Randomization Tests for Weak Null Hypotheses

Jason Wu and Peng Ding *

Abstract

The Fisher randomization test (FRT) is appropriate for any test statistic, under a sharp null hypothesis that can recover all missing potential outcomes. However, it is often of interest to test a weak null hypothesis that the treatment does not affect the units on average. To use the FRT for a weak null hypothesis, we must address two issues. First, we need to impute the missing potential outcomes although the weak null hypothesis cannot determine all of them. Second, we need to choose a suitable test statistic. For a general weak null hypothesis, we propose an approach to imputing missing potential outcomes under a compatible sharp null hypothesis. With this imputation scheme, we advocate using a studentized statistic. The resulting FRT has multiple desirable features. First, it is model-free. Second, it is finite-sample exact under the sharp null hypothesis that we use to impute the potential outcomes. Third, it preserves correct large-sample type I errors under the weak null hypothesis of interest. Therefore, our FRT is agnostic to treatment effect heterogeneity. We establish a unified theory for general factorial experiments. We also extend it to stratified and clustered experiments.

Key Words: Causal inference; Finite population asymptotics; Randomization-based inference; Robustness; Sharp null hypothesis; Studentization

*Jason Wu is a Ph.D. candidate and Peng Ding is an Assistant Professor, Department of Statistics, University of California, Berkeley. Corresponding author: Peng Ding, Email: pengdingpku@berkeley.edu, 425 Evans Hall, Berkeley, CA 94720 USA.

Introduction to the Fisher Randomization Test in Experiments

Literature Review

Randomization is a basis for statistical causal inference (Fisher 1935, Section II). It creates comparable treatment groups on average. More importantly, it justifies the Fisher randomization test (FRT). Under Fisher's sharp null hypothesis, the treatment does not affect any units whatsoever, and the distribution of any test statistic is known over all randomizations (Fisher 1935; Rubin 1980; Rosenbaum 2002a; Imbens and Rubin 2015). Therefore, the FRT gives a finite-sample exact p-value. In fact, many parametric or non-parametric tests are approximations to the FRT (Eden and Yates 1933; Pitman 1937; Kempthorne 1952; Box and Andersen 1955; Collier and Baker 1966; Bradley 1968; Lehmann 1975).

Another formulation of the FRT relies on exchangeability of outcomes under different treatments (Pitman 1937; Hoeffding 1952; Romano 1990). They called this formulation a "permutation test". Kempthorne and Doerfler (1969) highlighted the importance of the treatment assignment mechanism to justify the FRT, without assuming that the outcomes are exchangeable. Rubin (1980) extended the FRT using Neyman (1990)'s potential outcomes. He defined a null hypothesis to be sharp if it can determine all missing potential outcomes. He pointed out that any test statistic has a known distribution under a sharp null hypothesis, and therefore the FRT is finite-sample exact.

Randomized experiments are increasingly popular in the social sciences (Duflo et al. 2007; Gerber and Green 2012; Imbens and Rubin 2015; Athey and Imbens 2017). In such applications, testing sharp null hypotheses may not answer the questions of interest. Researchers often want to test weak null hypotheses that the treatment has zero effects on average. The ideal testing procedure must allow for treatment effect heterogeneity. Unfortunately, weak null hypotheses cannot determine all missing potential outcomes, even though the distributions of test statistics depend on them in general. Consequently, simple FRTs may not be directly applicable for testing weak null hypotheses.

It is challenging to use FRTs to test weak null hypotheses. Although sometimes we can still use the same FRTs, we need to modify the interpretations without sharp null hypotheses (Rosenbaum 1999, 2001, 2003; Caughey et al. 2017). Not all FRTs can preserve type I errors for weak null hypotheses even asymptotically. The famous Neyman–Fisher controversy is related to this issue for randomized block designs and Latin square designs (Neyman 1935; Sabbaghi and Rubin 2014). Gail et al. (1996) and Lin et al. (2017) gave empirical evidence based on simulation, and Ding and Dasgupta (2018) gave a theoretical analysis of the one-way layout. Two strategies exist for using FRTs to test weak null hypotheses. The first strategy relies on a simple observation that weak null hypotheses become sharp given appropriate nuisance parameters. It uses the maximum of the p-values over all values of the nuisance parameters or their confidence sets (Nolen and Hudgens 2011; Rigdon and Hudgens 2015; Li and Ding 2016; Ding et al. 2016). However, it can be computationally intensive and lacks power when the nuisance parameters are high dimensional. The second strategy uses conditional FRTs. It relies on partitioning the space of all randomizations, and in some subspaces, certain test statistics have known distributions under the weak null hypotheses (Athey et al. 2018; Basse et al. 2018). It can be restrictive and is not applicable in general settings.

Our Contributions

We propose a strategy for testing a general hypothesis in a completely randomized factorial experiment. The null hypothesis asserts that certain average factorial effects are zero, but cannot determine all missing potential outcomes. Our strategy has two components.

First, we specify a sharp null hypothesis. It must imply the weak null hypothesis of interest and be compatible with the observed data. It also implies treatment-unit additivity. In particular, it implies constant factorial effects of and beyond the weak null hypothesis. Under this sharp null hypothesis, we can impute all missing potential outcomes.

Second, we use the FRT with a studentized test statistic. Like other test statistics, the *randomization distribution* of this studentized statistic depends on unknown potential outcomes in general. Fortunately, its *permutation distribution* under the above sharp null hypothesis stochastically dominates its randomization distribution asymptotically. The former distribution is the actual distribution, but, due to its dependence on unknown quantities, is not computable in general. The latter distribution is a proxy under a special sharp null hypothesis, but it is computable using the permutation test.

The stochastic dominance relationship between them allows us to construct an asymptotically conservative test. Therefore, for testing the weak null hypothesis, we can use the FRT with the studentized statistic. Without studentization, the FRT may not control type I error even asymptotically. We examine several existing test statistics, and show that using them in FRTs can give wrong type I errors.

The idea of studentization already appears in the literature. Our theory has some new features. First, Neuhaus (1993), Janssen (1997), Janssen (1999), Janssen and Pauls (2003) and Chung and Romano (2013) used it in permutation tests, assuming that the outcomes are independent draws. In our fomulation, the random treatment assignment drives the statistical inference with fixed potential outcomes. We do not assume any exchangeability of outcomes. Asymptotically, the randomization distribution of the studentized statistic is not pivotal, but its permutation distribution is. In general, the FRT is conservative for the weak null hypothesis. This conservativeness is a feature of finite population causal inference (Neyman 1990; Imbens and Rubin 2015; Ding and Dasgupta 2018). It did not appear in the literature regarding permutation tests. Second, Pauly et al. (2015) used studentization in permutation tests for better empirical finite sample properties. It is more crucial in our setting because without studentization the FRT cannot control type I error in general. Third, Babu and Singh (1983) and Hall (1988) used it to achieve better second order accuracy in the bootstrap. Although the bootstrap has been a workhorse for many other statistical problems, Imbens and Menzel (2018) just started this direction of using it for finite population causal inference. The bootstrap is another resampling method for testing the weak null hypotheses. Compared to the bootstrap, FRTs have an additional advantage of being finite-sample exact under sharp null hypotheses.

Notation

Let 1_n and 0_n be vectors of n 1's and 0's, respectively. Let $1(\cdot)$ denote the indicator that an event happens. Let $A \succeq 0$ and $A \succ 0$ if A is positive semi-definite and positive definite, respectively. Write $A \succeq B$ if $A - B \succeq 0$. Let $\lambda_j(A)$ be the j-th largest eigenvalue of A. Let $\mathrm{diag}\{\cdot\}$ be a diagonal or block-diagonal matrix. If (X_N) is a sequence of random variables indexed by N, write $X_N \xrightarrow{\mathrm{d}} X$, $X_N \xrightarrow{\mathbb{P}} X$, $X_n \xrightarrow{\mathrm{a.s.}} X$ for convergence in distribution, probability, and almost surely, respectively. For random vectors or matrices, we use the same notation to denote convergence, entry by entry. Almost surely is also denoted "a.s." Let Π_N denote the set of permutations of $\{1, \dots, N\}$. Let π denote a generic element of Π_N, which is a mapping from $\{1, \dots, N\}$ to itself. Let $\mathrm{Unif}(\Pi_N)$ denote the uniform distribution over Π_N. Random variable B stochastically dominates A, written $A \leq_{\mathrm{st}} B$, if their cumulative distribution functions $F_A(x)$ and $F_B(x)$ satisfy $F_A(x) \geq F_B(x)$ for all x. Let $\xi_1, \xi_2, \dots$ be independent and identically distributed (i.i.d.) $\mathcal{N}(0,1)$ random variables.

FRTs Under the Potential Outcomes Framework

Notation for Completely Randomized Experiments

We use the potential outcomes framework of (Rubin 1974; Neyman 1990). Let $Y_i(j)$ be the outcome of unit i if it receives treatment j, where $i \in \{1, \dots, N\}$ and $j \in \{1, \dots, J\}$. Vectorize $Y_i = (Y_i(1), \dots, Y_i(J))^\top$. The means of the potential outcomes are $\bar{Y}(j) = N^{-1} \sum_{i=1}^N Y_i(j)$, vectorized as $\bar{Y} = (\bar{Y}(1), \dots, \bar{Y}(J))^\top$. The covariance between the potential outcomes is $S(j,k) = (N-1)^{-1} \sum_{i=1}^N \{Y_i(j) - \bar{Y}(j)\}\{Y_i(k) - \bar{Y}(k)\}$, which is a variance if $j = k$. The covariance matrix S has the (j,k)-th entry $S(j,k)$.

Let $W_i \in \{1, \dots, J\}$ represent the treatment that unit i receives, and define the indicator $W_i(j) = 1(W_i = j)$. We assume a completely randomized experiment (CRE). The experimenter picks $N_1, \dots, N_J \geq 2$ that sum to N, and assigns treatments randomly so that any realization satisfies $\sum_{i=1}^N W_i(j) = N_j$ for $j = 1, \dots, J$, and has probability $\prod_{j=1}^J N_j!/N!$. Unit i's observed outcome is $Y_i^{\mathrm{obs}} = Y_i(W_i) = \sum_{j=1}^J W_i(j) Y_i(j)$. So the observed means are $\hat{\bar{Y}}(j) = N_j^{-1} \sum_{i=1}^N W_i(j) Y_i^{\mathrm{obs}}$, vectorized as $\hat{\bar{Y}} = (\hat{\bar{Y}}(1), \dots, \hat{\bar{Y}}(J))^\top$. The observed variances are $\hat{S}(j,j) = (N_j - 1)^{-1} \sum_{i=1}^N W_i(j) \{Y_i^{\mathrm{obs}} - \hat{\bar{Y}}(j)\}^2$, which is the sample analog of $S(j,j)$. Because $Y_i(j)$ and $Y_i(k)$ are not jointly observable, there is no sample analog for $S(j,k)$. In general, we cannot estimate $S(j,k)$ consistently for $j \neq k$. For simplicity, we assume $S(j,j) > 0$ and $\hat{S}(j,j) > 0$ for all $W = (W_1, \dots, W_N)^\top$.

FRTs for Sharp and Weak Null Hypotheses

Fisher (1935) proposed the randomization test to analyze experimental data. Several formulations exist for the FRT (Pitman 1937; Hoeffding 1952; Basu 1980; Romano 1990). We use the formulation of Rubin (1980) based on potential outcomes.

Rubin (2005) called the potential outcome matrix $\{Y_i(j) : i = 1, \ldots, N; j = 1, \ldots, J\}$ the Science Table. He called a null hypothesis sharp if it, along with the observed data, can determine all the missing items in the Science Table. A test statistic is a function of the observed data and the null hypothesis. Under a sharp null hypothesis, any test statistic has a known randomization distribution. In particular, for every realization of W, we can first obtain the corresponding realization of observed data, and then compute the value of the test statistic. This gives the permutation distribution of the test statistic under a sharp null hypothesis. We can obtain the p-value by comparing the observed value of the test statistic to its permutation distribution. FRTs are therefore finite-sample exact for any test statistic and data generating process (Rosenbaum 2002a; Imbens and Rubin 2015).

Our main interest is to test

$$H_{0\mathrm{N}}(C, x) : C\bar{Y} = x, \tag{1}$$

where $x \in \mathbb{R}^m$ and $C \in \mathbb{R}^{m \times J}$ is a contrast matrix of full row rank m. A contrast matrix by definition satisfies $C1_J = 0_m$. We are particularly interested in $x = 0_m$, but study general x for completeness. The hypothesis $H_{0\mathrm{N}}(C, x)$ is not sharp by Rubin's definition. It only imposes restrictions on the averages of the potential outcomes. Hence, it is referred to as an average null, a weak null, or a Neyman null. In contrast, a sharp null, or a strong null, or a Fisher null, imposes restrictions on the individual potential outcomes.

We would like to test (1) with a statistic T. We cannot obtain the exact randomization distribution of T because it depends on unknown potential outcomes in general. To use the FRT, we need a sharp null hypothesis. We use

$$H_{0\mathrm{F}}(C, x, \tilde{C}, \tilde{x}) : \begin{pmatrix} C \\ \tilde{C} \end{pmatrix} Y_i = \begin{pmatrix} x \\ \tilde{x} \end{pmatrix} \text{ for } i = 1, \ldots, N, \tag{2}$$

where the matrix $(C^\top, \tilde{C}^\top, 1_J)$ is invertible. Given C and 1_J, we can construct $\tilde{C}$ using Gram-Schmidt orthogonalization. We also need to pick a value $\tilde{x} \in \mathbb{R}^{J-m-1}$. If $m = J - 1$, we do not need to construct $\tilde{C}$ or select $\tilde{x}$. Note $\tilde{x}$ is a nuisance parameter whose selection is necessary because the FRT demands an accompanying sharp null hypothesis. In the case with $x = 0_m$, we can simply choose $\tilde{x} = 0_{J-m-1}$ to get the classical sharp null of no individual effects whatsoever. The null hypothesis $H_{0\mathrm{F}}(C, x, \tilde{C}, \tilde{x})$ is sharp because it specifies all individual effects. It has two important features. First, it implies the weak null hypothesis of interest, $H_{0\mathrm{N}}(C, x)$. Second, it implies strict additivity, i.e., $Y_i(j) - Y_i(k)$ does not depend on the unit i, for $j, k = 1, \ldots, J$.

With the sharp null hypothesis (2), the FRT proceeds as follows.

FRT-1. Calculate T from $\{W_i, Y_i^{\mathrm{obs}} : i = 1, \ldots, N\}$.

FRT-2. Impute potential outcomes:

$$Y_i^* = \begin{pmatrix} Y_i^*(1) \\ \vdots \\ Y_i^*(J) \end{pmatrix} = z + (Y_i^{\text{obs}} - z_{W_i})1_J, \text{ where } z = \begin{pmatrix} z_1 \\ \vdots \\ z_J \end{pmatrix} = \begin{pmatrix} C \\ \tilde{C} \\ 1_J^\top \end{pmatrix}^{-1} \begin{pmatrix} x \\ \tilde{x} \\ 0 \end{pmatrix},$$

or, equivalently, $Y_i^*(j) = Y_i^{\text{obs}} + z_j - z_{W_i}$ for $j = 1, \ldots, J$.

FRT-3. For a permutation $\pi \in \Pi_N$, compute $Y_{\pi,i}^{\text{obs}} = \sum_{j=1}^{J} W_{\pi(i)}(j) Y_i^*(j)$ and calculate T_π from $\{W_{\pi(i)}, Y_{\pi,i}^{\text{obs}} : i = 1, \ldots, N\}$ the same way T was calculated.

FRT-4. The p-value is $(N!)^{-1} \sum_{\pi \in \Pi_N} 1(T_\pi \geq T)$.

As a sanity check, the imputed potential outcomes in FRT-2 satisfy $H_{0\text{F}}(C, x, \tilde{C}, \tilde{x})$ and $Y_i^*(W_i) = Y_i^{\text{obs}}$ for all i. In FRT-3, we permute the treatment labels. This differs from the usual "permutation tests" in which it is equivalent to permute the treatment labels or the observed outcomes. Our test reduces to the usual permutaton test if (2) asserts that $Y_i(j) = Y_i^{\text{obs}}$ for all $i = 1, \ldots, N$ and $j = 1, \ldots, J$. Implicitly, we use a larger value of T to denote a larger deviation from the null hypothesis. Therefore, the p-value in FRT-4 is the right-tail probability. If $N!$ is too large for computation, we can take iid draws from Π_N to approximate the p-value in FRT-4 subject to Monte Carlo error.

For any test statistic T, the p-value in (4) is valid under $H_{0\text{F}}(C, x, \tilde{C}, \tilde{x})$. Our central goal is to investigate whether the FRT can still control type I error for testing $H_{0\text{N}}(C, x)$. Roughly speaking, this turns out to be the case asymptotically with an appropriate test statistic T.

Basic Asymptotics for Finite Population Causal Inference

We have argued that the exact randomization distribution of T depends on unknown potential outcomes under $H_{0\text{N}}(C, x)$ in general. Finite-sample theory in this case is too challenging. Instead, we develop an asymptotic theory. Imagine a sequence of finite populations of potential outcomes. For each $N \geq 2J$, we fix in advance $N_1, \ldots, N_J \geq 2$. Independently across N, we generate W according to a CRE, from which we get Y_i^{obs} and calculate a test statistic. We use $(\cdot)$ or $(\cdot)_{N \geq 2J}$ to denote a sequence indexed by N with $N \to \infty$. Technically, we should index finite population quantities by N, and also index observed quantities by $N_1, \ldots, N_J$. For simplicity, and following the precedent of earlier authors, we usually drop these extra subscripts, unless we want to emphasize the dependence on N. We now state our assumptions on the sequence of potential outcomes.

Assumption 1. The sequence (N_j/N) converges to $p_j \in (0, 1)$ for all $j = 1, \ldots, J$. The sequences $(\bar{Y}_N)$ and (S_N) converge to $\bar{Y}_\infty < \infty$ and S_∞, where S_∞ has finite entries and positive main diagonal entries. Further,

$$\lim_{N \to \infty} \max_{j=1,\ldots,J} \max_{i=1,\ldots,N} \frac{1}{N} \{Y_i(j) - \bar{Y}(j)\}^2 = 0.$$

Assumption 2. Same as Assumption 1 with the last identity replaced by: there exists an $L < \infty$ such that

$$N^{-1}\sum_{i=1}^{N}\{Y_i(j) - \bar{Y}(j)\}^4 \le L$$

for all $j = 1, \dots, J$ and $N \ge 2J$.

Proposition 1. Assumption 2 implies Assumption 1.

The design of experiments often guarantees the existence of $p_j \in (0,1)$ because all treatment groups have comparable sizes in realistic cases. We can weaken the existence of $\bar{Y}_\infty$ and S_∞ by standardizing the potential outcomes. We might drop subscripts ∞ just as we drop N. For instance, S can mean the finite population covariance matrix or its limiting value, which will be clear from context. Intuitively, Assumption 1 requires more than two moments, and Assumption 2 requires four moments. Assumption 2 is thus stronger than Assumption 1. The following two results, as consequences of Li and Ding (2017), are our main asymptotic tools.

Proposition 2. Under Assumption 1, $\hat{\bar{Y}} - \bar{Y} \xrightarrow{\mathbb{P}} 0_J$, and $\hat{S}(j,j) \xrightarrow{\mathbb{P}} S(j,j)$ for $j = 1, \dots, J$.

Proposition 3. Under Assumption 1, $\sqrt{N}(\hat{\bar{Y}} - \bar{Y}) \xrightarrow{\mathrm{d}} \mathcal{N}(0_J, V)$, where

$$V = \lim_{N\to\infty} N \cdot \mathrm{Cov}(\hat{\bar{Y}}) = \lim_{N\to\infty} \begin{pmatrix} \frac{N-N_1}{N_1}S(1,1) & -S(1,2) & \cdots & -S(1,J) \\ -S(2,1) & \frac{N-N_2}{N_2}S(2,2) & \cdots & -S(2,J) \\ \vdots & \vdots & \ddots & \vdots \\ -S(J,1) & -S(J,2) & \cdots & \frac{N-N_J}{N_J}S(J,J) \end{pmatrix}. \tag{3}$$

The limiting distribution in Proposition 3 depends on unknown quantities. We need to estimate $N \cdot \mathrm{Cov}(\hat{\bar{Y}})$. The covariance, however, depends on $S(j,k)$ ($j \neq k$), which do not have unbiased estimators in general. Estimating the main diagonal is the best we can hope to do. Define

$$\hat{D} \equiv N \times \mathrm{diag}\left\{\hat{S}(1,1)/N_1, \cdots, \hat{S}(J,J)/N_J\right\} \succ 0.$$

Proposition 2 implies

$$\hat{D} \xrightarrow{\mathbb{P}} D \equiv \mathrm{diag}\left\{S(1,1)/p_1, \cdots, S(J,J)/p_J\right\} \succ 0. \tag{4}$$

Therefore, $V = D - S \preceq D$. We use $\hat{D}$ as an asymptotically conservative estimator for $N \cdot \mathrm{Cov}(\hat{\bar{Y}})$ in the sense that $\lim_{N\to\infty} N \cdot \mathrm{Cov}(\hat{\bar{Y}}) \preceq \mathrm{plim}_{N\to\infty} \hat{D}$. This is a main idea we will use repeatedly.

Test Statistics

Our main interest is whether the FRT with a test statistic T can control type I error when testing the weak null hypothesis $H_{0\mathrm{N}}(C, x)$. The next proposition makes precise what kind of T can accomplish this goal. Recall that T_π in FRT-3 is the permutational version of T, in which case we call T *suitable*.

Proposition 4. Consider testing $H_{0\mathrm{N}}(C,x)$. The FRT with test statistic T controls type I error at all levels if and only if $T \leq_{\mathrm{st}} T_\pi | W$ a.s. The FRT with test statistic T controls the asymptotic type I error at all levels if and only if the asymptotic distribution of $T_\pi|W$ stochastically dominates that of T for all almost all sequences of W.

Proposition 4 is intuitive. We use test statistic T, but use $T_\pi|W$ as the reference null distribution under $H_{0\mathrm{N}}(C,x)$. The p-value in FRT-4 is a right-tail probability that $T_\pi|W$ is at least the observed value of T. If $T_\pi|W$ stochastically dominates T, then any quantile of the asymptotic distribution of $T_\pi|W$ is at least that of T. Consequently, we have asymptotically conservative tests at any level.

Studentized Statistic

We advocate using the following statistic in the FRT:

$$X^2 = N(C\hat{\bar{Y}} - x)^\top (C\hat{D}C^\top)^{-1}(C\hat{\bar{Y}} - x). \tag{5}$$

X^2 is a Wald-type statistic with a conservative covariance estimator $C\hat{D}C^\top$ for $\sqrt{N}(C\hat{\bar{Y}} - x)$.

It is commonly called a studentized statistic. Their use in permutation tests appeared in a different formulation with independent samples. Romano (1990) pointed out the problem of un-studentized statistics in two sample tests. Janssen (1997) proposed using a studentized statistic to control the type I error without assuming equal distributions in the two sample problem. Chung and Romano (2013) studied the same phenomenon when the parameter of interest could be more general than the mean. Pauly et al. (2015) and Konietschke et al. (2015) used an equivalent studentized statistic in factorial experiments with independent samples. In those settings, studentization works because the test statistic is asymptotically pivotal.

In our case, X^2 is not asymptotically pivotal, but is instead stochastically dominated by a pivotal distribution. This is a key reason for its suitability based on Proposition 4. Now we state our main result that the FRT with X^2 is robust for two null hypotheses.

Theorem 1. If Assumption 1 holds, then under $H_{0\mathrm{N}}(C,x)$, $X^2 \xrightarrow{\mathrm{d}} \sum_{j=1}^m a_j \xi_j^2$, where $a_j \in [0,1]$ for $j = 1, \ldots, m$. If Assumption 2 holds and $\pi \sim \mathrm{Unif}(\Pi_N)$, then $X_\pi^2 | W \xrightarrow{\mathrm{d}} \chi_m^2$ a.s. In particular, the FRT with test statistic X^2 can asymptotically control type I error under $H_{0\mathrm{N}}(C,x)$ a.s.

Asymptotically, under $H_{0\mathrm{N}}(C,x)$, neither the distribution of X^2 nor that of $X_\pi^2|W$ depends on $\tilde{C}, \tilde{x}$, so the choice of $\tilde{x}$ does not matter. The distribution of X_π^2 also does not depend on $H_{0\mathrm{N}}(C,x)$. A violation of $H_{0\mathrm{N}}(C,x)$ is likely to increase the value of X^2 but not the values of X_π^2. An appealing consequence of these facts is that the FRT has power; see Pauly et al. (2015) and Chung and Romano (2013).

The FRT using X^2 controls the asymptotic type I error conservatively under $H_{0\mathrm{N}}(C, x)$. This holds by Theorem 1. This type of FRT is robust to weak null hypotheses and treatment effect heterogeneity. Our purpose of using the studentized statistic in the FRT is different from that in the literature. Pauly et al. (2015) and Konietschke et al. (2015) used it for better small sample performance. In those settings, the conservative issue did not exist.

Theorem 1 also justifies another asymptotically conservative test without using the FRT. We can reject $H_{0\mathrm{N}}(C, x)$ if $X^2 > \kappa_{m,\alpha}$, the $1-\alpha$ quantile of χ^2_m. This is computationally efficient without Monte Carlo. The FRT using X^2 has an additional property. It retains finite-sample exactness for testing $H_{0\mathrm{F}}(C, x, \tilde{C}, \tilde{x})$. This holds by the definition of the FRT in Section 2.2.

Box-Type Statistic

We now consider an alternative statistic, the Box-type statistic studied in Brunner et al. (1997). Because we will show it is not suitable in our context, we can restrict the discussion to $x = 0_m$. The test statistic is

$$B = N\hat{\bar{Y}}^\top M \hat{\bar{Y}} / \operatorname{tr}(M\hat{D}) \tag{6}$$

where $M = C^\top (CC^\top)^{-1} C$ is the projection matrix onto row space of C.

Under independent sampling, Brunner et al. (1997) approximated the asymptotic behavior of B by an F distribution using ideas from Box (1954), and called it a Box-type statistic. They advocated using B because of its superior empirical small sample properties under their framework.

Recall V in (3) and define $P = \operatorname{diag}\{p_1, \dots, p_J\}$. We have the following theorem.

Theorem 2. If Assumption 1 holds, then under $H_{0\mathrm{N}}(C, 0_m)$, $B \xrightarrow{\mathrm{d}} \sum_{j=1}^m \lambda_j(VM)\xi_j^2 / \operatorname{tr}(DM)$. If Assumption 2 holds and $\pi \sim \operatorname{Unif}(\Pi_N)$, then $B_\pi | W \xrightarrow{\mathrm{d}} \sum_{j=1}^m \lambda_j(P^{-1}M)\xi_j^2 / \operatorname{tr}(P^{-1}M)$ a.s.

From Theorem 2, the asymptotic mean of B is $\sum_{j=1}^m \lambda_j(VM) / \operatorname{tr}(DM) \leq 1$ because $V \preceq D$, and the asymptotic mean of $B_\pi | W$ is $\sum_{j=1}^m \lambda_j(P^{-1}M)\xi_j^2 / \operatorname{tr}(P^{-1}M) = 1$. Therefore, the former does not exceed the latter. This is necessary but not sufficient for the stochastic dominance condition of Proposition 4. Hence, the FRT with the Box-type statistic cannot control type I error in general, even asymptotically. We demonstrate this later by a simulation.

There are two situations where the B is suitable: equal variances, and testing a one-dimensional hypothesis.

Corollary 1. Under Assumption 2, if $S(1,1) = \cdots = S(J,J)$, then B meets the condition of Proposition 4 asymptotically. If C is a row vector, then $B = X^2$.

Statistics From the Ordinary Least Squares

It is common to analyze experimental data based on the ordinary least squares (OLS) fit of a (Normal) linear model (e.g., Morris 2010). The design matrix is a block diagonal matrix $\mathcal{X} = \operatorname{diag}\left\{1_{N_1}, \dots, 1_{N_J}\right\}$, and the response vector consists of the corresponding observed outcomes from treatment groups $1, \ldots, J$. The OLS coefficients are given in $\hat{\bar{Y}}$ with estimated covariance matrix $\hat{\sigma}^2(\mathcal{X}^\top \mathcal{X})^{-1}$, where $\hat{\sigma}^2 = (N-J)^{-1} \sum_{i=1}^N \sum_{j=1}^J W_i(j)\{Y_i^{\text{obs}} - \hat{\bar{Y}}(j)\}^2$ is the mean residual sum of squares. Based on these, the classical F statistic is

$$F = (C\hat{\bar{Y}})^\top \{\hat{\sigma}^2 C(\mathcal{X}^\top \mathcal{X})^{-1} C^\top\}^{-1} C\hat{\bar{Y}}/m \tag{7}$$

We do not make the usual assumptions of linear regression, but simply use it to get a test statistic F.

We first point out an interesting situation where F is identical to the Box-type statistic B. This result will be useful in our simulations.

Proposition 5. $B = F$ if $N_1 = \cdots = N_J$ and $M = C^\top(CC^\top)^{-1}C$ has the same entries along its main diagonal.

Except for the scaling by m and the use of $\hat{\sigma}^2$ in place of each $\hat{S}(j,j)$, F is identical to X^2. The use of a pooled variance estimate $\hat{\sigma}^2$ makes the F statistic problematic, as the following result formalizes.

Theorem 3. If Assumption 1 holds, then under $H_{0\text{N}}(C, 0_m)$, $m \times F \xrightarrow{\text{d}} \sum_{j=1}^m \lambda_j \big(CVC^\top(\bar{S}CP^{-1}C^\top)^{-1}\big)\xi_j^2$ where $\bar{S} = \sum_{j=1}^J p_j S(j,j)$. If Assumption 2 holds and $\pi \sim \text{Unif}(\Pi_N)$, then $m \times F_\pi | W \xrightarrow{\text{d}} \chi^2_m$ a.s.

The classical linear model implicitly assumes a constant treatment effect for all units (Kempthorne 1952). This implies equal variances under all treatment levels. In contrast, the potential outcomes framework does not assume such homogeneity. The assumptions underlying the F statistic are not compatible with the potential outcomes framework in general. If the potential outcomes do have equal variance, then it is not surprising that F is suitable.

Corollary 2. Under Assumption 2, if $S(1,1) = \cdots = S(J,J)$, then F meets the condition of Proposition 4 asymptotically.

A simple fix to the classic F statistic is to use the Huber–White covariance estimator for the least squares coefficients (Huber 1967; White 1980). It is popular in econometrics to use such an estimate of the covariance when the linear model is possibly misspecified or the error terms are heteroskedastic. Define the residual $\hat{\epsilon}_i = Y_i^{\text{obs}} - \hat{\bar{Y}}(W_i)$. Algebra shows that the Huber–White estimator for $N \cdot \text{Cov}(\hat{\bar{Y}})$ is

$$\begin{aligned} \hat{D}_{\text{HW}} &= N(\mathcal{X}^\top \mathcal{X})^{-1} \mathcal{X}^\top \text{diag}\left\{\hat{\epsilon}_1^2, \ldots, \hat{\epsilon}_N^2\right\} \mathcal{X}(\mathcal{X}^\top \mathcal{X})^{-1} \\ &= N \times \text{diag}\left\{\frac{N_1 - 1}{N_1^2}\hat{S}(1,1), \ldots, \frac{N_J - 1}{N_J^2}\hat{S}(J,J)\right\}. \end{aligned}$$

If we use $\hat{D}_{\text{HW}}$ in place of $\hat{\sigma}^2(\mathcal{X}^\top \mathcal{X})^{-1}$ in (7) and ignore the scaling by m, we get

$$X^2_{\text{HW}} = N(C\hat{\bar{Y}})^\top (C\hat{D}_{\text{HW}}C^\top)^{-1} C\hat{\bar{Y}}.$$

$\hat{D}_{\mathrm{HW}}$ is nearly identical to $\hat{D}$ if $N_j \approx N_j - 1$ for $j = 1, \ldots, J$. Therefore, X^2_{HW} and X^2 have the same asymptotic properties, and the Huber–White covariance estimator successfully salvages the F statistic.

Special Cases

Section 3 proposes a strategy for testing weak null hypotheses in general experiments. The results are directly applicable to many important examples.

One-Way Analysis of Variance

In the one-way analysis of variance (ANOVA), the goal is to test $H_{0\mathrm{N}} : \bar{Y}(1) = \cdots = \bar{Y}(J)$. It is a special case of the null hypothesis (1) with any contrast matrix $C \in \mathbb{R}^{(J-1)\times J}$ and $x = 0_{J-1}$, for instance $C = (1_{J-1}, -I_{J-1})$. In this case, we can impute potential outcomes in FRT-2 as $Y_i^*(j) = Y_i^{\mathrm{obs}}$ for $i = 1, \ldots, N$ and $j = 1, \ldots, J$ under $H_{0\mathrm{F}} : Y_i(1) = \cdots = Y_i(J)$, for $i = 1, \ldots, N$. To test $H_{0\mathrm{F}}$, Fisher (1925) proposed using the statistic

$$F = \frac{\sum_{j=1}^J N_j \{\hat{\bar{Y}}(j) - \bar{Y}_{\bullet}^{\mathrm{obs}}\}^2/(J-1)}{\sum_{j=1}^J (N_j - 1)\hat{S}(j,j)/(N-J)}, \text{ where } \bar{Y}_{\bullet}^{\mathrm{obs}} = \frac{1}{N}\sum_{i=1}^N Y_i^{\mathrm{obs}}. \tag{8}$$

He argued that $F_{J-1,N-J}$ approximates the randomization distribution of F. Ding and Dasgupta (2018) showed (8) is not suitable but

$$X^2 = \sum_{j=1}^J \frac{N_j}{\hat{S}(j,j)}\{\hat{\bar{Y}}(j) - \bar{Y}_S^{\mathrm{obs}}\}^2, \text{ where } \bar{Y}_S^{\mathrm{obs}} = \frac{\sum_{j=1}^J N_j \hat{\bar{Y}}(j)/\hat{S}(j,j)}{\sum_{j=1}^J N_j/\hat{S}(j,j)} \tag{9}$$

is suitable for testing $H_{0\mathrm{N}}$ with the FRT. See Schochet (2018) for a related discussion.

The next result shows that our framework encompasses these results as special cases.

Proposition 6. In the one-way ANOVA, the X^2 in (5) and (9) coincide, as do the F in (7) and (8).

Treatment-Control Experiments

In the treatment-control setting, $J = 2$, and unit i either receives the treatment (then $Y_i^{\mathrm{obs}} = Y_i(1)$) or control (then $Y_i^{\mathrm{obs}} = Y_i(2)$). A parameter of interest is the average treatment effect $\tau = \bar{Y}(1) - \bar{Y}(2)$. The weak null is $H_{0\mathrm{N}}(C, 0) : \tau = 0$ where $C = (1, -1)$ is a row vector. Thus, treatment-control is a special case of the one-way layout. A commonly-used statistic is $|\hat{\tau}|$, where $\hat{\tau} = \hat{\bar{Y}}(1) - \hat{\bar{Y}}(2)$ is the sample difference-in-means of outcomes. However, Ding and Dasgupta (2018) pointed out that $|\hat{\tau}|$ is not suitable for testing $H_{0\mathrm{N}}$ in general.

Corollary 3. In the treatment-control setting, for almost all sequences of W, $B = X^2$ can asymptotically control type I error, but F and $|\hat{\tau}|$ cannot, unless $N_1 = N_2$ or $S(1,1) = S(2,2)$.

From Corollary 1, the Box-type statistic B equals the studentized statistic X^2 in the treatment-control setting. Both are suitable. The fact that $|\hat{\tau}|$ is unsuitable is related to a "paradox" in Ding (2017); see also the comment of Loh et al. (2017). Corollary 3 asserts that a balanced design can salvage the suitability of the F and $|\hat{\tau}|$ statistics, even without homoskedasticity. Interestingly, however, this does not extend to $J > 2$, as our simulations later demonstrate.

2^K Factorial Designs

In 2^K factorial designs, our goal is to analyze K binary treatment factors simultaneously. In total, we have $J = 2^K$ possible treatment combinations. Dasgupta et al. (2015) advocated analyzing these designs using the potential outcomes framework. To do so, it is helpful to introduce the model matrix $G \in \{\pm 1\}^{(J-1)\times J}$. Let $*$ denote the component-wise product. Lu (2016b) defined the rows of G as $g_1^\top, \dots, g_{J-1}^\top$ as follows:

- for $j = 1, \dots, K$, let $g_j^\top$ be $-1_{2^{K-j}}^\top, 1_{2^{K-j}}^\top$ repeated 2^{j-1} times;
- the next $\binom{K}{2}$ values of g_j's are $g_{k(1)} * g_{k(2)}$ where $k(1) \neq k(2) \in \{1, \dots, K\}$;
- the next $\binom{K}{3}$ are component-wise products of triplets of distinct $g_1, \dots, g_K$, etc;
- finally, $g_{J-1} = g_1 * \cdots * g_K$.

The matrix G has rows orthogonal to each other and to 1_J, i.e., $GG^\top = J \times I_{J-1}$ and $G1_J = 0_{J-1}$. Let $\tilde{G} \in \{\pm 1\}^{K \times J}$ be the first K rows of G. Call its columns $z_1, \dots, z_J$, which are the possible treatment combinations. The following example illustrates the setup.

Example 1. When $K = 2$, we have

$$G = \begin{pmatrix} -1 & -1 & 1 & 1 \\ -1 & 1 & -1 & 1 \\ 1 & -1 & -1 & 1 \end{pmatrix} = \begin{pmatrix} g_1^\top \\ g_2^\top \\ g_3^\top \end{pmatrix} = \begin{pmatrix} \tilde{G} \\ g_3^\top \end{pmatrix} = \begin{pmatrix} z_1 & z_2 & z_3 & z_4 \\ 1 & -1 & -1 & 1 \end{pmatrix}.$$

The four possible treatment combinations are $z_1 = (-1, -1)^\top$, $z_2 = (-1, 1)^\top$, $z_3 = (1, -1)^\top$, and $z_4 = (1, 1)^\top$. We read these off from the first two rows of G. □

The rows of G define factorial effects. $g_1, \dots, g_K$ correspond to main effects, $g_{K+1}, \dots, g_{K+\binom{K}{2}}$ correspond to two-way interactions, etc, and g_{J-1} corresponds to the K-way interaction. Let $Y_i(j) = Y_i(z_j)$ be the response of unit i if it receives the treatment combination z_j. Then our previous notation carries over. The general factorial effect for unit i indexed by g_j is $\tau_{ij} = 2J^{-1} g_j^\top Y_i$, and the corresponding average factorial effect is $\bar{\tau}_{\bullet j} = N^{-1} \sum_{i=1}^N \tau_{ij} = 2J^{-1} g_j^\top \bar{Y}$. Vectorize these quantities: $\tau_i = (\tau_{i1}, \dots, \tau_{i,J-1})^\top = 2J^{-1} G Y_i$ and $\tau = (\bar{\tau}_{\bullet 1}, \dots, \bar{\tau}_{\bullet J-1})^\top = 2J^{-1} G \bar{Y}$.

We may perform inference on τ or any subset of its entries. Let $A = \{a(1), \dots, a(m)\} \subseteq \{1, \dots, J-1\}$ be the subset of interest, and let $C \in \{\pm 1\}^{m \times J}$ have rows $g_{a(1)}^\top, \dots, g_{a(m)}^\top$. Then $\tau_A = (\bar{\tau}_{\bullet a(1)}, \dots, \bar{\tau}_{\bullet a(m)})^\top = 2J^{-1} C \bar{Y}$. Testing whether $\tau_A = 2J^{-1} x$ is equivalent to testing $H_{0\mathrm{N}}(C, x)$.

We can use the FRT with X^2. The factorial design gives a natural choice of $\tilde{C}$ for the imputation step FRT-2. We simply let $g_j^\top$ be a row of $\tilde{C}$ whenever $j \notin A$.

Lu (2016b) discussed both randomization-based and regression-based inferences for 2^K factorial designs. He focused on point estimation and proposed using the Huber–White covariance estimator. We have also highlighted the importance of using the Huber–White covariance estimator to modify the F statistic in the FRT.

One-Sided Tests

We can also use FRTs for one-sided hypotheses. We first consider

$$\tilde{H}_{0\mathrm{N}}(C, x) : C\bar{Y} \geq x, \tag{10}$$

where $C \in \mathbb{R}^{1\times J}$ is a row vector with $C1_J = 0$, and $x \in \mathbb{R}$.

Example 2. In the two-sample problem with $J = 2$, we can test $\bar{Y}(2) - \bar{Y}(1) \geq 0$: whether the treatment level 1 results in smaller outcomes than treatment level 2 on average. In this case, $C = (-1, 1)$ and $x = 0$. □

Example 3. In a gold standard design for three arms, let level 1 be the placebo control, level 2 be the active control, and level 3 be the experimental treatment. Assume that smaller outcomes are more desirable, and we know that $\bar{Y}(2) > \bar{Y}(1)$ from previous studies. Given $\Delta > 0$, the goal is to test the hypothesis $\bar{Y}(1) - \bar{Y}(3) \leq \Delta\{\bar{Y}(1) - \bar{Y}(2)\}$, with $\Delta > 1$ corresponding to a superiority test and $\Delta \in (0, 1)$ corresponding to a non-inferiority test (Mutze et al. 2017). This null hypothesis is equivalent to $\tilde{H}_{0\mathrm{N}}(C, 0) : (\Delta - 1)\bar{Y}(1) - \Delta\bar{Y}(2) + \bar{Y}(3) \geq 0$ with $C = (\Delta - 1, -\Delta, 1)$. □

To impute the missing potential outcomes, we pretend that the null hypothesis is $H_{0\mathrm{N}}(C, x)$ and use the same procedure as before. We cannot use X^2 because it is intended for two-sided tests. For instance, X^2 can be large, even under $\tilde{H}_{0\mathrm{N}}(C, x)$. Instead we use a one-sided statistic $t_+ = \max(t, 0)$ where

$$t = \sqrt{N}(x - C\hat{\bar{Y}})/(C\hat{D}C^\top)^{1/2}.$$

Using t for the FRT also works for p-values at most 0.5. Mutze et al. (2017) used t for Example 3. We use t_+ so that Proposition 4 is directly applicable. We summarize the results below.

Corollary 4. Consider testing $\tilde{H}_{0\mathrm{N}}(C, x)$ in (10). If Assumption 1 holds, then under $\tilde{H}_{0\mathrm{N}}(C, x)$, $t \xrightarrow{\mathrm{d}} \mathcal{N}(0, a)$ for some $a \in [0, 1]$. If Assumption 2 holds and $\pi \sim \mathrm{Unif}(\Pi_N)$, then $t_\pi | W \xrightarrow{\mathrm{d}} \mathcal{N}(0, 1)$ for almost all sequences of W. In particular, for almost all sequences of W, the FRT with test statistic t_+ can asymptotically control type I error under $\tilde{H}_{0\mathrm{N}}(C, x)$.

When $C \in \mathbb{R}^{m\times J}$ and $x \in \mathbb{R}^m$ for $m > 1$, we can interpret (10) as component-wise inequalities. Neither X^2 nor t_+ can be used in this case. A simple fix is to test each component using t_+ with a Bonferroni correction.

Cluster-Randomized Experiments

In many cases, the N units are partitioned into L clusters (e.g., classrooms in educational studies, villages in public health studies). A cluster-randomized experiment assigns treatments to clusters. Units within a cluster must receive the same treatment. For $l = 1, \ldots, L$, let $\breve{W}_l \in \{1, \ldots, J\}$ represent the treatment that cluster l receives, and define the indicator $\breve{W}_l(j) = 1(\breve{W}_l = j)$. There are $L!/\prod_{j=1}^{J} L_j!$ possible realizations of $\{\breve{W}_1, \ldots, \breve{W}_L\}$. The mechanism of treatment assignment to clusters is identical to that to individuals in a CRE.

Middleton and Aronow (2015) pointed out that we cannot perform the same analysis as if we had a CRE on the N units. For instance, $\hat{\bar{Y}}(j)$ is no longer an unbiased estimator for $\bar{Y}(j)$ if the cluster sizes vary. Both Middleton and Aronow (2015) and Li and Ding (2017) advocated a CRE-like analysis. Let $X_i \in \{1, \ldots, L\}$ represent which cluster unit i is in. Define $A_l(j) = \sum_{i=1}^{N} I(X_i = l) Y_i(j)$, and view $\{A_l(j) : l = 1, \ldots, L, j = 1, \ldots, J\}$ as the aggregated potential outcomes at the cluster level. A cluster-randomized experiment is a CRE on the clusters. Define $A_l = (A_l(1), \ldots, A_l(J))^\top$, A_l^{obs}, $\bar{A} = (\bar{A}(1), \ldots, \bar{A}(J))^\top$, $\hat{\bar{A}} = (\hat{\bar{A}}(1), \ldots, \hat{\bar{A}}(J))^\top$ analogously to our previous notation for a CRE. First, $\bar{Y} = L\bar{A}/N$, so $L\hat{\bar{A}}/N$ is unbiased for $\bar{Y}$. Define $\hat{S}_A(j,j) = (L_j - 1)^{-1} \sum_{l=1}^{L} \breve{W}_l(j) \{A_l^{\text{obs}} - \hat{\bar{A}}(j)\}^2$ and $\hat{D}_A = L \times \operatorname{diag}\left\{\hat{S}_A(1,1)/L_1, \ldots, \hat{S}_A(J,J)/L_J\right\}$. We modify the X^2 statistic as

$$X_A^2 = L(C\hat{\bar{A}} - Nx/L)^\top (C\hat{D}_A C^\top)^{-1} (C\hat{\bar{A}} - Nx/L).$$

Then Theorem 1 tells us that X_A^2 is suitable for $H_{0\mathrm{N}}(C, x)$ as $L \to \infty$ if Assumption 2 holds for the aggregated potential outcomes.

Hodges–Lehmann estimation

Up to this point, we have focused on hypothesis testing. Using the duality between testing and estimation, we now extend previous results to estimation of $C\bar{Y}$. This strategy is sometimes referred to as Hodges–Lehmann estimation (Hodges and Lehmann 1963; Rosenbaum 2002a). For a fixed x, we can use the FRT to obtain a p-value for the null hypothesis $H_{0\mathrm{N}}(C, x)$. Let us denote this p-value by $p(x)$ to highlight its dependence on x.

The Hodges–Lehmann point estimator for $C\bar{Y}$ is the $x \in \mathbb{R}^m$ that results in the least significant p-value for testing $H_{0\mathrm{N}}(C, x)$: $\hat{\tau}_{\mathrm{HL}} \in \operatorname{argmax}_{x \in \mathbb{R}^m} p(x)$. Note that $x = C\bar{Y}^{\text{obs}}$ implies $X^2 = 0$, which in turn implies a p-value of 1. Thus $\hat{\tau}_{\mathrm{HL}} = C\bar{Y}^{\text{obs}}$, the usual unbiased estimator. Because X^2 is suitable, we have the following immediate corollary due to the duality between hypothesis testing and confidence sets.

Corollary 5. For $\alpha \in (0, 1)$ and almost all sequences of W, an asymptotically conservative $(1 - \alpha)$ confidence set for $C\bar{Y}$ is

$$\mathrm{CR}_\alpha = \left\{x \in \mathbb{R}^m : p(x) > \alpha\right\},$$

in the sense that $\lim_{N \to \infty} \mathbb{P}\{C\bar{Y} \in \mathrm{CR}_\alpha\} \geq 1 - \alpha$.

Getting a handle on CR_α can be computationally intensive, so it is useful to have the asymptotic approximation

$$\mathrm{CR}_\alpha \approx \left\{ x : N(C\hat{\bar{Y}} - x)^\top (C\hat{D}C^\top)^{-1}(C\hat{\bar{Y}} - x) \leq \kappa_{m,\alpha} \right\}, \tag{11}$$

where $\kappa_{m,\alpha}$ is the $1-\alpha$ quantile of χ^2_m. Because the X^2 statistic is a quadratic form, (11) is an ellipsoid centered at $C\hat{\bar{Y}}$. (11) can be used either directly as a $1-\alpha$ approximate confidence set or as an initial guess for searching the exact confidence region by inverting FRTs. We illustrate this later by a simulation.

Extensions

Stratified Randomized Experiment

We extend previous results to the stratified randomized experiment (SRE), also called the randomized block design. The overall setup from the CRE still applies, but each unit now also has an associated covariate $X_i \in \{1, \dots, H\}$. The treatment W does not affect the covariate X. Thus, our data are now $\{Y_i^{\mathrm{obs}}, X_i, W_i : i = 1, \dots, N\}$. The W_i's remain the sole source of randomness. For $h = 1, \dots, H$, the h-th stratum consists of all units i where $X_i = h$, with size $N_{[h]} = \sum_{i=1}^N 1(X_i = h)$ and proportion $\omega_{[h]} = N_{[h]}/N$. For $h = 1, \dots, H$ and $j = 1, \dots, J$, the sample sizes $N_{[h]j} = \sum_{i=1}^N 1(X_i = h, W_i = j) \geq 2$ are predetermined. In a SRE, we assign treatments within each stratum just as in a CRE (Imbens and Rubin 2015).

We extend our previous notation to each stratum. For $h = 1, \dots, H$, define the mean vector $\bar{Y}_{[h]} \in \mathbb{R}^J$ with the j-th entry $\bar{Y}_{[h]}(j) = N_{[h]}^{-1} \sum_{i=1}^N 1(X_i = h) Y_i(j)$. The covariance $S_{[h]}$ has the (j,k)-th entry $S_{[h]}(j,k) = (N_{[h]} - 1)^{-1} \sum_{i=1}^N 1(X_i = h)\{Y_i(j) - \bar{Y}_{[h]}(j)\}\{Y_i(k) - \bar{Y}_{[h]}(k)\}$. We need the following regularity condition, which essentially requires that Assumption 2 be true within all strata.

Assumption 3. For $h = 1, \dots, H$, (1) $\lim_{N\to\infty} N_{[h]}/N = \omega_{[h]} \geq 0$ and $\lim_{N\to\infty} N_{[h]j}/N_{[h]} = p_{[h]j} > 0$; (2) the sequences $(\bar{Y}_{[h]})$ and $(S_{[h]})$ converge to $\bar{Y}_{[h]\infty}$ and $S_{[h]\infty}$; (3) the matrix $S_{[h]\infty}$ has strictly positive main diagonal entries; (4) there exists an $L < \infty$ such that $N_{[h]}^{-1} \sum_{i=1}^N 1(X_i = h)\{Y_i(j) - \bar{Y}_{[h]}(j)\}^4 \leq L$ for all N and $j = 1, \dots, J$.

We do not distinguish Assumptions 1 and 2 in the SRE for simplicity. With a little abuse of notation, we use $\omega_{[h]}$ for both $N_{[h]}/N$ and its limit. The sample mean vector is $\hat{\bar{Y}}_{[h]} \in \mathbb{R}^J$ with the j-th entry $\hat{\bar{Y}}_{[h]}(j) = N_{[h]j}^{-1} \sum_{i=1}^N 1(X_i = h, W_i = j) Y_i^{\mathrm{obs}}$, and the sample variance is $\hat{S}_{[h]}(j,j) = (N_{[h]j} - 1)^{-1} \sum_{i=1}^N 1(X_i = h, W_i = j)\{Y_i^{\mathrm{obs}} - \hat{\bar{Y}}_{[h]}(j)\}^2$. Under Assumption 3, Proposition 3 implies that, within stratum h, the standardized stratum-wise sample mean $\sqrt{N_{[h]}}(\hat{\bar{Y}}_{[h]} - \bar{Y}_{[h]})$ is asymptotically Normal with mean 0 and covariance $V_{[h]}$. A conservative estimator for $V_{[h]}$ is

$$\hat{D}_{[h]} = N_{[h]} \mathrm{diag}\left\{ \hat{S}_{[h]}(1,1)/N_{[h]1}, \dots, \hat{S}_{[h]}(J,J)/N_{[h]J} \right\}.$$

An unbiased estimator for $\bar{Y}$ is $\check{\bar{Y}} = \sum_{h=1}^{H} \omega_{[h]} \hat{\bar{Y}}_{[h]}$. Because of the independence across strata, $\sqrt{N}(\check{\bar{Y}} - \bar{Y})$ is asymptotically Normal with mean 0 and covariance $\sum_{h=1}^{H} \omega_{[h]} V_{[h]}$. A conservative variance estimator is $\check{D} = \sum_{h=1}^{H} \omega_{[h]} \hat{D}_{[h]}$.

Based on these, we define a modification of X^2 that is suitable for use with the FRT in a SRE:

$$\begin{aligned} X^2 &= N(C\check{\bar{Y}} - x)^\top (C\check{D}C^\top)^{-1}(C\check{\bar{Y}} - x) \\ &= N\left(C\sum_{h=1}^{H} \omega_{[h]}\hat{\bar{Y}}_{[h]} - x\right)^\top \left(\sum_{h=1}^{H} \omega_{[h]} C\hat{D}_{[h]}C^\top\right)^{-1} \left(C\sum_{h=1}^{H} \omega_{[h]}\hat{\bar{Y}}_{[h]} - x\right) \end{aligned} \tag{12}$$

The special case $h = 1$ agrees with (5), justifying the abuse of notation. Besides the form of the test statistic, the FRT requires two more modifications in the case of an SRE. First, we can impute the potential outcomes stratum by stratum under the sharp null hypothesis

$$H_{0\mathrm{F}}(C, x_{[1]}, \ldots, x_{[H]}, \tilde{C}, \tilde{x}_{[1]}, \ldots, \tilde{x}_{[H]}) : \begin{pmatrix} C \\ \tilde{C} \end{pmatrix} Y_i^* = \begin{pmatrix} x_{[h]} \\ \tilde{x}_{[h]} \end{pmatrix}, \text{ whenever } X_i = h.$$

The above null hypothesis must satisfy $\sum_{h=1}^{H} \omega_{[h]} x_{[h]} = x$. In the case with $x = 0_m$, a natural choice is $x_{[h]} = 0_m$ and $\tilde{x}_{[h]} = 0_{J-m-1}$. Under the above sharp null hypothesis, we can impute all potential outcomes: for units with $X_i = h$,

$$Y_i^* = \begin{pmatrix} Y_i^*(1) \\ \vdots \\ Y_i^*(J) \end{pmatrix} = z_{[h]} + (Y_i^{\mathrm{obs}} - z_{[h],W_i})1_J, \text{ where } z_{[h]} = \begin{pmatrix} z_{[h],1} \\ \vdots \\ z_{[h],J} \end{pmatrix} = \begin{pmatrix} C \\ \tilde{C} \\ 1_J^\top \end{pmatrix}^{-1} \begin{pmatrix} x_{[h]} \\ \tilde{x}_{[h]} \\ 0 \end{pmatrix},$$

or, equivalently, $Y_i^*(j) = Y_i^{\mathrm{obs}} + z_{[h],j} - z_{[h],W_i}$. Second, we need to permute the treatment indicators within strata, independently across strata. Let $\Pi_{N,\mathrm{bl}} \subseteq \Pi_N$ be all such permutations from a SRE. The p-value is $\left(\prod_{h=1}^{H} N_{[h]}!\right)^{-1} \sum_{\pi \in \Pi_{N,\mathrm{bl}}} 1(X_\pi^2 \geq X^2)$.

Theorem 4. In a SRE, suppose Assumption 3 holds. Under $H_{0\mathrm{N}}(C, x)$, $X^2 \xrightarrow{\mathrm{d}} \sum_{j=1}^{m} a_j \xi_j^2$, where each $a_j \in [0, 1]$. If $\pi \sim \mathrm{Unif}(\Pi_{N,\mathrm{bl}})$, then $X_\pi^2 | W \xrightarrow{\mathrm{d}} \chi_m^2$ for almost all sequences of W. In particular, the FRT with test statistic X^2 can asymptotically control type I error because the condition of Proposition 4 holds.

Even if the original experiment is a CRE, with a discrete X, we can condition on the number of treated and control units within all strata. Then the treatment assignment is identical to a SRE. Therefore, in a CRE, we can still permute the treatment indicators within each stratum of X. This gives a conditional randomization test. Zheng and Zelen (2008) and Hennessy et al. (2016) demonstrated that conditional randomization tests often improve the power as long as the covariates are predictive of the outcomes. Holt and Smith (1979) and Miratrix et al. (2013) discussed post-stratification, the estimation analog of testing.

We have focused on the SRE with large strata, i.e., $N_{[h]} \to \infty$ for $1 \leq h \leq H$. Our theory does not cover SREs with many small strata, i.e., the $N_{[h]}$'s are bounded but $H \to \infty$ (Fogarty 2018a). Although we conjecture that similar results hold in such cases, we leave technical details to future research.

Multiple Outcomes and Multiple Testings

Our framework also admits an extension to the case where all potential outcomes $Y_i(j) \in \mathbb{R}^d$ are vectors. Define $\bar{Y}(j)$ and $\hat{\bar{Y}}(j) \in \mathbb{R}^d$ as before. It is convenient to gather these into long vectors

$$\bar{Y} = \begin{pmatrix} \bar{Y}(1) \\ \vdots \\ \bar{Y}(J) \end{pmatrix} \in \mathbb{R}^{dJ}, \qquad \hat{\bar{Y}} = \begin{pmatrix} \hat{\bar{Y}}(1) \\ \vdots \\ \hat{\bar{Y}}(J) \end{pmatrix} \in \mathbb{R}^{dJ}.$$

Covariances are now matrices: $S(j,k) = (N-1)^{-1}\sum_{i=1}^{N}\{Y_i(j) - \bar{Y}(j)\}\{Y_i(k) - \bar{Y}(k)\}^\top$, $\hat{S}(j,j) = (N_j - 1)^{-1}\sum_{i=1}^{N} W_i(j)\{Y_i^{\text{obs}} - \hat{\bar{Y}}(j)\}\{Y_i^{\text{obs}} - \hat{\bar{Y}}(j)\}^\top$, for $j,k = 1,\ldots,J$. The overall covariance matrix S has (j,k)-th block $S(j,k)$. Assume $S(j,j)$ and $\hat{S}(j,j)$ are both positive definite for all realizations of W.

Let $Y_i(j)_1,\ldots,Y_i(j)_d$ be the components of the potential outcomes $Y_i(j)$ for all i and j. We are interested in testing the weak null hypothesis

$$H_{0\text{N}}(C_1,\ldots,C_d,x_1,\ldots,x_d) : C_1 \begin{pmatrix} \bar{Y}(1)_1 \\ \vdots \\ \bar{Y}(J)_1 \end{pmatrix} = x_1, \ldots, C_d \begin{pmatrix} \bar{Y}(1)_d \\ \vdots \\ \bar{Y}(J)_d \end{pmatrix} = x_d, \tag{13}$$

where $C_1,\ldots,C_d$ are contrast matrices with J columns. Using the Kronecker product, we define

$$C = \begin{pmatrix} C_1 \otimes e_1^\top \\ \vdots \\ C_d \otimes e_d^\top \end{pmatrix}, \qquad x = \begin{pmatrix} x_1 \\ \vdots \\ x_d \end{pmatrix},$$

where $\{e_1,\ldots,e_d\}$ are the standard basis vectors of $\mathbb{R}^d$. We can write the above null hypothesis in a more compact form $H_{0\text{N}}(C,x) : C\bar{Y} = x$. It has the same form as (1), but C cannot be an arbitrary contrast matrix.

Example 4. To illustrate some possible contrast matrices, consider $J = 3$ and $d = 2$. The hypothesis $H_0 : \bar{Y}(1) = \bar{Y}(2) = \bar{Y}(3)$ has the contrast matrix

$$\begin{pmatrix} 1 & 0 & -1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & -1 \end{pmatrix} = \begin{pmatrix} C_1 \otimes e_1^\top \\ C_1 \otimes e_2^\top \end{pmatrix}, \text{ where } C_1 = \begin{pmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \end{pmatrix}$$

In this case, we test the same hypothesis entry by entry, and an equivalent contrast matrix is $C_1 \otimes I_2$. Our framework also allows for testing different hypotheses entry by entry, for instance $H_0 : \bar{Y}(1)_1 = \bar{Y}(2)_1$, $\bar{Y}(2)_2 = \bar{Y}(3)_2$. This hypothesis has the contrast matrix

$$\begin{pmatrix} 1 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & -1 \end{pmatrix} = \begin{pmatrix} C_1 \otimes e_1^\top \\ C_2 \otimes e_2^\top \end{pmatrix}, \text{ where } C_1 = (1, -1, 0) \text{ and } C_2 = (0, 1, -1). \quad \square$$

Our framework does not allow comparison of different entries under different treatments, for instance $H_0 : \bar{Y}(1)_1 = \bar{Y}(2)_2$. This null hypothesis does not have a clear causal interpretation under the potential outcomes framework. Under iid sampling, Friedrich et al. (2017) allow for a general contrast matrix C, and even allow the length of $Y_i(j)$ to depend on treatment j. We impose a restriction on C as Example 4 illustrates.

Under i.i.d. sampling, Chung and Romano (2016b) use permutation tests for the two-sample problem with vector outcomes. Srivastava and Kubokawa (2013), Konietschke et al. (2015) and Friedrich and Pauly (2018) use bootstrap tests for general hypotheses. We will use the FRT for (13). It is not a sharp null hypothesis, but we impute the missing potential outcomes under

$$H_{0\mathrm{F}} : \begin{pmatrix} C_1 \\ \tilde{C}_1 \end{pmatrix} \begin{pmatrix} Y_i(1)_1 \\ \vdots \\ Y_i(J)_1 \end{pmatrix} = \begin{pmatrix} x_1 \\ \tilde{x}_1 \end{pmatrix}, \ldots, \begin{pmatrix} C_d \\ \tilde{C}_d \end{pmatrix} \begin{pmatrix} Y_i(1)_d \\ \vdots \\ Y_i(J)_d \end{pmatrix} = \begin{pmatrix} x_d \\ \tilde{x}_d \end{pmatrix}, \text{ for } i = 1, \ldots, N, \tag{14}$$

where the matrix $(C_1^\top, \tilde{C}_1^\top, 1_J)$ is invertible. We construct the $\tilde{C}$'s and $\tilde{x}$'s for each component of the outcome in the same way as the scalar case. In the hypothesis $H_{0\mathrm{F}}$, we suppress the dependence on the C's, $\tilde{C}$'s, x's and $\tilde{x}$'s. For the first component, we impute the potential outcomes as

$$\begin{pmatrix} Y_i^*(1)_1 \\ \vdots \\ Y_i^*(J)_1 \end{pmatrix} = z_1 + \{(Y_i^{\mathrm{obs}})_1 - z_{1W_i}\} 1_J, \text{ where } z_1 = \begin{pmatrix} z_{11} \\ \vdots \\ z_{1J} \end{pmatrix} = \begin{pmatrix} C_1 \\ \tilde{C}_1 \\ 1_J^\top \end{pmatrix}^{-1} \begin{pmatrix} x_1 \\ \tilde{x}_1 \\ 0 \end{pmatrix} \tag{15}$$

and similarly for the second through the d-th entries, replacing all subscripts 1 by $2, \ldots, d$.

An appropriate modification of X^2 in (5) is

$$X^2 = N(C\hat{\bar{Y}} - x)^\top (C\hat{D}C^\top)^{-1} (C\hat{\bar{Y}} - x),$$

where the block diagonal matrix $\hat{D} = N \cdot \mathrm{diag}\left\{\hat{S}(1,1)/N_1, \ldots, \hat{S}(J,J)/N_J\right\}$ is an asymptotically conservative estimator of $N \cdot \mathrm{Cov}(\hat{\bar{Y}})$. This mirrors (4). Using the FRT with X^2 can control the asymptotic type I error under $H_{0\mathrm{N}}(C_1, \ldots, C_d, x_1, \ldots, x_d)$. We first give the asymptotic conditions and then extend Theorem 1 to the vector case. Let $|\cdot|$ be the Euclidean norm, note it reduces to the usual absolute value for scalars.

Assumption 4. The sequence (N_j/N) converges to $p_j \in (0,1)$ for all $j = 1, \ldots, J$. The sequences

$(\bar{Y}_N)$ and (S_N) converge to $\bar{Y}_\infty$ and S_∞, where $|\bar{Y}_\infty| < \infty$, S_∞ is positive semi-definite, and $S_\infty(j,j)$ is positive definite for all $j = 1, \dots, J$. Further, $\lim_{N\to\infty} \max_{j=1,\dots,J} \max_{i=1,\dots,N} N^{-1}|Y_i(j) - \bar{Y}(j)|^2 = 0$.

Assumption 5. Same as Assumption 4 with the last identity replaced by: there exists an $L < \infty$ such that $N^{-1}\sum_{i=1}^{N} |Y_i(j) - \bar{Y}(j)|^4 \le L$ for all $j = 1, \dots, J$ and $N = (d+1)J, \dots, \infty$.

Proposition 7. Assumption 5 implies Assumption 4.

Theorem 5. If Assumption 4 holds, then under $H_{0\mathrm{N}}(C, x)$, $X^2 \xrightarrow{\mathrm{d}} \sum_{j=1}^{m} a_j \xi_j^2$, where each $a_j \in [0, 1]$. If Assumption 5 holds and $\pi \sim \mathrm{Unif}(\Pi_N)$, then $X_\pi^2 | W \xrightarrow{\mathrm{d}} \chi_m^2$ for almost all sequences of W. In particular, for almost all sequences of W, the FRT with test statistic X^2 can asymptotically control type I error.

Theorem 5 provides a basis for a single FRT for multiple outcomes. Following Chung and Romano (2016b, Section 4), we can use Theorem 5 with the closure procedure for multiple testings. We omit the details.

We require all realizations of $\hat{S}(j,j)$ to be invertible, for which it is necessary that $N_j \ge d+1$. Friedrich and Pauly (2018) instead advocate $\tilde{X}^2 = N(C\hat{\bar{Y}} - x)^\top (C\tilde{D}C^\top)^{-1}(C\hat{\bar{Y}} - x)$ with a bootstrap, where $\tilde{D}$ is a diagonal matrix with the same main diagonal as $\hat{D}$. However, $\tilde{X}^2$ is not suitable for the FRT because the asymptotic distribution of $\tilde{X}_\pi^2 | W$ is not pivotal. So it will fail for the same reason the Box type statistic B in (6) does. We will explore FRTs with large d in future research.

Simulations

Breaking the Box-Type Statistic

Previous sections show that using X^2 in the FRT is suitable, but using B or F is not. To complement the asymptotic theory, we would like to observe their performance in finite samples through simulation.

Here we choose the ANOVA and factorial setups with balanced design $N_j = 20$ for all j. In all cases, we take $\bar{Y}(1) = \cdots = \bar{Y}(J) = 0$ so in particular $H_{0\mathrm{N}}(C, 0_m)$ holds. Lemma 5 implies that B and F are equivalent. For the ANOVA, we take $J = 3$, and generate the potential outcomes with covariance matrix $S = uu^\top$, where $u^\top = (u_1, u_2, u_3) = (1, 2, 3)$. For the factorial case, we take $K = 2$ and $J = 2^2 = 4$, and generate the potential outcomes with covariance matrix $S = uu^\top$, where $u^\top = (u_1, u_2, u_3, u_4) = (3, 1, 1, 3)$. Explicitly, we first generate $Y_i(1) \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$ for $i = 1, \dots, N = 20J$, center them, and scale $Y_i(j) = u_j Y_i(1)$ for $j \ne 1$. In all cases, we generate 2000 different realizations of the observed outcomes, and for each of them run the FRT with both X^2 and B, using 2500 permutations per p-value calculation.

We can use our theorems to show that the permutation distributions of X^2 and $2B$ are asymptotically χ_2^2 in both the ANOVA and factorial designs, but their asymptotic distributions under

H_{0N} are

$$
\begin{aligned}
X^2 &\xrightarrow{\mathrm{d}} \xi_1^2 + 0.758\xi_2^2, \qquad 2B \xrightarrow{\mathrm{d}} 1.423\xi_1^2 + 0.434\xi_2^2, \qquad && \text{(ANOVA)}, \\
X^2 &\xrightarrow{\mathrm{d}} \xi_1^2 + \xi_2^2 \overset{\mathrm{d}}{=} \chi_2^2, \qquad 2B \xrightarrow{\mathrm{d}} 1.8\xi_1^2 + 0.2\xi_2^2, \qquad && \text{(Factorial)}.
\end{aligned} \tag{16}
$$

Each weight for X^2 is at most 1, while the weights for $2B$ are only at most 1 on average. In our factorial design example, the FRT with X^2 is not conservative because X^2 has the same χ_2^2 asymptotic distribution as its permutation distribution.

Figure 1 shows the simulation results. In the ANOVA case, B fails to control type I error, illustrated most visibly at level 0.02. In the factorial design, B fails more dramatically. The focus on balanced designs is to highlight that, when $J > 2$, the balanced design does not guarantee the suitability of B or F as it does in treatment-control experiments (see Corollary 3). Of course, forgoing balanced designs can make both B and F fail more extremely. The comparison of X^2 and F in such cases is investigated through extensive simulation in Ding and Dasgupta (2018).

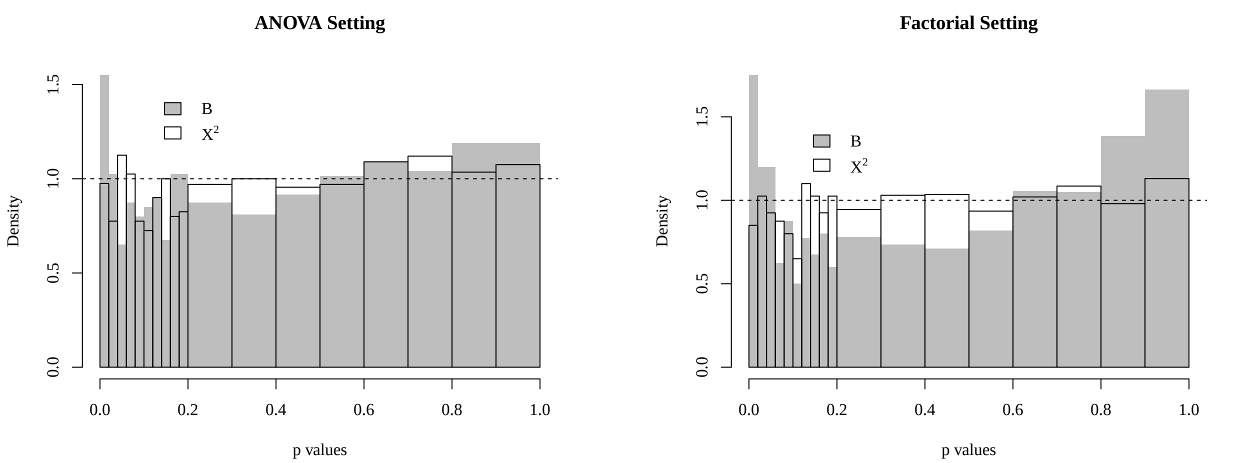

Figure 1: p-values from FRTs using X^2 and $B = F$. We use a finer resolution for smaller p-values because most hypothesis tests are conducted at levels in that interval. The dashed lines are the Uniform$(0, 1)$ densities. In both ANOVA and factorial designs, the FRT with B fails to control type I error, and dramatically so at level 0.02.

A natural extension of the simulation just performed can be made to SREs. Morris (2010), for instance, suggests testing $H_{0N} : \bar{Y}(1) = \cdots = \bar{Y}(J)$ with the F statistic from a linear regression of the observed response on stratum and treatment indicators, i.e. $J + H$ predictors. Although Morris (2010) has emphasized the usual OLS assumptions under which such an F test is appropriate, practitioners do not always check such assumptions. We therefore would like to compare X^2 and F under a potential outcomes framework with an SRE.

We reconsider the ANOVA setup in an SRE with $H = 2$ strata. Let us make the first stratum of potential outcomes identical to those of the ANOVA simulation above, and the second stratum identical to the first. In particular, we have a balanced design with each $N_{[h]j} = 20$, and $J =$

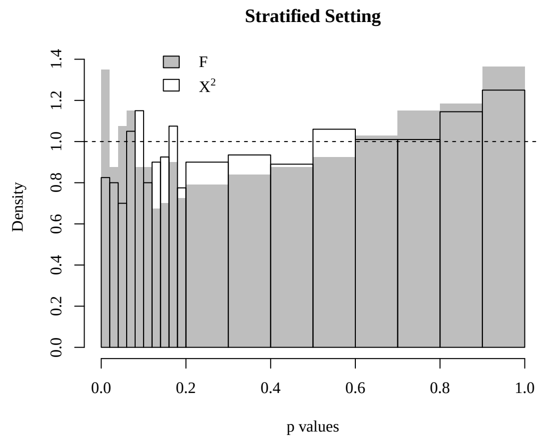

Figure 2: p-values from FRTs using X^2 and $B = F$, when data are simulated according to a SRE. The failure of $B = F$ to control type I error at level 0.02 is again visible, though not extreme.

3. Figure 2 gives the results of the simulation of an SRE. It demonstrates just as the previous simulation that the F statistic is unsuitable in general, though the failure is far from catastrophic.

Confidence Regions

We investigate the confidence regions based on Corollary 5. We consider a balanced factorial design with $K = 2$, $J = 2^2 = 4$, and $N_j = 10$ $(j = 1, 2, 3, 4)$. We are interested in inferring the main effects (τ_1, τ_2), both individually and jointly. We again design the potential outcomes so that $\bar{Y}(1) = \cdots = \bar{Y}(4) = 0$. Take $Y_i(j) \overset{\text{i.i.d.}}{\sim} U^2 - 1/3$ where $U \sim \text{Unif}(0, 1)$, then center so that each treatment j has mean 0, and finally multiply each Y_i by the same 4×4 matrix

$$\begin{pmatrix} 2 & 1 & 3/2 & 1 \\ 0 & \sqrt{5} & \sqrt{5}/2 & 2/\sqrt{5} \\ 0 & 0 & 3/\sqrt{2} & 1/\sqrt{2} \\ 0 & 0 & 0 & \sqrt{3.7} \end{pmatrix}$$

to make the potential outcomes correlated.

Using (11), we first compute an asymptotic confidence region for (τ_1, τ_2). Then we find a grid of points centered at $(\hat{\tau}_1, \hat{\tau}_2)$ that encompasses the asymptotic region. At each point (x_1, x_2) of this grid, we use the FRT with X^2 to test $(\tau_1 = x_1, \tau_2 = x_2)$, both individually and jointly. We include the point in the confidence region if and only if the p-value exceeds $\alpha = 0.05$.

Figure 3 shows the results for the marginal hypothesis tests. The behavior is very regular: the p-value is highest near $\hat{\tau}_1$ or $\hat{\tau}_2$, and decreases monotonically in each direction. The FRT and its asymptotic approximation give nearly identical confidence intervals.

Figure 4 shows the result for the joint test. The left graph shows the FRT confidence region is again close to its asymptotic approximation, but not as close as in the one-dimensional case.

In particular, the former is noticeably larger. The right graph explain this and shows that the p-values from the FRT tend to be larger than those from the χ^2 approximation.

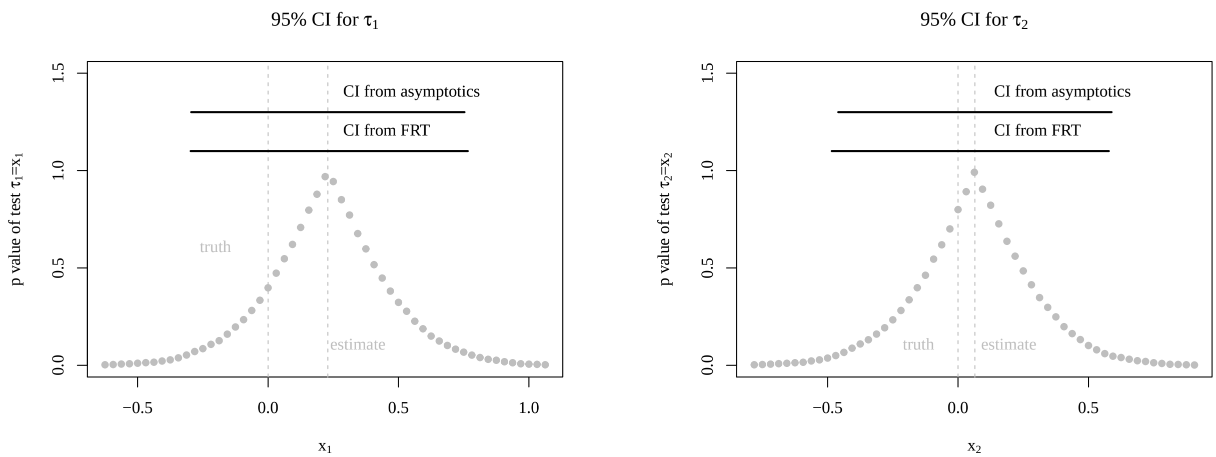

Figure 3: For τ_1 and τ_2 individually, the FRT and asymptotic approximation give nearly identical confidence intervals. For the second main effect, the FRT confidence interval is shifted due to the discrete resolution.

Applications

Financial Incentives for Exercise

Charness and Gneezy (2009) were interested in whether financial incentives caused college students to exercise more. They randomly assigned 40 students to one of three possible treatments: no financial incentive (control), a small one, or a large one. We henceforth index these groups by $j = 1, 2, 3$, respectively. For each student, the response was the average number of weekly gym visits after the study minus that before the study. Let $Y_i(j)$ denote this quantity for the i-th student, if s/he received treatment j. For many students, $Y_i^{\text{obs}} = 0$, which would be problematic for the FRT with X^2 if, after a certain permutation, all permuted observations in a group were 0. To preclude this, we added a very small amount of random noise to the observations that were 0. For this dataset, the sample means are $-0.02937, 0.05414, 0.6398$, and the sample variances are $0.1523, 0.3859, 1.489$, for groups $j = 1, 2, 3$, respectively. Just from inspection of these numbers, it seems a large financial incentive has a positive effect while a small one does not. It is also clear that the data are heteroskedastic.

We test the following four hypotheses at level 1% : $2\bar{Y}(1) = \bar{Y}(2) + \bar{Y}(3)$, $\bar{Y}(1) = \bar{Y}(2, 3)$ (here we collapse treatment levels $j = 2, 3$ to one), $\bar{Y}(1) = \bar{Y}(2) = \bar{Y}(3)$, and $\bar{Y}(1) = \bar{Y}(2)$ (here we ignore the $j = 3$ observations). We use the X^2 and F statistics, and get p-values both by comparing against the quantiles from appropriate asymptotic distributions and from FRTs. As we discussed earlier, p-values from FRTs are also exact for testing Fisher's sharp null. Table 1

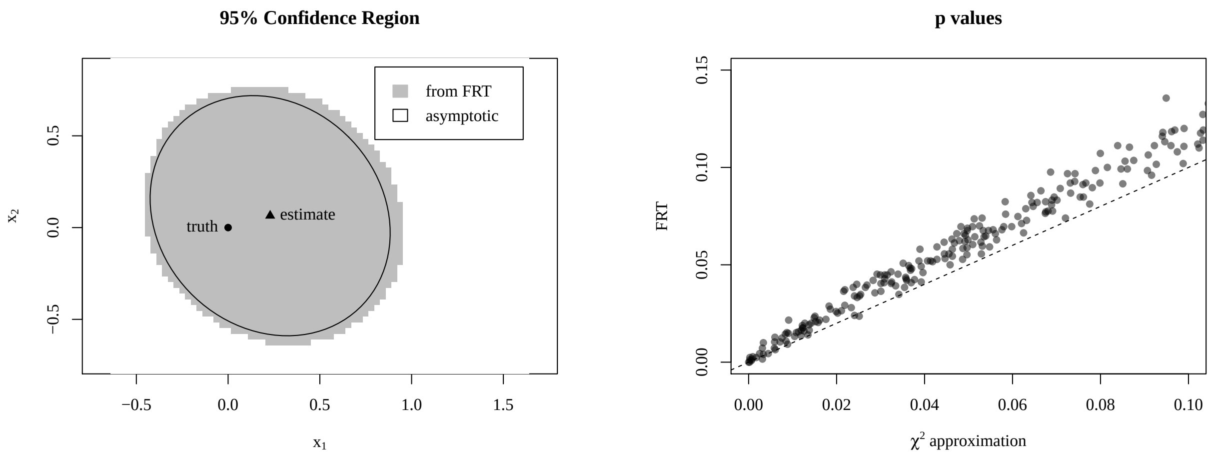

Figure 4: The left graph shows the FRT confidence region is again close to its asymptotic approximation, but the former is noticeably larger. The right graph is a scatter plot of p-values from testing $\tau_1 = \tau_2 = 0$ repeatedly from the original set of potential outcomes, focusing on the region where they are less than 0.1.

Table 1: Analyzing Charness and Gneezy (2009)'s data. We report p-values as percents, and calculate the FRT p-values using 10^4 Monte Carlo simulations.

Hypothesis	$X^2 \xrightarrow{d} \chi^2_m$	FRT using X^2	$F \xrightarrow{d} F_{m,N-J}$	FRT using F
$2\bar{Y}(1) = \bar{Y}(2) + \bar{Y}(3)$	0.2522	0.27	1.970	1.59
$\bar{Y}(1) = \bar{Y}(2) = \bar{Y}(3)$	0.4189	0.49	0.06198	0.01
$\bar{Y}(1) = \bar{Y}(2,3)$	0.3353	0.49	2.454	2.34
$\bar{Y}(1) = \bar{Y}(2)$	47.15	47.93	47.37	47.93

gives the results.

Testing the first two hypotheses, it seems that financial incentives have a statistically significant impact on gym attendance. As suggested by Theorems 1 and 3, we should trust the p-values from X^2 more than those from F. The latter in this case seems to have overly conservative behavior. Testing the third hypothesis suggests that one group has different behavior from another in a statistically significant way.

With evidence that financial incentives might be helpful, we finally test the fourth hypothesis only comparing the control and small incentive groups, and get insignificant p-values. Note in this case $X^2 = F$ because it is a balanced treatment-control design. We have thus confirmed the findings in Charness and Gneezy (2009), that large financial incentives seem to induce people to visit the gym more often, but not small ones.

Academic Support Services and Financial Incentives for Grades

Angrist et al. (2009) were interested in whether academic support services and/or financial incentives caused college students to improve their grades. Their data consisted of student grades for

Table 2: Analyzing Angrist et al. (2009)'s data. We report p-values as percents, and calculate the FRT p-values using 10^4 Monte Carlo simulations.

Hypothesis	$X^2 \xrightarrow{d} \chi^2_m$	FRT using X^2	$F \xrightarrow{d} F_{m,N-J}$	FRT using F
No effect from services	72.84	72.34	73.92	73.58
No effect from incentives	1.192	1.43	1.602	1.80
No effects from either	3.652	3.99	5.262	5.28
No interaction	99.53	99.47	99.55	99.5
$\bar{Y}(1) = \bar{Y}(2) = \bar{Y}(3) = \bar{Y}(4)$	3.880	4.31	5.849	5.71

a certain semester on a 100 point scale. In that semester, students were either in a control group, offered a fellowship to improve their grade, offered services, or both. We thus have a 2^2 factorial experiment, and henceforth index these treatment groups by $j = 1, 2, 3, 4$, respectively. For this dataset, the sample means are $63.9, 65.8, 64.1, 66.1$, and the sample variances are $145, 124, 160, 114$, for groups $j = 1, 2, 3, 4$, respectively.

We test the following five hypotheses at level 1%: financial services have no effect, services have no effect, neither has an effect, no interactions, and that all group means are the same. In symbols, these are $\bar{Y}(1) + \bar{Y}(2) = \bar{Y}(3) + \bar{Y}(4)$, $\bar{Y}(1) + \bar{Y}(3) = \bar{Y}(2) + \bar{Y}(4)$, both of the previous two, $\bar{Y}(1) + \bar{Y}(4) = \bar{Y}(2) + \bar{Y}(3)$, and $\bar{Y}(1) = \bar{Y}(2) = \bar{Y}(3) = \bar{Y}(4)$. We again use the X^2 and F statistics, and get p-values both by comparing against the quantiles from appropriate asymptotic distributions and from FRTs. As we discussed earlier, p-values from FRTs are also exact for testing Fisher's sharp null. Table 2 shows the results.

We cannot reject any of these null hypotheses at level 1%. From the second and fourth hypotheses, the data do not seem to suggest services have any effect, or that there is a non-additive effect from combining incentives and services. On the other hand, we nearly reject the hypothesis of no effect from incentives alone, with p-values just over 1%.

Our finding that the effect of incentives is more significant than the others agrees with the conclusions of Angrist et al. (2009). Angrist et al. (2009) additionally conducted subgroup analysis, and noticed that the observed effects on grades come nearly entirely from female students.

Discussion

We have proposed a strategy for using the FRT to test a weak null hypothesis. It imputes the missing potential outcomes under a compatible sharp null hypothesis, and then uses the studentized statistic in the FRT. It furthers the current literature in two directions. First, it complements the tests based on asymptotic distributions. Our FRT is also finite-sample exact under the sharp null hypothesis. Second, it gives guidance for choosing test statistics for the sharp null hypothesis. Although the finite-sample exactness property of the FRT holds for any test statistic, the p-values are sensitive to the choices of test statistics. For example, all the p-values in Tables 1 and 2 are valid for Fisher's sharp null hypothesis. Unfortunately, these p-values range below and above the significance level. This can be confusing in practice. Therefore, it is important to consider weak null hypotheses and then use studentized statistics. Our FRTs can control asymptotic type I error under weak null hypotheses and have power under corresponding alternative hypotheses.

Our theory ignores covariates. The analysis of covariance is a classical topic (Fisher 1935) and still attracts attention (Lin 2013; Lu 2016a; Fogarty 2018b,a; Middleton 2018). Tukey (1993) and Rosenbaum (2002b) discussed strategies for testing sharp null hypotheses. It is important to extend the theory to test weak null hypotheses with covariate adjustment. We have focused on completely randomized factorial experiments and extended the theory to stratified and clustered experiments. We conjecture that the strategy is also applicable for experiments with general treatment assignment mechanisms (Mukerjee et al. 2018). Fogarty (2016) also used the idea of studentization in sensitivity analysis of matched observational studies.

Rank statistics are attractive for the FRT because their distributions do not depend on the outcome values under Fisher's sharp null hypothesis (Lehmann 1975; Rosenbaum 2002a). They are useful for testing some weak null hypotheses (Brunner and Puri 2001; Chung and Romano 2016a; Brunner et al. 2017; Umlauft et al. 2017). However, it is not entirely clear what causal effects they are testing under the potential outcomes framework. We leave this to future work.

Acknowledgments

We thank Guilaume Basse, Joel Middleton, and Zach Branson for helpful comments.

References

J. Angrist, D. Lang, and P. Oropoulos. Incentives and services for college achievement: Evidence from a randomized trial. *American Economic Journal: Applied Economics*, 1:136–63, 2009.

S. Athey and G. W. Imbens. *The Econometrics of Randomized Experiments*, volume 1 of *Handbook of Economic Field Experiments*, chapter 3, pages 73–140. Elsevier B.V, 2017.

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

G. J. Babu and K. Singh. Inference on means using the bootstrap. *The Annals of Statistics*, 11: 999–1003, 1983.

G. Basse, A. Feller, and P. Toulis. Exact tests for two-stage randomized designs in the presence of interference. *Biometrika*, page in press, 2018.

D. Basu. Randomization analysis of experimental data: The Fisher randomization test. *Journal of the American Statistical Association*, 75:575–582, 1980.

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.

G. Box. Some theorems on quadratic forms applied in the study of analysis of variance problems. *Annals of Mathematical Statistics*, 25:290–302, 1954.

G. E. P. Box and S. L. Andersen. Permutation theory in the derivation of robust criteria and the study of departures from assumption. *Journal of the Royal Statistical Society. Series B (Methodological)*, 17:1–34, 1955.

J. V. Bradley. *Distribution-free statistical tests*. Upper Saddle River, NJ: Prentice Hall, 1968.

E. Brunner and M. L. Puri. Nonparametric methods in factorial designs. *Statistical Papers*, 42: 1–52, 2001.

E. Brunner, H. Dette, and A. Munk. Box-type approximations in nonparametric factorial designs. *Journal of the American Statistical Association*, 92:1494–502, 1997.

E. Brunner, F. Konietschke, M. Pauly, and M. L. Puri. Rank-based procedures in factorial designs: hypotheses about non-parametric treatment effects. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79:1463–1485, 2017.

D. Caughey, A. Dafoe, and L. Miratrix. Beyond the sharp null: Randomization inference, bounded null hypotheses, and confidence intervals for maximum effects. *arXiv preprint arXiv:1709.07339*, 2017.

G. Charness and U. Gneezy. Incentives to exercise. *Econometrica*, 77:909–31, 2009.

E. Chung and J. 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, 2016a.

E. Chung and J. P. Romano. Multivariate and multiple permutation tests. *Journal of Econometrics*, 193:76–91, 2016b.

R. O. Collier and F. B. Baker. Some monte carlo results on the power of the f-test under permutation in the simple randomized block design. *Biometrika*, 53:199–203, 1966.

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: 727–53, 2015.

P. Ding. A paradox from randomization-based causal inference (with discussion). *Statistical Science*, 32:331–45, 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.

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

E. Duflo, R. Glennerster, and M. Kremer. *Using randomization in development economics research: A toolkit*, volume 4, chapter 61, pages 3895–3962. Elsevier, 2007.

T. Eden and F. Yates. On the validity of fisher's z test when applied to an actual example of non-normal data. *The Journal of Agricultural Science*, 23:6–17, 1933.

R. A. Fisher. *Statistical Methods for Research Workers*. Edinburgh: Oliver and Boyd, 1925.

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

C. B. Fogarty. Studentized sensitivity analysis for the sample average treatment effect in paired observational studies. *arXiv preprint arXiv:1609.02112*, 2016.

C. B. Fogarty. On mitigating the analytical limitations of finely stratified experiments. *Journal of the Royal Statistical Society. Series B (Methodological)*, page in press, 2018a.

C. B. Fogarty. Regression assisted inference for the average treatment effect in paired experiments. *Biometrika*, page in press, 2018b.

S. Friedrich and M. Pauly. MATS: Inference for potentially singular and heteroscedastic MANOVA. *Journal of Multivariate Analysis*, 165:166–179, 2018.

S. Friedrich, E. Brunner, and M. Pauly. Permuting longitudinal data in spite of the dependencies. *Journal of Multivariate Analysis*, 153:255–265, 2017.

M. H. Gail, S. D. Mark, R. J. Carroll, S. B. Green, and D. Pee. On design considerations and randomization-based inference for community intervention trials. *Statistics in Medicine*, 15: 1069–1092, 1996.

A. S. Gerber and D. P. Green. *Field Experiments: Design, Analysis, and Interpretation*. New York: W. W. Norton and Company, 2012.

P. Hall. Theoretical comparison of bootstrap confidence intervals. *The Annals of Statistics*, 16: 927–953, 1988.

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.

J. L. Hodges and E. L. Lehmann. Estimates of location based on rank tests. *The Annals of Mathematical Statistics*, 34:598–611, 1963.

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

D. Holt and T. M. F. Smith. Post stratification. *Journal of the Royal Statistical Society: Series A (General)*, 142:33–46, 1979.

P. J. Huber. The behavior of maximum likelihood estimates under nonstandard conditions. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1:221–33, 1967.

G. W. Imbens and K. Menzel. A causal bootstrap. Technical report, National Bureau of Economic Research, 2018.

G. W. Imbens and D. B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge: Cambridge University Press, 2015.

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

A. Janssen. Testing nonparametric statistical functionals with applications to rank tests. *Journal of Statistical Planning and Inference*, 81:71–93, 1999.

A. Janssen and T. Pauls. How do bootstrap and permutation tests work? *Annals of Statistics*, 31: 768–806, 2003.

O. Kempthorne. *The Design and Analysis of Experiments*. New York: John Wiley and Sons, 1952.

O. Kempthorne and T. E. Doerfler. The behaviour of some significance tests under experimental randomization. *Biometrika*, 56:231–248, 1969.

F. Konietschke, A. C. Bathke, S. W. Harrar, and M. Pauly. Parametric and nonparametric bootstrap methods for general MANOVA. *Journal of Multivariate Analysis*, 140:291–301, 2015.

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

X. Li and P. Ding. Exact confidence intervals for the average causal effect on a binary outcome. *Statistics in Medicine*, 35:957–960, 2016.

X. Li and P. Ding. General forms of finite population central limit theorems with applications to causal inference. *Journal of the American Statistical Association*, 112:1759–69, 2017.

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

W. Lin, S. D. Halpern, M. Prasad Kerlin, and D. S. Small. A "placement of death" approach for studies of treatment effects on ICU length of stay. *Statistical Methods in Medical Research*, 26: 292–311, 2017.

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

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

J. Lu. On randomization-based and regression-based inferences for 2^K factorial designs. *Statistics and Probability Letters*, 112:72–78, 2016b.

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.

L. Miratrix, J. Sekhon, and B. Yu. Adjusting treatment effect estimates by post-stratification in randomized experiments. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 75:369–96, 2013.

M. Morris. *Design of Experiments: An Introduction Based on Linear Models*. London: Chapman and Hall/CRC, 2010.

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*, page in press, 2018.

T. Mutze, F. Konietschke, A. Munk, and T. Friede. A studentized permutation test for three-arm trials in the gold standard design. *Statistics in Medicine*, 36:883–98, 2017.

G. Neuhaus. Conditional rank tests for the two-sample problem under random censorship. *The Annals of Statistics*, 21:1760–1779, 1993.

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

J. Neyman. On the application of probability theory to agricultural experiments. *Statistical Science*, 5:465–472, 1990.

T. L. Nolen and M. G. Hudgens. Randomization-based inference within principal strata. *Journal of the American Statistical Association*, 106:581–593, 2011.

M. Pauly, E. Brunner, and F. Konietschke. Asymptotic permutation tests in general factorial designs. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 77:461–73, 2015.

E. J. G. Pitman. Significance tests which may be applied to samples from any populations. *Supplement to the Journal of the Royal Statistical Society*, 4:119–130, 1937.

J. Rigdon and M. G. Hudgens. Randomization inference for treatment effects on a binary outcome. *Statistics in Medicine*, 34:924–935, 2015.

J. P. Romano. On the behavior of randomization tests without a group invariance assumption. *Journal of the American Statistical Association*, 85:686–92, 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. Effects attributable to treatment: Inference in experiments and observational studies with a discrete pivot. *Biometrika*, 88:219–231, 2001.

P. R. Rosenbaum. *Observational Studies*. New York: Springer, 2 edition, 2002a.

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

P. R. Rosenbaum. Exact Confidence Intervals for Nonconstant Effects by Inverting the Signed Rank Test. *The American Statistician*, 57:132–138, May 2003.

D. B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66:688–701, 1974.

D. B. Rubin. Comment on D. Basu. *Journal of the American Statistical Association*, 75:591–593, 1980.

D. B. Rubin. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100:322–31, 2005.

A. Sabbaghi and D. B. Rubin. Comments on the Neyman–Fisher controversy and its consequences. *Statistical Science*, 29:267–284, 2014.

P. Z. Schochet. Multi-armed RCTs: A design-based framework. *Journal of Educational and Behavioral Statistics*, page in press, 2018.

M. S. Srivastava and T. Kubokawa. Tests for multivariate analysis of variance in high dimension under non-normality. *Journal of Multivariate Analysis*, 115:204–216, 2013.

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

M. Umlauft, F. Konietschke, and M. Pauly. Rank-based permutation approaches for non-parametric factorial designs. *British Journal of Mathematical and Statistical Psychology*, 70:368–390, 2017.

H. White. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48:817–38, 1980.

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

Supplementary Material for "Randomization Tests for Weak Null Hypotheses"

by Jason Wu and Peng Ding

Let $X_N \xrightarrow{\text{d}} X$, $X_N \xrightarrow{\text{a.s.}} X$ and $X_N \xrightarrow{\mathbb{P}} X$ denote convergence in distribution, almost surely, and in probability, respectively. For random vectors or matrices, we use the same notation to denote convergence, entry by entry. For convergence in probability, we may also write $\text{plim}_{N\to\infty} X_N = X$. Let $\xi_1, \xi_2, \dots \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$. Let $\lambda_j(A)$ be the j-th largest eigenvalue of matrix A. Let $|\cdot|$ be the absolute value of a real number or the Euclidean norm of a vector. Let $\|\cdot\|_F$ be the Frobenius norm of a matrix. For $A, B \in \mathbb{R}^{m\times n}$, let $A * B$ be the component-wise product of A and B: $(AB)_{ij} = A_{ij}B_{ij}$. Let $\max_i$, $\max_j$, and $\max_{i,j}$ denote the maximums over $\{i = 1, \dots, n\}$, $\{j = 1, \dots, J\}$, and both. Let $a \vee b = \max(a, b)$ be the maximum value of a and b.

Appendix A1 gives several useful lemmas and their proofs. Appendix A2 gives the proofs of the main theorems. Appendix A3 gives the proofs of other corollaries and propositions.

Lemmas

Lemma A1. (i) If $X \sim \mathcal{N}(0_J, A)$, then $X^\top B X \overset{\text{d}}{=} \sum_{j=1}^J \lambda_j(AB)\xi_j^2$. If A is a projection matrix, then $\lambda_j(AB) \le \lambda_1(B)$ for all j.

(ii) If $A \succeq 0, B \succeq 0$ and B is a correlation matrix, then $\lambda_1(A * B) \le \lambda_1(A)$.

(iii) If $X_n \xrightarrow{\text{d}} \mathcal{N}(0_m, A)$ and $B_n \xrightarrow{\mathbb{P}} B \succ 0$, then $X_n^\top B_n^{-1} X_n \xrightarrow{\text{d}} \sum_{j=1}^m \lambda_j(AB^{-1})\xi_j^2$. If $B \succeq A$, then $\lambda_j(AB^{-1}) \in [0, 1]$ for all j.

Proof of Lemma A1. (i) and (ii) come from Ding and Dasgupta (2018). We prove (iii). The Continuous Mapping Theorem implies $B_n^{-1} \xrightarrow{\mathbb{P}} B^{-1}$, and Slutsky's Theorem then implies $X_n^\top B_n^{-1} X_n \xrightarrow{\text{d}} X^\top B^{-1} X$. By (i), $X^\top B^{-1} X$ has the same distribution as $\sum_{j=1}^m \lambda_j(AB^{-1})\xi_j^2$. If $B \succeq A$, then each $\lambda_j(AB^{-1}) \in [0, 1]$. $\square$

Lemma A2. A finite population $(Y_1, \dots, Y_N)$ has mean $\bar{Y}_N$ and variance $S_N = (N-1)^{-1}\sum_{i=1}^N (Y_i - \bar{Y}_N)^2$. Let $\mathcal{A} \subseteq \{1, \dots, N\}$ be a simple random sample of size N_1, and $\hat{\bar{Y}}_N = N_1^{-1}\sum_{i\in\mathcal{A}} Y_i$. Then for $t \ge 0$,

$$\mathbb{P}(\hat{\bar{Y}}_N - \bar{Y}_N \ge t) \vee \mathbb{P}(\hat{\bar{Y}}_N - \bar{Y}_N \le -t) \le \exp\left\{-\frac{Np_{N,1}^2 t^2}{C_N S}\right\} \le \exp\left\{-\frac{Np_{N,1}^2 t^2}{CS}\right\},$$

where

$$p_{N,1} = N_1/N, \quad C_N = \left[1 + \min\left\{1, 9p_{N,1}^2, 9(1-p_{N,1})^2\right\}/70\right]^2, \quad C = (71/70)^2.$$

Proof of Lemma A2. Bloniarz et al. (2016) prove the first inequality. The second inequality follows from $C_N \leq C$. $\square$

Lemma A2 is crucial for our proof of almost sure convergence for sampling without replacement. We now state the almost sure convergence result.

Lemma A3. Let $\big(\{Y_{N,i} : i = 1, \ldots, N\}\big)$ be a sequence of populations with means $(\bar{Y}_N)$ and variances (S_N). Suppose we take a simple random sample from each population of size $N_1 \geq 2$ with sample mean $\hat{\bar{Y}}_N$ and variance $\hat{S}_N$. Assume $\lim_{N\to\infty} N_1/N = p_1 > 0$.

(i) If the sequence $(S_N)_{N\geq 2}$ is bounded above by $S_{\max} < \infty$, then $|\hat{\bar{Y}}_N - \bar{Y}_N| \overset{\text{a.s.}}{\to} 0$. If we also have $\lim_{N\to\infty} \bar{Y}_N = \bar{Y}_\infty$, then $\hat{\bar{Y}}_N \overset{\text{a.s.}}{\to} \bar{Y}_\infty$. Assumption 1 implies these results.

(ii) If there is $L < \infty$ such that $N^{-1}\sum_{i=1}^N (Y_{N,i} - \bar{Y}_N)^4 \leq L$ for all N, then $|\hat{S}_N - S_N| \overset{\text{a.s.}}{\to} 0$. If we also have $\lim_{N\to\infty} S_N = S_\infty$, then $\hat{S}_N \overset{\text{a.s.}}{\to} S_\infty$. Assumption 2 implies these results.

Proof of Lemma A3. (i) Let $p_{N,1} = N_1/N$. Because $p_{N,1} \to p_1$, we can pick a positive integer N^* such that $N \geq N^*$ implies $p_{N,1} > p_1/2$. Then by Lemma A2, there is a universal constant $C \in (0, \infty)$ independent of N such that for $N \geq N^*$ and $t \geq 0$,

$$\mathbb{P}(|\hat{\bar{Y}}_N - \bar{Y}_N| \geq t) \leq 2\exp\left\{-\frac{Np_{N,1}^2}{CS_N}t^2\right\} \leq 2\exp\left\{-\frac{p_1^2}{4CS_{\max}}Nt^2\right\}$$

$$\Longrightarrow \sum_{N\geq N^*} \mathbb{P}(|\hat{\bar{Y}}_N - \bar{Y}_N| \geq t) \leq 2\sum_{N\geq N^*} \exp\left\{-\frac{p_1^2}{4CS_{\max}}Nt^2\right\} < \infty.$$

So by the Borel–Cantelli Lemma, $|\hat{\bar{Y}}_N - \bar{Y}_N| \overset{\text{a.s.}}{\to} 0$.

(ii) First, by the Cauchy–Schwarz Inequality, we have that for all N

$$S_N = \frac{1}{N-1}\sum_{i=1}^N (Y_{N,i} - \bar{Y}_N)^2 \leq \frac{N}{N-1}\left\{\frac{1}{N}\sum_{i=1}^N (Y_{N,i} - \bar{Y}_N)^4\right\}^{1/2} \leq \frac{N}{N-1}\sqrt{L},$$

which is bounded above as $N \to \infty$. So by (i), $|\hat{\bar{Y}}_N - \bar{Y}_N| \overset{\text{a.s.}}{\to} 0$.

Second, let $W_{N,i}$ be the indicator for Y_i being in the simple random sample. Define as an intermediate quantity $\tilde{S}_N = (N_1 - 1)^{-1}\sum_{i=1}^N W_{N,i}(Y_{N,i} - \bar{Y}_N)^2$, which differs from $\hat{S}_N$ by an almost surely zero quantity as $N \to \infty$:

$$\begin{aligned}
\hat{S}_N - \tilde{S}_N =& \frac{1}{N_1 - 1}\sum_{i=1}^N W_{N,i}\{(Y_{N,i} - \hat{\bar{Y}}_N)^2 - (Y_{N,i} - \bar{Y}_N)^2\} \\
=& \frac{1}{N_1 - 1}\{2(\bar{Y}_N - \hat{\bar{Y}}_N)\sum_{i=1}^N W_{N,i}Y_{N,i} + N_1((\hat{\bar{Y}}_N)^2 - \bar{Y}_N^2)\} \\
=& \frac{N_1}{N_1 - 1}\{2(\bar{Y}_N - \hat{\bar{Y}}_N)\hat{\bar{Y}}_N + (\hat{\bar{Y}}_N)^2 - \bar{Y}_N^2\} \\
=& \frac{-N_1}{N_1 - 1}(\hat{\bar{Y}}_N - \bar{Y}_N)^2 \overset{\text{a.s.}}{\to} 0.
\end{aligned}$$

Third, we note that the following variance is bounded from above for all N:

$$\mathrm{Var}\{(Y_{N,i}-\bar{Y}_N)^2 : i=1,\dots,N\} \le \frac{1}{N-1}\sum_{i=1}^{N}(Y_{N,i}-\bar{Y}_N)^4 \le \frac{N}{N-1}L.$$

So by (i), $\left|N_1^{-1}\sum_{i=1}^{N} W_{N,i}(Y_{N,i}-\bar{Y}_N)^2 - N^{-1}\sum_{i=1}^{N}(Y_{N,i}-\bar{Y}_N)^2\right| \overset{\text{a.s.}}{\to} 0$, and therefore

$$\begin{aligned}
|\tilde{S}_N - S_N| &= \left|\frac{N_1}{N_1-1}\frac{1}{N_1}\sum_{i=1}^{N} W_{N,i}(Y_{N,i}-\bar{Y}_N)^2 - \frac{N_1}{N_1-1}\frac{1}{N}\sum_{i=1}^{N}(Y_{N,i}-\bar{Y}_N)^2 + \frac{N-N_1}{(N-1)(N_1-1)}\frac{N-1}{N}S_N\right| \\
&\le \frac{N_1}{N_1-1}\left|\frac{1}{N_1}\sum_{i=1}^{N} W_{N,i}(Y_{N,i}-\bar{Y}_N)^2 - \frac{1}{N}\sum_{i=1}^{N}(Y_{N,i}-\bar{Y}_N)^2\right| + \frac{N-N_1}{(N-1)(N_1-1)}\frac{N-1}{N}S_N \\
&\le \frac{N_1}{N_1-1}\left|\frac{1}{N_1}\sum_{i=1}^{N} W_{N,i}(Y_{N,i}-\bar{Y}_N)^2 - \frac{1}{N}\sum_{i=1}^{N}(Y_{N,i}-\bar{Y}_N)^2\right| + \frac{1}{N_1-1}L^{1/2} \overset{\text{a.s.}}{\to} 0.
\end{aligned}$$

We now finally have $|\hat{S}_N - S_N| \le |\hat{S}_N - \tilde{S}_N| + |\tilde{S}_N - S_N| \overset{\text{a.s.}}{\to} 0$. □

An important quantity in the subsequent results is $\bar{Y}_{\bullet}^{\text{obs}} := N^{-1}\sum_{i=1}^{N} Y_i^{\text{obs}}$.

Lemma A4. Under Assumption 1 and for all sequences of W, the imputed potential outcomes in FRT-2 satisfy $\lim_{N\to\infty} N^{-1}\max_{i,j}\{Y_i^*(j) - \bar{Y}^*(j)\}^2 = 0$.

Proof of Lemma A4. Because $(\bar{Y}(j))$ converges for all $j=1,\dots,J$, and the z_j's do not depend on N, we may pick $Y_{\max} \in \mathbb{R}$ such that for all N,

$$\max_j |\bar{Y}(j)| \vee \max_j |z_j - \bar{z}| \le Y_{\max}. \tag{A1}$$

Put $L_N = \max_{i,j}\{Y_i(j) - \bar{Y}(j)\}^2$, which is $o(N)$ by Assumption 1. Then

$$\max_{i,j}|Y_i(j) - \bar{Y}(j)| = \left[\max_{i,j}\{Y_i(j) - \bar{Y}(j)\}^2\right]^{1/2} \le L_N^{1/2}. \tag{A2}$$

Next,

$$\max_i |Y_i^{\text{obs}}| \le \max_{i,j}|Y_i(j)| \le \max_{i,j}|Y_i(j) - \bar{Y}(j)| + \max_j |\bar{Y}(j)| \le L_N^{1/2} + Y_{\max}.$$

Hence

$$|\bar{Y}_{\bullet}^{\text{obs}}| \le \max_i |Y_i^{\text{obs}}| \le L_N^{1/2} + Y_{\max}, \quad \max_i |Y_i^{\text{obs}} - \bar{Y}_{\bullet}^{\text{obs}}| \le \max_i |Y_i^{\text{obs}}| + |\bar{Y}_{\bullet}^{\text{obs}}| \le 2(L_N^{1/2} + Y_{\max}).$$

Using the above bounds and the additional bound $(a+b)^2 \le 2(a^2+b^2)$, we have

$$\max_i (Y_i^{\text{obs}} - \bar{Y}_{\bullet}^{\text{obs}})^2 = \left(\max_i |Y_i^{\text{obs}} - \bar{Y}_{\bullet}^{\text{obs}}|\right)^2 \le 4(L_N^{1/2} + Y_{\max})^2 \le 8(L_N + Y_{\max}^2).$$

Incorporating the z's, we have

$$\begin{aligned}\max_{i,j}\{Y_i^*(j)-\bar{Y}^*(j)\}^2 =& \max_i(Y_i^{\text{obs}}-z_{W_i}-\bar{Y}_\bullet^{\text{obs}}+\bar{z})^2\\ \le& 2\{\max_i(Y_i^{\text{obs}}-\bar{Y}_\bullet^{\text{obs}})^2+\max_i(z_{W_i}-\bar{z})^2\}\\ \le& 16(L_N+Y_{\max}^2)+2Y_{\max}^2,\end{aligned}$$

which is $o(N)$ as desired. □

Now we state the vector versions of Lemmas A3 and A4.

Lemma A5. Let $(\{Y_{N,i}: i=1,\dots,N\})$ be a sequence of populations with means and covariances $\bar{Y}_N \in \mathbb{R}^d$ and S_N. Suppose we take a simple random sample from each population of size $N_1 \ge d+1$ such that $\lim_{N\to\infty} N_1/N = p_1 > 0$, with sample mean and covariance $\hat{\bar{Y}}_N$ and $\hat{S}_N$.

(i) If the sequence $(\|S_N\|_F)_{N\ge d+1}$ is bounded above by $S_{\max} < \infty$, then $|\hat{\bar{Y}}_N - \bar{Y}_N| \overset{\text{a.s.}}{\to} 0$. If we also have $\lim_{N\to\infty} \bar{Y}_N = \bar{Y}_\infty$, then $\hat{\bar{Y}}_N \overset{\text{a.s.}}{\to} \bar{Y}_\infty$. Assumption 4 implies these results.

(ii) If there is $L<\infty$ such that $N^{-1}\sum_{i=1}^N |Y_{N,i}-\bar{Y}_N|^4 \le L$ for all N, then $\|\hat{S}_N - S_N\|_F \overset{\text{a.s.}}{\to} 0$. If we also have $\lim_{N\to\infty} S_N = S_\infty$, then $\hat{S}_N \overset{\text{a.s.}}{\to} S_\infty$. Assumption 5 implies these results.

Proof of Lemma A5. (i) Note that each component of Y_{Ni} meets Lemma A3, so $|\hat{\bar{Y}}_N - \bar{Y}_N| \overset{\text{a.s.}}{\to} 0$ holds component by component.

(ii) Because each component of Y_{Ni} meets Lemma A3, each entry on the main diagonal of $\hat{S}_N - S_N$ converges almost surely to 0. It is thus enough to show convergence of the $(1,2)$th entry, for then identical logic will show convergence of an arbitrary off-diagonal entry. Let Y_{1Ni} and Y_{2Ni} be the first and second entries of Y_{Ni}.

First, note that $\|S_N\|_F$ is bounded above:

$$\begin{aligned}\|S_N\|_F &= \frac{1}{N-1}\Big\|\sum_{i=1}^N (Y_{Ni}-\bar{Y}_N)(Y_{Ni}-\bar{Y}_N)^\top\Big\|_F\\ &\le \frac{1}{N-1}\sum_{i=1}^N |Y_{Ni}-\bar{Y}_N|^2 \le \frac{\sqrt{N}}{N-1}\left(\sum_{i=1}^N |Y_{Ni}-\bar{Y}_N|^4\right)^{1/2} \le \frac{N\sqrt{L}}{N-1},\end{aligned}$$

where the first inequality follows from the Triangle Inequality and $\|ab^\top\|_F = |a|\cdot|b|$ for two vectors a and b, and the second inequality by the Cauchy–Schwarz Inequality. By (i), we have $|\hat{\bar{Y}}_N - \bar{Y}_N| \overset{\text{a.s.}}{\to} 0$.

Second, let W_{Ni} be the indicator for Y_{Ni} being in the simple random sample. Put $\tilde{S}_{12N} =$

$(N_1-1)^{-1}\sum_{i=1}^{N} W_{Ni}(Y_{1Ni}-\bar{Y}_{1N})(Y_{2Ni}-\bar{Y}_{2N})$, then

$$\begin{aligned}\hat{S}_{12N}-\tilde{S}_{12N} =&\frac{1}{N_1-1}\sum_{i=1}^{N} W_{Ni}\{(Y_{1Ni}-\hat{\bar{Y}}_{1N})(Y_{2Ni}-\hat{\bar{Y}}_{2N})-(Y_{1Ni}-\bar{Y}_{1N})(Y_{2Ni}-\bar{Y}_{2N})\}\\ =&\frac{1}{N_1-1}\sum_{i=1}^{N} W_{Ni}\{(\bar{Y}_{1N}-\hat{\bar{Y}}_{1N})Y_{2Ni}+(\bar{Y}_{2N}-\hat{\bar{Y}}_{2N})Y_{1Ni}+\hat{\bar{Y}}_{1N}\hat{\bar{Y}}_{2N}-\bar{Y}_{1N}\bar{Y}_{2N}\}\\ =&\frac{N_1}{N_1-1}\{(\bar{Y}_{1N}-\hat{\bar{Y}}_{1N})\hat{\bar{Y}}_{2N}+(\bar{Y}_{2N}-\hat{\bar{Y}}_{2N})\hat{\bar{Y}}_{1N}+\hat{\bar{Y}}_{1N}\hat{\bar{Y}}_{2N}-\bar{Y}_{1N}\bar{Y}_{2N}\}\\ =&\frac{-N_1}{N_1-1}(\bar{Y}_{1N}-\hat{\bar{Y}}_{1N})(\bar{Y}_{2N}-\hat{\bar{Y}}_{2N})\overset{\text{a.s.}}{\to}0\end{aligned}$$

Third, we note that the following variance is bounded above for all N:

$$\begin{aligned}&\mathrm{Var}\{(Y_{1Ni}-\bar{Y}_{1N})(Y_{2Ni}-\bar{Y}_{2N}):i=1,\dots,N\}\\ \le&\frac{1}{N-1}\sum_{i=1}^{N}(Y_{1Ni}-\bar{Y}_{1N})^2(Y_{2Ni}-\bar{Y}_{2N})^2\\ \le&\frac{1}{N-1}\left\{\sum_{i=1}^{N}(Y_{1Ni}-\bar{Y}_{1N})^4\sum_{i=1}^{N}(Y_{2Ni}-\bar{Y}_{2N})^4\right\}^{1/2}\\ \le&\frac{1}{N-1}\left\{\sum_{i=1}^{N}(Y_{1Ni}-\bar{Y}_{1N})^4\vee\sum_{i=1}^{N}(Y_{2Ni}-\bar{Y}_{2N})^4\right\}\\ \le&\frac{NL}{N-1}.\end{aligned}$$

So by (i), $\left|N_1^{-1}\sum_{i=1}^{N} W_{N,i}(Y_{1Ni}-\bar{Y}_{1N})(Y_{2Ni}-\bar{Y}_{2N})-N^{-1}\sum_{i=1}^{N}(Y_{1Ni}-\bar{Y}_{1N})(Y_{2Ni}-\bar{Y}_{2N})\right|\overset{\text{a.s.}}{\to}0$. In addition, $S_{12N}\le\|S_N\|_F$ is bounded from above. These imply that

$$\begin{aligned}|\tilde{S}_{12N}-S_{12N}| =&\Big|\frac{N_1}{N_1-1}\left\{\frac{1}{N_1}\sum_{i=1}^{N} W_{N,i}(Y_{1Ni}-\bar{Y}_{1N})(Y_{2Ni}-\bar{Y}_{2N})-\frac{1}{N}\sum_{i=1}^{N}(Y_{1Ni}-\bar{Y}_{1N})(Y_{2Ni}-\bar{Y}_{2N})\right\}\\ &+(\frac{N_1}{(N_1-1)N}-\frac{1}{N-1})\sum_{i=1}^{N}(Y_{1Ni}-\bar{Y}_{1N})(Y_{2Ni}-\bar{Y}_{2N})\Big|\\ \le&\frac{N_1}{N_1-1}\Big|\frac{1}{N_1}\sum_{i=1}^{N} W_{N,i}(Y_{1Ni}-\bar{Y}_{1N})(Y_{2Ni}-\bar{Y}_{2N})-\frac{1}{N}\sum_{i=1}^{N}(Y_{1Ni}-\bar{Y}_{1N})(Y_{2Ni}-\bar{Y}_{2N})\Big|\\ &+\frac{N-N_1}{(N-1)(N_1-1)}\frac{N-1}{N}S_{12N}\overset{\text{a.s.}}{\to}0.\end{aligned}$$

We now finally have $|\hat{S}_{12N}-S_{12N}|\le|\hat{S}_{12N}-\tilde{S}_{12N}|+|\tilde{S}_{12N}-S_{12N}|\overset{\text{a.s.}}{\to}0$. □

Lemma A6. Under Assumption 4 and for all sequences of W, the imputed potential outcomes in (15) satisfy $\lim_{N\to\infty}N^{-1}\max_{i,j}|Y_i^*(j)-\bar{Y}^*(j)|^2=0$.

Proof of Lemma A6. From (15), we obtain $\{Y_i^*(j)_1:i=1,\dots,N,j=1,\dots,J\}$ from $\{W_i,(Y_i^{\text{obs}})_1:i=1,\dots,N\}$ in the same way as FRT-2. So by Lemma A4, we have $\lim_{N\to\infty}N^{-1}\max_{i,j}\{Y_i^*(j)_1-$

$\bar{Y}^*(j)_1\}^2 = 0$. Doing the same for the other $d-1$ entries gives the desired result. $\square$

Proofs of the Main Theorems

We extend the notation to handle the permutation distributions in Theorems 1–3. The observed values satisfy the following ANOVA identities:

$$\bar{Y}^{\text{obs}}_{\bullet} = \frac{1}{N}\sum_{i=1}^{N} Y_i^{\text{obs}} = \frac{1}{N}\sum_{j=1}^{J} N_j \hat{\bar{Y}}(j),$$
$$\sum_{i=1}^{N}(Y_i^{\text{obs}} - \bar{Y}^{\text{obs}}_{\bullet})^2 = \sum_{j=1}^{J}(N_j - 1)\hat{S}(j,j) + \sum_{j=1}^{J} N_j\{\hat{\bar{Y}}(j) - \bar{Y}^{\text{obs}}_{\bullet}\}^2.$$

The imputed potential outcomes in FRT-2 are strictly additive, as $Y_i^*(j) - Y_i^*(k) = z_j + Y_i^{\text{obs}} - z_{W_i} - (z_k + Y_i^{\text{obs}} - z_{W_i}) = z_j - z_k$ does not depend on the unit i. Define $\bar{z} = N^{-1}\sum_{j=1}^{J} N_j z_j$. The means and variances of the imputed potential outcomes are

$$\bar{Y}^*(j) = \frac{1}{N}\sum_{i=1}^{N}(Y_i^{\text{obs}} + z_j - z_{W_i}) = \bar{Y}^{\text{obs}}_{\bullet} + z_j - \bar{z},$$
$$S^*(j,k) = \frac{1}{N-1}\sum_{i=1}^{N}\{Y_i^*(j) - \bar{Y}^*(j)\}\{Y_i^*(k) - \bar{Y}^*(k)\} = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i^{\text{obs}} - z_{W_i} - \bar{Y}^{\text{obs}}_{\bullet} + \bar{z})^2 = s^*.$$

As a consequence of strict additivity, $S^*(j,k)$ does not depend on (j,k), so we call their common value s^*. The covariance structure of the imputed potential outcomes is thus $S^* = s^* 1_{J\times J}$.

The analogs of $\hat{D}$ and V for imputed potential outcomes are

$$\hat{D}_\pi = N \times \text{diag}\left\{\hat{S}_\pi(1,1)/N_1, \ldots, \hat{S}_\pi(J,J)/N_J\right\}, \qquad V_\pi = s^*(P^{-1} - 1_{J\times J}).$$

Similar to (4), we have that, conditional on W, $\hat{D}_\pi - s^* P^{-1} \xrightarrow{\mathbb{P}} 0$.

Proof of Theorems 1–3. We first prove the asymptotic distribution under $H_{0\text{N}}(C,x)$, and then prove the permutation distributions.

Asymptotic distributions of X^2, F, and B. Under Assumption 1, we assume $H_{0\text{N}}(C,x)$. We have $\sqrt{N}(C\hat{\bar{Y}} - x) \xrightarrow{\text{d}} \mathcal{N}(0_m, CVC^\top)$ by Proposition 3, $C\hat{D}C^\top \xrightarrow{\mathbb{P}} CDC^\top \succ 0$ and $CDC^\top \succeq CVC^\top$ by (4). Hence, by Lemma A1

$$X^2 = \sqrt{N}(C\hat{\bar{Y}} - x)^\top (C\hat{D}C^\top)^{-1}\sqrt{N}(C\hat{\bar{Y}} - x) \xrightarrow{\text{d}} \sum_{j=1}^{m} a_j \xi_j^2, \quad \text{with } a_j \in [0,1] \quad (j = 1, \ldots, m).$$

We deal with B and F similarly. Assume $x = 0_m$. By (4) and the Continuous Mapping Theorem,

$\mathrm{tr}(M\hat{D})CC^\top \xrightarrow{\mathbb{P}} \mathrm{tr}(MD)CC^\top = \mathrm{tr}(DM)CC^\top$. By Lemma A1,

$$\begin{aligned}B =&\sqrt{N}(C\hat{\bar{Y}})^\top(\mathrm{tr}(M\hat{D})CC^\top)^{-1}\sqrt{N}C\hat{\bar{Y}} \xrightarrow{\mathrm{d}} \sum_{j=1}^{m}\lambda_j\big(CVC^\top(\mathrm{tr}(DM)CC^\top)^{-1}\big)\xi_j^2\\ \overset{\mathrm{d}}{=}&\sum_{j=1}^{m}\frac{1}{\mathrm{tr}(DM)}\lambda_j(VC^\top(CC^\top)^{-1}C)\xi_j^2 \overset{\mathrm{d}}{=} \frac{\sum_{j=1}^{m}\lambda_j(VM)\xi_j^2}{\mathrm{tr}(DM)}.\end{aligned}$$

Recall the definitions of $\mathcal{X}$ and $\hat{\sigma}^2$ in F, we have $\hat{\sigma}^2 \xrightarrow{\mathbb{P}} \sum_{j=1}^{J} p_j S(j,j) = \bar{S}$ by Proposition 2, $(N_j-1)/(N-J) \to p_j$, and

$$\left(N^{-1}\mathcal{X}^\top\mathcal{X}\right)^{-1} = \mathrm{diag}\left\{N_1/N,\ldots,N_J/N\right\}^{-1} \xrightarrow{\mathbb{P}} P^{-1}.$$

Therefore, by Lemma A1,

$$mF = \sqrt{N}(C\hat{\bar{Y}})^\top\left\{\hat{\sigma}^2 C\left(N^{-1}\mathcal{X}^\top\mathcal{X}\right)^{-1}C^\top\right\}^{-1}\sqrt{N}C\hat{\bar{Y}} \xrightarrow{\mathrm{d}} \sum_{j=1}^{m}\lambda_j\big(CVC^\top(\bar{S}CP^{-1}C^\top)^{-1}\big)\xi_j^2.$$

Permutation distributions. We first show, for almost all realizations of the sequence of treatment assignments W, that Assumption 1 holds for $\{V_i^*(j) : i = 1,\ldots,N, j = 1,\ldots,J\}$ where $V_i^*(j) := \{Y_i^*(j) - \bar{Y}^*(j)\}/\sqrt{s^*}$ are the standardized imputed potential outcomes. Clearly they always have mean 0 and variance 1, so it is enough to verify that, almost surely

$$\lim_{N\to\infty}\max_{i,j}\frac{1}{N}\{V_i^*(j) - \bar{V}^*(j)\}^2 = \lim_{N\to\infty}\max_{i,j}\frac{\{Y_i^*(j) - \bar{Y}^*(j)\}^2}{Ns^*} = 0 \tag{A3}$$

Recalling the definition of s^*, we have

$$\begin{aligned}s^* =&\frac{1}{N-1}\sum_{i=1}^{N}(Y_i^{\mathrm{obs}} - z_{W_i} - \bar{Y}_\bullet^{\mathrm{obs}} + \bar{z})^2\\ =&\frac{1}{N-1}\sum_{j=1}^{J}(N_j-1)\hat{S}(j,j) + \frac{N}{N-1}\sum_{j=1}^{J}\frac{N_j}{N}\{\hat{\bar{Y}}(j) - z_j - \bar{Y}_\bullet^{\mathrm{obs}} + \bar{z}\}^2\\ \geq&\frac{N_1-1}{N-1}\hat{S}(1,1) \xrightarrow{\mathrm{a.s.}} p_1 S(1,1)\end{aligned}$$

where the last step is by Lemma A3. This shows the sequence $(s^*)_{N\geq 2J}$ is bounded away from 0, as $p_1, S(1,1) > 0$. Now we also have $\lim_{N\to\infty} N^{-1}\max_{i,j}\{Y_i^*(j) - \bar{Y}^*(j)\}^2 = 0$, no matter what the realization of the sequence W is, by Lemma A4. These two facts together show (A3).

Because $\hat{S}(1,1) \xrightarrow{\mathrm{a.s.}} S(1,1)$ by Lemma A3, we for the rest of the proof fix a sequence $\{W\}_{N=1}^{\infty}$ along which $\hat{S}(1,1) \to S(1,1)$, so the only remaining randomness comes from $\pi \sim \mathrm{Unif}(\Pi_N)$. Note for $i = 1,\ldots,N$ that $CV_i^* = C(Y_i^* - \bar{Y}^*)/\sqrt{s^*} = 0_m$ because $CY_i^* = x$ by (2). In particular, the standardized imputed potential outcomes satisfy $H_{0\mathrm{N}}(C, 0_m)$, ie $C\bar{V}^* = 0_m$. Hence, by

Proposition 3, we have

$$\sqrt{\frac{N}{s^*}}(C\hat{\bar{Y}}_\pi - x) = \sqrt{N}C(\hat{\bar{Y}}_\pi - \bar{Y}^*)/\sqrt{s^*} = \sqrt{N}C\hat{\bar{V}}_\pi \xrightarrow{\text{d}} \mathcal{N}\big(0_m, C(P^{-1} - 1_J 1_J^\top)C^\top\big) \overset{\text{d}}{=} \mathcal{N}(0_m, CP^{-1}C^\top),$$

because the standardized imputed potential outcomes have covariance structure $1_J 1_J^\top$ and $C1_J = 0_m$. Next, for $j = 1, \dots, J$, we have

$$\frac{\hat{S}_\pi(j,j)}{s^*} = \frac{1}{N-1}\sum_{i=1}^N W_{\pi(i)}(j)\frac{\{Y_i^*(j) - \bar{Y}^*(j)\}^2}{s^*} = \frac{1}{N-1}\sum_{i=1}^N W_{\pi(i)}(j)V_i^*(j)^2 \xrightarrow{\mathbb{P}} 1$$

by Proposition 2 and because the standardized imputed potential outcomes have group variances 1. Then

$$\hat{D}_\pi/s^* \xrightarrow{\mathbb{P}} P^{-1}, \quad \hat{\sigma}_\pi^2/s^* = (N-J)^{-1}\sum_{j=1}^J (N_j - 1)\hat{S}_\pi(j,j)/s^* \xrightarrow{\mathbb{P}} 1, \quad \text{tr}(M\hat{D}_\pi)/s^* \xrightarrow{\mathbb{P}} \text{tr}(MP^{-1}).$$

We thus finally have by Lemma A1

$$X_\pi^2 = \sqrt{\frac{N}{s^*}}(C\hat{\bar{Y}}_\pi - x)^\top (C\hat{D}_\pi C^\top/s^*)^{-1}\sqrt{\frac{N}{s^*}}(C\hat{\bar{Y}}_\pi - x) \xrightarrow{\text{d}} \sum_{j=1}^m \lambda_j\big(CP^{-1}C^\top(CP^{-1}C^\top)^{-1}\big)\xi_j^2 \overset{\text{d}}{=} \chi_m^2,$$

with $x = 0_m$ for the B and F statistics:

$$\begin{aligned}
B_\pi =& \sqrt{\frac{N}{s^*}}(C\hat{\bar{Y}}_\pi)^\top\{\text{tr}(M\hat{D}_\pi)CC^\top/s^*\}^{-1}\sqrt{\frac{N}{s^*}}C\hat{\bar{Y}}_\pi \\
&\xrightarrow{\text{d}} \sum_{j=1}^m \lambda_j\big(CP^{-1}C^\top(\text{tr}(MP^{-1})CC^\top)^{-1}\big)\xi_j^2 \overset{\text{d}}{=} \frac{\sum_{j=1}^m \lambda_j(MP^{-1})\xi_j^2}{\text{tr}(MP^{-1})}, \\
mF_\pi =& \sqrt{\frac{N}{s^*}}(C\hat{\bar{Y}}_\pi)^\top\left\{\frac{\hat{\sigma}_\pi^2}{s^*}C\left(\frac{1}{N}X^\top X\right)^{-1}C^\top\right\}^{-1}\sqrt{\frac{N}{s^*}}C\hat{\bar{Y}}_\pi \\
&\xrightarrow{\text{d}} \sum_{j=1}^m \lambda_j\big(CP^{-1}C^\top(CP^{-1}C^\top)^{-1}\big)\xi_j^2 \overset{\text{d}}{=} \chi_m^2. \qquad \square
\end{aligned}$$

Extending Theorem 1 to the case of stratified experiments or vector potential outcomes is straightforward. We also supply their proofs for completeness.

Proof of Theorem 4. We first discuss the asymptotic distribution of X^2 under $H_{0\text{N}}(C, x)$ and then discuss its permutation distribution.

Asymptotic distribution of X^2. For $h = 1, \dots, H$, we have $\mathbb{E}\hat{\bar{Y}}_{[h]} = \bar{Y}_{[h]}$, and Assumption 1 holds in each stratum h. By Proposition 3,

$$\sqrt{N_{[h]}}C(\hat{\bar{Y}}_{[h]} - \bar{Y}_{[h]}) \xrightarrow{\text{d}} \mathcal{N}(0_m, CV_{[h]}C^\top), \text{ where } V_{[h]} = \underset{N\to\infty}{\text{plim}}\,\hat{D}_{[h]} - S_{[h]}.$$

Under $H_{0N}(C, x)$, we have $x = C\bar{Y} = N^{-1}\sum_{h=1}^{H} N_{[h]} C\bar{Y}_{[h]}$. Because $(\hat{\bar{Y}}_{[1]}, \ldots, \hat{\bar{Y}}_{[H]})$ are mutually independent in SRE, we have

$$\begin{aligned}
\sqrt{N}(C\check{\bar{Y}} - x) &= \sum_{h=1}^{H} \sqrt{\frac{N_{[h]}}{N}} \sqrt{N_{[h]}} C(\hat{\bar{Y}}_{[h]} - \bar{Y}_{[h]}) \\
&\xrightarrow{\mathrm{d}} \sum_{h=1}^{H} \sqrt{\omega_{[h]}} \mathcal{N}\left(0_m, CV_{[h]}C^\top\right) \stackrel{\mathrm{d}}{=} \mathcal{N}\left(0_m, \sum_{h=1}^{H} \omega_{[h]} CV_{[h]}C^\top\right).
\end{aligned}$$

Next, note $\operatorname{plim}_{N\to\infty} \hat{D}_{[h]} \succeq V_{[h]}$ implies

$$\operatorname*{plim}_{N\to\infty} \sum_{h=1}^{H} \frac{N_{[h]}}{N} C\hat{D}_{[h]}C^\top \succeq \sum_{h=1}^{H} \omega_{[h]} CV_{[h]}C^\top,$$

so by Lemma A1, we have

$$X^2 = \sqrt{N}(C\check{\bar{Y}} - x)^\top \left(C \sum_{h=1}^{H} \frac{N_{[h]}}{N} \hat{D}_{[h]} C^\top\right)^{-1} \sqrt{N}(C\check{\bar{Y}} - x) \xrightarrow{\mathrm{d}} \sum_{j=1}^{m} a_j \xi_j^2.$$

Permutation distribution of X^2. We first show Assumption 1 holds almost surely within each stratum for the centered imputed potential outcomes $Y_i^*(j) - \bar{Y}_{[h]}^*(j)$. Note that these potential outcomes have mean 0, so the means automatically converge. Because the original potential outcomes satisfy Assumption 1 in each stratum, Lemma A4 gives

$$\lim_{N\to\infty} N_{[h]}^{-1} \max_{j} \max_{i:X_i=h} \{Y_i^*(j) - \bar{Y}_{[h]}^*(j)\}^2 = 0.$$

Extending previous notation to the SRE, put $\bar{Y}_{[h],\bullet}^{\mathrm{obs}} = N_{[h]}^{-1}\sum_{i=1}^{N} I(X_i = h)Y_i^{\mathrm{obs}}$, $\bar{z}_{[h]} := N_{[h]}^{-1}\sum_{j=1}^{J} N_{[h]j} z_{[h],j}$. In stratum h, the covariance structure is $s_{[h]}^* 1_J 1_J^\top$, where

$$\begin{aligned}
s_{[h]}^* =& \frac{1}{N_{[h]} - 1} \sum_{i=1}^{N} I(X_i = h)(Y_i^{\mathrm{obs}} - z_{[h],W_i} - \bar{Y}_{[h],\bullet}^{\mathrm{obs}} + \bar{z}_{[h]})^2 \\
=& \frac{1}{N_{[h]} - 1} \sum_{j=1}^{J} (N_{[h]j} - 1)\hat{S}_{[h]}(j,j) + \frac{N_{[h]}}{N_{[h]} - 1} \sum_{j=1}^{J} \frac{N_{[h]j}}{N_{[h]}} \{\hat{\bar{Y}}_{[h]}(j) - z_{[h],j} - \bar{Y}_{[h],\bullet}^{\mathrm{obs}} + \bar{z}_{[h]}\}^2 \\
\xrightarrow{\text{a.s.}}& \sum_{j=1}^{J} \omega_{[h]j} \left[S_{[h]}(j,j) + \left\{\bar{Y}_{[h]}(j) - z_{[h],j} - \sum_{k=1}^{J} \omega_{[h]k} \bar{Y}_{[h]}(k) + \bar{z}_{[h]}\right\}^2\right]
\end{aligned}$$

where in the second line we use the ANOVA identities with $Y_i^{\mathrm{obs}} \leftarrow Y_i^{\mathrm{obs}} - z_{[h],W_i}$, and the almost sure convergence is due to Lemma A3, applicable because of Assumption 3. This shows Assumption 1 holds within each stratum almost surely.

For the rest of the proof, fix a sequence (W) along which $(s_{[h]}^*)_{N \geq 2HJ}$ converges. Because each

$CY_i^* = x_{[h]}$ whenever $X_i = h$, we have $C\bar{Y}^*_{[h]} = x_{[h]}$, and by Proposition 3,

$$\sqrt{N_{[h]}}C(\hat{\hat{Y}}_{[h],\pi} - \bar{Y}^*_{[h]}) \xrightarrow{\text{d}} \mathcal{N}\big(0_m, s^*_{[h]}C(P^{-1} - 1_J1_J^\top)C^\top\big) \stackrel{\text{d}}{=} \mathcal{N}(0_m, s^*_{[h]}CP^{-1}C^\top).$$

This is because the imputed potential outcomes have covariance structure $s^*_{[h]}1_J1_J^\top$ within each stratum. Since $x = N^{-1}\sum_{h=1}^H N_{[h]}x_{[h]} = N^{-1}\sum_{h=1}^H N_{[h]}C\bar{Y}^*_{[h]}$, it follows that

$$\begin{aligned}\sqrt{N}(C\breve{Y}_\pi - x) &= \sum_{h=1}^H \sqrt{\frac{N_{[h]}}{N}}\sqrt{N_{[h]}}C(\hat{\hat{Y}}_{[h],\pi} - \bar{Y}^*_{[h]}) \\ &\xrightarrow{\text{d}} \sum_{h=1}^H \sqrt{\omega_{[h]}}\mathcal{N}\left(0_m, s^*_{[h]}CP^{-1}C^\top\right) \stackrel{\text{d}}{=} \mathcal{N}\left(0_m, \sum_{h=1}^H \omega_{[h]}s^*_{[h]}CP^{-1}C^\top\right).\end{aligned}$$

again because, conditioning on W, $(\hat{\hat{Y}}_{[1],\pi}, \dots, \hat{\hat{Y}}_{[H],\pi})$ are mutually independent. Next, from Proposition 2, we have $\hat{D}_{[h],\pi} \xrightarrow{\mathbb{P}} s^*_{[h]}P^{-1}$, so $N^{-1}C\sum_{h=1}^H N_{[h]}\hat{D}_{[h],\pi}C^\top \xrightarrow{\mathbb{P}} \sum_{h=1}^H \omega_{[h]}s^*_{[h]}CP^{-1}C^\top$, and we finally have from Lemma A1

$$X^2_\pi = \sqrt{N}(C\breve{Y}_\pi - x)^\top\left(C\sum_{h=1}^H \frac{N_{[h]}}{N}\hat{D}_{[h],\pi}C^\top\right)^{-1}\sqrt{N}(C\breve{Y}_\pi - x) \xrightarrow{\text{d}} \chi^2_m. \qquad \square$$

Before proving Theorem 5, we make an auxiliary observation. The group means of the imputed potential outcomes converge almost surely under Assumption 4. Indeed, note that $\hat{\hat{Y}}(j) \xrightarrow{\text{a.s.}} \bar{Y}(j)$ under Assumption 4 by Lemma A5. We focus on $\bar{Y}^*(1)_2$: the same arguments work for any component of any $\bar{Y}^*(j)$. We define $\bar{z}_2 = N^{-1}\sum_{j=1}^J N_j z_{2j}$, and then have

$$\bar{Y}^*(1)_2 = \frac{1}{N}\sum_{i=1}^N Y_i^*(1)_2 = \frac{1}{N}\sum_{i=1}^N (z_{21} + Y^{\text{obs}}_{i,2} - z_{2,W_i}) = z_{21} + \frac{1}{N}\sum_{j=1}^J N_j\hat{\hat{Y}}(j)_2 - \bar{z}_2 \xrightarrow{\text{a.s.}} z_{21} + \sum_{j=1}^J p_j\bar{Y}(j)_2 - \bar{z}_2.$$

Proof of Theorem 5. We first discuss the asymptotic distribution of X^2 under $H_{0\text{N}}(C, x)$ and then discuss its permutation distribution.

Asymptotic distribution of X^2**.** Under Assumption 4 and $H_{0\text{N}}(C, x)$, we use Li and Ding (2017) to prove the following results in parallel with Propositions 2 and 3. First, $\hat{\hat{Y}} \xrightarrow{\mathbb{P}} \bar{Y}$ and $\hat{S}(j, j) \xrightarrow{\mathbb{P}} S(j, j)$ for $j = 1, \dots, J$. Second, $\sqrt{N}(C\hat{\hat{Y}} - x) \xrightarrow{\text{d}} \mathcal{N}(0_m, CVC^\top)$, where

$$V = \lim_{N\to\infty} N \cdot \text{Cov}(\hat{\hat{Y}}) = \lim_{N\to\infty}\begin{pmatrix} \frac{N-N_1}{N_1}S(1,1) & -S(1,2) & \cdots & -S(1,J) \\ -S(2,1) & \frac{N-N_2}{N_2}S(2,2) & \cdots & -S(2,J) \\ \vdots & \vdots & \ddots & \vdots \\ -S(J,1) & -S(J,2) & \cdots & \frac{N-N_J}{N_J}S(J,J)\end{pmatrix}. \qquad \text{(A4)}$$

Because $C\hat{D}C^\top \xrightarrow{\mathbb{P}} C(V+S)C^\top \succeq CVC^\top$, it follows from Lemma A1 that $X^2 = N(C\hat{\bar{Y}} - x)^\top(C\hat{D}C^\top)^{-1}(C\hat{\bar{Y}} - x) \xrightarrow{\mathrm{d}} \sum_{j=1}^m a_j\chi_j^2$.

Permutation distribution of X^2. We first show Assumption 4 holds almost surely for the imputed potential outcomes $Y_i^*(j)$. Their means $\bar{Y}^*(j)$ converge almost surely by the auxiliary remark. We have

$$\lim_{N\to\infty} N^{-1} \max_{i,j} |Y_i^*(j) - \bar{Y}^*(j)|^2 = 0$$

by Lemma A6, applicable because the original potential outcomes satisfy Assumption 4. It suffices to show that the (j,k)th block of the covariance matrix of the imputed potential outcomes converges almost surely. Note that for $j,k = 1,\ldots,J$,

$$\begin{aligned} S^*(j,k) &= \frac{1}{N-1}\sum_{i=1}^N \{Y_i^*(j) - \bar{Y}^*(j)\}\{Y_i^*(k) - \bar{Y}^*(k)\}^\top & \text{(A5)} \\ &= \frac{1}{N-1}\sum_{j=1}^J\sum_{i=1}^N W_i(j)\{Y_i^*(j) - \bar{Y}^*(j)\}\{Y_i^*(j) - \bar{Y}^*(j)\}^\top & \text{(A6)} \\ &= \frac{1}{N-1}\sum_{j=1}^J (N_j - 1)\hat{S}(j,j) + \frac{1}{N-1}\sum_{j=1}^J N_j\{\hat{\bar{Y}}(j) - \bar{Y}^*(j)\}\{\hat{\bar{Y}}(j) - \bar{Y}^*(j)\}^\top, & \text{(A7)} \end{aligned}$$

where (A5) is the definition, (A6) holds because $Y_i^*(j) - \bar{Y}^*(j)$ does not depend on j due to strict additivity and $\sum_{j=1}^J W_i(j) = 1$, and (A7) follows from the bias-variance decomposition and noting $Y_i^*(j) = Y_i^{\mathrm{obs}}$ when $W_i = j$. It converges almost surely because all quantities used to define it do. $\bar{Y}^*(j)$ converges almost surely by our auxiliary remark. $\hat{\bar{Y}}(j)$, and $\hat{S}(j,j)$ converge almost surely because of Lemma A5, applicable because of Assumption 5. This shows Assumption 4 holds almost surely.

For the rest of the proof, fix a sequence W along which Assumption 4 is met. The limit of $S^*(1,1)$ must be invertible because the above calculation shows $S^*(1,1) \succeq (N-1)^{-1}(N_1 - 1)S(1,1) \succ 0$. Because each $CY_i^* = x$ and the imputed potential outcomes have covariance structure $(1_J 1_J^\top) \otimes S^*(1,1)$, (A4) gives us

$$\begin{aligned} \sqrt{N}(C\hat{\bar{Y}}_\pi - x) = \sqrt{N}C(\hat{\bar{Y}}_\pi - \bar{Y}^*) &\xrightarrow{\mathrm{d}} \mathcal{N}\big(0_m, C\{(P^{-1} - 1_J 1_J^\top) \otimes S^*(1,1)\}C^\top\big) \\ &\overset{\mathrm{d}}{=} \mathcal{N}\big(0_m, C\{P^{-1} \otimes S^*(1,1)\}C^\top\big) \end{aligned}$$

The cancellation in the last line occurred, for instance because the $(1,2)$th block of $C\{(1_J 1_J^\top) \otimes S^*(1,1)\}C^\top$ is $(C_1 \otimes e_1^\top)\{(1_J 1_J^\top) \otimes S^*(1,1)\}(C_2 \otimes e_2^\top)^\top = (C_1 1_J 1_J^\top C_2^\top) \otimes \{e_1^\top S^*(1,1)e_2\}$, which vanishes because C_1, C_2 are themselves contrast matrices. Next,

$$\hat{D}_\pi \xrightarrow{\mathbb{P}} \mathrm{diag}\left\{\frac{S^*(1,1)}{p_1}, \ldots, \frac{S^*(1,1)}{p_J}\right\} = P^{-1} \otimes S^*(1,1)$$

so $C\hat{D}_\pi C^\top \xrightarrow{\mathbb{P}} C\{P^{-1} \otimes S^*(1,1)\}C^\top$, and we finally have from Lemma A1 that $X^2_\pi = N(C\hat{\bar{Y}}_\pi - x)^\top(C\hat{D}_\pi C^\top)^{-1}(C\hat{\bar{Y}}_\pi - x) \xrightarrow{\mathrm{d}} \chi^2_m$. □

Proofs of other results

Proof of Proposition 1. The conclusion follows from

$$\max_{i,j} \frac{1}{N}\{Y_i(j) - \bar{Y}(j)\}^2 = \frac{1}{N}\left[\max_{i,j}\{Y_i(j) - \bar{Y}(j)\}^4\right]^{1/2} \leq \frac{1}{N}\left[\max_j \sum_{i=1}^N \{Y_i(j) - \bar{Y}(j)\}^4\right]^{1/2} \leq \sqrt{\frac{L}{N}} \to 0.$$

□

Proof of Proposition 2. It follows from Theorem 1 and Proposition 1 of Li and Ding (2017). □

Proof of Proposition 3. It follows from Theorem 5 of of Li and Ding (2017). □

Proof of Proposition 4. Assume $H_{0\mathrm{N}}(C,x)$ throughout. Define

$$F(x) = \mathbb{P}(T \leq x), \quad G(x) = \mathbb{P}(T < x), \quad F_W(x) = \mathbb{P}(T_\pi \leq x|W), \quad G_W(x) = \mathbb{P}(T_\pi < x|W).$$

Let $U \sim \mathrm{Unif}(0,1)$. We will show that

$$\mathbb{P}\left\{(N!)^{-1}\sum_{\pi\in\Pi_N} 1(T_\pi \geq T) \leq \alpha\right\} \leq \alpha \text{ for all } \alpha \in (0,1) \Longleftrightarrow T \leq_{\mathrm{st}} T_\pi|W \text{ for all sequence of } W.$$

"$\Leftarrow$" Fix $\alpha \in (0,1)$, then

$$\mathbb{P}\left\{\frac{1}{N!}\sum_{\pi\in\Pi_N} 1(T_\pi \geq T) \leq \alpha\right\} = \mathbb{P}\{1 - G_W(T) \leq \alpha\} \leq \mathbb{P}\{G(T) \geq 1-\alpha\} \leq \mathbb{P}(U \geq 1-\alpha) = \alpha$$

where we have used $T \leq_{\mathrm{st}} T_\pi|W$ if and only if $G_W(x) \leq G(x)$ for $x \in \mathbb{R}$ and $G(T) \leq_{\mathrm{st}} U$.

"$\Rightarrow$" If for some W it is not true that $T \leq_{\mathrm{st}} T_\pi|W$, then there exists $x \in \mathbb{R}$ such that $F(x) < G_W(x)$ (this is because $F = G$, $F_W = G_W$, Lebesgue almost everywhere), pick $\alpha \in (1-F(x), 1-G_W(x))$. Then we fail to control type I error because

$$\mathbb{P}\left\{\frac{1}{N!}\sum_{\pi\in\Pi_N} 1(T_\pi \geq T) \leq \alpha\right\} \geq 1 - \mathbb{P}\{G_W(T) \leq 1-\alpha\} = 1 - F\big(\sup\{t : G_W(t) \leq 1-\alpha\}\big) \geq 1 - F(x) > \alpha$$

where the second "=" follows because $\{t : G_W(t) \leq 1-\alpha\}$ is closed (due to the left continuity of G_W), hence its measure under the distribution of T is F evaluated at its right endpoint. The second $\geq$ follows because $G_W(t) \leq 1-\alpha < G_W(x)$ implies $t \leq x$ (as G_W is nondecreasing), so $\sup\{t : G_W(t) \leq 1-\alpha\} \leq x$. □

Proof of Corollary 1. First, if $S(1,1) = \cdots = S(J,J) = \bar{S} = \sum_{j=1}^{J} p_j S(j,j)$, then $D = \bar{S}P^{-1}$ based on the definition in (4). Recall that $\lambda_j(VM) \le \lambda_j(DM)$ because $V \preceq D$. Therefore, under $H_{0\mathrm{N}}(C, 0_m)$, Theorem 2 implies that

$$B \xrightarrow{\mathrm{d}} \frac{\sum_{j=1}^{m} \lambda_j(VM)\xi_j^2}{\mathrm{tr}(DM)} \le_{\mathrm{st}} \frac{\sum_{j=1}^{m} \lambda_j(DM)\xi_j^2}{\mathrm{tr}(DM)} = \frac{\sum_{j=1}^{m} \bar{S}\lambda_j(P^{-1}M)\xi_j^2}{\bar{S}\,\mathrm{tr}(P^{-1}M)} \stackrel{\mathrm{d}}{=} \frac{\sum_{j=1}^{m} \lambda_j(P^{-1}M)\xi_j^2}{\mathrm{tr}(P^{-1}M)}.$$

The right-hand side of the above equation is the asymptotic distribution of $B_\pi | W$.

Second, if C is a row vector, then $M = C^\top C / CC^\top$. Therefore,

$$B = \frac{\hat{\bar{Y}}^\top C^\top C\hat{\bar{Y}}/CC^\top}{\mathrm{tr}(C^\top C\hat{D})/CC^\top} = \frac{(C\hat{\bar{Y}})^\top C\hat{\bar{Y}}}{C\hat{D}C^\top} = (C\hat{\bar{Y}})^\top (C\hat{D}C^\top)^{-1} C\hat{\bar{Y}} = X^2. \qquad \square$$

Proof of Proposition 5. Under a balanced design, $N_1 = \cdots = N_J = N/J$, $\mathcal{X}^\top\mathcal{X} = N_1 \times I_J$, and $\hat{\sigma}^2 = J^{-1}\sum_{j=1}^{J} \hat{S}(j,j)$. Thus, $F = N_1 \hat{\bar{Y}}^\top M \hat{\bar{Y}}/(m\hat{\sigma}^2)$.

If the projection matrix M has the same values on its main diagonal, then $\mathrm{tr}(M) = \mathrm{rank}(M)$ implies each value is in fact m/J. This further implies

$$\frac{N}{\mathrm{tr}(\hat{D}M)} = \frac{N}{\sum_{j=1}^{J} \frac{N}{N_j}\hat{S}(j,j)\frac{m}{J}} = \frac{N}{m\sum_{j=1}^{J}\hat{S}(j,j)} = \frac{N_1}{m\hat{\sigma}^2} \Longrightarrow B = \frac{N\hat{\bar{Y}}^\top M\hat{\bar{Y}}}{\mathrm{tr}(\hat{D}M)} = \frac{N_1\hat{\bar{Y}}^\top M\hat{\bar{Y}}}{m\hat{\sigma}^2} = F. \qquad \square$$

Proof of Corollary 2. If $S(1,1) = \cdots = S(J,J)$, then $\bar{S} = \sum_{j=1}^{J} p_j S(j,j) = S(1,1)$ and $D = \bar{S} \times P^{-1}$. Therefore, $0 \le \lambda_j\big(CVC^\top(\bar{S}CP^{-1}C^\top)^{-1}\big) = \lambda_j\big(CVC^\top(CDC^\top)^{-1}\big) \le 1$ because $V \preceq D$. By Theorem 3, under $H_{0\mathrm{N}}(C, 0_m)$, we have

$$m \times F \xrightarrow{\mathrm{d}} \sum_{j=1}^{m} \lambda_j\big(CVC^\top(\bar{S}CP^{-1}C^\top)^{-1}\big)\xi_j^2 \le_{\mathrm{st}} \chi^2_m, \qquad m \times F_\pi | W \xrightarrow{\mathrm{d}} \chi^2_m. \qquad \square$$

Proof of Proposition 6. The conclusions follow from simple linear algebra facts. They seem to be known, but we give a proof for completeness.

We first equate the X^2. As stated, in the ANOVA setting, $C = (1_{J-1},\ -I_{J-1})$ and $x = 0_{J-1}$. Put $Q_j = N_j/\hat{S}(j,j)$ and $Q = \sum_{j=1}^{J} Q_j$. Then by block matrix multiplication

$$\frac{1}{N}C\hat{D}C^\top = (1_{J-1},\ -I_{J-1})\begin{pmatrix} 1/Q_1 & & 0 \\ & \ddots & \\ 0 & & 1/Q_J \end{pmatrix}\begin{pmatrix} 1_{J-1}^\top \\ -I_{J-1} \end{pmatrix} = \frac{1}{Q_1}1_{J-1}1_{J-1}^\top + \begin{pmatrix} 1/Q_2 & & 0 \\ & \ddots & \\ 0 & & 1/Q_J \end{pmatrix}$$

Thus using the Sherman–Morrison formula, we have

$$
\begin{aligned}
(\frac{1}{N}C\hat{D}C^\top)^{-1} &= \begin{pmatrix} Q_2 & & 0 \\ & \ddots & \\ 0 & & Q_J \end{pmatrix} - \left\{ \frac{1}{Q_1} \begin{pmatrix} Q_2 \\ \vdots \\ Q_J \end{pmatrix} (Q_2, \ldots, Q_J) \right\} \Big/ \left\{ 1 + \frac{1}{Q_1} \sum_{j=2}^{J} Q_j \right\} \\
&= \begin{pmatrix} Q_2 & & 0 \\ & \ddots & \\ 0 & & Q_J \end{pmatrix} - \frac{1}{Q} \begin{pmatrix} Q_2 \\ \vdots \\ Q_J \end{pmatrix} (Q_2, \ldots, Q_J)
\end{aligned}
$$

Thus from (5) we have

$$
\begin{aligned}
X^2 &= \left(\hat{\bar{Y}}(1) - \hat{\bar{Y}}(2), \ldots, \hat{\bar{Y}}(1) - \hat{\bar{Y}}(J)\right) \left\{ \begin{pmatrix} Q_2 & & 0 \\ & \ddots & \\ 0 & & Q_J \end{pmatrix} - \frac{1}{Q} \begin{pmatrix} Q_2 \\ \vdots \\ Q_J \end{pmatrix} (Q_2, \ldots, Q_J) \right\} \begin{pmatrix} \hat{\bar{Y}}(1) - \hat{\bar{Y}}(2) \\ \vdots \\ \hat{\bar{Y}}(1) - \hat{\bar{Y}}(J) \end{pmatrix} \\
&= \sum_{j=2}^{J} Q_j \{\hat{\bar{Y}}(1) - \hat{\bar{Y}}(j)\}^2 - \frac{1}{Q} \left[\sum_{j=2}^{J} Q_j \{\hat{\bar{Y}}(1) - \hat{\bar{Y}}(j)\} \right]^2 \qquad \text{(A8)}
\end{aligned}
$$

Now we recognize the expression in (9) as Q times the variance of $\{\hat{\bar{Y}}(1), \ldots, \hat{\bar{Y}}(J)\}$ under the probabilities $Q_1/Q, \ldots, Q_J/Q$. But variance is unaffected by switching signs, and then adding the constant $\hat{\bar{Y}}(1)$ to all quantities, so (9) is Q times the variance of $\{0, \hat{\bar{Y}}(1) - \hat{\bar{Y}}(2), \ldots, \hat{\bar{Y}}(1) - \hat{\bar{Y}}(J)\}$ under the same probabilities, which is precisely what X^2 is in (A8).

Now we equate the F. Recall that $m = J - 1$. It is thus enough to show

$$
(C\hat{\bar{Y}})^\top \{C(\mathcal{X}^\top \mathcal{X})^{-1} C^\top\}^{-1} C\hat{\bar{Y}} = \sum_{j=1}^{J} N_j \{\hat{\bar{Y}}(j) - \bar{Y}^{\text{obs}}_{\bullet}\}^2.
$$

This follows an identical argument to showing the X^2 coincide, with N_j, N in place of Q_j, Q. □

Proof of Corollary 3. Because $C = (1, -1)$ is a row vector, Corollary 1 implies $B = X^2$, which is suitable. Ding and Dasgupta (2018) have proved the rest of the corollary. □

Proof of Corollary 4. Under Assumption 1 and the null hypothesis with a row vector C, we have $\sqrt{N}(C\hat{\bar{Y}} - x) \xrightarrow{\text{d}} \mathcal{N}(0, CVC^\top)$ by Theorem 3, $C\hat{D}C^\top \xrightarrow{\mathbb{P}} CDC^\top > 0$ and $CDC^\top \geq CVC^\top$ by (4). Hence,

$$
t = \frac{\sqrt{N}(x - C\hat{\bar{Y}})}{(C\hat{D}C^\top)^{1/2}} \xrightarrow{\text{d}} \mathcal{N}(0, a), \text{ where } a = \frac{CVC^\top}{CDC^\top} \in [0, 1].
$$

To show the permutation distribution, we have $\hat{S}(1,1) \xrightarrow{\text{a.s.}} S(1,1)$ by Lemma A3, so fix a sequence of W along which $\hat{S}(1,1) \to S(1,1)$. Then $\sqrt{N/s^*}(C\hat{\bar{Y}}_\pi - x) \xrightarrow{\text{d}} \mathcal{N}(0, CP^{-1}C^\top)$ and $\hat{D}_\pi/s^* \xrightarrow{\mathbb{P}} P^{-1}$

(these are intermediate steps in the proof of Theorem 1), so

$$T_\pi | W = \frac{\sqrt{N}(x - C\hat{\hat{Y}}_\pi)}{(C\hat{D}_\pi C^\top)^{1/2}} = \sqrt{\frac{N}{s^*}} \frac{x - C\hat{\hat{Y}}_\pi}{(C\hat{D}_\pi C^\top)^{1/2}} \xrightarrow{\mathrm{d}} \mathcal{N}(0, 1).$$

It finally holds that t_+ is suitable by Lemma 4 and the fact that $\mathcal{N}(0, a)_+ \leq_{\mathrm{st}} \mathcal{N}(0, 1)_+$. □

Proof of Proposition 7. We omit it because it is similar to the proof of Proposition 1. □