathcal{Z}$ such that the randomization distribution $T^{\pi}$ no longer equals the sampling distribution $T$. Consequently, $p_{\text{FRT}}$ loses its finite-sample exactness as the basis for controlling the type one error rates in general. Wu and Ding (2020) gave examples in which FRT yields invalid type one error rates under $H_{0\text{N}}$ even asymptotically.

The discrepancy between the distributions of $T$ and $T^{\pi}$ in the absence of $H_{0\text{F}}$ is at the heart of the loss of finite-sample exactness when applying FRT to hypotheses other than $H_{0\text{F}}$. The sampling distribution of $T$, on the one hand, is induced by the randomization of $Z$ based on the *true finite population* of $\{(Y_i(0), Y_i(1), x_i\}_{i=1}^N$. The FRT procedure, on the other hand, assumes the observed outcomes $Y = (Y_1, \ldots, Y_N)^{\text{T}}$ remain unchanged over all possible assignments, and is essentially using $\{(Y_i'(0), Y_i'(1), x_i\}_{i=1}^N$, where $Y_i'(0) = Y_i'(1) = Y_i$, as the *pseudo finite population* to generate the randomization distribution of $T$, represented by $T^{\pi}$. This ends up mixing $\{Y_i(0)\}_{i=1}^N$ and $\{Y_i(1)\}_{i=1}^N$ with proportions $p_0$ and $p_1$ in the absence of $H_{0\text{F}}$, and thereby results in the different distributions of $T$ and $T^{\pi}$.

Despite the possibly liberal type one error rates in general, sensible choice of the test statistic restores the validity of FRT for testing $H_{0\text{N}}$ at least asymptotically. Wu and Ding (2020) showed that FRT preserves the correct type one error rates asymptotically with a class of robustly studentized statistics. We extend their discussion to the setting with covariates, and propose a general strategy for covariate-adjusted FRT that ensures both asymptotic validity and higher power for testing $H_{0\text{N}}$.

Assume the finite-population asymptotic framework that embeds $\mathcal{S} = \{Y_i(0), Y_i(1), x_i\}_{i=1}^N$ and $Z = (Z_i)_{i=1}^N$ into a sequence of finite populations and assignments for $N = 1, \ldots, \infty$. Technically, all quantities depend on $N$, but we omit the subscript $N$ for simplicity.

**Definition 1.** A test statistic $T$ is *proper* for testing $H_{0\text{N}}$ if under $H_{0\text{N}}$,

$$\lim_{N\to\infty} \text{pr}(p_{\text{FRT}} \leq \alpha) \leq \alpha \qquad \text{for all } \alpha \in (0, 1)$$

holds for any $\mathcal{S}$.

Assume a one-sided test and $p$-value as the right-tail probability as in (1). A statistic $T$ is proper for testing $H_{0\text{N}}$ if under $H_{0\text{N}}$, the sampling distribution of $T$ is stochastically dominated by

its randomization distribution for almost all sequences of $Z$ as $N \to \infty$.

### 2.2. Two strategies for covariate-adjusted FRT and twelve test statistics

We review two general strategies for covariate adjustment in FRT. We focus on test statistics based on estimators of $\tau$ to accommodate both $H_{0\text{F}}$ and $H_{0\text{N}}$, and unify them under the OLS formulation for easy implementation.

Let $\hat{Y}(z) = N_z^{-1} \sum_{i:Z_i=z} Y_i$ be the sample average of the outcomes under treatment $z$. The difference-in-means estimator $\hat{\tau}_{\text{N}} = \hat{Y}(1) - \hat{Y}(0)$ is unbiased for $\tau$ under complete randomization (Neyman 1923/1990), and affords a natural statistic for testing both $H_{0\text{F}}$ and $H_{0\text{N}}$. Algebraically, it equals the coefficient of $Z_i$ from the OLS fit of $Y_i$ on $(1, Z_i)$. It is also common to use $\hat{\tau}_{\text{N}}/\hat{\text{se}}_{\text{N}}$ or $\hat{\tau}_{\text{N}}/\tilde{\text{se}}_{\text{N}}$ as the test statistic, where $\hat{\text{se}}_{\text{N}}$ and $\tilde{\text{se}}_{\text{N}}$ are the classic and robust standard errors from the same OLS fit:

$$\hat{\text{se}}_{\text{N}}^2 = \frac{N(N_1-1)}{(N-2)N_1N_0}\hat{S}_1^2 + \frac{N(N_0-1)}{(N-2)N_1N_0}\hat{S}_0^2 \approx \frac{\hat{S}_1^2}{N_0} + \frac{\hat{S}_0^2}{N_1}, \quad \tilde{\text{se}}_{\text{N}}^2 = \frac{N_1-1}{N_1^2}\hat{S}_1^2 + \frac{N_0-1}{N_0^2}\hat{S}_0^2 \approx \frac{\hat{S}_1^2}{N_1} + \frac{\hat{S}_0^2}{N_0}$$

with $\hat{S}_z^2 = (N_z-1)^{-1}\sum_{i:Z_i=z}\{Y_i - \hat{Y}(z)\}^2$ for $z = 0, 1$ (Angrist and Pischke 2009). This yields three unadjusted test statistics as the baseline for discussing the possible improvement via covariate adjustment. The randomization distributions can then be generated by replacing $Z_i$ with $Z_{\pi(i)}$ in the above OLS fit over all $\pi \in \Pi$. In particular, we can compute $\hat{\tau}_{\text{N}}^{\pi}$ as the coefficient of $Z_{\pi(i)}$ from the OLS fit of $Y_i$ on $(1, Z_{\pi(i)})$ with $(\hat{\tau}_{\text{N}}/\hat{\text{se}}_{\text{N}})^{\pi} = \hat{\tau}_{\text{N}}^{\pi}/\hat{\text{se}}_{\text{N}}^{\pi}$ and $(\hat{\tau}_{\text{N}}/\tilde{\text{se}}_{\text{N}})^{\pi} = \hat{\tau}_{\text{N}}^{\pi}/\tilde{\text{se}}_{\text{N}}^{\pi}$, where $\tilde{\text{se}}_{\text{N}}^{\pi}$ and $\hat{\text{se}}_{\text{N}}^{\pi}$ are the corresponding classic and robust standard errors. The same intuition extends to the covariate-adjusted variants below as we shall introduce in a minute. Chung and Romano (2013) and Wu and Ding (2020) showed that randomization tests with the robustly studentized $\hat{\tau}_{\text{N}}/\tilde{\text{se}}_{\text{N}}$ are asymptotically valid for $H_{0\text{N}}$ under the super- and finite-population frameworks, respectively. We thus also consider studentization in covariate adjustment.

The first strategy for covariate adjustment is to fit an outcome model with covariates alone and use the residuals as the fixed, covariate-adjusted pseudo outcomes for conducting FRT. This appears to be the dominating approach advocated by Rosenbaum (2002); see also Gail et al. (1988), Raz (1990), Tukey (1993) and Ottoboni et al. (2018). Let $e = (e_1, \ldots, e_N)^{\text{T}}$ be the residuals from the OLS fit of $Y_i$ on $(1, x_i)$, which can be viewed as pseudo outcomes unaffected by the treatment under $H_{0\text{F}}$. The difference in means of the residuals, $\hat{\tau}_{\text{R}} = \hat{e}(1) - \hat{e}(0)$, equals the coefficient of $Z_i$ from the OLS fit of $e_i$ on $(1, Z_i)$ and affords an intuitive estimator of $\tau$ after adjusting for the covariates. Similar to the discussion of $\hat{\tau}_{\text{N}}$, we can use $\hat{\tau}_{\text{R}}$, $\hat{\tau}_{\text{R}}/\hat{\text{se}}_{\text{R}}$, and $\hat{\tau}_{\text{R}}/\tilde{\text{se}}_{\text{R}}$ as the test statistics for testing $H_{0\text{F}}$ or $H_{0\text{N}}$ by FRT, where $\hat{\text{se}}_{\text{R}}$ and $\tilde{\text{se}}_{\text{R}}$ are the classic and robust standard errors from the OLS fit that yields $\hat{\tau}_{\text{R}}$. We regress $Y_i$ on $(1, x_i)$ to form the residuals $e$ whereas Rosenbaum (2002) regressed $Y_i$ on $x_i$ alone without the intercept. The difference does not affect $\hat{\tau}_{\text{R}}$ with centered covariates.

The second strategy for covariate adjustment is to directly fit an outcome model with both the treatment and covariates, and use the coefficient or $t$-values of the treatment as the test statistic for FRT. Fisher (1935) suggested an estimator $\hat{\tau}_{\text{F}}$ for $\tau$, which equals the coefficient of $Z_i$ from the OLS

Table 1: Twelve test statistics where $\hat{\text{se}}$ and $\tilde{\text{se}}$ denote the classic and robust standard errors.

| | Neyman (1923/1990) | Rosenbaum (2002) | Fisher (1935) | Lin (2013) |
|---|---|---|---|---|
| unstudentized | $\hat{\tau}_{\text{N}}$ | $\hat{\tau}_{\text{R}}$ | $\hat{\tau}_{\text{F}}$ | $\hat{\tau}_{\text{L}}$ |
| studentized by $\hat{\text{se}}$ | $\hat{\tau}_{\text{N}}/\hat{\text{se}}_{\text{N}}$ | $\hat{\tau}_{\text{R}}/\hat{\text{se}}_{\text{R}}$ | $\hat{\tau}_{\text{F}}/\hat{\text{se}}_{\text{F}}$ | $\hat{\tau}_{\text{L}}/\hat{\text{se}}_{\text{L}}$ |
| studentized by $\tilde{\text{se}}$ | $\hat{\tau}_{\text{N}}/\tilde{\text{se}}_{\text{N}}$ | $\hat{\tau}_{\text{R}}/\tilde{\text{se}}_{\text{R}}$ | $\hat{\tau}_{\text{F}}/\tilde{\text{se}}_{\text{F}}$ | $\hat{\tau}_{\text{L}}/\tilde{\text{se}}_{\text{L}}$ |

fit of $Y_i$ on $(1, Z_i, x_i)$. Lin (2013) recommended an improved estimator, $\hat{\tau}_{\text{L}}$, as the coefficient of $Z_i$ from the OLS fit of $Y_i$ on $\{1, Z_i, x_i - \bar{x}, Z_i(x_i - \bar{x})\}$ with centered covariates and treatment-covariates interactions. These two covariate-adjusted estimators, along with their respective studentized variants, afford six additional test statistics, namely $\hat{\tau}_*$, $\hat{\tau}_*/\hat{\text{se}}_*$, and $\hat{\tau}_*/\tilde{\text{se}}_*$ $(* = \text{F}, \text{L})$, for testing $H_{0\text{F}}$ or $H_{0\text{N}}$ by FRT, where $\hat{\text{se}}_*$ and $\tilde{\text{se}}_*$ are the classic and robust standard errors from the respective OLS fits.

This gives us a total of twelve test statistics, three unadjusted and nine adjusted, for testing the treatment effects via FRT. Table 1 summarizes them, with the subscripts N, R, F, and L indicating Neyman (1923/1990), Rosenbaum (2002), Fisher (1935), and Lin (2013), respectively. All twelve statistics are finite-sample exact for testing $H_{0\text{F}}$ irrespective of whether the models are correctly specified or not. Our goal is to evaluate their abilities to preserve the correct type one error rates under $H_{0\text{N}}$. Without loss of generality, we assume two-sided FRT for the rest of the text, or, equivalently, we use the absolute values of the test statistics in Table 1 to compute the $p_{\text{FRT}}$ in (1).

The two strategies for covariate adjustment unify nicely under the OLS formulation yet differ materially with regard to the role of covariates under the permutations induced by the FRT procedure. The first strategy, on the one hand, adjusts for the covariates only once to form the pseudo outcomes $e$ and proceeds with the permutations in a covariate-free fashion. The second strategy, on the other hand, adjusts for the covariates in each of the $N!$ permutations of $Z$.

Before giving the formal results on the finite-population asymptotics, we unify below the three covariate-adjusted estimators as the differences in means of distinct adjusted outcomes. Let $S_x^2 = (N-1)^{-1}\sum_{i=1}^N x_i x_i^{\text{T}}$ and $\hat{S}_{xY} = (N-1)^{-1}\sum_{i=1}^N x_i Y_i$ be the finite-population covariance matrices of the centered $(x_i)_{i=1}^N$ with itself and $(Y_i)_{i=1}^N$, respectively. Let $\hat{\tau}_x = \hat{x}(1) - \hat{x}(0)$ be the difference in means of the covariates under treatment and control, where $\hat{x}(z) = N_z^{-1}\sum_{i:Z_i=z} x_i$. Let $\hat{S}^2_{x(z)} = (N_z - 1)^{-1}\sum_{i:Z_i=z}\{x_i - \hat{x}(z)\}\{x_i - \hat{x}(z)\}^{\text{T}}$ and $\hat{S}_{xY(z)} = (N_z-1)^{-1}\sum_{i:Z_i=z}\{x_i - \hat{x}(z)\}\{Y_i - \hat{Y}(z)\}$ be the sample covariance matrices of $x_i$ with itself and $Y_i$ under treatment $z$. Let $\hat{\gamma}_{\text{R}}$ and $\hat{\gamma}_{\text{F}}$ be the coefficients of $x_i$ from the OLS fits of $Y_i$ on $(1, x_i)$ and $(1, Z_i, x_i)$, respectively. Let $\hat{\gamma}_{\text{L}} = p_0\hat{\gamma}_{\text{L},1} + p_1\hat{\gamma}_{\text{L},0}$, where $\hat{\gamma}_{\text{L},z}$ is the coefficient of $x_i$ from the OLS fit of $Y_i$ on $(1, x_i)$ over the units under treatment $z$.

**Proposition 1.** We have

$$\hat{\tau}_* \;=\; N_1^{-1}\sum_{i:Z_i=1}(Y_i - x_i^{\text{T}}\hat{\gamma}_*) - N_0^{-1}\sum_{i:Z_i=0}(Y_i - x_i^{\text{T}}\hat{\gamma}_*) = \hat{\tau}_{\text{N}} - \hat{\tau}_x^{\text{T}}\hat{\gamma}_*, \quad (* = \text{R}, \text{F})$$

$$\hat{\tau}_{\text{L}} \;=\; N_1^{-1}\sum_{i:Z_i=1}(Y_i - x_i^{\text{T}}\hat{\gamma}_{\text{L},1}) - N_0^{-1}\sum_{i:Z_i=0}(Y_i - x_i^{\text{T}}\hat{\gamma}_{\text{L},0}) = \hat{\tau}_{\text{N}} - \hat{\tau}_x^{\text{T}}\hat{\gamma}_{\text{L}},$$

where $\hat{\gamma}_{\text{R}} = (S_x^2)^{-1}\hat{S}_{xY}$, $\hat{\gamma}_{\text{F}} = \hat{\gamma}_{\text{R}} - (1-1/N)^{-1}p_1p_0\hat{\tau}_{\text{F}}(S_x^2)^{-1}\hat{\tau}_x$, and $\hat{\gamma}_{\text{L},z} = (\hat{S}^2_{x(z)})^{-1}\hat{S}_{xY(z)}$.

Proposition 1 entails only the algebraic properties of the OLS fits and holds under arbitrary data generating process. It unifies $\hat{\tau}_*$ ($* = \text{R}, \text{F}, \text{L}$) as the difference-in-means estimators defined on the adjusted outcomes, or, equivalently, as $\hat{\tau}_{\text{N}}$ with corrections based on the imbalance in means of the covariates.

Under $H_{0\text{N}}$, FRT with $\hat{\tau}_{\text{N}}$ does not preserve the correct type one error rates but FRT with $\hat{\tau}_{\text{N}}/\tilde{\text{se}}_{\text{N}}$ does (Ding and Dasgupta 2018). In the next section, we will extend the result to the nine covariate-adjusted test statistics in Table 1 and establish the properness of the four robustly studentized $t$-statistics, namely $\hat{\tau}_*/\tilde{\text{se}}_*$ ($* = \text{N}, \text{R}, \text{F}, \text{L}$), for testing $H_{0\text{N}}$. Refer to them as the *robust $t$-statistics* hence. We will further show that among them, $\hat{\tau}_{\text{L}}/\tilde{\text{se}}_{\text{L}}$ delivers the highest power under alternative hypotheses.

**Remark 2.** Inspired by the distinction between $\hat{\tau}_{\text{F}}$ and $\hat{\tau}_{\text{L}}$ under the second covariate adjustment strategy, an alternative way to implement the first, pseudo-outcome-based strategy is to fit two separate OLS regressions of $Y_i$ on $x_i$ for the treated and control units, respectively, both without the intercept, and then use the resulting residuals for conducting FRT. Despite the computational advantage of this approach in that it adjusts for the covariates only once, the resulting tests lead to distinct sampling and randomization distributions even under $H_{0\text{F}}$, and are thus not finite-sample exact for testing $H_{0\text{F}}$. We see the finite-sample exactness under $H_{0\text{F}}$ the first criterion for a test to qualify as FRT, and thus do not pursue this direction. In particular, properness under $H_{0\text{N}}$ can be a rather weak requirement without the finite-sample exactness under $H_{0\text{F}}$. See Section **??** in the supplementary material for more examples of permutation tests of this type that we do not recommend in general.

# 3. Asymptotic theory for FRTs for testing $\tau = 0$

## 3.1. Limiting distributions under complete randomization

We will develop in Theorems 1–4 the limiting distributions of the twelve test statistics in Table 1 under the finite-population framework conditioning on $\mathcal{S} = \{Y_i(0), Y_i(1), x_i\}_{i=1}^N$. The theorems assume neither $H_{0\text{F}}$ nor $H_{0\text{N}}$ but hold for arbitrary $\mathcal{S}$ that satisfies the regularity conditions specified in Condition 1 below. Applying them to finite populations that actually satisfy $H_{0\text{N}}$ elucidates the asymptotic validity and power of FRT for testing $H_{0\text{N}}$ in Sections 3.2 and 3.3. Let $\bar{Y}(z) = N^{-1}\sum_{i=1}^N Y_i(z)$ and $S_z^2 = (N-1)^{-1}\sum_{i=1}^N\{Y_i(z)-\bar{Y}(z)\}^2$ be the finite-population mean and variance of $\{Y_i(z)\}_{i=1}^N$, respectively. Let $S_\tau^2 = (N-1)^{-1}\sum_{i=1}^N(\tau_i-\tau)^2$ be the finite-population variance of $(\tau_i)_{i=1}^N$. Let $S_{xY(z)} = (N-1)^{-1}\sum_{i=1}^N x_iY_i(z)$ be the finite-population covariance matrix of $\{x_i, Y_i(z)\}_{i=1}^N$. Let $w_i(z) = (S_x^2)^{-1}x_iY_i(z)$ with $\bar{w}(z) = N^{-1}\sum_{i=1}^N w_i(z) = (1-1/N)(S_x^2)^{-1}S_{xY(z)}$.

**Condition 1.** As $N \to \infty$, for $z = 0, 1$, (i) $p_z$ has a limit in $(0, 1)$, (ii) the first two moments of $\{Y_i(0), Y_i(1), x_i\}_{i=1}^N$ have finite limits; $S_x^2$ and its limit are both positive definite; $S_z^2 - S_{xY(z)}^{\mathrm{T}}(S_x^2)^{-1}S_{xY(z)}$ has a finite positive limit; the second moments of $\{w_i(0), w_i(1)\}_{i=1}^N$ have finite limits, and (iii) there exists a $c_0 < \infty$ independent of $N$ such that $N^{-1}\sum_{i=1}^N Y_i^4(z) \leq c_0$, $N^{-1}\sum_{i=1}^N \|x_i\|_4^4 \leq c_0$, and $N^{-1}\sum_{i=1}^N \|w_i(z)\|_4^4 \leq c_0$.

Condition 1(ii) ensures $S_\tau^2$ has a finite limit. We also use $p_z$, $\bar{Y}(z)$, $\tau$, $S_z^2$, $S_x^2$, $S_{xY(z)}$, and $S_\tau^2$ to denote their limiting values without introducing new symbols. The exact meaning should be clear from the context.

Denote by $P_Z$-a.s. a statement that holds for almost all sequences of $Z$. We review in Theorem 1 the asymptotic distributions of the three unadjusted test statistics from Ding and Dasgupta (2018), and extend them to the covariate-adjusted cases in Theorems 2–4.

**Theorem 1.** Assume Condition 1 and complete randomization.

(a) $\sqrt{N}(\hat{\tau}_{\mathrm{N}} - \tau) \rightsquigarrow \mathcal{N}(0, v_{\mathrm{N}})$, and $\sqrt{N}\hat{\tau}_{\mathrm{N}}^{\pi} \rightsquigarrow \mathcal{N}(0, v_{\mathrm{N0}})$ $P_Z$-a.s., where $v_{\mathrm{N}} = p_1^{-1}S_1^2 + p_0^{-1}S_0^2 - S_\tau^2$ and $v_{\mathrm{N0}} = p_0^{-1}S_1^2 + p_1^{-1}S_0^2 + \tau^2$.

(b) $(\hat{\tau}_{\mathrm{N}} - \tau)/\hat{\mathrm{se}}_{\mathrm{N}} \rightsquigarrow \mathcal{N}(0, c'_{\mathrm{N}})$, and $(\hat{\tau}_{\mathrm{N}}/\hat{\mathrm{se}}_{\mathrm{N}})^{\pi} \rightsquigarrow \mathcal{N}(0, 1)$ $P_Z$-a.s., where $c'_{\mathrm{N}} = v_{\mathrm{N}}/(v_{\mathrm{N0}} - \tau^2)$.

(c) $(\hat{\tau}_{\mathrm{N}} - \tau)/\tilde{\mathrm{se}}_{\mathrm{N}} \rightsquigarrow \mathcal{N}(0, c_{\mathrm{N}})$, and $(\hat{\tau}_{\mathrm{N}}/\tilde{\mathrm{se}}_{\mathrm{N}})^{\pi} \rightsquigarrow \mathcal{N}(0, 1)$ $P_Z$-a.s., where $c_{\mathrm{N}} = v_{\mathrm{N}}/(v_{\mathrm{N}} + S_\tau^2) \leq 1$.

Theorem 1 gives the asymptotic sampling and randomization distributions of $\hat{\tau}_{\mathrm{N}}$, $\hat{\tau}_{\mathrm{N}}/\hat{\mathrm{se}}_{\mathrm{N}}$, and $\hat{\tau}_{\mathrm{N}}/\tilde{\mathrm{se}}_{\mathrm{N}}$. Building up intuitions for Theorem 1 helps to understand Theorems 2–4 for the covariate-adjusted cases below.

First, it clarifies $\sqrt{N}\hat{\tau}_{\mathrm{N}}$ and $\sqrt{N}\hat{\tau}_{\mathrm{N}}^{\pi}$ as both asymptotically normal with asymptotic variances $v_{\mathrm{N}}$ and $v_{\mathrm{N0}}$ that are in general not equal. Recall from the definition of randomization distribution that the distribution of $\hat{\tau}_{\mathrm{N}}^{\pi}$ is always conditional on the assignment vector $Z$. The fact that $\sqrt{N}\hat{\tau}_{\mathrm{N}}^{\pi} \rightsquigarrow \mathcal{N}(0, v_{\mathrm{N0}})$ $P_Z$-a.s. suggests this randomization distribution is asymptotically identical for almost all sequences of $Z$.

Second, the component $S_\tau^2$ in $v_{\mathrm{N}}$ is unique to the finite-population inference and cannot be estimated consistently from the observed outcomes. The resulting inferences are thus necessarily conservative unless $\tau_i = \tau$ for all $i = 1, \ldots, N$ (Neyman 1923/1990). The classic and robust standard errors afford two convenient estimators of $\mathrm{cov}(\hat{\tau}_{\mathrm{N}})$ with

$$N\hat{\mathrm{se}}_{\mathrm{N}}^2 = v_{\mathrm{N0}} - \tau^2 + o_P(1), \quad N\tilde{\mathrm{se}}_{\mathrm{N}}^2 = v_{\mathrm{N}} + S_\tau^2 + o_P(1) \tag{2}$$

by Lemma **??** in the supplementary material. This implies $\tilde{\mathrm{se}}_{\mathrm{N}}^2$ is asymptotically conservative for $\mathrm{cov}(\hat{\tau}_{\mathrm{N}})$ whereas $\hat{\mathrm{se}}_{\mathrm{N}}^2$ in general is not, entailing the asymptotic sampling distributions of the classic and robust $t$-statistics with variances $c'_{\mathrm{N}}$ and $c_{\mathrm{N}}$, respectively. In contrast, the super-population framework has no analog of $S_\tau^2$ such that $\tilde{\mathrm{se}}_{\mathrm{N}}^2$ is consistent for the true sampling variance (Chung and Romano 2013).

Further, Theorem 1 does not assume $H_{0\text{N}}$ or $H_{0\text{F}}$ but holds for arbitrary $\mathcal{S}$ that satisfies Condition 1. The null hypotheses $H_{0\text{N}}$ and $H_{0\text{F}}$ impose restrictions on $\mathcal{S}$, and thereby enable more informative comparisons between the asymptotic sampling and randomization distributions. The weak null hypothesis $H_{0\text{N}}$, on the one hand, ensures $\tau = 0$ such that all six normal distributions center at zero with $v_{\text{N}0} = p_0^{-1}S_1^2 + p_1^{-1}S_0^2$ and $c'_{\text{N}} = v_{\text{N}}/v_{\text{N}0}$. Compare the expressions of $v_{\text{N}}$ and $v_{\text{N}0}$ to see that the asymptotic distributions of $\hat{\tau}_{\text{N}}$ and $\hat{\tau}^{\pi}_{\text{N}}$ still differ unless $(p_1 - p_0)(S_1^2 - S_0^2) + S_\tau^2 = 0$. This demonstrates that $p_{\text{FRT}}$ can be invalid even asymptotically. Compare the distributions of $\hat{\tau}_{\text{N}}/\tilde{\text{se}}_{\text{N}}$ and $(\hat{\tau}_{\text{N}}/\tilde{\text{se}}_{\text{N}})^{\pi}$ to see $|\hat{\tau}_{\text{N}}/\tilde{\text{se}}_{\text{N}}| \leq_{\text{st}} |(\hat{\tau}_{\text{N}}/\tilde{\text{se}}_{\text{N}})^{\pi}|$ asymptotically $P_Z$-a.s.. This underpins the asymptotic validity of FRT with $\hat{\tau}_{\text{N}}/\tilde{\text{se}}_{\text{N}}$ for testing $H_{0\text{N}}$ as we shall elaborate in more detail in Section 3.2.

The strong null hypothesis $H_{0\text{F}}$, on the other hand, further entails $S_0^2 = S_1^2$ and $S_\tau^2 = 0$ such that $v_{\text{N}} = v_{\text{N}0} = (p_0 p_1)^{-1}S^2$ and $c_{\text{N}} = c'_{\text{N}} = 1$, where $S^2$ denotes the common value of $S_0^2$ and $S_1^2$. The resulting identical asymptotic distributions of $\hat{\tau}_{\text{N}}$ and $\hat{\tau}^{\pi}_{\text{N}}$ are a trivial consequence of their exact equivalence in finite samples. From (2), the strong null hypothesis also ensures the asymptotic equivalence of the classic and robust standard errors.

Last but not least, the way in which FRT is conducted ensures $(\hat{\tau}_{\text{N}}/\hat{\text{se}}_{\text{N}})^{\pi}$ and $(\hat{\tau}_{\text{N}}/\tilde{\text{se}}_{\text{N}})^{\pi}$ converge in distribution to standard normal. Recall the pseudo finite population $\{Y'_i(0), Y'_i(1), x_i\}_{i=1}^N$ with $Y'_i(0) = Y'_i(1) = Y_i$ from which the FRT procedure generates the randomization distributions of all three test statistics. It satisfies $H_{0\text{F}}$, so the same intuition from the last paragraph extends here and ensures the consistency of both $N(\hat{\text{se}}^{\pi}_{\text{N}})^2$ and $N(\tilde{\text{se}}^{\pi}_{\text{N}})^2$ for estimating the asymptotic variance of $\sqrt{N}\hat{\tau}^{\pi}_{\text{N}}$. This guarantees the convergence of $(\hat{\tau}_{\text{N}}/\hat{\text{se}}_{\text{N}})^{\pi}$ and $(\hat{\tau}_{\text{N}}/\tilde{\text{se}}_{\text{N}})^{\pi}$ to the standard normal irrespective of the true value of $\tau$. The same intuition carries over to the nine covariate-adjusted test statistics with the original potential outcomes replaced by the adjusted counterparts. We formalize the idea in Theorems 2–4.

Let $\gamma_z = (S_x^2)^{-1}S_{xY(z)}$ be the coefficient of $x_i$ from the OLS fit of $Y_i(z)$ on $(1, x_i)$. Let $a_i(z) = Y_i(z) - \bar{Y}(z) - x_i^{\text{T}}(p_1\gamma_1 + p_0\gamma_0)$ be the adjusted potential outcomes under treatment $z$, with finite-population mean zero and variance $S^2_{a(z)}$. Theorem 2 gives the asymptotic distributions of $\hat{\tau}_{\text{R}}$, $\hat{\tau}_{\text{R}}/\hat{\text{se}}_{\text{R}}$, and $\hat{\tau}_{\text{R}}/\tilde{\text{se}}_{\text{R}}$ from the first, pseudo-outcome-based strategy for covariate adjustment.

**Theorem 2.** Assume Condition 1 and complete randomization.

(a) $\sqrt{N}(\hat{\tau}_{\text{R}} - \tau) \rightsquigarrow \mathcal{N}(0, v_{\text{R}})$, and $\sqrt{N}\hat{\tau}^{\pi}_{\text{R}} \rightsquigarrow \mathcal{N}(0, v_{\text{R}0})$ $P_Z$-a.s., where $v_{\text{R}} = p_1^{-1}S^2_{a(1)} + p_0^{-1}S^2_{a(0)} - S_\tau^2$ and $v_{\text{R}0} = p_0^{-1}S^2_{a(1)} + p_1^{-1}S^2_{a(0)} + \tau^2$.

(b) $(\hat{\tau}_{\text{R}} - \tau)/\hat{\text{se}}_{\text{R}} \rightsquigarrow \mathcal{N}(0, c'_{\text{R}})$, and $(\hat{\tau}_{\text{R}}/\hat{\text{se}}_{\text{R}})^{\pi} \rightsquigarrow \mathcal{N}(0, 1)$ $P_Z$-a.s., where $c'_{\text{R}} = v_{\text{R}}/(v_{\text{R}0} - \tau^2)$.

(c) $(\hat{\tau}_{\text{R}} - \tau)/\tilde{\text{se}}_{\text{R}} \rightsquigarrow \mathcal{N}(0, c_{\text{R}})$, and $(\hat{\tau}_{\text{R}}/\tilde{\text{se}}_{\text{R}})^{\pi} \rightsquigarrow \mathcal{N}(0, 1)$ $P_Z$-a.s., where $c_{\text{R}} = v_{\text{R}}/(v_{\text{R}} + S_\tau^2) \leq 1$.

Interestingly, Theorem 2 also holds for $\hat{\tau}_{\text{F}}, \hat{\tau}_{\text{F}}/\hat{\text{se}}_{\text{F}}$, and $\hat{\tau}_{\text{F}}/\tilde{\text{se}}_{\text{F}}$ from the second, model-output-based strategy. This echos the numeric result from Proposition 1, which implies that the difference between $\hat{\gamma}_{\text{F}}$ and $\hat{\gamma}_{\text{R}}$ is of higher order under complete randomization.

**Theorem 3.** Theorem 2 holds if we replace all the subscripts R with F.

The asymptotic equivalence of $\hat{\tau}_{\text{R}}$ and $\hat{\tau}_{\text{F}}$ is perhaps no surprise after all, despite the distinction in procedure. Both statistics use a common coefficient, namely $\hat{\gamma}_{\text{R}}$ and $\hat{\gamma}_{\text{F}}$, to adjust the observed outcomes under both treatment and control, and estimate this coefficient using the pooled data. Such practice, despite expeditious, can be problematic in experiments with unequal group sizes and heterogeneous treatment effects with respect to covariates (Freedman 2008).

Lin (2013)'s estimator, on the other hand, accommodates separate adjustments for outcomes under treatment and control evident from Proposition 1. Let $b_i(z) = Y_i(z) - \bar{Y}(z) - x_i^{\text{T}}\gamma_z$ be the adjusted potential outcomes under treatment-specific coefficient $\gamma_z$, with mean zero and finite-population variance $S^2_{b(z)}$. Let $S^2_\xi$ be the finite-population variance of $\xi_i = b_i(1) - b_i(0)$ for $i = 1, \ldots, N$. Theorem 4 gives the asymptotic distributions of $\hat{\tau}_{\text{L}}$, $\hat{\tau}_{\text{L}}/\hat{\text{se}}_{\text{L}}$, and $\hat{\tau}_{\text{L}}/\tilde{\text{se}}_{\text{L}}$.

**Theorem 4.** Assume Condition 1 and complete randomization.

(a) $\sqrt{N}(\hat{\tau}_{\text{L}} - \tau) \rightsquigarrow \mathcal{N}(0, v_{\text{L}})$, and $\sqrt{N}\hat{\tau}^{\pi}_{\text{L}} \rightsquigarrow \mathcal{N}(0, v_{\text{L}0})$ $P_Z$-a.s., where $v_{\text{L}} = p_1^{-1}S^2_{b(1)} + p_0^{-1}S^2_{b(0)} - S^2_\xi$ and $v_{\text{L}0} = v_{\text{R}0} = v_{\text{F}0} = p_0^{-1}S^2_{a(1)} + p_1^{-1}S^2_{a(0)} + \tau^2$.

(b) $(\hat{\tau}_{\text{L}} - \tau)/\hat{\text{se}}_{\text{L}} \rightsquigarrow \mathcal{N}(0, c'_{\text{L}})$, and $(\hat{\tau}_{\text{L}}/\hat{\text{se}}_{\text{L}})^{\pi} \rightsquigarrow \mathcal{N}(0, 1)$ $P_Z$-a.s., where $c'_{\text{L}} = v_{\text{L}}/(p_0^{-1}S^2_{b(1)} + p_1^{-1}S^2_{b(0)})$.

(c) $(\hat{\tau}_{\text{L}} - \tau)/\tilde{\text{se}}_{\text{L}} \rightsquigarrow \mathcal{N}(0, c_{\text{L}})$, and $(\hat{\tau}_{\text{L}}/\tilde{\text{se}}_{\text{L}})^{\pi} \rightsquigarrow \mathcal{N}(0, 1)$ $P_Z$-a.s., where $c_{\text{L}} = v_{\text{L}}/(v_{\text{L}} + S^2_\xi) \leq 1$.

The asymptotic variance of $\hat{\tau}_{\text{L}}$ is less than or equal to $v_{\text{F}} = v_{\text{R}}$ (Lin 2013), but those of the randomization distributions are all equal, $v_{\text{L}0} = v_{\text{R}0} = v_{\text{F}0}$. Similar to the comments on the pseudo finite population after Theorem 1, this is due to the mixing of the treated and control outcomes in the FRT procedure, which effectively results in covariate adjustment based on the pooled data even in constructing Lin (2013)'s estimator. In fact, Lemma **??** in the supplementary material gives a stronger result that $\hat{\tau}^{\pi}_*$ $(* = \text{R}, \text{F}, \text{L})$ are all asymptotically equivalent.

The asymptotic sampling distributions in Theorems 3 and 4 are not new (Freedman 2008; Lin 2013), but the randomization distributions are. Both the asymptotic sampling and randomization distributions of $\hat{\tau}_{\text{R}}$, $\hat{\tau}_{\text{R}}/\hat{\text{se}}_{\text{R}}$, and $\hat{\tau}_{\text{R}}/\tilde{\text{se}}_{\text{R}}$ in Theorem 2 are new. The analysis of the randomization distributions builds upon the existing sampling distributions but requires additional technical tools, such as the finite-population strong law of large numbers. We unify the existing and new results in the above four theorems to facilitate discussions on the asymptotic validity.

Technically, Condition 1 requires more moments than the usual asymptotic analysis of $\hat{\tau}_*$ $(* = \text{N}, \text{F}, \text{L})$. This is due to the strong statement of the almost sure convergence of the randomization distributions in Theorems 1–4. This is sufficient but unnecessary for showing that FRT controls the asymptotic type one error rates, which only requires that the quantiles of the asymptotic randomization distribution are greater than or equal to those of the asymptotic sampling distribution. We can use the *subsequence argument*, a standard proving device for the bootstrap (van der vaart and Wellner 1996), to relax the moment conditions. However, we keep the current version of Condition 1 to simplify the statements of the theorems and their proofs.

**Remark 3.** All results in Theorems 1–4 extend to the super-population framework under independent treatment assignments with minor modifications. A key distinction is that the robust

standard error $\tilde{\text{se}}_{\text{L}}$ must be modified to ensure consistency (Berk et al. 2013; Negi and Wooldridge 2021). Motivated by the similarity in procedure, we also evaluate the design-based properties of five existing permutation tests originally for linear models and show the superiority of the proposed FRT for testing the treatment effects (DiCiccio and Romano 2017; Freedman and Lane 1983; Kennedy 1995; ter Braak 1992; Manly 1997). We relegate the details to Section **??** in the supplementary material.

### 3.2. Asymptotic validity for testing $\tau = 0$

Theorems 1–4 establish the sampling and randomization distributions for all twelve test statistics in Table 1 as asymptotically normal. A statistic as such is proper under two-sided FRT if under $H_{0\text{N}}$, the asymptotic variance of its randomization distribution is greater than or equal to that of its sampling distribution. In general, $v_*/v_{*0}$ and $c'_*$ can be either greater or less than 1, suggesting the improperness of $\hat{\tau}_*$ and $\hat{\tau}_*/\hat{\text{se}}_*$ for $* = \text{N}, \text{R}, \text{F}, \text{L}$. On the other hand, $c_* \leq 1$, ensuring the properness of $\hat{\tau}_*/\tilde{\text{se}}_*$ for $* = \text{N}, \text{R}, \text{F}, \text{L}$.

**Corollary 1.** Assume Condition 1 and complete randomization. The robust $t$-statistics $\hat{\tau}_*/\tilde{\text{se}}_*$ ($* = \text{N}, \text{R}, \text{F}, \text{L}$) are the only test statistics in Table 1 proper for testing $H_{0\text{N}}$ via FRT.

Corollary 1 highlights the necessity of robust studentization in constructing asymptotically valid FRT for testing $H_{0\text{N}}$. The other eight test statistics may also preserve the correct type one error rates asymptotically with additional conditions on $(p_0, p_1)$ or $\mathcal{S}$. The former is within the control of the designer whereas the latter is not.

**Corollary 2.** Assume Condition 1 and complete randomization. As $N$ goes to infinity,

(a) all twelve test statistics in Table 1 preserve the correct type one error rates if $p_0 = p_1 = 1/2$ or $\tau_i = \tau$ for all $i = 1, \ldots, N$;

(b) $\hat{\tau}_{\text{N}}$ and $\hat{\tau}_{\text{N}}/\hat{\text{se}}_{\text{N}}$ preserve the correct type one error rates if $S_0^2 = S_1^2$; $\hat{\tau}_{\text{R}}$, $\hat{\tau}_{\text{R}}/\hat{\text{se}}_{\text{R}}$, $\hat{\tau}_{\text{F}}$, and $\hat{\tau}_{\text{F}}/\hat{\text{se}}_{\text{F}}$ do so if $S^2_{a(0)} = S^2_{a(1)}$; $\hat{\tau}_{\text{L}}$ and $\hat{\tau}_{\text{L}}/\hat{\text{se}}_{\text{L}}$ do so if $S^2_{b(0)} = S^2_{b(1)}$.

Corollary 2 states the properness of the unstudentized coefficients and classic $t$-statistics when $p_0 = p_1 = 1/2$. The result echos the fact that for the usual two-sample $t$-test, one may use either the pooled or unpooled estimate of the variance whenever the ratio of the two sample sizes tends to one. Freedman (2008) and Lin (2013) discovered this result in complete randomization under the finite-population inference; Bugni et al. (2018) discovered parallel results in covariate-adaptive randomization under the super-population inference.

### 3.3. Insights for power under alternative hypotheses

A natural next question is the relative power of FRT with the four robust $t$-statistics under alternative hypotheses. Recall that Theorems 1–4 hold for arbitrary $\tau$. A deviation from $H_{0\text{N}}$ shifts the center of $\hat{\tau}_*/\tilde{\text{se}}_*$ while leaving its asymptotic randomization distribution intact at $\mathcal{N}(0, 1)$. With

$|\tau|/\tilde{\text{se}}_*$ tending to $\infty$ for any fixed $\tau \neq 0$ as $N$ goes to infinity, all four statistics would have power converging to 1 under the alternative hypothesis for fixed $\tau \neq 0$, with the relative power determined by $|\hat{\tau}_*/\tilde{\text{se}}_*|$ as the distance between the observed value of the test statistics, $\hat{\tau}_*/\tilde{\text{se}}_*$, and the center of the reference distribution, namely 0. With $\hat{\tau}_*$ converging to $\tau$ in probability for all $* = \text{N}, \text{R}, \text{F}, \text{L}$, a heuristic argument is that the smaller the robust standard error, the higher the asymptotic relative power. We give in Corollary 3 the relative order of the robust standard errors as $N$ goes to infinity, and demonstrate its heuristic relation with the power via simulation in Section 5. Rigorous quantification of the power entails specification of the alternative distributions (Lehmann 1975; Rosenbaum 2010). We leave the technical details to future work.

**Corollary 3.** Assume complete randomization and Condition 1. We have

$$\frac{\tilde{\text{se}}_*^2}{\tilde{\text{se}}_{\text{N}}^2} - \frac{p_1^{-1}S_{a(1)}^2 + p_0^{-1}S_{a(0)}^2}{p_1^{-1}S_1^2 + p_0^{-1}S_0^2} = o(1) \quad \text{for } * = \text{R and F}, \qquad \frac{\tilde{\text{se}}_{\text{L}}^2}{\tilde{\text{se}}_{\text{N}}^2} - \frac{p_1^{-1}S_{b(1)}^2 + p_0^{-1}S_{b(0)}^2}{p_1^{-1}S_1^2 + p_0^{-1}S_0^2} = o(1)$$

hold $P_Z$-a.s., with the limiting values of $\tilde{\text{se}}_{\text{L}}^2/\tilde{\text{se}}_*^2$ less than or equal to 1 for $* = \text{N}, \text{R}, \text{F}$.

With $S_{b(z)}^2 \leq S_{a(z)}^2$ and $S_{b(z)}^2 \leq S_z^2$ for $z = 0, 1$, Corollary 3 ensures $\tilde{\text{se}}_{\text{L}}$ has the smallest limiting value among $\tilde{\text{se}}_*$ ($* = \text{N}, \text{R}, \text{F}, \text{L}$), and thereby ensures $\hat{\tau}_{\text{L}}/\tilde{\text{se}}_{\text{L}}$ has the highest power asymptotically. The limiting values of $\tilde{\text{se}}_{\text{R}}$ and $\tilde{\text{se}}_{\text{F}}$, on the other hand, can be even greater than that of $\tilde{\text{se}}_{\text{N}}$. This mirrors the asymptotic efficiency theory of point estimation and suggests $\hat{\tau}_{\text{R}}/\tilde{\text{se}}_{\text{R}}$ and $\hat{\tau}_{\text{F}}/\tilde{\text{se}}_{\text{F}}$ can be even less powerful than $\hat{\tau}_{\text{N}}/\tilde{\text{se}}_{\text{N}}$ despite the extra use of covariates (Freedman 2008; Lin 2013). See also Fogarty (2018a, Corollary 1) for analogous results in the context of finely stratified experiments.

FRT with $\hat{\tau}_{\text{L}}/\tilde{\text{se}}_{\text{L}}$, as a result, is finite-sample exact for testing $H_{0\text{F}}$, asymptotically valid for testing $H_{0\text{N}}$, and enjoys the highest power under alternatives, all irrespective of whether the linear model that generates it is correctly specified or not. It is thus our final recommendation for testing both $H_{0\text{F}}$ and $H_{0\text{N}}$ by FRT.

### 3.4. Confidence interval by inverting FRTs

We next extend the theory from testing hypotheses to constructing confidence intervals. This is conceptually straightforward given their duality. Consider using FRT to test $H_{0\text{N}}(c) : \tau = c$. We can pretend to be testing a strong null hypothesis of constant effect, $H_{0\text{F}}(c) : \tau_i = c$ for all $i = 1, \ldots, N$, and compute the $p$-value, denoted by $p_{\text{FRT}}(c)$, by using $Y - cZ$ as the fixed outcomes for FRT. Inverting a sequence of such FRTs on a bounded set of the possible values of $c$ yields $\text{CI}_{\text{FRT},\alpha}$ as a tentative interval estimator for the average treatment effect $\tau$. By duality, it is an asymptotic $1-\alpha$ confidence interval for $\tau$ if we use the robust $t$-statistics to perform the FRTs. Duality further suggests the one based on $\hat{\tau}_{\text{L}}/\tilde{\text{se}}_{\text{L}}$ to have the smallest width asymptotically.

Alternatively, the robust Wald-type confidence intervals $(\hat{\tau}_* - q_{1-\alpha/2}\tilde{\text{se}}_*, \hat{\tau}_* + q_{1-\alpha/2}\tilde{\text{se}}_*)$ ($* = \text{N}, \text{F}, \text{R}, \text{L}$) cover $\tau$ with probability approaching $1-\alpha$ as $N$ goes to infinity, where $q_{1-\alpha/2}$ is the $(1-\alpha/2)$th quantile of the standard normal. These confidence intervals are asymptotically identical to $\text{CI}_{\text{FRT},\alpha}$ based on the robust $t$-statistics. They are convenient approximations for $\text{CI}_{\text{FRT},\alpha}$ which

can be used as initial values in the grid search over $c$. We recommend using $\text{CI}_{\text{FRT},\alpha}$ based on $\hat{\tau}_{\text{L}}/\widetilde{\text{se}}_{\text{L}}$ because of its multiple guarantees: it has finite-sample exact coverage rate when $\tau_i = \tau$ for all $i = 1, \ldots, N$, has correct asymptotic coverage rate when $\tau_i$'s vary, and has smaller width compared to the confidence interval without covariate adjustment.

# 4. Extensions to other experimental designs

## 4.1. Cluster randomization

Consider $N$ units nested in $M$ clusters of sizes $n_i$ ($i = 1, \ldots, M$; $\sum_{i=1}^{M} n_i = N$). The average cluster size is $\bar{n} = N/M$. Cluster randomization randomly assigns $M_1$ clusters to receive the treatment and the rest $M_0 = M - M_1$ clusters to receive the control. Let $x_{ij}$ and $\{Y_{ij}(z) : z = 0, 1\}$ be the covariate and potential outcomes for the $j$th unit in cluster $i$ ($i = 1, \ldots, M$; $j = 1, \ldots, n_i$). The average treatment effect equals

$$\tau = N^{-1} \sum_{i=1}^{M} \sum_{j=1}^{n_i} \{Y_{ij}(1) - Y_{ij}(0)\} = M^{-1} \sum_{i=1}^{M} \{\tilde{Y}_{i\cdot}(1) - \tilde{Y}_{i\cdot}(0)\}, \tag{3}$$

where $\tilde{Y}_{i\cdot}(z) = \sum_{j=1}^{n_i} Y_{ij}(z)/\bar{n}$ is the cluster total of potential outcomes scaled by $1/\bar{n}$.

Let $Z_i$ be the treatment level received by cluster $i$. The observed outcome for unit $ij$ is $Y_{ij} = Z_i Y_{ij}(1) + (1 - Z_i) Y_{ij}(0)$. Let $\tilde{Y}_{i\cdot} = \sum_{j=1}^{n_i} Y_{ij}/\bar{n}$ and $\tilde{x}_{i\cdot} = \sum_{j=1}^{n_i} x_{ij}/\bar{n}$ be the scaled cluster totals of observed outcomes and covariates in cluster $i$. Then $\tilde{Y}_{i\cdot} = Z_i \tilde{Y}_{i\cdot}(1) + (1 - Z_i)\tilde{Y}_{i\cdot}(0)$ gives the observed analog of $\tilde{Y}_{i\cdot}(z)$ under cluster randomization. This, together with the expression of $\tau$ from (3), suggests the equivalence of $(Z_i, \tilde{Y}_{i\cdot}, \tilde{x}_{i\cdot})_{i=1}^{M}$ to data from a complete randomization with potential outcomes $\{\tilde{Y}_{i\cdot}(0), \tilde{Y}_{i\cdot}(1)\}_{i=1}^{M}$ and average treatment effect $\tau$ (Middleton and Aronow 2015; Li and Ding 2017), and allows us to derive results in parallel with Theorems 1–4 under a modified version of Condition 1 on cluster totals by replacing $Y_i(z)$ with $\tilde{Y}_{i\cdot}(z)$, $x_i$ with $\tilde{x}_{i\cdot}$, and $(N, N_0, N_1)$ with $(M, M_0, M_1)$ as $M$ goes to infinity. This requires a large number of clusters to ensure the accuracy of the asymptotic approximation. See Su and Ding (2021) for more subtle requirements on the cluster sizes when the regularity conditions are given in terms of the individual potential outcomes.

## 4.2. Stratified randomization

Consider $N$ units in $K$ strata of sizes $N_{[k]}$ ($k = 1, \ldots, K$; $\sum_{k=1}^{K} N_{[k]} = N$). Stratified randomization conducts an independent complete randomization in each stratum, and assigns at complete random $N_{[k]z}$ units to receive treatment $z$ in stratum $k$ ($k = 1, \ldots, K$; $z = 0, 1$). Denote by $x_{[k]i}$, $\{Y_{[k]i}(z) : z = 0, 1\}$, and $Z_{[k]i}$ the covariate, potential outcomes, and treatment indicator for the $i$th unit in stratum $k$ ($k = 1, \ldots, K$; $i = 1, \ldots, N_{[k]}$). The randomization scheme independently draws $Z_{[k]} = (Z_{[k]1}, \ldots, Z_{[k]N_{[k]}})^{\text{T}}$ as a random permutation of $N_{[k]1}$ 1's and $N_{[k]0}$ 0's for $k = 1, \ldots, K$.

Equivalently, it draws $Z = (Z_{[1]}^{\mathrm{T}}, \ldots, Z_{[K]}^{\mathrm{T}})^{\mathrm{T}}$ uniformly from the set

$$\mathcal{Z}_{\mathrm{str}} = \left\{ \boldsymbol{z} = (z_{[k]i}) \in \{0,1\}^N : \sum_{i=1}^{N_{[k]}} z_{[k]i} = N_{[k]1} \text{ for } k = 1, \ldots, K \right\}$$

subject to the stratum-wise treatment size restriction. Let $Y$ and $X$ be the vectorization and concatenation of $Y_{[k]i}(z)$'s and $x_{[k]i}(z)$'s, respectively. For arbitrary test statistic $T(Z, Y, X)$, a two-sided FRT under stratified randomization permutes the treatment vector $Z$ within $\mathcal{Z}_{\mathrm{str}}$, and computes the $p$-value as

$$p_{\mathrm{FRT,str}} = |\mathcal{Z}_{\mathrm{str}}|^{-1} \sum_{\pi: Z_\pi \in \mathcal{Z}_{\mathrm{str}}} 1\{|T(Z_\pi, Y, X)| \geq |T(Z, Y, X)|\}.$$

The finite-population average treatment effect equals

$$\tau = N^{-1} \sum_{k=1}^{K} \sum_{i=1}^{N_{[k]}} \{Y_{[k]i}(1) - Y_{[k]i}(0)\} = \sum_{k=1}^{K} \omega_{[k]} \tau_{[k]},$$

where $\omega_{[k]} = N_{[k]}/N$ and $\tau_{[k]} = N_{[k]}^{-1} \sum_{i=1}^{N_{[k]}} \{Y_{[k]i}(1) - Y_{[k]i}(0)\}$ define the relative size and stratum-wise average treatment effect of stratum $k$, respectively. Of interest is the choice of the test statistic that ensures valid test of $H_{0\mathrm{N}} : \tau = 0$ via FRT.

Assume $\hat{\tau}_{*[k]}$ and $\tilde{\mathrm{se}}_{*[k]}$ as the basic estimator and robust standard error obtained from stratum $k$, where $*$ can be N, R, F, and L. The weighted average $\hat{\tau}_* = \sum_{k=1}^{K} \omega_{[k]} \hat{\tau}_{*[k]}$, with a slight abuse of notation, affords an intuitive estimator of $\tau$ with squared robust standard error $\tilde{\mathrm{se}}_*^2 = \sum_{k=1}^{K} \omega_{[k]}^2 \tilde{\mathrm{se}}_{*[k]}^2$. The abuse of notation causes little confusion because $\hat{\tau}_*$ and $\tilde{\mathrm{se}}_*$ reduce to their definitions under complete randomization when $K = 1$. This suggests $\hat{\tau}_*/\tilde{\mathrm{se}}_*$ ($* = \mathrm{N}, \mathrm{R}, \mathrm{F}, \mathrm{L}$) as four intuitive choices of the test statistic for testing both $H_{0\mathrm{F}}$ and $H_{0\mathrm{N}}$ under stratified randomization. The properness of $\hat{\tau}_*/\tilde{\mathrm{se}}_*$ for testing $H_{0\mathrm{N}}$ is a direct application of Theorems 1–4.

**Corollary 4.** Assume stratified randomization and Condition 1 holds within all strata $k = 1, \ldots, K$. We have $(\hat{\tau}_* - \tau)/\tilde{\mathrm{se}}_* \rightsquigarrow \mathcal{N}(0, c_*)$, and $(\hat{\tau}_*/\tilde{\mathrm{se}}_*)^\pi \rightsquigarrow \mathcal{N}(0,1)$ $P_Z$-a.s., where $c_* \leq 1$, for $* = \mathrm{N}, \mathrm{R}, \mathrm{F}, \mathrm{L}$.

Bugni et al. (2018) focused on covariate-adaptive experiments in which the proportions of the treatment are homogeneous across strata. They showed that in those covariate-adaptive experiments, one can form a simpler estimator from the OLS fit of the outcome on the treatment and the stratum indicators. When the proportions of treatment vary across strata, this simple OLS fit yields inconsistent estimator for $\tau$. To address this issue, Bugni et al. (2019) proposed to use the weighted average of the stratum-specific estimators, and Liu and Yang (2020) and Ye et al. (2020) discussed covariate adjustment allowing for additional covariates beyond the stratum indicators. We further their theory to FRTs, and use the weighted average of the stratum-specific estimators to allow for the proportions of treatment to vary across strata.

Even if the original experiment is completely randomized, if a discrete covariate $X$ is available, we can condition on the numbers of treated and control units landing in each stratum. The resulting assignment mechanism is identical to stratified randomization, such that we can permute the subvector of $Z$ within each stratum of $X$ as if the original experiment were stratified. This is known as the *conditional randomization test*. Zheng and Zelen (2008) and Hennessy et al. (2016) observed that they typically enhance the power if the covariates are predictive of the outcomes.

Among the four robust $t$-statistics, $\hat{\tau}_{\text{N}}/\widetilde{\text{se}}_{\text{N}}$ is the simplest and $\hat{\tau}_{\text{L}}/\widetilde{\text{se}}_{\text{L}}$ is the most powerful. Corollary 4 assumes $N_{[k]}$ goes to infinity for each $k$. With a large number of small strata, we need to modify the test statistic and the asymptotic scheme (Liu and Yang 2020). Since this involves different technical tools, we defer the technical details to future work.

## 4.3. Rerandomization

### 4.3.1. FRT with rerandomization

Rerandomization, termed by Cox (1982) and Morgan and Rubin (2012), samples the treatment indicators under covariate balance constraints. Bruhn and McKenzie (2009) reviewed several field experiments in economics and suggested that rerandomization is widespread although often poorly documented. Banerjee et al. (2020) discussed the pros and cons of such experimental design. Although rerandomization can improve covariate balance, it also imposes challenges for the subsequent data analysis. We focus here on a special rerandomization that uses the Mahalanobis distance between covariate means as the balance criterion, known as ReM. Although it might not be the exact rerandomization used in field experiments in economics, it has nice statistical properties that allow for simple analysis of the experimental data.

The designer of ReM accepts a treatment vector $Z$ if and only if

$$\mathcal{A}: \quad \hat{\tau}_x^{\text{T}}\{\text{cov}(\hat{\tau}_x)\}^{-1}\hat{\tau}_x < a \tag{4}$$

for a predetermined constant $a$. Let $\mathcal{Z}_a = \{\boldsymbol{z} : \boldsymbol{z} \in \mathcal{Z} \text{ satisfies (4)}\}$ be the set of acceptable assignments under threshold $a$. The sampling distribution of the test statistic $T$ is uniform over $\{T(\boldsymbol{z}, Y(\boldsymbol{z}), X) : \boldsymbol{z} \in \mathcal{Z}_a\}$. FRT under ReM proceeds by permuting $Z$ within $\mathcal{Z}_a$ and computes the $p$-value as

$$p_{\text{FRT},\mathcal{A}} = |\mathcal{Z}_a|^{-1} \sum_{\pi: Z_\pi \in \mathcal{Z}_a} 1\{|T(Z_\pi, Y, X)| \geq |T(Z, Y, X)|\}. \tag{5}$$

It compares the observed value of $T$ to its randomization distribution under ReM, denoted by $T^{\pi|\mathcal{A}}$. Under ReM in (4), $p_{\text{FRT},\mathcal{A}}$ is finite-sample exact for $H_{0\text{F}}$ for arbitrary $T$. Of interest is its large-sample validity for testing $H_{0\text{N}}$, which depends on the stochastic dominance relation between the asymptotic distributions of the test statistic. Theorem 5 summarizes the results based on the additional notation below.

Let $\epsilon \sim \mathcal{N}(0,1)$ and $\mathcal{L} \sim D_1 \mid (\|D\|_2^2 \leq a)$, where $D = (D_1, \ldots, D_J)^{\text{T}} \sim \mathcal{N}(0_J, I_J)$, be

independent standard and truncated normals, respectively, and let $r_{J,a} = P(\chi^2_{J+2} \leq a)/P(\chi^2_J \leq a) \in (0,1]$ be the variance of $\mathcal{L}$. Let $\mathcal{U}(\rho) = (1-\rho^2)^{1/2} \cdot \epsilon + \rho \cdot \mathcal{L}$ be a linear combination of $\epsilon$ and $\mathcal{L}$ for $\rho \in [0,1]$ with mean 0 and variance $v(\rho) = 1 - (1 - r_{J,a})\rho^2$. Recall $v_*$ and $v_{*0}$ in Theorems 1–4 as the asymptotic variances of $\sqrt{N}\hat{\tau}_*$ and $\sqrt{N}\hat{\tau}^\pi_*$ under complete randomization, with $v_{\text{R}} = v_{\text{F}}$ and $v_{\text{R}0} = v_{\text{F}0}$. Let $\rho^2_* = 1 - v_{\text{L}}/v_*$ and $\rho^2_{*0} = 1 - v_{\text{L}0}/v_{*0}$ for $* = \text{N}, \text{R}, \text{F}$, with $\rho_{\text{R}0} = \rho_{\text{F}0} = 0$.

**Theorem 5.** Assume Condition 1, ReM in design, and $p_{\text{FRT},\mathcal{A}}$ in (5) in analysis.

(a) $\sqrt{N}(\hat{\tau}_{\text{N}} - \tau) \rightsquigarrow v_{\text{N}}^{1/2} \cdot \mathcal{U}(\rho_{\text{N}})$, $(\hat{\tau}_{\text{N}} - \tau)/\hat{\text{se}}_{\text{N}} \rightsquigarrow (c'_{\text{N}})^{1/2} \cdot \mathcal{U}(\rho_{\text{N}})$, and $(\hat{\tau}_{\text{N}} - \tau)/\tilde{\text{se}}_{\text{N}} \rightsquigarrow c_{\text{N}}^{1/2} \cdot \mathcal{U}(\rho_{\text{N}})$; $\sqrt{N}\hat{\tau}^{\pi|\mathcal{A}}_{\text{N}} \rightsquigarrow v_{\text{N}0}^{1/2} \cdot \mathcal{U}(\rho_{\text{N}0})$, $(\hat{\tau}_{\text{N}}/\hat{\text{se}}_{\text{N}})^{\pi|\mathcal{A}} \rightsquigarrow \mathcal{U}(\rho_{\text{N}0})$, and $(\hat{\tau}_{\text{N}}/\tilde{\text{se}}_{\text{N}})^{\pi|\mathcal{A}} \rightsquigarrow \mathcal{U}(\rho_{\text{N}0})$ hold $P_Z$-a.s..

(b) $\sqrt{N}(\hat{\tau}_* - \tau) \rightsquigarrow v_*^{1/2} \cdot \mathcal{U}(\rho_*)$, $(\hat{\tau}_* - \tau)/\hat{\text{se}}_* \rightsquigarrow (c'_*)^{1/2} \cdot \mathcal{U}(\rho_*)$, and $(\hat{\tau}_* - \tau)/\tilde{\text{se}}_* \rightsquigarrow c_*^{1/2} \cdot \mathcal{U}(\rho_*)$; $\sqrt{N}\hat{\tau}^{\pi|\mathcal{A}}_* \rightsquigarrow \mathcal{N}(0, v_{*0})$, $(\hat{\tau}_*/\hat{\text{se}}_*)^{\pi|\mathcal{A}} \rightsquigarrow \mathcal{N}(0,1)$, and $(\hat{\tau}_*/\tilde{\text{se}}_*)^{\pi|\mathcal{A}} \rightsquigarrow \mathcal{N}(0,1)$ hold $P_Z$-a.s. $(* = \text{R}, \text{F})$.

(c) $\hat{\tau}_{\text{L}}$, $\hat{\tau}_{\text{L}}/\hat{\text{se}}_{\text{L}}$, and $\hat{\tau}_{\text{L}}/\tilde{\text{se}}_{\text{L}}$ have identical sampling and randomization distributions as under complete randomization in Theorem 4.

Compare Theorem 5 under ReM with Theorems 1–4 under complete randomization. The asymptotic sampling and randomization distributions of $\hat{\tau}_{\text{N}}$, $\hat{\tau}_{\text{N}}/\hat{\text{se}}_{\text{N}}$, and $\hat{\tau}_{\text{N}}/\tilde{\text{se}}_{\text{N}}$ change to non-normal. The asymptotic sampling distributions of $\hat{\tau}_*$, $\hat{\tau}_*/\hat{\text{se}}_*$, and $\hat{\tau}_*/\tilde{\text{se}}_*$ $(* = \text{R}, \text{F})$ change to non-normal, whereas their asymptotic randomization distributions remain the same. ReM does not affect these two sets of asymptotic randomization distributions because of the asymptotic independence between $\hat{\tau}^\pi_*$ $(* = \text{R}, \text{F})$ and $\hat{\tau}^\pi_x$. The asymptotic sampling and randomization distributions of $\hat{\tau}_{\text{L}}$, $\hat{\tau}_{\text{L}}/\hat{\text{se}}_{\text{L}}$, and $\hat{\tau}_{\text{L}}/\tilde{\text{se}}_{\text{L}}$ all remain unchanged. ReM does not affect them because of the asymptotic independence between $\hat{\tau}_{\text{L}}$ and $\hat{\tau}_x$ and that between $\hat{\tau}^\pi_{\text{L}}$ and $\hat{\tau}^\pi_x$.

In the case of symmetric yet non-normal limiting distributions as those of $\hat{\tau}_*$, $\hat{\tau}_*/\hat{\text{se}}_*$, and $\hat{\tau}_*/\tilde{\text{se}}_*$ for $* = \text{N}, \text{R}, \text{F}$, determination of properness entails comparisons of not only the variances but also all the central quantile ranges. A test statistic $T$ is proper under a two-sided FRT if $T$ has wider or equal central quantile ranges than $T^{\pi|\mathcal{A}}$ for all quantiles.

**Corollary 5.** Assume Condition 1, ReM in design, and $p_{\text{FRT},\mathcal{A}}$ in (5) in analysis. The covariate-adjusted robust $t$-statistics $\hat{\tau}_*/\tilde{\text{se}}_*$ $(* = \text{R}, \text{F}, \text{L})$ are the only test statistics in Table 1 proper for testing $H_{0\text{N}}$ via FRT.

Compare Corollary 5 with Corollary 1 to see that the unadjusted $\hat{\tau}_{\text{N}}/\tilde{\text{se}}_{\text{N}}$, whereas proper under complete randomization, is no longer proper under ReM due to the non-normal limiting distribution of $\hat{\tau}^\pi_{\text{N}}$. Cohen and Fogarty (2020) also noticed this phenomenon and gave a numeric example. They proposed a prepivoting approach to improve studentization. We do not pursue that direction given $\hat{\tau}_{\text{N}}/\tilde{\text{se}}_{\text{N}}$ is inferior to $\hat{\tau}_{\text{L}}/\tilde{\text{se}}_{\text{L}}$ even under complete randomization. The three covariate-adjusted robust $t$-statistics, namely $\hat{\tau}_{\text{R}}/\tilde{\text{se}}_{\text{R}}$, $\hat{\tau}_{\text{F}}/\tilde{\text{se}}_{\text{F}}$, and $\hat{\tau}_{\text{L}}/\tilde{\text{se}}_{\text{L}}$, are the only options in Table 1 proper for testing $H_{0\text{N}}$ under ReM. Covariate adjustment is thus essential for securing properness under ReM in addition to robust studentization. The same reasoning as that leads to Corollary 3 ensures FRT with $\hat{\tau}_{\text{L}}/\tilde{\text{se}}_{\text{L}}$

delivers the highest power among the three proper statistics. It is thus our recommendation for conducting FRT under ReM.

**Remark 4.** When covariates have different levels of importance for the outcomes, Morgan and Rubin (2012) proposed using ReM with differing criteria for different tiers of covariates. The resulting FRT permutes the treatment vector $Z$ within the subset of $\mathcal{Z}$'s that satisfy the tiered balance criteria. The sampling and randomization distributions of the twelve test statistics in Table 1 parallel those in Theorem 5, with $\hat{\tau}_{\text{L}}/\tilde{\text{se}}_{\text{L}}$ being the most powerful among the proper options. It is thus our recommendation for this extension as well. We omit the technical details due to its repetitiveness.

#### 4.3.2. FRT in the case of designer-analyzer information discrepancy

Discussion so far assumes the analyzer and the designer use the same covariates $(x_i)_{i=1}^N$ and threshold $a$ for doing ReM in the design and analysis stages, respectively. An interesting question, also a real concern in practice, is what if the designer and the analyzer do not communicate? Bruhn and McKenzie (2009), Heckman and Karapakula (2021), and Heckman et al. (2020) gave examples arising in field experiments in economics. Li and Ding (2020) discussed optimal covariate adjustment based on estimation precision.

A relatively easy case is that the analyzer has access to additional covariates beyond those used in the design of ReM. Using FRT under this ReM with $\hat{\tau}_{\text{L}}/\tilde{\text{se}}_{\text{L}}$ is again our recommendation in this case. A more challenging case is that the analyzer is either unaware of the ReM in the design stage or does not have access to all covariates used in the ReM. In the absence of full information about the design, Heckman and Karapakula (2021) proposed to use the maximum $p$-value from the worst-case FRT over a set of designs consistent with the available information. Without completely specifying these designs, an alternative option is to use $p_{\text{FRT}}$ in (1) such that the analysis coincides with that under complete randomization. Under $H_{0\text{F}}$, the finite-sample exactness is lost unless the original experiment is indeed completely randomized. Of interest is how such information discrepancy further affects the test's properness for testing $H_{0\text{N}}$.

Keep $x_i$ as the covariates the analyzer uses in the analysis stage, and let $d_i$ be the covariates the designer used for conducting ReM in the design stage, possibly different from $x_i$. The designer accepts an allocation if $\hat{\tau}_d^{\text{T}}\{\text{cov}(\hat{\tau}_d)\}^{-1}\hat{\tau}_d < a$ with $\hat{\tau}_d$ being the difference in means of $d_i$. The analyzer, on the other hand, uses $X = (x_1, \ldots, x_N)^{\text{T}}$ in addition to $Y$ and $Z$ to form the test statistic, and proceeds with the standard, unrestricted FRT that permutes $Z$ over all possible permutations in $\mathcal{Z}$. Of interest is whether the resulting $p$-value, namely $p_{\text{FRT}}$ in (1), can still preserve the correct type one error rates under $H_{0\text{N}}$ despite the information discrepancy.

Focus on the twelve test statistics in Table 1 for the rest of the discussion. The key is, again, the comparison of the stochastic dominance relations between their respective sampling and randomization distributions when only the weak null hypothesis holds. The randomization distributions, on the one hand, are readily available from Theorems 1–4 given the analysis is based on the unrestricted FRT. The possible discrepancy between $d_i$ and $x_i$, on the other hand, causes the sampling

distributions to deviate from those in Theorem 5. We furnish this missing piece in Proposition 2, and state the sampling distributions of the twelve test statistics under ReM using$d_i$'s for arbitrary $\mathcal{S}' = \{Y_i(0), Y_i(1), x_i, d_i\}_{i=1}^N$ that satisfies the regularity conditions.

Let $S^2_{z|d}$, $S^2_{a(z)|d}$, $S^2_{b(z)|d}$, $S^2_{\tau|d}$, and $S^2_{\xi|d}$ be the finite-population variances of the linear projections of $Y_i(z)$, $a_i(z)$, $b_i(z)$, $\tau_i$, and $\xi_i$ onto $d_i$, respectively, for $z = 0, 1$. Let

$$\rho^2_{\mathrm{N}|d} = \frac{p_1^{-1}S^2_{1|d} + p_0^{-1}S^2_{0|d} - S^2_{\tau|d}}{p_1^{-1}S^2_1 + p_0^{-1}S^2_0 - S^2_\tau},$$
$$\rho^2_{\mathrm{R}|d} = \rho^2_{\mathrm{F}|d} = \frac{p_1^{-1}S^2_{a(1)|d} + p_0^{-1}S^2_{a(0)|d} - S^2_{\tau|d}}{p_1^{-1}S^2_{a(1)} + p_0^{-1}S^2_{a(0)} - S^2_\tau},$$
$$\rho^2_{\mathrm{L}|d} = \frac{p_1^{-1}S^2_{b(1)|d} + p_0^{-1}S^2_{b(0)|d} - S^2_{\xi|d}}{p_1^{-1}S^2_{b(1)} + p_0^{-1}S^2_{b(0)} - S^2_\xi}$$

be the squared multiple correlations between $\hat{\tau}_*$ ($* = \mathrm{N}, \mathrm{R}, \mathrm{F}, \mathrm{L}$) and $\hat{\tau}_d$.

**Proposition 2.** Assume Condition 1 holds for $\{Y_i(0), Y_i(1), x_i'\}_{i=1}^N$ with $x_i'$ being the union of $x_i$ and $d_i$, and ReM using $d_i$'s. For $* = \mathrm{N}, \mathrm{R}, \mathrm{F}, \mathrm{L}$, we have

$$\sqrt{N}(\hat{\tau}_* - \tau) \rightsquigarrow v_*^{1/2} \cdot \mathcal{U}(\rho_{*|d}), \quad (\hat{\tau}_* - \tau)/\hat{\mathrm{se}}_* \rightsquigarrow (c_*')^{1/2} \cdot \mathcal{U}(\rho_{*|d}), \quad (\hat{\tau}_* - \tau)/\tilde{\mathrm{se}}_* \rightsquigarrow c_*^{1/2} \cdot \mathcal{U}(\rho_{*|d}).$$

Assuming the actual assignment is conducted by ReM using $d_i$'s that are possibly different from $x_i$'s, Proposition 2 is a special case of Li and Ding (2020) and generalizes the sampling distributions in Theorem 5 to allow for distinct covariates for the design and analysis stages, respectively. The resulting sampling distributions are in general scaled $\mathcal{U}$ distributions as linear combinations of independent standard and truncated normals. In particular, $\rho_{\mathrm{L}|d} = 0$ if $x_i$ can linearly represent $d_i$, rendering the limiting distributions of $\hat{\tau}_\mathrm{L}$, $\hat{\tau}_\mathrm{L}/\hat{\mathrm{se}}_\mathrm{L}$, and $\hat{\tau}_\mathrm{L}/\tilde{\mathrm{se}}_\mathrm{L}$ identical to those under complete randomization in Theorem 4. The following corollary holds by comparing Proposition 2 with Theorems 1–4.

**Corollary 6.** Assume Condition 1 holds for $\{Y_i(0), Y_i(1), x_i'\}_{i=1}^N$ with $x_i'$ being the union of the $x_i$ and $d_i$, ReM using $d_i$'s in design, and $p_{\mathrm{FRT}}$ in (1) in analysis. The robust $t$-statistics $\hat{\tau}_*/\tilde{\mathrm{se}}_*$ ($* = \mathrm{N}, \mathrm{R}, \mathrm{F}, \mathrm{L}$) are the only test statistics in Table 1 proper for testing $H_{0\mathrm{N}}$ via FRT.

The four robust $t$-statistics thus ensure $p_{\mathrm{FRT}}$ in (1) remains asymptotically valid under ReM even if the analyzer has only partial information on the covariates the designer used to form the balance criterion. Ironically, a less informed analysis restores the properness of $\hat{\tau}_\mathrm{N}/\tilde{\mathrm{se}}_\mathrm{N}$ under ReM by restoring its asymptotic randomization distribution back to the standard normal. Nevertheless, this properness comes at the cost of being overly conservative.

Further, the asymptotic randomization distributions of $\hat{\tau}_*/\tilde{\mathrm{se}}_*$ ($* = \mathrm{R}, \mathrm{F}, \mathrm{L}$) remain unchanged in computing $p_{\mathrm{FRT}}$ in (1) and $p_{\mathrm{FRT},\mathcal{A}}$ in (5). It might thus be tempting to ignore the rerandomization and conduct unrestricted FRT in the analysis stage whatsoever, even when exact information is

available. We do not encourage such practice given its lack of finite-sample exactness under $H_{0\text{F}}$ in the first place.

# 5. Simulation

We examine in this section the validity and power of the proposed method for testing the weak null hypothesis via simulation. We conducted the simulation under complete randomization, stratified randomization, and rerandomization, respectively, and summarized the $p$-values over 1,000 independent repetitions to approximate the error rates. The patterns are almost identical for the three design types, highlighting the importance of robust studentization and efficient covariate adjustment for securing large-sample validity and additional power, respectively. To avoid repetitiveness, we present below the results from the stratified randomization.

We first examine the large-sample validity of the twelve test statistics under $H_{0\text{N}}$. Consider a finite population of $N = 500$ units, $i = 1, \ldots, N$, with a univariate covariate, $(x_i)_{i=1}^N$, as i.i.d. Unif$(-1, 1)$. We generate the potential outcomes as $Y_i(1) \sim \mathcal{N}(x_i^3, 1)$ and $Y_i(0) \sim \mathcal{N}(-x_i^3, 0.5^2)$, and center $Y_i(1)$'s and $Y_i(0)$'s respectively to ensure $\tau = 0$.

We divide the units into $K = 3$ strata by the values of their covariates at cutoffs $-0.3$ and 0.3. The resulting strata consist of units with $x_i$'s in $[-1, -0.3]$, $x_i$'s in $(-0.3, 0.3]$, and $x_i$'s in $(0.3, 1]$, respectively. Denote by $N_{[k]}$ the number of units in stratum $k$, and set $N_{[k]1} = [0.2N_k]$ and $N_{[k]0} = N_{[k]} - N_{[k]1}$ as the corresponding stratum-wise treatment sizes. We fix $\{Y_i(0), Y_i(1), x_i\}_{i=1}^N$ in the simulation, and draw a random permutation of $N_{[k]1}$ 1's and $N_{[k]0}$ 0's within stratum $k$ for $k = 1, 2, 3$ to obtain the observed outcomes and conduct FRTs.

The procedure is repeated 1,000 times, with the $p$-values approximated by 500 independent permutations of the treatment vector in each replication. Figure 1(a) shows the $p$-values under $H_{0\text{N}}$. The four robust $t$-statistics, as shown in the last row, are the only ones that preserve the correct type one error rates. In fact, they are conservative, which is coherent with Corollary 4. All the other eight statistics yield type one error rates greater than the nominal levels and are thus not proper for testing $H_{0\text{N}}$.

We then evaluate the power of the four proper test statistics when $\tau \neq 0$. Take $Y_i(1) \sim \mathcal{N}(0.1 + x_i, 0.4^2)$ and $Y_i(0) \sim \mathcal{N}(-x_i, 0.1^2)$ for an alternative with $\tau$ close to 0.1, and inherit the rest of the settings from the last two paragraphs. Figure 1(b) shows the $p$-values of the four proper test statistics under the alternative. The theoretically most powerful $\hat{\tau}_{\text{L}}/\tilde{\text{se}}_{\text{L}}$ indeed delivers the highest power among the four proper options. The tests based on $\hat{\tau}_{\text{F}}/\tilde{\text{se}}_{\text{F}}$ and $\hat{\tau}_{\text{R}}/\tilde{\text{se}}_{\text{R}}$, on the other hand, show even lower power than the unadjusted $\hat{\tau}_{\text{N}}/\tilde{\text{se}}_{\text{N}}$. This is coherent with the theoretical results from Corollary 3 and concludes $\hat{\tau}_{\text{L}}/\tilde{\text{se}}_{\text{L}}$ as our final recommendation for conducting FRT under stratified randomization.

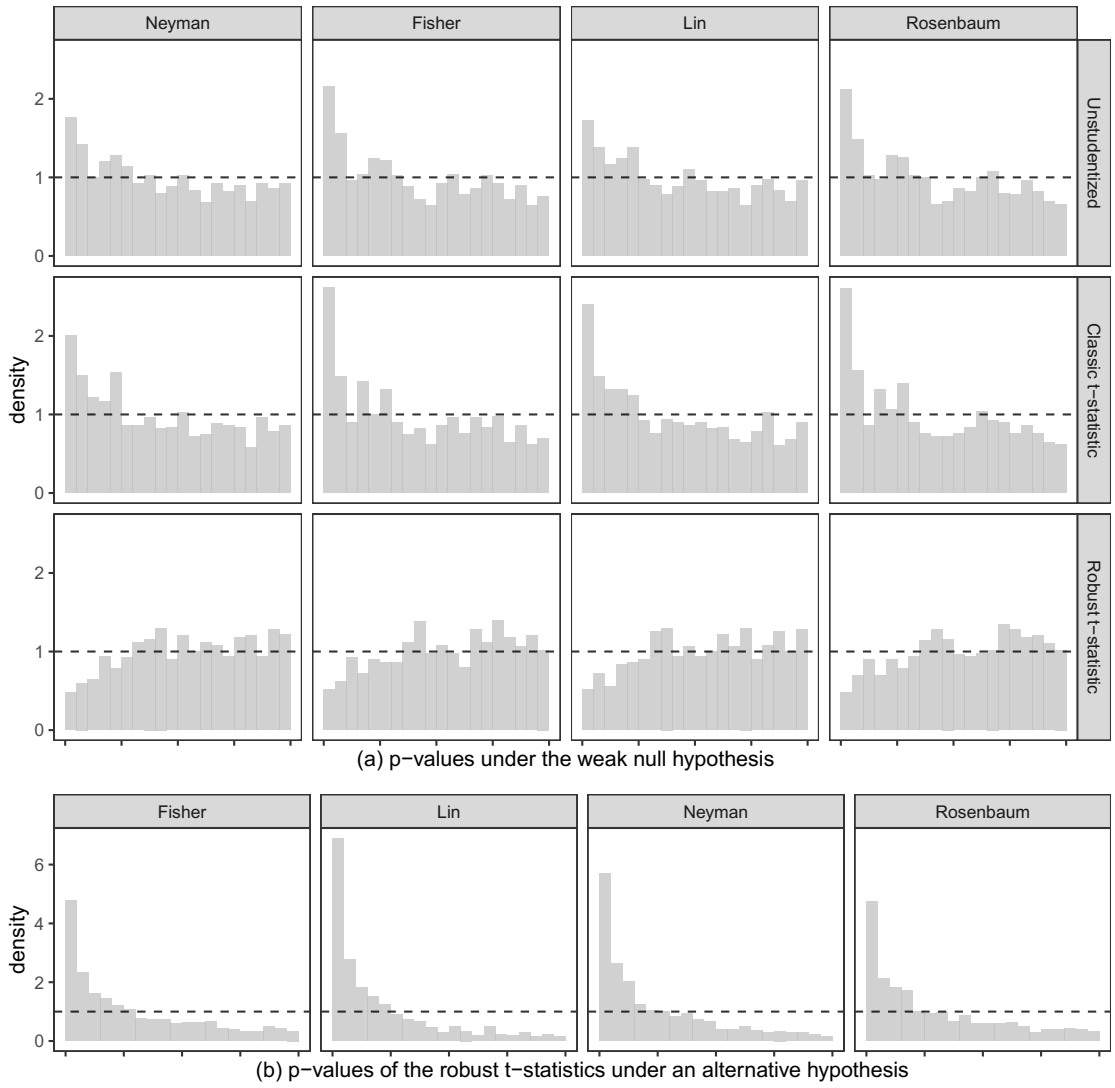


Figure 1: Empirical histograms of the $p_{\text{FRT,str}}$'s with 20 bins in $(0, 1)$.

# 6. Application

Chong et al. (2016) conducted a randomized experiment on 219 students of a rural secondary school in the Cajamarca district of Peru during the 2009 school year. They first provided the village with free iron supplements and trained the local staffs to distribute one free iron pill to any adolescent who requested one in person. They then randomly assigned the students to three arms with three different types of videos: in the first video, a popular soccer player was encouraging the use of iron supplements to maximize energy ("soccer" arm); in the second video, a physician was encouraging the use of iron supplements to improve overall health ("physician" arm); the third video did not mention iron and served as the control ("control" arm). The experiment was stratified by the class

Table 2: Re-analyzing the data from Chong et al. (2016). "N" denotes the unadjusted estimators and tests, and "L" denotes the covariate-adjusted estimators and tests. The "$p_{\text{FRT}}$" values for the overall comparisons in the last two rows are all $p_{\text{FRT,str}}$.

(a) soccer versus control

| | est | s.e. | $p_{\text{normal}}$ | $p_{\text{FRT}}$ |
|---|---|---|---|---|
| class 1 | | | | |
| N | 0.051 | 0.502 | 0.919 | 0.924 |
| L | 0.050 | 0.489 | 0.919 | 0.929 |
| class 2 | | | | |
| N | −0.158 | 0.451 | 0.726 | 0.722 |
| L | −0.176 | 0.452 | 0.698 | 0.700 |
| class 3 | | | | |
| N | 0.005 | 0.403 | 0.990 | 0.989 |
| L | −0.096 | 0.385 | 0.803 | 0.806 |
| class 4 | | | | |
| N | −0.492 | 0.447 | 0.271 | 0.288 |
| L | −0.511 | 0.447 | 0.253 | 0.283 |
| class 5 | | | | |
| N | 0.390 | 0.369 | 0.291 | 0.314 |
| L | 0.443 | 0.318 | 0.164 | 0.186 |
| all | | | | |
| N | −0.051 | 0.204 | 0.802 | 0.800 |
| L | −0.074 | 0.200 | 0.712 | 0.712 |

(b) physician versus control

| | est | s.e. | $p_{\text{normal}}$ | $p_{\text{FRT}}$ |
|---|---|---|---|---|
| class 1 | | | | |
| N | 0.567 | 0.426 | 0.183 | 0.192 |
| L | 0.588 | 0.418 | 0.160 | 0.174 |
| class 2 | | | | |
| N | 0.193 | 0.438 | 0.659 | 0.666 |
| L | 0.265 | 0.409 | 0.517 | 0.523 |
| class 3 | | | | |
| N | 1.305 | 0.494 | 0.008 | 0.012 |
| L | 1.501 | 0.462 | 0.001 | 0.003 |
| class 4 | | | | |
| N | −0.273 | 0.413 | 0.508 | 0.515 |
| L | −0.313 | 0.417 | 0.454 | 0.462 |
| class 5 | | | | |
| N | −0.050 | 0.379 | 0.895 | 0.912 |
| L | −0.067 | 0.279 | 0.811 | 0.816 |
| all | | | | |
| N | 0.406 | 0.202 | 0.045 | 0.047 |
| L | 0.463 | 0.190 | 0.015 | 0.017 |

level from 1 to 5. The treatment group sizes within classes are shown in the matrix below:

$$\begin{array}{c} \\ \text{soccer} \\ \text{physician} \\ \text{control} \end{array} \begin{array}{c} \begin{array}{ccccc} \text{class 1} & \text{class 2} & \text{class 3} & \text{class 4} & \text{class 5} \end{array} \\ \begin{pmatrix} 16 & 19 & 15 & 10 & 10 \\ 17 & 20 & 15 & 11 & 10 \\ 15 & 19 & 16 & 12 & 10 \end{pmatrix} \end{array}.$$

One outcome of interest is the average grade in the third and fourth quarters of 2009, and an important background covariate is the anemia status at baseline. We make pairwise comparisons of the "soccer" arm versus the "control" arm and the "physician" arm versus the "control" arm. We also compare FRTs with and without adjusting for the covariate of baseline anemia status. We use their data set to illustrate FRTs under complete randomization and stratified randomization. The ten subgroup analyses use FRTs for complete randomization within each class level. The two overall analyses use FRTs for stratified randomization averaging over all class levels.

Table 2 shows the point estimators, the robust standard errors, the $p$-values based on large-sample approximations of the robust $t$-statistics, and the $p$-values based on FRTs. In most strata, covariate adjustment decreases the standard errors since the baseline anemia status is predictive of the outcome. Two exceptions are the pairwise comparison of the "soccer" arm versus the "control" arm within class 2 and the pairwise comparison of the "physician" arm versus the "control" arm

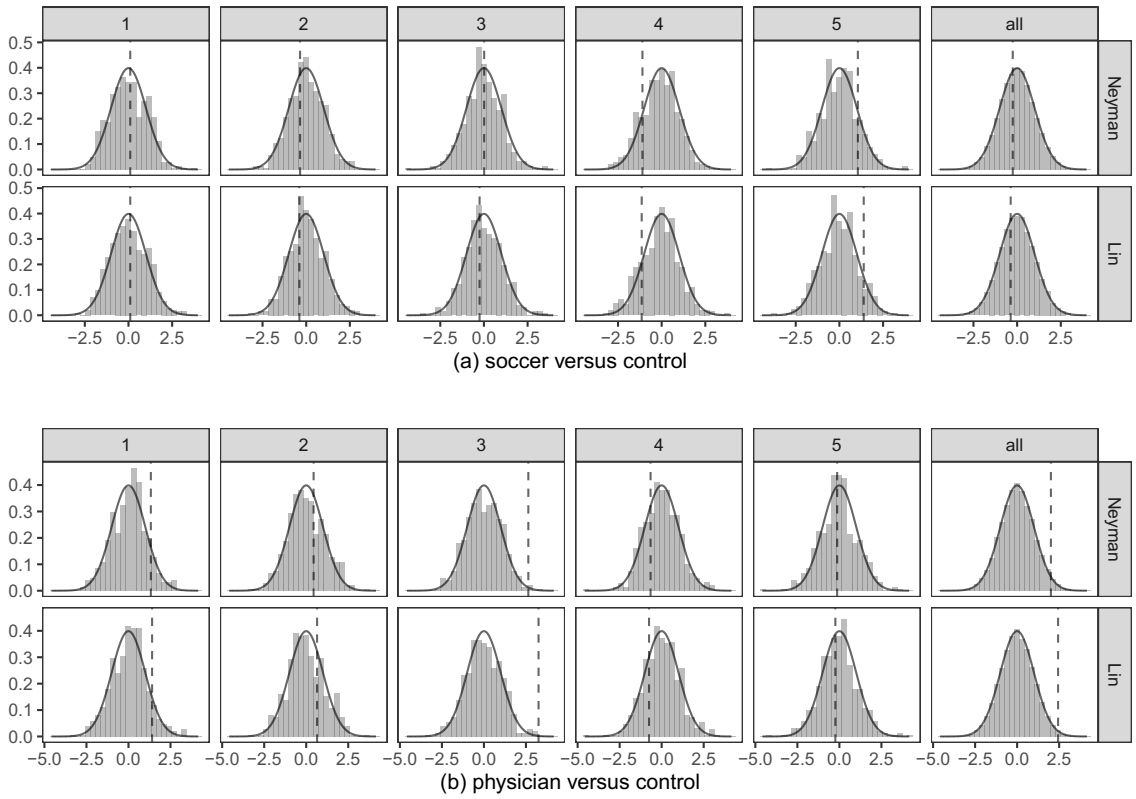


Figure 2: Randomization distributions based on $5 \times 10^4$ Monte Carlo simulations versus $\mathcal{N}(0, 1)$.

within class 4, with differences both in the third digit after the decimal point. This is likely due to the small group sizes within these strata, leaving the asymptotic approximations dubious. The $p$-values from the large-sample approximations and FRTs are close with the latter being slightly larger in most cases. Based on the theory, the $p$-values based on FRTs should be trusted more given their additional guarantee of finite-sample exactness under the strong null hypothesis. This becomes important in this example given the relatively small group sizes within strata.

Bind and Rubin (2020) suggested reporting not only the $p$-values but also the randomization distributions of the test statistics when conducting FRT. Echoing their recommendation, we show in Figure 2 the histograms of the randomization distributions of the robust $t$-statistics alongside the asymptotic approximations. The discrepancy is quite clear in the subgroup analyses yet becomes unnoticeable after averaged over all class levels. Overall, the $p$-values based on large-sample approximations do not differ substantially from those based on FRTs in this application. The two approaches yield coherent conclusions: the video with a physician telling the benefits of iron supplements improved the academic performance and the effect was most significant among students in class 3; in contrast, the video with a popular soccer player telling the benefits did not have any significant effect.

Table 3: Final recommendations for FRT and test statistic $\hat{\tau}_*/\tilde{\text{se}}_*$ in different experiments.

| design | presence of covariates | | other comments |
|---|---|---|---|
| | no | yes | |
| complete randomization | $* = \text{N}$ | $* = \text{L}$ | |
| cluster randomization | $* = \text{N}$ | $* = \text{L}$ | use cluster total outcomes |
| stratified randomization | $* = \text{N}$ | $* = \text{L}$ | weighted average over strata |
| ReM, complete design information | | $* = \text{L}$ | |
| ReM, incomplete design information | $* = \text{N}$ | $* = \text{L}$ | use $p_{\text{FRT}}$ not $p_{\text{FRT},\mathcal{A}}$ |

# 7. Discussion

Echoing Fisher (1935), Proschan and Dodd (2019), Young (2019), and Bind and Rubin (2020), we believe FRT should be the default choice for analyzing experimental data given its flexibility to accommodate complex randomization schemes and arbitrary outcome generating processes. We established in this paper the theory for covariate adjustment in FRT under complete randomization, cluster randomization, stratified randomization, and rerandomization using the Mahalanobis distance, respectively, with final recommendations of the test statistics summarized in Table 3. Equipped with the finite-sample exactness under the strong null hypothesis, the recommended FRTs promise an additional guarantee under the weak null hypothesis and strictly dominate the counterparts based on large-sample approximations. A key point to note is that robust studentization is necessary for the resulting FRT to retain asymptotic validity when only the weak null hypothesis holds. A casual choice of the test statistic is likely to lead to misleading conclusions.

We conjecture that the strategy of appropriately studentizing an efficient, covariate-adjusted estimator works for FRT in general experiments as well (e.g., Dasgupta et al. 2015; Lu 2016; Mukerjee et al. 2018; Middleton 2018; Fogarty 2018a,b). This strategy works for estimators with normal limiting distributions and may also work for estimators with non-normal limiting distributions as shown in the asymptotic theory of rerandomization. Cohen and Fogarty (2020)'s prepivoting approach may work more broadly but we leave the general theory to future research.

We focused on procedures based on OLS. It is of great interest to extend the theory to high dimensional settings (Bloniarz et al. 2016; Lei and Ding 2020), nonlinear models (Zhang et al. 2008; Moore and van der Laan 2009; Moore et al. 2011; Jiang et al. 2019; Guo and Basse 2020), and even estimators based on machine learning algorithms (Wager et al. 2016; Wu and Gagnon-Bartsch 2018; Farrell et al. 2021; Chen et al. 2020).

If the main parameter of interest is the average treatment effect, the asymptotic theory inevitably involves some moment conditions. Without these conditions, the inference becomes challenging (Bahadur and Savage 1956), and FRT may not control type one error rates even asymptotically with heavy-tailed outcomes. An alternative class of FRTs use rank statistics to gain robustness with respect to outliers (Lehmann 1975; Rosenbaum 2002). Although different rank statistics always work under the strong null hypothesis, they in general target parameters other than the average treatment effect (e.g., Rosenbaum 1999, 2003; Chung and Romano 2016). Chung and Ro-

mano (2016) proposed to studentize the Wilcoxon statistic in a permutation test, shedding light on the general theory of FRT with rank statistics.

## References


J. D. Angrist and J.-S. Pischke. *Mostly Harmless Econometrics.* Princeton University Press, 2009.

S. Athey, D. Eckles, and G. W. Imbens. Exact p-values for network interference. *Journal of the American Statistical Association*, 113:230–240, 2018.

R. R. Bahadur and L. J. Savage. The nonexistence of certain statistical procedures in nonparametric problems. *Annals of Mathematical Statistics*, 27:1115–1122, 1956.

A. V. Banerjee, S. Chassang, S. Montero, and E. Snowberg. A theory of experimenters: Robustness, randomization, and balance. *American Economic Review*, 110:1206–1230, 2020.

G. Basse, P. Ding, A. Feller, and P. Toulis. Randomization tests for peer effects in group formation experiments. *arXiv*, page 1904.02308, 2019.

R. Berk, E. Pitkin, L. Brown, A. Buja, E. George, and L. Zhao. Covariance adjustments for the analysis of randomized field experiments. *Evaluation Review*, 37:170–196, 2013.

M. A. C. Bind and D. B. Rubin. When possible, report a Fisher-exact P value and display its underlying null randomization distribution. *Proceedings of the National Academy of Sciences of the United States of America*, 117:19151–19158, 2020.

A. Bloniarz, H. Liu, C. Zhang, J. Sekhon, and B. Yu. Lasso adjustments of treatment effect estimates in randomized experiments. *Proceedings of the National Academy of Sciences of the United States of America*, 113:7383–7390, 2016.

D. R. Brillinger, L. V. Jones, and J. W. Tukey. The management of weather resources. Technical report, US Government Printing Office, Washington, DC, 1978.

M. Bruhn and D. McKenzie. In pursuit of balance: Randomization in practice in development field experiments. *American Economic Journal: Applied Economics*, 1:200–232, 2009.

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

F. A. Bugni, I. A. Canay, and A. M. Shaikh. Inference under covariate-adaptive randomization with multiple treatments. *Quantitative Economics*, 10:1747–1785, 2019.

I. A. Canay, J. P. Romano, and A. M. Shaikh. Randomization tests under an approximate symmetry assumption. *Econometrica*, 85:1013–1030, 2017.

M. D. Cattaneo, B. R. Frandsen, and R. Titiunik. Randomization inference in the regression discontinuity design: An application to party advantages in the US Senate. *Journal of Causal Inference*, 3:1–24, 2015.

X. Chen, Y. Liu, S. Ma, and Z. Zhang. Efficient estimation of general treatment effects using neural networks with a diverging number of confounders. *arXiv preprint arXiv:2009.07055*, 2020.

A. Chong, I. Cohen, E. Field, E. Nakasone, and M. Torero. Iron deficiency and schooling attainment in Peru. *American Economic Journal: Applied Economics*, 8:222–55, 2016.

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

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

P. L. Cohen and C. B. Fogarty. Gaussian prepivoting for finite population causal inference. *https://arxiv.org/abs/2002.06654*, 2020.

D. R. Cox. Randomization and concomitant variables in the design of experiments. In P. R. Krishnaiah G. Kallianpur and J. K. Ghosh, editors, *Statistics and Probability: Essays in Honor of C. R. Rao*, pages 197–202. North-Holland, Amsterdam, 1982.

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

C. J. DiCiccio and J. P. Romano. Robust permutation tests for correlation and regression coefficients. *Journal of the American Statistical Association*, 112:1211–1220, 2017.

P. Ding and T. Dasgupta. A randomization-based perspective of analysis of variance: a test statistic robust to treatment effect heterogeneity. *Biometrika*, 105:45–56, 2018.

F. Eicker. Limit theorems for regressions with unequal and dependent errors. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 59–82. Berkeley, CA: University of California Press, 1967.

M. H. Farrell, T. Liang, and S. Misra. Deep neural networks for estimation and inference. *Econometrica*, 89:181–213, 2021.

R. A. Fisher. *The Design of Experiments*. Edinburgh, London: Oliver and Boyd, 1st edition, 1935.

C. B. Fogarty. On mitigating the analytical limitations of finely stratified experiments. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80:1035–1056, 2018a.

C. B. Fogarty. Regression assisted inference for the average treatment effect in paired experiments. *Biometrika*, 105:994–1000, 2018b.

D. Freedman and D. Lane. A nonstochastic interpretation of reported significance levels. *Journal of Business and Economic Statistics*, 1:292–298, 1983.

D. A. Freedman. On regression adjustments to experimental data. *Advances in Applied Mathematics*, 40:180–193, 2008.

M. H. Gail, W. Y. Tan, and S. Piantadosi. Tests for no treatment effect in randomized clinical trials. *Biometrika*, 75:57–64, 1988.

P. Ganong and S. Jäger. A permutation test for the regression kink design. *Journal of the American Statistical Association*, 113:494–504, 2018.

K. Guo and G. Basse. The generalized Oaxaca–Blinder estimator. *arXiv*, page 2004.11615, 2020.

J. J. Heckman and G. Karapakula. Using a satisficing model of experimenter decision-making to guide finite-sample inference for compromised experiments. *Econometrics Journal*, page in press, 2021.

J. J. Heckman, R. Pinto, and A. M. Shaikh. Inference with imperfect randomization: The case of the Perry preschool program. Working paper, University of Chicago, 2020.

J. Hennessy, T. Dasgupta, L. Miratrix, C. Pattanayak, and P. Sarkar. A conditional randomization test to account for covariate imbalance in randomized experiments. *Journal of Causal Inference*, 4:61–80, 2016.

W. Hoeffding. The large-sample power of tests based on permutations of observations. *The Annals of Mathematical Statistics*, 23:169–192, 1952.

P. J. Huber. The behavior of maximum likelihood estimates under nonstandard conditions. In Lucien M. Le Cam and Jerzy Neyman, editors, *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 221–233. Berkeley, California: University of California Press, 1967.

A. Janssen. Studentized permutation tests for non-iid hypotheses and the generalized Behrens–Fisher problem. *Statistics and Probability Letters*, 36:9–21, 1997.

F. Jiang, L. Tian, H. Fu, T. Hasegawa, and L. J. Wei. Robust alternatives to ANCOVA for estimating the treatment effect via a randomized comparative study. *Journal of the American Statistical Association*, 114:1854–1864, 2019.

P. E. Kennedy. Randomization tests in econometrics. *Journal of Business and Economic Statistics*, 13:85–94, 1995.

S. Lee and A. M. Shaikh. Multiple testing and heterogeneous treatment effects: Re-evaluating the effect of progresa on school enrollment. *Journal of Applied Econometrics*, 29:612–626, 2014.

E. L. Lehmann. *Nonparametrics: Statistical Methods Based on Ranks*. San Francisco: Holden-Day, Inc., 1975.

E. L. Lehmann and J. P. Romano. *Testing Statistical Hypotheses*. New York: Springer, 3rd edition, 2005.

L. Lei and P. Ding. Regression adjustment in completely randomized experiments with a diverging number of covariates. *Biometrika*, page in press, 2020.

X. Li and P. Ding. General forms of finite population central limit theorems with applications to causal inference. *Journal of the American Statistical Association*, 112:1759–1169, 2017.

X. Li and P. Ding. Rerandomization and regression adjustment. *Journal of the Royal Statistical Society, Series B (Methodological)*, 82:241–268, 2020.

W. Lin. Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique. *Annals of Applied Statistics*, 7:295–318, 2013.

H. Liu and Y. Yang. Regression-adjusted average treatment effect estimates in stratified randomized experiments. *Biometrika*, 107:935–948, 2020.

J. Lu. Covariate adjustment in randomization-based causal inference for $2^K$ factorial designs. *Statistics and Probability Letters*, 119:11–20, 2016.

J. G. MacKinnon and M. D. Webb. Randomization inference for difference-in-differences with few treated clusters. *Journal of Econometrics*, 218:435–450, 2020.

B. F. J. Manly. *Randomization, Bootstrap and Monte Carlo Methods in Biology*. Chapman & Hall, 1997.

J. A. Middleton. A unified theory of regression adjustment for design-based inference. *arXiv preprint arXiv:1803.06011*, 2018.

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

K. L. Moore and M. J. van der Laan. Covariate adjustment in randomized trials with binary outcomes: targeted maximum likelihood estimation. *Statistics in Medicine*, 28:39–64, 2009.

K. L. Moore, R. Neugebauer, T. Valappil, and M. J. van der Laan. Robust extraction of covariate information to improve estimation efficiency in randomized trials. *Statistics in Medicine*, 30: 2389–2408, 2011.

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

R. Mukerjee, T. Dasgupta, and D. B. Rubin. Using standard tools from finite population sampling to improve causal inference for complex experiments. *Journal of the American Statistical Association*, 113:868–881, 2018.

A. Negi and J. M. Wooldridge. Revisiting regression adjustment in experiments with heterogeneous treatment effects. *Econometric Reviews*, 40:504–534, 2021.

J. Neyman. On the application of probability theory to agricultural experiments (with discussion). *Statistical Science*, 5:465–472, 1923/1990.

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

K. Ottoboni, F. Lewis, and L. Salmaso. An empirical comparison of parametric and permutation tests for regression analysis of randomized experiments. *Statistics in Biopharmaceutical Research*, 10:264–273, 2018.

M. Pauly, E. Brunner, and F. Konietschke. Asymptotic permutation tests in general factorial designs. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 77:461–473, 2015.

M. A. Proschan and L. E. Dodd. Re-randomization tests in clinical trials. *Statistics in Medicine*, 38:2292–2302, 2019.

J. Raz. Testing for no effect when estimating a smooth function by nonparametric regression: a randomization approach. *Journal of the American Statistical Association*, 85:132–138, 1990.

J. P. Romano. On the behavior of randomization tests without a group invariance assumption. *Journal of the American Statistical Association*, 85:686–692, 1990.

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

P. R. Rosenbaum. Covariance adjustment in randomized experiments and observational studies. *Statistical Science*, 17:286–327, 2002.

P. R. Rosenbaum. Exact confidence intervals for nonconstant effects by inverting the signed rank test. *American Statistician*, 57:132–138, 2003.

P. R. Rosenbaum. *Design of Observational Studies*. New York: Springer, 2nd edition, 2010.

A. J. Stephens, E. J. Tchetgen Tchetgen, and V. De Gruttola. Flexible covariate-adjusted exact tests of randomized treatment effects with application to a trial of HIV education. *Annals of Applied Statistics*, 7:2106–2137, 2013.

F. Su and P. Ding. Model-assisted analyses of cluster-randomized experiments. *arXiv*, page 2104.04647, 2021.

C. J. F. ter Braak. Permutation versus bootstrap significance tests in multiple regression and ANOVA. In K.H. Jöckel, G. Rothe, and W. Sendler, editors, *Bootstrapping and Related Techniques*, pages 79–85. Berlin: Springer-Verlag, 1992.

J. W. Tukey. Tightening the clinical trial. *Controlled Clinical Trials*, 14:266–285, 1993.

A. W. van der vaart and J. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. New York: Springer Verlag, 1996.

S. Wager, W. Du, J. Taylor, and R. J. Tibshirani. High-dimensional regression adjustments in randomized experiments. *Proceedings of the National Academy of Sciences of the United States of America*, 113:12673–12678, 2016.

H. White. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48:817–838, 1980.

E. Wu and J. A. Gagnon-Bartsch. The LOOP estimator: Adjusting for covariates in randomized experiments. *Evaluation Review*, 42:458–488, 2018.

J. Wu and P. Ding. Randomization tests for weak null hypotheses in randomized experiments. *Journal of American Statistical Association*, 105:in press, 2020.

T. Ye, Y. Yi, and Q. Zhao. Inference on average treatment effect under minimization and other covariate-adaptive randomization methods. *arXiv preprint arXiv:2007.09576*, 2020.

A. Young. Channeling Fisher: Randomization tests and the statistical insignificance of seemingly significant experimental results. *Quarterly Journal of Economics*, 134:557–598, 2019.

M. Zhang, A. A. Tsiatis, and M. Davidian. Improving efficiency of inferences in randomized clinical trials using auxiliary covariates. *Biometrics*, 64:707–715, 2008.

L. Zheng and M. Zelen. Multi-center clinical trials: Randomization and ancillary statistics. *Annals of Applied Statistics*, 2:582–600, 2008.