Award:

R305D150040 - PRESIDENT AND FELLOWS OF HARVARD COLLEGE

Paper Title:

A randomization-based perspective of analysis of variance: A test statistic robust to treatment effect heterogeneity.

Authors:

Peng Ding and Tirthankar Dasgupta

Publication Date:

Jul 24, 2017 (arXiv)

Oct 31, 2017 (online first, journal); March, 2018 (print)

DOI:

https://doi.org/10.1093/biomet/asx059

Biometrika, *pp*. 1–25

Printed in Great Britain

A randomization-based perspective of analysis of variance: a test statistic robust to treatment effect heterogeneity

BY PENG DING

Department of Statistics, University of California, Berkeley, 425 Evans Hall, Berkeley, California 94720, U.S.A.

pengdingpku@berkeley.edu

AND TIRTHANKAR DASGUPTA

Department of Statistics and Biostatistics, Rutgers University, 110 Frelinghuysen Road, Piscataway, New Jersey 08901, U.S.A.

tirthankar.dasgupta@rutgers.edu

SUMMARY

Fisher randomization tests for Neyman's null hypothesis of no average treatment effects are considered in a finite population setting associated with completely randomized experiments with more than two treatments. The consequences of using the F statistic to conduct such a test are examined both theoretically and computationally, and it is argued that under treatment effect heterogeneity, use of the F statistic in the Fisher randomization test can severely inflate the type I error under Neyman's null hypothesis. An alternative test statistic is proposed, its asymptotic distributions under Fisher's and Neyman's null hypotheses are derived, and its advantages demonstrated.

Some key words: Additivity; Fisher randomization test; Null hypothesis; One-way layout

1. INTRODUCTION

One-way analysis of variance (Fisher, 1925) is arguably the most commonly used tool to analyze completely randomized experiments with more than two treatments. The standard F test for testing equality of mean treatment effects can be justified either by assuming a linear additive super population model with identically and independently distributed normal error terms, or by using the asymptotic randomization distribution of the F statistic. As observed by many experts, units in most real-life experiments are rarely random samples from a super population, making a finite population randomization-based perspective on inference important (e.g. Rosenbaum, 2010; Imbens & Rubin, 2015; Dasgupta et al., 2015). Fisher randomization tests are useful tools for such inference, because they pertain to a finite population of units, and assess the statistical significance of treatment effects without making any assumptions about the underlying distribution of the outcome.

In causal inference from finite population, two types of hypotheses are of interest: Fisher's sharp null hypothesis of no treatment effect on any experimental unit (Fisher, 1935; Rubin, 1980), and Neyman's null hypothesis of no average treatment effect (Neyman, 1923, 1935). These hypotheses are equivalent without treatment effect heterogeneity (Ding et al., 2016) or equivalently under the assumption of strict additivity of treatment effects, i.e., the same treatment effect for each unit (Kempthorne, 1952). In the context of a multi-treatment completely randomized experiment, Neyman's null hypothesis allows for treatment effect heterogeneity, which is weaker than Fisher's null hypothesis and is of greater interest. We find that the Fisher randomization test using the F statistic can inflate the type I error under Neyman's null hypothesis, when the sample sizes and variances of the outcomes under different treatment levels are negatively associated. We propose to use the X^2 statistic defined in §5, a statistic robust to treatment effect heterogeneity, because the resulting Fisher randomization test is exact under Fisher's null hypothesis and controls asymptotic type I error under Neyman's null hypothesis.

2. COMPLETELY RANDOMIZED EXPERIMENT WITH J TREATMENTS

Consider a finite population of N experimental units, each of which can be exposed to any one of J treatments. Let $Y_i(j)$ denote the potential outcome (Neyman, 1923) of unit i when assigned to treatment level j $(i = 1, \ldots, N; j = 1, \ldots, J)$. For two different treatment levels j and j', we define the unit-level treatment effect as $\tau_i(j, j') = Y_i(j) - Y_i(j')$, and the population-level treatment effect as

$$\tau(j, j') = N^{-1} \sum_{i=1}^{N} \tau_i(j, j') = N^{-1} \sum_{i=1}^{N} \{Y_i(j) - Y_i(j')\} \equiv \bar{Y}_{\cdot}(j) - \bar{Y}_{\cdot}(j'),$$

where $\bar{Y}_{\cdot}(j) = N^{-1} \sum_{i=1}^{N} Y_i(j)$ is the average of the N potential outcomes for treatment j.

The treatment assignment mechanism can be represented by the binary random variable $W_i(j)$, which equals 1 if the ith unit is assigned to treatment j, and 0 otherwise. Equivalently, it can be represented by the discrete random variable $W_i = \sum_{j=1}^{J} j W_i(j) \in \{1, \ldots, J\}$, the treatment received by unit i. Let $(W_1, \ldots, W_N)$ be the treatment assignment vector, and let $(w_1, \ldots, w_N)$ denote its realization. For the $N = \sum_{j=1}^{J} N_j$ units, $(N_1, \ldots, N_J)$ are assigned at random to treatments $(1, \ldots, J)$ respectively, the treatment assignment mechanism satisfies $\mathrm{pr}\{(W_1, \ldots, W_N) = (w_1, \ldots, w_N)\} = \prod_{j=1}^{J} N_j!/N!$ if $\sum_{i=1}^{N} W_i(j) = N_j$, and 0 otherwise. The observed outcomes are deterministic functions of the treatment received and the potential outcomes, given by $Y_i^{\mathrm{obs}} = \sum_{j=1}^{J} W_i(j) Y_i(j)$ $(i = 1, \ldots, N)$.

3. THE FISHER RANDOMIZATION TEST UNDER THE SHARP NULL HYPOTHESIS

Fisher (1935) was interested in testing the following sharp null hypothesis of zero individual treatment effects:

$$H_0(\text{Fisher}) : Y_i(1) = \cdots = Y_i(J), \quad (i = 1, \ldots, N).$$

Under H_0(Fisher), all the J potential outcomes $Y_i(1), \ldots, Y_i(J)$ are equal to the observed outcome Y_i^{obs}, for all units $i = 1, \ldots, N$. Thus any possible realization of the treatment assignment vector would generate the same vector of observed outcomes. This means, under H_0(Fisher) and given any realization $(W_1, \ldots, W_N) = (w_1, \ldots, w_N)$, the observed outcomes are fixed. Consequently, the randomization distribution or null distribution of any test statistic, which is a function of the observed outcomes and treatment assignment vector, is its distribution over all possible realizations of the treatment assignment. The p-value is the tail probability measuring the extremeness of the test statistic with respect to its randomization distribution. Computationally, we can enumerate or simulate a subset of all possible randomizations to obtain this randomization distribution of any test statistic and thus perform the Fisher randomization test (Fisher, 1935; Imbens & Rubin, 2015). Fisher (1925) suggested using the F statistic to test the departure from H_0(Fisher). Define $\bar{Y}_{\cdot}^{\mathrm{obs}}(j) = N_j^{-1} \sum_{i=1}^{N} W_i(j) Y_i^{\mathrm{obs}}$ as the sample average of the observed outcomes within treatment level j, and $\bar{Y}_{\cdot}^{\text{obs}} = N^{-1}\sum_{i=1}^{N} Y_i^{\text{obs}}$ as the sample average of all the observed outcomes. Define $s_{\text{obs}}^2(j) = (N_j - 1)^{-1}\sum_{i=1}^{N} W_i(j)\{Y_i^{\text{obs}} - \bar{Y}_{\cdot}^{\text{obs}}(j)\}^2$ and $s_{\text{obs}}^2 = (N-1)^{-1}\sum_{i=1}^{N}(Y_i^{\text{obs}} - \bar{Y}_{\cdot}^{\text{obs}})^2$ as the corresponding sample variances with divisors $N_j - 1$ and $N - 1$, respectively. Let

$$\text{SSTre} = \sum_{j=1}^{J} N_j\{\bar{Y}_{\cdot}^{\text{obs}}(j) - \bar{Y}_{\cdot}^{\text{obs}}\}^2$$

be the treatment sum of squares, and let

$$\text{SSRes} = \sum_{j=1}^{J} \sum_{i:W_i(j)=1} \{Y_i^{\text{obs}} - \bar{Y}_{\cdot}^{\text{obs}}(j)\}^2 = \sum_{j=1}^{J}(N_j - 1)s_{\text{obs}}^2(j)$$

be the residual sum of squares. The treatment and residual sums of squares sum up to the total sum of squares $\sum_{i=1}^{N}(Y_i^{\text{obs}} - \bar{Y}_{\cdot}^{\text{obs}})^2 = (N-1)s_{\text{obs}}^2$. The F statistic

$$F = \frac{\text{SSTre}/(J-1)}{\text{SSRes}/(N-J)} \equiv \frac{\text{MSTre}}{\text{MSRes}} \tag{1}$$

is defined as the ratio of the mean squares of treatment $\text{MSTre} = \text{SSTre}/(J-1)$ to the mean squares of residual $\text{MSRes} = \text{SSRes}/(N-J)$.

The distribution of (1) under H_0(Fisher) can be well approximated by an $F_{J-1,N-J}$ distribution with degrees of freedom $J-1$ and $N-J$, as is often used in the analysis of variance table obtained from fitting a normal linear model. Whereas it is relatively easy to show that (1) follows $F_{J-1,N-J}$ if the observed outcomes follows a normal linear model drawn from a super population, arriving at such a result using a purely randomization-based argument is non-trivial. Below, we state a known result on the approximate randomization distribution of (1), in which we use the notation $A_N \dot{\sim} B_N$ to represent two sequences of random variables $\{A_N\}_{N=1}^{\infty}$ and $\{B_N\}_{N=1}^{\infty}$ that have the same asymptotic distribution as $N \to \infty$. Throughout our discussion, we assume the following regularity conditions required by the finite population central limit theorem for causal inference (Li & Ding, 2017).

Condition 1. As $N \to \infty$, for all j, N_j/N has a positive limit, $\bar{Y}_{\cdot}(j)$ and $S_{\cdot}^2(j)$ have finite limits, and $N^{-1}\max_{1\le i\le N}|Y_i(j) - \bar{Y}_{\cdot}(j)|^2 \to 0$.

THEOREM 1. *Assume* H_0(Fisher). *Over repeated sampling of* $(W_1,\ldots,W_N)$, *the expectations of the residual and treatment sums of squares are* $E(\text{SSTre}) = (J-1)s_{\text{obs}}^2$ *and* $E(\text{SSRes}) = (N-J)s_{\text{obs}}^2$, *and as* $N \to \infty$, *the asymptotic distribution of* (1) *is*

$$F \dot{\sim} \frac{\chi_{J-1}^2/(J-1)}{\{(N-1) - \chi_{J-1}^2\}/(N-J)} \dot{\sim} F_{J-1,N-J}.$$

Remark 1. As $N \to \infty$, both the statistic F and random variable $F_{J-1,N-J}$ are asymptotically $\chi_{J-1}^2/(J-1)$. The original F approximation for randomization inference for a finite population was derived by cumbersome moment matching between statistic (1) and the corresponding $F_{J-1,N-J}$ distribution (Welch, 1937; Pitman, 1938; Kempthorne, 1952). Similar to Silvey (1954), we provide a simpler proof based on the finite population central limit theorem in the Supplementary Material.

Remark 2. Under H_0(Fisher), the total sum of squares is fixed, but its components SSTre and SSRes are random through the treatment assignment $(W_1,\ldots,W_N)$, and their expectations are calculated with respect to the distribution of the treatment assignment. Also, the ratio of expectations of the numerator MSTre and denominator MSRes of (1) is 1 under H_0(Fisher).

4. SAMPLING PROPERTIES OF THE F STATISTIC UNDER NEYMAN'S NULL HYPOTHESIS

In Section 3, we discussed the randomization distribution, i.e., the sampling distribution under H_0(Fisher), of the F statistic in (1). However, the sampling distribution of the F statistic under Neyman's null hypothesis of zero average treatment effect (Neyman, 1923, 1935), i.e.,

$$H_0(\text{Neyman}) : \bar{Y}_\cdot(1) = \cdots = \bar{Y}_\cdot(J),$$

is often of major interest but is under-investigated (Imbens & Rubin, 2015). H_0(Neyman) imposes weaker restrictions on the potential outcomes than H_0(Fisher), making it impossible to compute the exact, or even approximate distribution of the F statistic under H_0(Neyman). However, analytical expressions for E(SSTre) and E(SSRes) can be derived under H_0(Neyman) along the lines of Theorem 1, and can be used to gain insights about the consequences of testing H_0(Neyman) using the Fisher randomization test with the F statistic in (1).

For treatment level $j = 1, \ldots, J$, define $p_j = N_j/N$ as the proportion of the units, and $S_\cdot^2(j) = (N-1)^{-1}\sum_{i=1}^N\{Y_i(j) - \bar{Y}_\cdot(j)\}^2$ as the finite population variances of potential outcomes. Let $\bar{Y}_\cdot(\cdot) = \sum_{j=1}^J p_j\bar{Y}_\cdot(j)$ and $S^2 = \sum_{j=1}^J p_j S_\cdot^2(j)$ be the weighted averages of the finite population means and variances. The sampling distribution of the F statistic in (1) depends crucially on the finite population variance of the unit-level treatment effects

$$S_\cdot^2(j\text{-}j') = (N-1)^{-1}\sum_{i=1}^N\{\tau_i(j,j') - \tau(j,j')\}^2.$$

DEFINITION 1. *The potential outcomes* $\{Y_i(j) : i = 1, \ldots, N,\ j = 1, \ldots, J\}$ *have strictly additive treatment effects if for all* $j \neq j'$, *the unit-level treatment effects* $\tau_i(j,j')$ *are the same for* $i = 1, \ldots, N$, *or equivalently,* $S_\cdot^2(j\text{-}j') = 0$ *for all* $j \neq j'$.

Kempthorne (1955) obtained the following result on the sampling expectations of SSRes and SSTre for balanced designs with $p_j = 1/J$ under the assumption of strict additivity:

$$E(\text{SSRes}) = (N-J)S^2, \quad E(\text{SSTre}) = \frac{N}{J}\sum_{j=1}^J\{\bar{Y}_\cdot(j) - \bar{Y}_\cdot(\cdot)\}^2 + (J-1)S^2. \tag{2}$$

This result implies that with balanced treatment assignments and strict additivity, $E(\text{MSRes} - \text{MSTre}) = 0$ under H_0(Neyman), and provides a heuristic justification for testing H_0(Neyman) using the Fisher randomization test with the F statistic. However, strict additivity combined with H_0(Neyman) implies H_0(Fisher), for which this result is already known by Theorem 1. We will now derive results that do not require the assumption of strict additivity, and thus are more general than those in Kempthorne (1955). For this purpose, we introduce a measure of deviation from additivity. Let

$$\Delta = \sum\sum_{j<j'} p_j p_{j'} S_\cdot^2(j\text{-}j')$$

be a weighted average of the variances of unit-level treatment effects. By Definition 1, $\Delta = 0$ under strict additivity. If strict additivity does not hold, i.e., there is treatment effect heterogeneity (Ding et al., 2016), then $\Delta \neq 0$. Thus Δ is a measure of deviation from additivity and plays a crucial role in the following results on the sampling distribution of the F statistic.

THEOREM 2. *Over repeated sampling of* $(W_1, \ldots, W_N)$*, the expectation of the residual sum of squares is* $E(\text{SSRes}) = \sum_{j=1}^{J}(N_j - 1)S_{\cdot}^2(j)$*, and the expectation of the treatment sum of squares is*

$$E(\text{SSTre}) = \sum_{j=1}^{J} N_j \left\{ \bar{Y}_{\cdot}(j) - \bar{Y}_{\cdot}(\cdot) \right\}^2 + \sum_{j=1}^{J}(1 - p_j)S_{\cdot}^2(j) - \Delta,$$

which reduces to $E(\text{SSTre}) = \sum_{j=1}^{J}(1 - p_j)S_{\cdot}^2(j) - \Delta$ *under* $H_0(\text{Neyman})$.

COROLLARY 1. *Under* $H_0(\text{Neyman})$ *with strict additivity in Definition* 1*, or, equivalently, under* $H_0(\text{Fisher})$*, the above results reduce to* $E(\text{SSRes}) = (N - J)S^2$ *and* $E(\text{SSTre}) = (J - 1)S^2$*, which coincide with Theorem* 1.

COROLLARY 2. *For a balanced design with* $p_j = 1/J$*,*

$$E(\text{SSRes}) = (N - J)S^2, \quad E(\text{SSTre}) = \frac{N}{J}\sum_{j=1}^{J}\{\bar{Y}_{\cdot}(j) - \bar{Y}_{\cdot}(\cdot)\}^2 + (J - 1)S^2 - \Delta.$$

Furthermore, under $H_0(\text{Neyman})$*,* $E(\text{SSRes}) = (N - J)S^2$ *and* $E(\text{SSTre}) = (J - 1)S^2 - \Delta$*, implying that the difference between the mean squares of the residual and the treatment is* $E(\text{MSRes} - \text{MSTre}) = \Delta/(J - 1) \geq 0$.

The result in (2) is a special case of Corollary 2 for $\Delta = 0$. Corollary 2 implies that, for balanced designs, if the assumption of strict additivity does not hold, then testing $H_0(\text{Neyman})$ using the Fisher randomization test with the F statistic may be conservative, in a sense that it may reject a null hypothesis less often than the nominal level. However, for unbalanced designs, the conclusion is not definite, as will be seen from the following result.

COROLLARY 3. *Under* $H_0(\text{Neyman})$*, the difference between the mean squares of the residual and the treatment is*

$$E(\text{MSRes} - \text{MSTre}) = \frac{(N - 1)J}{(J - 1)(N - J)}\sum_{j=1}^{J}(p_j - J^{-1})S_{\cdot}^2(j) + \frac{\Delta}{J - 1}.$$

Corollary 3 shows that the mean square of the residual may be bigger or smaller than that of the treatment, depending on the balance or lack thereof of the experiment and the variances of the potential outcomes. Under $H_0(\text{Neyman})$, when the p_j's and $S_{\cdot}^2(j)$'s are positively associated, the Fisher randomization test using F tends to be conservative; when the p_j's and $S_{\cdot}^2(j)$'s are negatively associated, the Fisher randomization test using F may not control correct type I error.

5. A TEST STATISTIC THAT CONTROLS TYPE I ERROR MORE PRECISELY THAN F

To address the failure of the F statistic to control type I error of the Fisher randomization test under $H_0(\text{Neyman})$ in unbalanced experiments, we propose to use the following X^2 test statistic for the Fisher randomization test. Define $\hat{Q}_j = N_j / s_{\text{obs}}^2(j)$, and define the weighted average of the sample means as $\bar{Y}_w^{\text{obs}} = \sum_{j=1}^{J}\hat{Q}_j\bar{Y}_{\cdot}^{\text{obs}}(j) / \sum_{j=1}^{J}\hat{Q}_j$. Define the X^2 test statistic as

$$X^2 = \sum_{j=1}^{J}\hat{Q}_j\left\{\bar{Y}_{\cdot}^{\text{obs}}(j) - \bar{Y}_w^{\text{obs}}\right\}^2, \tag{3}$$

which can be obtained from weighted least squares. This test statistic has been exploited in the classical analysis of variance literature (e.g., James, 1951; Welch, 1951; Johansen, 1980; Rice & Gaines, 1989; Weerahandi, 1995; Krishnamoorthy et al., 2007) based on the normal linear model with heteroskedasticity, and a similar idea called studentization has been adopted in the permutation test literature (e.g., Neuhaus, 1993; Janssen, 1997, 1999; Janssen & Pauls, 2003; Chung & Romano, 2013; Pauly et al., 2015).

Clearly, replacing the F statistic by the X^2 statistic does not affect the validity of the Fisher randomization test for testing $H_0(\text{Fisher})$, because we always have an exact test for $H_0(\text{Fisher})$ no matter which test statistic we use. Moreover, we derive a new result showing that the Fisher randomization test using X^2 as the test statistic can also control the asymptotic type I error for testing $H_0(\text{Neyman})$. This means that the Fisher randomization test using X^2 as the test statistic can control the type I error under both $H_0(\text{Fisher})$ and $H_0(\text{Neyman})$ asymptotically, making X^2 a more attractive choice than the classical F statistic for conducting the Fisher randomization test. Below, we formally state this new result.

THEOREM 3. *Under $H_0(\text{Fisher})$, the asymptotic distribution of X^2 is χ^2_{J-1} as $N \to \infty$. Under $H_0(\text{Neyman})$, the asymptotic distribution of X^2 is stochastically dominated by χ^2_{J-1}, i.e., for any constant $a > 0$, $\lim_{N\to\infty} \text{pr}(X^2 \geq a) \leq \text{pr}(\chi^2_{J-1} \geq a)$.*

Remark 3. Under $H_0(\text{Fisher})$, the randomization distribution of $\text{SSTre}/s^2_{\text{obs}}$ follows χ^2_{J-1} asymptotically as shown in the Supplementary Material. Under $H_0(\text{Neyman})$, however, the asymptotic distribution of $\text{SSTre}/s^2_{\text{obs}}$ is not χ^2_{J-1}, and the asymptotic distribution of F is not $F_{N-J,J-1}$ as suggested by Corollary 3. Fortunately, if we weight each treatment square by the inverse of the sample variance of the outcomes, the resulting X^2 statistic preserves the asymptotic χ^2_{J-1} randomization distribution under $H_0(\text{Fisher})$, and has an asymptotic distribution that is stochastically dominated by χ^2_{J-1} under $H_0(\text{Neyman})$.

Therefore, under $H_0(\text{Neyman})$, the type I error of the Fisher randomization test using X^2 does not exceed the nominal level. Although we can perform the Fisher randomization test by enumerating or simulating from all possible realizations of the treatment assignment, Theorem 3 suggests that an asymptotic rejection rule against $H_0(\text{Fisher})$ or $H_0(\text{Neyman})$ is $X^2 > x_{1-\alpha}$, the $1-\alpha$ quantile of the χ^2_{J-1} distribution. Because the asymptotic distribution of X^2 under $H_0(\text{Neyman})$ is stochastically dominated by χ^2_{J-1}, its true $1-\alpha$ quantile is asymptotically smaller than $x_{1-\alpha}$, and the corresponding Fisher randomization test is conservative in the sense of having smaller type I error than the nominal level asymptotically.

Remark 4. This asymptotic conservativeness is not particular to our test statistic, but rather a feature of finite population inference (Neyman, 1923; Aronow et al., 2014; Imbens & Rubin, 2015). It distinguishes Theorem 3 from previous results in the permutation test literature (e.g., Chung & Romano, 2013; Pauly et al., 2015), where the conservativeness did not appear and the correlation between the potential outcomes played no role in the theory.

The form of X^2 in (3) suggests its difference from F when the potential outcomes have different variances under different treatment levels. Otherwise we show that they are asymptotically equivalent in the following sense.

COROLLARY 4. *If $S^2_{\cdot}(1) = \cdots = S^2_{\cdot}(J)$, then $(J-1)F \dot{\sim} X^2$.*

Under treatment effect additivity in Definition 1, the condition $S^2_{\cdot}(1) = \cdots = S^2_{\cdot}(J)$ holds, and the equivalence between $(J-1)F$ and X^2 guarantees that the Fisher randomization tests using F and X^2 have the same asymptotic type I error and power. However, Corollary 4 is a large-sample result, and we evaluate it in finite samples in the Supplementary Material.

6. SIMULATION

6·1. *Type I error of the Fisher randomization test using F*

In this subsection, we use simulation to evaluate the finite sample performance of the Fisher randomization test using F under H_0(Neyman). We consider the following three cases, where $\mathcal{N}(\mu, \sigma^2)$ denotes a normal distribution with mean μ and variance σ^2. We choose significance level 0.05 for all tests.

Case 1. For balanced experiments with sample sizes $N = 45$ and $N = 120$, we generate potential outcomes under two cases: (1.1) $Y_i(1) \sim \mathcal{N}(0, 1)$, $Y_i(2) \sim \mathcal{N}(0, 1.2^2)$, $Y_i(3) \sim \mathcal{N}(0, 1.5^2)$; and (1.2) $Y_i(1) \sim \mathcal{N}(0, 1)$, $Y_i(2) \sim \mathcal{N}(0, 2^2)$, $Y_i(3) \sim \mathcal{N}(0, 3^2)$. These potential outcomes are independently generated, and standardized to have zero means.

Case 2. For unbalanced experiments with sample sizes $(N_1, N_2, N_3) = (10, 20, 30)$ and $(N_1, N_2, N_3) = (20, 30, 50)$, we generate potential outcomes under two cases: (2.1) $Y_i(1) \sim \mathcal{N}(0, 1)$, $Y_i(2) = 2Y_i(1)$, $Y_i(3) = 3Y_i(1)$; and (2.2) $Y_i(1) \sim \mathcal{N}(0, 1)$, $Y_i(2) = 3Y_i(1)$, $Y_i(3) = 5Y_i(1)$. These potential outcomes are standardized to have zero means. In this case, $p_1 < p_2 < p_3$ and $S^2_\cdot(1) < S^2_\cdot(2) < S^2_\cdot(3)$.

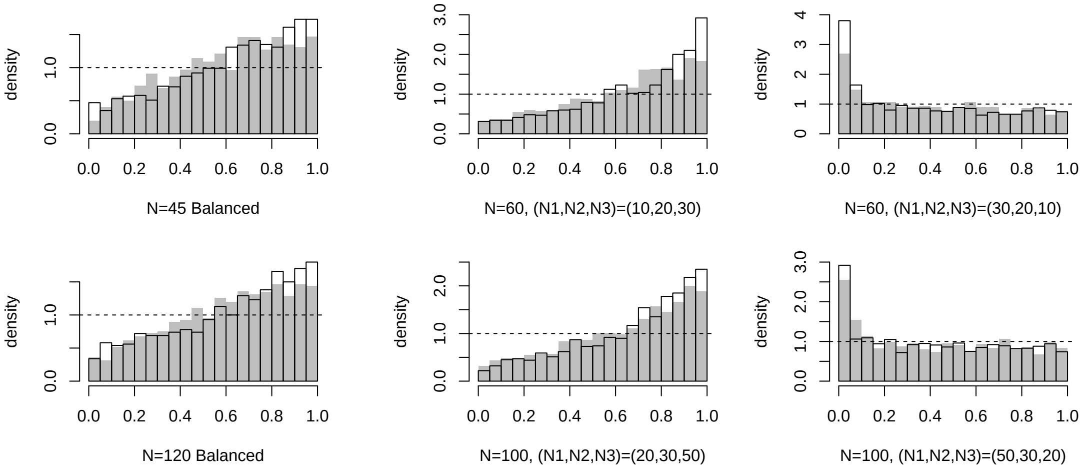
Case 3. For unbalanced experiments with sample sizes $(N_1, N_2, N_3) = (30, 20, 10)$ and $(N_1, N_2, N_3) = (50, 30, 20)$, we generate potential outcomes under two cases: (3.1) $Y_i(1) \sim \mathcal{N}(0, 1)$, $Y_i(2) = 2Y_i(1)$, $Y_i(3) = 3Y_i(1)$; and (3.2) $Y_i(1) \sim \mathcal{N}(0, 1)$, $Y_i(2) = 3Y_i(1)$, $Y_i(3) = 5Y_i(1)$. These potential outcomes are standardized to have zero means. In this case, $p_1 > p_2 > p_3$ and $S^2_\cdot(1) < S^2_\cdot(2) < S^2_\cdot(3)$.

Once generated, the potential outcomes are treated as fixed constants. Over 2000 simulated randomizations, we calculate the observed outcomes, and then perform the Fisher randomization test using F to approximate the p-values by 2000 draws of the treatment assignment. The histograms of the p-values are shown in Figures 1(a)–1(c) corresponding to cases 1–3 above. We also report the rejection rates associated with these cases along with their standard errors in the next few paragraphs.

In Figure 1(a), the Fisher randomization test using F is conservative with p-values distributed towards 1. With larger heterogeneity in the potential outcomes, the histograms of the p-values have larger masses near 1. For case (1.1), the rejection rates are 0.010 and 0.018, and for case (1.2), the rejection rates are 0.023 and 0.016, for sample sizes $N = 45$ and $N = 120$ respectively, with all Monte Carlo standard errors no larger than 0.003.

In Figure 1(b), the sample sizes under each treatment level are increasing in the variances of the potential outcomes. The Fisher randomization test using F is conservative with p-values distributed towards 1. Similar to Figure 1(a), with larger heterogeneity in the potential outcomes, the p-values have larger masses near 1. For case (2.1), the rejection rates are 0.016 and 0.014, and for case (2.2), the rejection rates are 0.015 and 0.011, for sample sizes $N = 45$ and $N = 120$ respectively, with all Monte Carlo standard errors no larger than 0.003.

In Figure 1(c), the sample sizes under different treatment levels are decreasing in the variances of the potential outcomes. For case (3.1), the rejection rates are 0.133 and 0.126, and for case (3.2), the rejection rates are 0.189 and 0.146, for sample sizes $N = 45$ and $N = 120$ respectively, with all Monte Carlo standard errors no larger than 0.009. The Fisher randomization test using F does not preserve correct type I error with p-values distributed towards 0. With larger heterogeneity in the potential outcomes, the p-values have larger masses near 0.



(a) Balanced experiments, case 1 (b) Unbalanced experiments, case 2 (c) Unbalanced experiments, case 3

Fig. 1: Histograms of the p-values under $H_0(\text{Neyman})$ based on the Fisher randomization tests using F, with grey histogram and white histograms for the first and second sub-cases.

These empirical findings agree with our theory in Section 4, i.e., if the sample sizes under different treatment levels are decreasing in the sample variances of the observed outcomes, then the Fisher randomization test using F may not yield correct type I error under $H_0(\text{Neyman})$.

6·2. *Type I error of the Fisher randomization test using* X^2

Figure 2 shows the same simulation as Figure 1, but with test statistic X^2.

Figure 2(a) shows a similar pattern as Figure 1(a). For case (1.1), the rejection rates are 0.016 and 0.012, and for case (1.2), the rejection rates are 0.014 and 0.010, for sample sizes $N = 45$ and $N = 120$ respectively, with all Monte Carlo standard errors no larger than 0.003.

Figure 2(b) shows better performance of the Fisher randomization test using X^2 than Figure 1(b), with p-values closer to uniform. For case (2.1), the rejection rates are 0.032 and 0.038, and for case (2.2), the rejection rates are 0.026 and 0.030, for sample sizes $N = 45$ and $N = 120$ respectively, with all Monte Carlo standard errors no larger than 0.004.

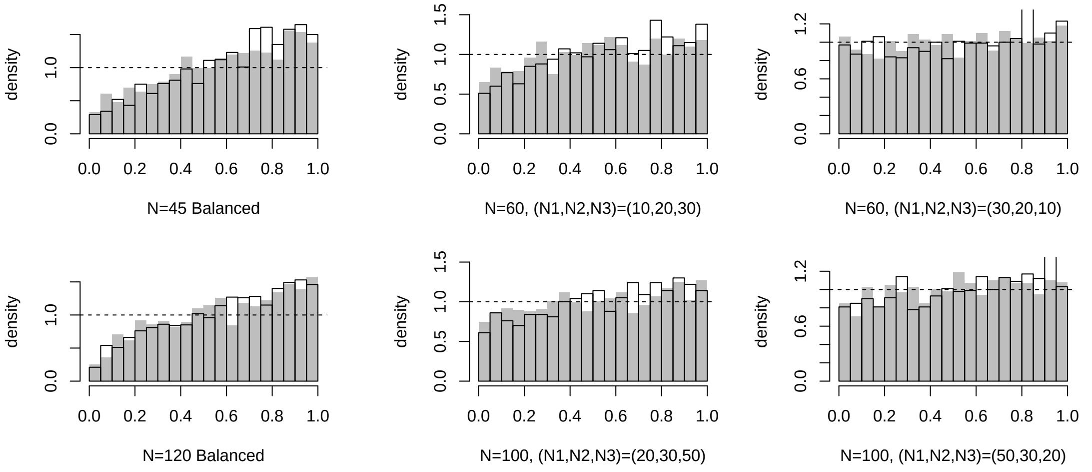
Figure 2(c) shows much better performance of the Fisher randomization test using X^2 than Figure 1(c), because the p-values are much closer to uniform. For case (3.1), the rejection rates are 0.052 and 0.042, and for case (3.2), the rejection rates are 0.048 and 0.040, for sample sizes $N = 45$ and $N = 120$ respectively, with all Monte Carlo standard errors no larger than 0.005. This agrees with our theory that the Fisher randomization test using X^2 can control the asymptotic type I error under $H_0(\text{Neyman})$.

6·3. *Power comparison of the Fisher randomization tests using* F *and* X^2

In this subsection, we compare the powers of the Fisher randomization tests using F and X^2 under alternative hypotheses. We consider the following cases.

Case 4. For balanced experiments with sample sizes $N = 30$ and $N = 45$, we generate potential outcomes from $Y_i(1) \sim \mathcal{N}(0, 1)$, $Y_i(2) \sim \mathcal{N}(0, 2^2)$, $Y_i(3) \sim \mathcal{N}(0, 3^2)$. These potential outcomes are independently generated, and shifted to have means $(0, 1, 2)$.

Case 5. For unbalanced experiments with sample sizes $(N_1, N_2, N_3) = (10, 20, 30)$ and $(N_1, N_2, N_3) = (20, 30, 50)$, we first generate $Y_i(1) \sim \mathcal{N}(0, 1)$ and standardize them to have



(a) Balanced experiments, case 1 (b) Unbalanced experiments, case 2 (c) Unbalanced experiments, case 3

Fig. 2: Histograms of the p-values under H_0(Neyman) based on the Fisher randomization tests using X^2, with grey histogram and white histograms for the first and second sub-cases.

mean zero, and we then generate $Y_i(2) = 3Y_i(1) + 1$ and $Y_i(3) = 5Y_i(1) + 2$. In this case, $p_1 < p_2 < p_3$ and $S^2_\cdot(1) < S^2_\cdot(2) < S^2_\cdot(3)$.

Case 6. For unbalanced experiments with sample sizes $(N_1, N_2, N_3) = (30, 20, 10)$ and $(N_1, N_2, N_3) = (50, 30, 20)$, we generate potential outcomes the same as the above case 5. In this case, $p_1 > p_2 > p_3$ and $S^2_\cdot(1) < S^2_\cdot(2) < S^2_\cdot(3)$.

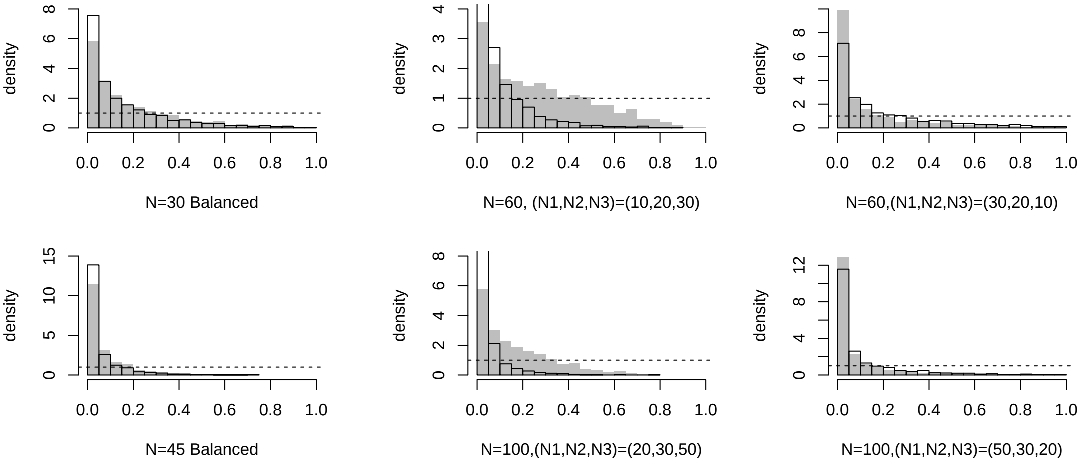
Over 2000 simulated data sets, we perform the Fisher randomization test using F and X^2 and obtain the p-values by 2000 draws of the treatment assignment. The histograms of the p-values, in Figures 3(a)–3(c), correspond to cases 4–6 above. The Monte Carlo standard errors for the rejection rates below are all close but no larger than 0.011.

For case 4, the rejection rates using X^2 and F are 0.290 and 0.376 respectively with sample size $N = 30$, and 0.576 and 0.692 respectively with sample size $N = 45$. For case 5, the powers using X^2 and F are 0.178 and 0.634 respectively with sample size $N = 60$, and 0.288 and 0.794 respectively with sample size $N = 100$. Therefore, when the experiments are balanced or when the sample sizes are positively associated with the variances of the potential outcomes, the Fisher randomization test using F has larger power than that using X^2.

For case 6, the rejection rates using X^2 and F are 0.494 and 0.355 respectively with sample size $N = 60$, and 0.642 and 0.576 respectively with sample size $N = 100$. Therefore, when the sample sizes are negatively associated with the variances of the potential outcomes, the Fisher randomization test using F has smaller power than that using X^2.

6·4. *Simulation studies under other distributions and practical suggestions*

In the Supplementary Material, we give more numerical examples. First, we conduct simulation studies in parallel with §§6·1–6·3 with outcomes generated from exponential distributions. The conclusions are nearly identical to those in §§6·1–6·3, because the finite population central limit theorems holds under mild moment conditions without imposing any distributional assumptions.



(a) Balanced experiments, case 4 (b) Unbalanced experiments, case 5 (c) Unbalanced experiments, case 6

Fig. 3: Histograms of the p-values under alternative hypotheses based on the Fisher randomization tests using F and X^2, with grey histograms for X^2 and white histograms for F.

Second, we use two numerical examples to illustrate the conservativeness issue in Theorem 3. Third, we compare different behaviors of the Fisher randomization tests using F and X^2 in two real-life examples.

7. DISCUSSION

As shown in the proofs of Theorems 1 and 3 in the Supplementary Material, we need to analyze the eigenvalues of the covariance matrix of $\{\bar{Y}_\cdot^{\text{obs}}(1), \ldots, \bar{Y}_\cdot^{\text{obs}}(J)\}$ to obtain the properties of F and X^2 for general $J > 2$. Moreover, we consider the case with $J = 2$ to gain more insights and to make connections with existing literature. For $j \neq j'$, an unbiased estimator for $\tau(j, j')$ is $\hat{\tau}(j, j') = \bar{Y}_\cdot^{\text{obs}}(j) - \bar{Y}_\cdot^{\text{obs}}(j')$, which has sampling variance $\text{var}\{\hat{\tau}(j, j')\} = S_\cdot^2(j)/N_j + S_\cdot^2(j')/N_{j'} - S_\cdot^2(j\text{-}j')/(N_j + N_{j'})$ and an conservative variance estimator $s_{\text{obs}}^2(j)/N_j + s_{\text{obs}}^2(j')/N_{j'}$ (Neyman, 1923).

COROLLARY 5. *When $J = 2$, the F and X^2 statistics reduce to*

$$F \approx \frac{\hat{\tau}^2(1,2)}{s_{\text{obs}}^2(1)/N_2 + s_{\text{obs}}^2(2)/N_1}, \quad X^2 = \frac{\hat{\tau}^2(1,2)}{s_{\text{obs}}^2(1)/N_1 + s_{\text{obs}}^2(2)/N_2},$$

where the approximation of F is due to ignoring the difference between N and $N-2$ and the difference between N_j and $N_j - 1$ ($j = 1, 2$). Under $H_0(\text{Fisher})$, $F \overset{\cdot}{\sim} \chi_1^2$ and $X^2 \overset{\cdot}{\sim} \chi_1^2$. Under $H_0(\text{Neyman})$, $F \overset{\cdot}{\sim} C_1\chi_1^2$ and $X^2 \overset{\cdot}{\sim} C_2\chi_1^2$, where

$$C_1 = \lim_{N \to +\infty} \frac{\text{var}\{\hat{\tau}(1,2)\}}{S_\cdot^2(1)/N_2 + S_\cdot^2(2)/N_1}, \quad C_2 = \lim_{N \to +\infty} \frac{\text{var}\{\hat{\tau}(1,2)\}}{S_\cdot^2(1)/N_1 + S_\cdot^2(2)/N_2} \leq 1. \tag{4}$$

Depending on the sample sizes and the finite population variances, C_1 can be either larger than or smaller than 1. Consequently, using F in the Fisher randomization test can be conservative or anti-conservative under $H_0(\text{Neyman})$. In contrast, C_2 is always no larger than 1, and therefore using X^2 in the Fisher randomization test is conservative for testing $H_0(\text{Neyman})$. Neyman

(1923) proposed to use the square root of X^2 to test $H_0(\text{Neyman})$ based on a normal approximation, which is asymptotically equivalent to the Fisher randomization test using X^2. Both are conservative unless the unit-level treatments are constant.

In practice, for treatment-control experiments, the difference-in-means statistic $\hat{\tau}(1,2)$ was widely used in the Fisher randomization test (Imbens & Rubin, 2015), which, however, can be conservative or anti-conservative for testing $H_0(\text{Neyman})$ as shown in Gail et al. (1996), Lin et al. (2017) and Ding (2017) using numerical examples. We formally state this result below, recognizing the equivalence between $\hat{\tau}(1,2)$ and F in a two-sided test.

COROLLARY 6. *When $J = 2$, the two-sided Fisher randomization test using $\hat{\tau}(1,2)$ is equivalent to using*

$$T^2 = \frac{\hat{\tau}^2(1,2)}{N s^2_{\text{obs}}/(N_1 N_2)} \approx \frac{\hat{\tau}^2(1,2)}{s^2_{\text{obs}}(1)/N_2 + s^2_{\text{obs}}(2)/N_1 + \hat{\tau}^2(1,2)/N},$$

where the approximation is due to ignoring the difference between $(N, N_1 - 1, N_2 - 1)$ and (N, N_1, N_2). Under $H_0(\text{Fisher})$, $T^2 \dot{\sim} F \dot{\sim} \chi^2_1$, and under $H_0(\text{Neyman})$, $T^2 \dot{\sim} F \dot{\sim} C_1\chi^2_1$ with C_1 defined in (4).

Remark 5. Analogously, under the super population model, Romano (1990) showed that the Fisher randomization test using $\hat{\tau}(1,2)$ can be conservative or anti-conservative for testing the hypothesis of equal means of two samples. Janssen (1997, 1999) and Chung & Romano (2013) suggested using the studentized statistic, or equivalently X^2, to remedy the problem of possibly inflated type I error, which is asymptotically exact under the super population model.

After rejecting either $H_0(\text{Fisher})$ or $H_0(\text{Neyman})$, it is often of interest to test pairwise hypotheses, i.e., for $j \neq j'$, $Y_i(j) = Y_i(j')$ for all i, or $\bar{Y}_{\cdot}(j) = \bar{Y}_{\cdot}(j')$. According to Corollaries 5 and 6, we recommend using the Fisher randomization test with test statistic $\hat{\tau}^2(j,j')/\{s^2_{\text{obs}}(j)/N_j + s^2_{\text{obs}}(j')/N_{j'}\}$, which will yield conservative type I error even if the experiment is unbalanced and the variances of the potential outcomes vary across treatment groups.

The analogue between our finite population theory and Chung & Romano (2013)'s super population theory suggests that similar results may also hold for layouts of higher order and other test statistics (Pauly et al., 2015; Chung & Romano, 2016a,b; Friedrich et al., 2017). In more complex experimental designs, often multiple effects are of interest simultaneously, raising the problem of multiple testings (Chung & Romano, 2016b). We leave these to future work.

ACKNOWLEDGEMENT

The research of Peng Ding was partially funded by Institute of Education Sciences, U.S.A. The authors thank Xinran Li, Zhichao Jiang, Lo-Hua Yuan and Robin Gong for suggestions for early versions of the paper. We thank a reviewer and the Associate Editor for helpful comments.

REFERENCES

ANGRIST, J., LANG, D. & OREOPOULOS, P. (2009). Incentives and services for college achievement: Evidence from a randomized trial. *American Economic Journal: Applied Economics* **1**, 136–163.

ARONOW, P. M., GREEN, D. P. & LEE, D. K. (2014). Sharp bounds on the variance in randomized experiments. *The Annals of Statistics* **42**, 850–871.

CHUNG, E. & ROMANO, J. P. (2013). Exact and asymptotically robust permutation tests. *The Annals of Statistics* **41**, 484–507.

CHUNG, E. & ROMANO, J. P. (2016a). Asymptotically valid and exact permutation tests based on two-sample U-statistics. *Journal of Statistical Planning and Inference* **168**, 97–105.

CHUNG, E. & ROMANO, J. P. (2016b). Multivariate and multiple permutation tests. *Journal of Econometrics* **193**, 76–91.

DASGUPTA, T., PILLAI, N. S. & RUBIN, D. B. (2015). Causal inference from 2^K factorial designs using the potential outcomes model. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* **74**, 727–753.

DING, P. (2017). A paradox from randomization-based causal inference (with discussion). *Statistical Science* , in press.

DING, P., FELLER, A. & MIRATRIX, L. (2016). Randomization inference for treatment effect variation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **78**, 655–671.

FISHER, R. A. (1925). *Statistical Methods for Research Workers*. Edinburgh: Oliver & Boyd.

FISHER, R. A. (1935). *The Design of Experiments*. Edinburgh: Oliver & Boyd.

FRIEDRICH, S., BRUNNER, E. & PAULY, M. (2017). Permuting longitudinal data in spite of the dependencies. *Journal of Multivariate Analysis* **153**, 255–265.

GAIL, M. H., MARK, S. D., CARROLL, R. J., GREEN, S. B. & PEE, D. (1996). On design considerations and randomization-based inference for community intervention trials. *Statistics in Medicine* **15**, 1069–1092.

IMBENS, G. W. & RUBIN, D. B. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. New York: Cambridge University Press.

JAMES, G. (1951). The comparison of several groups of observations when the ratios of the population variances are unknown. *Biometrika* **38**, 324–329.

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

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

JANSSEN, A. & PAULS, T. (2003). How do bootstrap and permutation tests work? *Annals of Statistics* **31**, 768–806.

JOHANSEN, S. (1980). The Welch–James approximation to the distribution of the residual sum of squares in a weighted linear regression. *Biometrika* **67**, 85–92.

KEMPTHORNE, O. (1952). *The Design and Analysis of Experiments*. London: Chapman & Hall.

KEMPTHORNE, O. (1955). The randomization theory of experimental inference. *Journal of the American Statistical Association* **50**, 946–967.

KRISHNAMOORTHY, K., LU, F. & MATHEW, T. (2007). A parametric bootstrap approach for ANOVA with unequal variances: Fixed and random models. *Computational Statistics & Data Analysis* **51**, 5731–5742.

LI, X. & DING, P. (2017). General forms of finite population central limit theorems with applications to causal inference. *Journal of the American Statistical Association* , in press.

LIN, W., HALPERN, S. D., PRASAD KERLIN, M. & SMALL, D. S. (2017). A "placement of death" approach for studies of treatment effects on ICU length of stay. *Statistical Methods in Medical Research* **26**, 292–311.

MONTGOMERY, D. C. (2000). *Design and Analysis of Experiments (5th Edition)*. New York: John Wiley & Sons.

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

NEYMAN, J. (1923). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statistical Science* **5**, 465–472.

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

PAULY, M., BRUNNER, E. & KONIETSCHKE, F. (2015). Asymptotic permutation tests in general factorial designs. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **77**, 461–473.

PITMAN, E. J. (1938). Significance tests which may be applied to samples from any populations: III. the analysis of variance test. *Biometrika* **29**, 322–335.

RICE, W. R. & GAINES, S. D. (1989). One-way analysis of variance with unequal variances. *Proceedings of the National Academy of Sciences* **86**, 8183–8184.

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

ROSENBAUM, P. R. (2010). *Design of Observational Studies*. New York: Springer.

RUBIN, D. B. (1980). Comment on "Randomization analysis of experimental data: The Fisher randomization test" by D. Basu. *Journal of the American Statistical Association* **75**, 591–593.

SCHUR, J. (1911). Bemerkungen zur Theorie der beschränkten Bilinearformen mit unendlich vielen Veränderlichen. *Journal für die reine und Angewandte Mathematik* **140**, 1–28.

SILVEY, S. D. (1954). The asymptotic distributions of statistics arising in certain non-parametric tests. *Glasgow Mathematical Journal* **2**, 47–51.

STYAN, G. P. (1973). Hadamard products and multivariate statistical analysis. *Linear Algebra and Its Applications* **6**, 217–240.

WEERAHANDI, S. (1995). ANOVA under unequal error variances. *Biometrics* **51**, 589–599.

WELCH, B. (1937). On the z-test in randomized blocks and Latin squares. *Biometrika* **29**, 21–52.

WELCH, B. (1951). On the comparison of several mean values: An alternative approach. *Biometrika* **38**, 330–336.

Supplementary Materials

§S8 presents the proofs, §S9 contains examples, and §S10 gives additional simulation.

S8. PROOFS

To prove the theorems, we need the following lemmas about completely randomized experiments.

LEMMA S1. *The treatment assignment indicator $W_i(j)$ is a Bernoulli random variable with mean $p_j = N_j/N$ and variance $p_j(1-p_j)$. The covariances of the treatment assignment indicators are*

$$\begin{aligned}
\operatorname{cov}\{W_i(j), W_{i'}(j)\} &= -p_j(1-p_j)/(N-1), && (i \neq i') \\
\operatorname{cov}\{W_i(j), W_i(j')\} &= -p_j p_{j'}, && (j \neq j') \\
\operatorname{cov}\{W_i(j), W_{i'}(j')\} &= p_j p_{j'}/(N-1), && (i \neq i', j \neq j').
\end{aligned}$$

*Proof of Lemma S*1. The proof is straightforward. □

LEMMA S2. *Assume $(c_1, \ldots, c_N)$ and $(d_1, \ldots, d_N)$ are two fixed vectors with means $\bar{c}$ and $\bar{d}$, finite population variances S_c^2 and S_d^2. The finite population covariance is $S_{cd} = (S_c^2 + S_d^2 - S_{c\text{-}d}^2)/2$, where $S_{c\text{-}d}^2$ is the finite population variance of $(c_1 - d_1, \ldots, c_N - d_N)$. For $j \neq j'$,*

$$\operatorname{var}\left\{\frac{1}{N_j}\sum_{i=1}^N W_i(j)c_i\right\} = \frac{1-p_j}{N_j}S_c^2, \quad \operatorname{cov}\left\{\frac{1}{N_j}\sum_{i=1}^N W_i(j)c_i, \frac{1}{N_{j'}}\sum_{i=1}^N W_i(j')d_i\right\} = -\frac{S_{cd}}{N}.$$

*Proof of Lemma S*2. Lemma S2 is known, and its special forms appeared in Kempthorne (1955). We give an elementary proof for completeness. Applying Lemma S1, we have

$$\begin{aligned}
&\operatorname{var}\left\{\frac{1}{N_j}\sum_{i=1}^N W_i(j)c_i\right\} \\
&= \frac{1}{N_j^2}\operatorname{var}\left\{\sum_{i=1}^N W_i(j)(c_i - \bar{c})\right\} \\
&= \frac{1}{N_j^2}\left\{\sum_{i=1}^N \operatorname{var}\{W_i(j)\}(c_i - \bar{c})^2 - \sum\sum_{i \neq i'} \operatorname{cov}\{W_i(j), W_{i'}(j)\}(c_i - \bar{c})(c_{i'} - \bar{c})\right\} \\
&= \frac{1}{N_j^2}\left\{\sum_{i=1}^N p_j(1-p_j)(c_i - \bar{c})^2 - \sum\sum_{i \neq i'} \frac{p_j(1-p_j)}{N-1}(c_i - \bar{c})(c_{i'} - \bar{c})\right\} \\
&= \frac{1}{N_j^2}\left\{p_j(1-p_j)\sum_{i=1}^N (c_i - \bar{c})^2 + \frac{p_j(1-p_j)}{N-1}\sum_{i=1}^N (c_i - \bar{c})^2\right\} \\
&= \frac{1-p_j}{N_j}S_c^2.
\end{aligned}$$

For $j \neq j'$, applying Lemma S1 again, we have

$$
\begin{aligned}
&\operatorname{cov}\left\{ \frac{1}{N_j}\sum_{i=1}^N W_i(j)c_i, \frac{1}{N_j'}\sum_{i=1}^N W_i(j')d_i \right\} \\
&= \frac{1}{N_jN_{j'}}\operatorname{cov}\left\{ \sum_{i=1}^N W_i(j)(c_i-\bar c), \sum_{i=1}^N W_i(j')(d_i-\bar d) \right\} \\
&= \frac{1}{N_jN_{j'}}\left\{ \sum_{i=1}^N \operatorname{cov}\{W_i(j), W_i(j')\}(c_i-\bar c)(d_i-\bar d) \right. \\
&\qquad\qquad \left. + \sum\sum_{i\neq i'} \operatorname{cov}\{W_i(j), W_{i'}(j')\}(c_i-\bar c)(d_{i'}-\bar d) \right\} \\
&= \frac{1}{N_jN_{j'}}\left\{ -\sum_{i=1}^N p_jp_{j'}(c_i-\bar c)(d_i-\bar d) + \sum\sum_{i\neq i'} \frac{p_jp_{j'}}{N-1}(c_i-\bar c)(d_{i'}-\bar d) \right\} \\
&= -\frac{1}{N_jN_{j'}}\left\{ p_jp_{j'}\sum_{i=1}^N (c_i-\bar c)(d_i-\bar d) + \frac{p_jp_{j'}}{N-1}\sum_{i=1}^N (c_i-\bar c)(d_i-\bar d) \right\} \\
&= -S_{cd}/N. \qquad\qquad \square
\end{aligned}
$$

Proof of Theorem 1. Under H_0(Fisher), $\{Y_i^{\text{obs}} : i = 1, \ldots, N\}$ and SSTot $= (N-1)s^2_{\text{obs}}$ are fixed. Because $\{Y_i^{\text{obs}} : W_i(j) = 1\}$ is a simple random sample from the finite population $\{Y_i^{\text{obs}} : i = 1, \ldots, N\}$, the sample mean $\bar Y_\cdot^{\text{obs}}(j)$ is unbiased for the population mean $\bar Y_\cdot^{\text{obs}}$, and the sample variance $s^2_{\text{obs}}(j)$ is unbiased for the population variance s^2_{obs}. Therefore,

$$
E(\text{SSRes}) = \sum_{j=1}^J E\left\{(N_j-1)s^2_{\text{obs}}(j)\right\} = \sum_{j=1}^J (N_j-1)s^2_{\text{obs}} = (N-J)s^2_{\text{obs}},
$$

which further implies that

$$
E(\text{SSTre}) = \text{SSTot} - E(\text{SSRes}) = (N-1)s^2_{\text{obs}} - (N-J)s^2_{\text{obs}} = (J-1)s^2_{\text{obs}}.
$$

Applying Lemma S2, we have

$$
\operatorname{var}\{\bar Y_\cdot^{\text{obs}}(j)\} = \frac{1-p_j}{N_j}s^2_{\text{obs}}, \quad \operatorname{cov}\{\bar Y_\cdot^{\text{obs}}(j), \bar Y_\cdot^{\text{obs}}(j')\} = -\frac{s^2_{\text{obs}}}{N}. \tag{S5}
$$

Therefore, the finite population central limit theorem (Li & Ding, 2017, Theorem 5), coupled with the variance and covariance formulae in (S5), implies

$$
V \equiv \begin{bmatrix} N_1^{1/2}\{\bar Y_\cdot^{\text{obs}}(1) - \bar Y_\cdot^{\text{obs}}\} \\ N_2^{1/2}\{\bar Y_\cdot^{\text{obs}}(2) - \bar Y_\cdot^{\text{obs}}\} \\ \vdots \\ N_J^{1/2}\{\bar Y_\cdot^{\text{obs}}(J) - \bar Y_\cdot^{\text{obs}}\} \end{bmatrix} \mathrel{\dot\sim} \mathcal{N}_J \left[0, s^2_{\text{obs}} \begin{pmatrix} 1-p_1 & -p_1^{1/2}p_2^{1/2} & \cdots & -p_1^{1/2}p_J^{1/2} \\ -p_2^{1/2}p_1^{1/2} & 1-p_2 & \cdots & -p_2^{1/2}p_J^{1/2} \\ \vdots & & & \vdots \\ -p_J^{1/2}p_1^{1/2} & -p_J^{1/2}p_2^{1/2} & \cdots & 1-p_J \end{pmatrix} \right],
$$

where $\mathcal{N}_J$ denotes a J-dimensional normal random vector. The above asymptotic covariance matrix can be simplified as $s^2_{\text{obs}}(I_J - qq^{\mathrm{T}}) \equiv s^2_{\text{obs}}P$, where I_J is the $J \times J$ identity matrix, and $q = (p_1^{1/2}, \ldots, p_J^{1/2})^{\mathrm{T}}$. The matrix P is a projection matrix of rank $J-1$, which is orthogonal to the vector q. Consequently, the treatment sum of squares can be represented as $\text{SSTre} = V^{\mathrm{T}} V \dot\sim \chi^2_{J-1} s^2_{\text{obs}}$, and the F statistic can be represented as

$$
\begin{aligned}
F = \frac{\text{SSTre}/(J-1)}{\{(N-1)s^2_{\text{obs}} - \text{SSTre}\}/(N-J)} &\dot\sim \frac{\chi^2_{J-1} s^2_{\text{obs}}/(J-1)}{\{(N-1)s^2_{\text{obs}} - \chi^2_{J-1} s^2_{\text{obs}}\}/(N-J)} \\
= \frac{\chi^2_{J-1}/(J-1)}{\{(N-1) - \chi^2_{J-1}\}/(N-J)} &\dot\sim F_{J-1,N-J} \dot\sim \chi^2_{J-1}/(J-1). \qquad \square
\end{aligned}
$$

Proof of Theorem 2. First, because $\bar{Y}^{\text{obs}}_{\cdot}(j) = \sum_{i=1}^{N} W_i(j) Y_i(j)/N_j$, Lemma S2 implies that $\bar{Y}^{\text{obs}}_{\cdot}(j)$ has mean $\bar{Y}_{\cdot}(j)$ and variance $(1-p_j)S^2_{\cdot}(j)/N_j$, and

$$
\begin{aligned}
\text{cov}\{\bar{Y}^{\text{obs}}_{\cdot}(j), \bar{Y}^{\text{obs}}_{\cdot}(j')\} &= \text{cov}\left\{ \frac{1}{N_j} \sum_{i=1}^{N} W_i(j) Y_i(j), \frac{1}{N_{j'}} \sum_{i=1}^{N} W_i(j') Y_i(j') \right\} \\
&= -\frac{1}{2N} \{ S^2_{\cdot}(j) + S^2_{\cdot}(j') - S^2_{\cdot}(j\text{-}j') \}.
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
\text{var}(\bar{Y}^{\text{obs}}_{\cdot}) &= \sum_{j=1}^{J} p_j^2 \text{var}\{\bar{Y}^{\text{obs}}_{\cdot}(j)\} + \sum\sum_{j\neq j'} p_j p_{j'} \text{cov}\{\bar{Y}^{\text{obs}}_{\cdot}(j), \bar{Y}^{\text{obs}}_{\cdot}(j')\} \\
&= \sum_{j=1}^{J} p_j^2 \frac{1-p_j}{N_j} S^2_{\cdot}(j) - \sum\sum_{j\neq j'} p_j p_{j'} \frac{1}{2N} \{ S^2_{\cdot}(j) + S^2_{\cdot}(j') - S^2_{\cdot}(j\text{-}j') \} \\
&= \frac{1}{N} \left\{ \sum_{j=1}^{J} p_j(1-p_j) S^2_{\cdot}(j) \right. \\
&\qquad \left. -\frac{1}{2} \sum\sum_{j\neq j'} p_j p_{j'} S^2_{\cdot}(j) - \frac{1}{2} \sum\sum_{j\neq j'} p_j p_{j'} S^2_{\cdot}(j') + \frac{1}{2} \sum\sum_{j\neq j'} p_j p_{j'} S^2_{\cdot}(j\text{-}j') \right\}.
\end{aligned}
$$

Because

$$
\begin{aligned}
\sum\sum_{j\neq j'} p_j p_{j'} S^2_{\cdot}(j) &= \sum_{j=1}^{J} p_j(1-p_j) S^2_{\cdot}(j), \\
\sum\sum_{j\neq j'} p_j p_{j'} S^2_{\cdot}(j') &= \sum_{j=1}^{J} p_{j'}(1-p_{j'}) S^2_{\cdot}(j') = \sum_{j=1}^{J} p_j(1-p_j) S^2_{\cdot}(j),
\end{aligned}
$$

the variance of $\bar{Y}^{\text{obs}}_{\cdot}$ reduces to

$$
\text{var}(\bar{Y}^{\text{obs}}_{\cdot}) = (2N)^{-1} \sum\sum_{j\neq j'} p_j p_{j'} S^2_{\cdot}(j\text{-}j') = \Delta/N.
$$

Second,

$$
\begin{aligned}
\text{cov}\{\bar{Y}^{\text{obs}}_{\cdot}(j), \bar{Y}^{\text{obs}}_{\cdot}\} &= p_j \text{var}\{\bar{Y}^{\text{obs}}_{\cdot}(j)\} + \sum_{j'\neq j} p_{j'} \text{cov}\{\bar{Y}^{\text{obs}}_{\cdot}(j), \bar{Y}^{\text{obs}}_{\cdot}(j')\} \\
&= \frac{1}{N}(1-p_j) S^2_{\cdot}(j) - \frac{1}{2N} \sum_{j'\neq j} p_{j'} \{ S^2_{\cdot}(j) + S^2_{\cdot}(j') - S^2_{\cdot}(j\text{-}j') \}.
\end{aligned}
$$

We further define $\sum_{j'\neq j} p_{j'} S_{\cdot}^2(j\text{-}j') = \Delta_j$. Because

$$\sum_{j'\neq j} p_{j'} S_{\cdot}^2(j) = (1-p_j)S_{\cdot}^2(j), \quad \sum_{j'\neq j} p_{j'} S_{\cdot}^2(j') = S^2 - p_j S_{\cdot}^2(j),$$

the covariance between $\bar{Y}_{\cdot}^{\text{obs}}(j)$ and $\bar{Y}_{\cdot}^{\text{obs}}$ reduces to

$$\begin{aligned}\text{cov}\{\bar{Y}_{\cdot}^{\text{obs}}(j), \bar{Y}_{\cdot}^{\text{obs}}\} &= (2N)^{-1}\left\{2(1-p_j)S_{\cdot}^2(j) - (1-p_j)S_{\cdot}^2(j) - S^2 + p_j S_{\cdot}^2(j) + \Delta_j\right\} \\ &= (2N)^{-1}\left\{S_{\cdot}^2(j) - S^2 + \Delta_j\right\}.\end{aligned}$$

Third, $\bar{Y}_{\cdot}^{\text{obs}}(j) - \bar{Y}_{\cdot}^{\text{obs}}$ has mean $\bar{Y}_{\cdot}(j) - \sum_{j=1}^J p_j \bar{Y}_{\cdot}(j)$ and variance

$$\begin{aligned}\text{var}\{\bar{Y}_{\cdot}^{\text{obs}}(j) - \bar{Y}_{\cdot}^{\text{obs}}\} &= \text{var}\{\bar{Y}_{\cdot}^{\text{obs}}(j)\} + \text{var}(\bar{Y}_{\cdot}^{\text{obs}}) - 2\text{cov}\{\bar{Y}_{\cdot}^{\text{obs}}(j), \bar{Y}_{\cdot}^{\text{obs}}\} \\ &= \frac{1}{N}\left\{\frac{1-p_j}{p_j}S_{\cdot}^2(j) + \Delta - S_{\cdot}^2(j) + S^2 - \Delta_j\right\}.\end{aligned}$$

Finally, the expectation of the treatment sum of squares is

$$\begin{aligned}E(\text{SSTre}) &= E\left[\sum_{j=1}^J N_j\{\bar{Y}_{\cdot}^{\text{obs}}(j) - \bar{Y}_{\cdot}^{\text{obs}}\}^2\right] \\ &= \sum_{j=1}^J N_j\left\{\bar{Y}_{\cdot}(j) - \sum_{j=1}^J p_j\bar{Y}_{\cdot}(j)\right\}^2 + \sum_{j=1}^J p_j\left\{\frac{1-p_j}{p_j}S_{\cdot}^2(j) + \Delta - S_{\cdot}^2(j) + S^2 - \Delta_j\right\},\end{aligned}$$

which follows from the mean and variance formulas of $\bar{Y}_{\cdot}^{\text{obs}}(j) - \bar{Y}_{\cdot}^{\text{obs}}$. Some algebra gives

$$\begin{aligned}E(\text{SSTre}) &= \sum_{j=1}^J N_j\left\{\bar{Y}_{\cdot}(j) - \sum_{j=1}^J p_j\bar{Y}_{\cdot}(j)\right\}^2 + \sum_{j=1}^J(1-p_j)S_{\cdot}^2(j) + \Delta - S^2 + S^2 - 2\Delta \\ &= \sum_{j=1}^J N_j\left\{\bar{Y}_{\cdot}(j) - \sum_{j=1}^J p_j\bar{Y}_{\cdot}(j)\right\}^2 + \sum_{j=1}^J(1-p_j)S_{\cdot}^2(j) - \Delta.\end{aligned}$$

Under H_0(Neyman), i.e., $\bar{Y}_{\cdot}(1) = \cdots = \bar{Y}_{\cdot}(J)$, or, equivalently, $\bar{Y}_{\cdot}(j) - \sum_{j=1}^J p_j\bar{Y}_{\cdot}(j) = 0$ for all j, the expectation of the treatment sum of squares further reduces to

$$E(\text{SSTre}) = \sum_{j=1}^J(1-p_j)S_{\cdot}^2(j) - \Delta.$$

Because $\{Y_i^{\text{obs}} : W_i(j) = 1\}$ is a simple random sample from $\{Y_i(j) : i = 1, 2, \ldots, N\}$, the sample variance is unbiased for the population variance, i.e., $E\{s_{\text{obs}}^2(j)\} = S_{\cdot}^2(j)$. Therefore, the mean of the residual sum of squares is

$$E(\text{SSRes}) = E\left\{(N_j - 1)s_{\text{obs}}^2(j)\right\} = \sum_{j=1}^J(N_j - 1)S_{\cdot}^2(j).$$

This completes the proof. □

Proof of Corollary 1. Additivity implies $S^2 = S_{\cdot}^2(j)$ for all j and $\Delta = 0$, and the conclusions follow. □

Proof of Corollary 2. For balanced designs, $p_j = 1/J, N_j = N/J$ and $S^2 = \sum_{j=1}^J S^2_{\cdot}(j)/J$, and therefore Theorem 2 implies

$$E(\text{SSRes}) = \frac{N-J}{J}\sum_{j=1}^{J} S^2_{\cdot}(j) = (N-J)S^2,$$

$$E(\text{SSTre}) = \frac{N}{J}\sum_{j=1}^{J}\{\bar{Y}_{\cdot}(j) - \bar{Y}_{\cdot}(\cdot)\}^2 + (J-1)S^2 - \Delta.$$

Moreover, under H_0(Neyman), $E(\text{SSRes})$ is unchanged, and $E(\text{SSTre}) = (J-1)S^2 - \Delta$. Therefore, the expectation of the mean treatment squares is no larger than the expectation of the mean residual squares, because $E(\text{MSRes}) - E(\text{MSTre}) = \Delta/(J-1) \geq 0$. □

Proof of Corollary 3. Under H_0(Neyman),

$$\begin{aligned}
E(\text{MSRes}) - E(\text{MSTre}) &= \sum_{j=1}^{J}\left(\frac{N_j - 1}{N-J} - \frac{1-p_j}{J-1}\right) S^2_{\cdot}(j) + \frac{\Delta}{J-1} \\
&= \frac{(N-1)J}{(J-1)(N-J)}\sum_{j=1}^{J}(p_j - J^{-1})S^2_{\cdot}(j) + \frac{\Delta}{J-1}.
\end{aligned}$$
□

To prove Theorem 3, we need the following two lemmas: the first is about the quadratic form of the multivariate normal distribution, and the second, due to Schur (1911), provides an upper bound for the largest eigenvalue of the element-wise product of two matrices. The proof of the first follows from straightforward linear algebra, and the proof of the second can be found in Styan (1973, Corollary 3). Below we use $A * B$ to denote the element-wise product of A and B, i.e, the (i,j)-th element of $A * B$ is the product of the (i,j)-th elements of A and B, $A_{ij}B_{ij}$.

LEMMA S3. *If $X \sim \mathcal{N}_J(0, A)$, then $X^{\mathrm{T}}BX \sim \sum_{j=1}^J \lambda_j \xi_j$, where the ξ_j's are iid χ^2_1, and the λ_j's are eigenvalues of BA.*

LEMMA S4. *If A is positive semidefinite and B is a correlation matrix, then the maximum eigenvalue of $A * B$ does not exceed the maximum eigenvalue of A.*

Proof of Theorem 3. We first prove the result under H_0(Neyman), and then view the result under H_0(Fisher) as a special case.

Let $Q_j = N_j/S^2_{\cdot}(j)$ for $j = 1, \ldots, J$, and $Q = \sum_{j=1}^J Q_j$ be their sum. Define $q_w^{\mathrm{T}} = (Q_1^{1/2}, \ldots, Q_J^{1/2})/Q^{1/2}$, and $P_w = I_J - q_w q_w^{\mathrm{T}}$ is a projection matrix of rank $J-1$. Let $\bar{Y}^{\text{obs}}_{w0} = Q^{-1}\sum_{j=1}^J Q_j \bar{Y}^{\text{obs}}_{\cdot}(j)$ be a weighted average of the means of the observed outcomes. According to Li & Ding (2017, Proposition 3), $s^2_{\text{obs}}(j) - S^2_{\cdot}(j) \to 0$ in probability $(j = 1, \ldots, J)$. By Slutsky's Theorem, X^2 has the same asymptotic distribution as

$$X_0^2 = \sum_{j=1}^{J} Q_j\left\{\bar{Y}^{\text{obs}}_{\cdot}(j) - \bar{Y}^{\text{obs}}_{w0}\right\}^2.$$

Define ρ_{jk} as the finite population correlation coefficient of potential outcomes $\{Y_i(j)\}_{i=1}^N$ and $\{Y_i(k)\}_{i=1}^N$, and R as the corresponding correlation matrix with (j,k)-th element ρ_{jk}. The finite population central limit theorem (Li & Ding, 2017, Theorem 5) implies

$$V_0 \equiv \begin{bmatrix} Q_1^{1/2}\{\bar{Y}^{\mathrm{obs}}_{\cdot}(1) - \bar{Y}_{\cdot}(1)\} \\ Q_2^{1/2}\{\bar{Y}^{\mathrm{obs}}_{\cdot}(2) - \bar{Y}_{\cdot}(2)\} \\ \vdots \\ Q_J^{1/2}\{\bar{Y}^{\mathrm{obs}}_{\cdot}(J) - \bar{Y}_{\cdot}(J)\} \end{bmatrix}$$

$$\dot{\sim} \mathcal{N}_J \left[0, \begin{pmatrix} 1-p_1 & -p_1^{1/2}p_2^{1/2}\rho_{12} & \cdots & -p_1^{1/2}p_J^{1/2}\rho_{1J} \\ -p_2^{1/2}p_1^{1/2}\rho_{21} & 1-p_2 & \cdots & -p_2^{1/2}p_J^{1/2}\rho_{2J} \\ \vdots & & & \vdots \\ -p_J^{1/2}p_1^{1/2}\rho_{J1} & -p_J^{1/2}p_2^{1/2}\rho_{J2} & \cdots & 1-p_J \end{pmatrix} = P * R \right],$$

where $P = I_J - qq^{\mathrm{T}}$ is the projection matrix defined in the proof of Theorem 1. In the above, the mean and covariance matrix of the random vector V_0 follow directly from Lemmas S1 and S2.

Under H_0(Neyman) with $\bar{Y}_{\cdot}(1) = \cdots = \bar{Y}_{\cdot}(J)$, we can verify that

$$X_0^2 = \sum_{j=1}^{J} Q_j\{\bar{Y}^{\mathrm{obs}}_{\cdot}(j) - \bar{Y}_{\cdot}(j)\}^2 - \frac{1}{Q}\left[\sum_{j=1}^{J} Q_j\{\bar{Y}^{\mathrm{obs}}_{\cdot}(j) - \bar{Y}_{\cdot}(j)\}\right]^2,$$

which can be further rewritten as a quadratic form (cf. Chung & Romano, 2013)

$$X_0^2 = V_0^{\mathrm{T}}(I_J - q_w q_w^{\mathrm{T}})V_0 \equiv V_0^{\mathrm{T}} P_w V_0.$$

According to Lemma S3, X_0^2 has asymptotic distribution $\sum_{j=1}^{J-1}\lambda_j\xi_j$, where the λ_j's are the $J-1$ nonzero eigenvalues of $P_w(P*R)$. The summation is from $j=1$ to $J-1$ because $P_w(P*R)$ has rank at most $J-1$. The eigenvalues $(\lambda_1, \ldots, \lambda_{J-1})$ are all smaller than or equal to the largest eigenvalue of $P*R$, because P_w is a projection matrix. According to Lemma S4, the maximum eigenvalue of the element-wise product $P*R$ is no larger than the maximum eigenvalue of P, which is 1. Therefore, $X_0^2 \dot{\sim} \sum_{j=1}^{J-1}\lambda_j\xi_j$, where $\lambda_j \leq 1$ for all j. Because the χ^2_{J-1} can be represented as $\xi_1 + \cdots + \xi_{J-1}$, it is clear that the asymptotic distribution of X_0^2 is stochastically dominated by χ^2_{J-1}.

When performing the Fisher randomization test, we treat all observed outcomes as fixed, and consequently, the randomization distribution is essentially the repeated sampling distribution of X^2 under $Y_i(1) = \cdots = Y_i(J) = Y_i^{\mathrm{obs}}$. This restricts $S^2_{\cdot}(j)$ to be constant, and the correlation coefficients between potential outcomes to be 1. Correspondingly, $P_w = P, R = 1_J 1_J^{\mathrm{T}}$, and the asymptotic covariance matrix of V_0 is P. Applying Lemma S3 again, we know that the asymptotic randomization distribution of X^2 is χ^2_{J-1}, because $PP = P$ has $J-1$ nonzero eigenvalues and all of them are 1.

Mathematically, the randomization distribution under H_0(Fisher) is the same as the permutation distribution. Therefore, applying Chung & Romano (2013) yields the same result for X^2 under H_0(Fisher). □

Proof of Corollary 4. As shown in the proof of Theorem 3, X^2 is asymptotically equivalent to X_0^2, and therefore we need only to show the equivalence between $(J-1)F$ and X_0^2. If $S^2_{\cdot}(1) = \cdots = S^2_{\cdot}(J) = S^2$, then $\bar{Y}^{\mathrm{obs}}_{w0} = \bar{Y}^{\mathrm{obs}}_{\cdot}$, and

$$X_0^2 = \frac{\sum_{j=1}^{J}\{\bar{Y}^{\mathrm{obs}}_{\cdot}(j) - \bar{Y}^{\mathrm{obs}}_{\cdot}\}^2}{S^2} = \frac{\mathrm{SSTre}}{S^2}.$$

Because MSRes $= \sum_{j=1}^{J}(N_j-1)s^2_{\text{obs}}(j)/(N-J)$ converges to S^2 in probability (Li & Ding, 2017, Proposition 3), Slutsky's Theorem implies

$$(J-1)F = \frac{\text{SSTre}}{\text{MSRes}} \dot{\sim} \frac{\text{SSTre}}{S^2}.$$

Therefore, $(J-1)F \dot{\sim} X_0^2 \dot{\sim} X^2$. □

Proof of Corollary 5. First, we discuss F. Because $\bar{Y}_{\cdot}^{\text{obs}} = p_1\bar{Y}_{\cdot}^{\text{obs}}(1) + p_2\bar{Y}_{\cdot}^{\text{obs}}(2)$, we have

$$\bar{Y}_{\cdot}^{\text{obs}}(1) - \bar{Y}_{\cdot}^{\text{obs}} = p_2\hat{\tau}(1,2), \qquad \bar{Y}_{\cdot}^{\text{obs}}(2) - \bar{Y}_{\cdot}^{\text{obs}} = -p_1\hat{\tau}(1,2).$$

The treatment sum of squares reduces to

$$\text{SSTre} = N_1\left\{\bar{Y}_{\cdot}^{\text{obs}}(1) - \bar{Y}_{\cdot}^{\text{obs}}\right\}^2 + N_2\left\{\bar{Y}_{\cdot}^{\text{obs}}(2) - \bar{Y}_{\cdot}^{\text{obs}}\right\}^2 = Np_1p_2\hat{\tau}^2(1,2),$$

and the residual sum of squares reduces to SSRes $= (N_1-1)s^2_{\text{obs}}(1) + (N_2-1)s^2_{\text{obs}}(2)$. Therefore, the F statistic reduces to

$$F = \frac{\text{SSTre}}{\text{SSRes}/(N-2)} = \frac{\hat{\tau}^2(1,2)}{\frac{N(N_1-1)}{(N-2)N_1N_2}s^2_{\text{obs}}(1) + \frac{N(N_2-1)}{(N-2)N_1N_2}s^2_{\text{obs}}(2)} \approx \frac{\hat{\tau}^2(1,2)}{s^2_{\text{obs}}(1)/N_2 + s^2_{\text{obs}}(2)/N_1},$$

where the approximation follows from ignoring the difference between N and $N-2$ and the difference between N_j and N_j-1 $(j=1,2)$. Following from Theorem 1 or proving it directly, we know that $F \dot{\sim} F_{1,N-2} \dot{\sim} \chi^2_1$ under H_0(Fisher). However, under H_0(Neyman), Neyman (1923), coupled with the finite population central limit theorem (Li & Ding, 2017, Theorem 5), imply

$$\frac{\hat{\tau}(1,2)}{\left\{\frac{S_{\cdot}^2(1)}{N_1} + \frac{S_{\cdot}^2(2)}{N_2} - \frac{S_{\cdot}^2(1\text{-}2)}{N}\right\}^{1/2}} \dot{\sim} \mathcal{N}(0,1),$$

and $s^2_{\text{obs}}(j) \to S_{\cdot}^2(j)$ in probability $(j=1,2)$. Therefore, the asymptotic distribution of F under H_0(Neyman) is $F \dot{\sim} C_1\chi^2_1$, where

$$C_1 = \lim_{N\to+\infty}\frac{S_{\cdot}^2(1)/N_1 + S_{\cdot}^2(2)/N_2 - S_{\cdot}^2(1\text{-}2)/N}{S_{\cdot}^2(1)/N_2 + S_{\cdot}^2(2)/N_1}.$$

Second, we discuss X^2. Because

$$\bar{Y}_w^{\text{obs}} = \left\{\frac{N_1}{s^2_{\text{obs}}(1)}\bar{Y}_{\cdot}^{\text{obs}}(1) + \frac{N_2}{s^2_{\text{obs}}(2)}\bar{Y}_{\cdot}^{\text{obs}}(2)\right\}\Big/\left\{\frac{N_1}{s^2_{\text{obs}}(1)} + \frac{N_2}{s^2_{\text{obs}}(2)}\right\},$$

we have

$$\begin{aligned}
\bar{Y}_{\cdot}^{\text{obs}}(1) - \bar{Y}_w^{\text{obs}} &= \frac{N_2}{s^2_{\text{obs}}(2)}\hat{\tau}^2(1,2)\Big/\left\{\frac{N_1}{s^2_{\text{obs}}(1)} + \frac{N_2}{s^2_{\text{obs}}(2)}\right\},\\
\bar{Y}_{\cdot}^{\text{obs}}(2) - \bar{Y}_w^{\text{obs}} &= -\frac{N_1}{s^2_{\text{obs}}(1)}\hat{\tau}^2(1,2)\Big/\left\{\frac{N_1}{s^2_{\text{obs}}(1)} + \frac{N_2}{s^2_{\text{obs}}(2)}\right\}.
\end{aligned}$$

Therefore, the X^2 statistic reduces to

$$\begin{aligned}
X^2 &= \left\{\frac{N_1}{s^2_{\text{obs}}(1)}\frac{N_2^2}{s^4_{\text{obs}}(2)}\hat{\tau}^2(1,2) + \frac{N_2}{s^2_{\text{obs}}(2)}\frac{N_1^2}{s^4_{\text{obs}}(1)}\hat{\tau}^2(1,2)\right\}\Big/\left\{\frac{N_1}{s^2_{\text{obs}}(1)} + \frac{N_2}{s^2_{\text{obs}}(2)}\right\}^2\\
&= \frac{\hat{\tau}^2(1,2)}{s^2_{\text{obs}}(1)/N_1 + s^2_{\text{obs}}(2)/N_2}.
\end{aligned}$$

Following from Theorem 3 or proving it directly, we know that $X^2 \dot\sim \chi^2_1$ under H_0(Fisher). However, under H_0(Neyman), we can use an argument similar to that for F and obtain $X^2 \dot\sim C_2\chi^2_1$, where

$$C_2 = \lim_{N\to+\infty} \frac{S^2_\cdot(1)/N_1 + S^2_\cdot(2)/N_2 - S^2_\cdot(1\text{-}2)/N}{S^2_\cdot(1)/N_1 + S^2_\cdot(2)/N_2} \leq 1.$$

The constant C_2 is smaller than or equal to 1 with equality holding if the limit of $S^2_\cdot(1\text{-}2)$ is zero, i.e., the unit-level treatment effects are constant asymptotically. □

Proof of Corollary 6. In the Fisher randomization test, s^2_{obs} is fixed, and therefore using $\hat\tau(1,2)$ is equivalent to using T^2. Using simple algebra similar to Ding (2017), we have the following decomposition

$$(N-1)s^2_{\text{obs}} = (N_1-1)s^2_{\text{obs}}(1) + (N_2-1)s^2_{\text{obs}}(2) + N_1N_2\hat\tau(1,2)/N,$$

which implies the equivalent formula of T^2 in Corollary 6. Under H_0(Fisher) or H_0(Neyman), $\hat\tau(1,2) \to 0$ in probability, which coupled with Slutsky's Theorem, implies the asymptotic equivalence $T^2 \dot\sim F$. □

S9. NUMERICAL EXAMPLES

*Example S*1. We consider $J = 3$, sample sizes $N_1 = 120, N_2 = 80$ and $N_3 = 40$. We generate the first set of potential outcomes from

$$Y_i(1) \sim \mathcal{N}(0,1), Y_i(2) = 3Y_i(1), Y_i(3) = 5Y_i(1), \tag{S6}$$

and the second set of potential outcomes from

$$Y_i(1) \sim \mathcal{N}(0,1), Y_i(2) \sim \mathcal{N}(0,3^2), Y_i(3) \sim \mathcal{N}(0,5^2). \tag{S7}$$

After generating the potential outcomes, we center the $Y_i(j)$'s by subtracting the mean to make $\bar Y_\cdot(j) = 0$ for all j so that H_0(Neyman) holds. Figure S4 shows the distributions of X^2 over repeated sampling of the treatment assignment vector $(W_1,\dots,W_N)$ for potential outcomes generated from (S6) and (S7). The true sampling distributions under both cases are stochastically dominated by χ^2_2. Under (S6), the correlation coefficients between the potential outcomes are 1; whereas under (S7), the correlation coefficients are 0. With less correlated potential outcomes, the gap between the true distribution and χ^2_2 becomes larger.

*Example S*2. We use an example from Montgomery (2000, Exercise 3.15) with 4 treatment levels. The sample variances and the sample sizes differ for the four treatment levels, as shown in Table S1. The p-values of the Fisher randomization test using F and X^2 are 0.003 and 0.010, respectively. If we choose a stringent size, say $\alpha = 0.01$, then the evidence against the null is strong from the first test, but the evidence is weak from the second test. If our interest is H_0(Neyman), then the different strength of evidence may be due to the different variances and sample sizes of the treatment groups. Because of this, we recommend making decision based on the Fisher randomization test using X^2.

*Example S*3. We reanalyze the data from Angrist et al. (2009), which contain a control group and 3 treatment groups designed to improve academic performance among college freshmen. Table S2 summaries the sample sizes, means and variances of the final grades under 4 treatment groups. The p-values of the Fisher randomization test using F and X^2 are 0.058 and

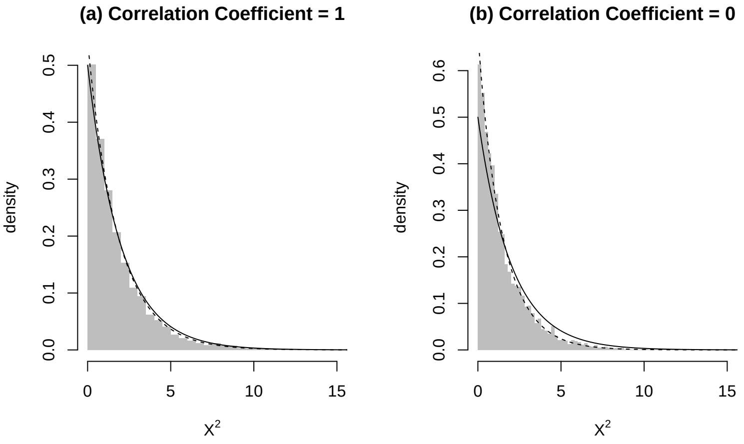


Fig. S4: Distributions of X^2. The histograms are the sampling distributions, the dotted lines are the asymptotic distributions, and the solid lines are the χ^2_2 distribution.

Table S1: A randomized experiment with $J = 4$

	1	2	3	4
observed outcome	58.2	56.3	50.1	52.9
	57.2	54.5	54.2	49.9
	58.4	57.0	55.4	50.0
	55.8	55.3		51.7
	54.9			
sample size	5	4	3	4
mean	56.9	55.8	53.2	51.1
variance	2.3	1.2	7.7	2.1

Table S2: A randomized experiment with $J = 4$, where control, sfp, ssp and sfsp denote the four treatment groups.

	control	sfp	ssp	sfsp
sample size	854	219	212	119
mean	63.86	65.83	64.13	66.10
variance	144.97	124.45	159.76	114.33

0.045, respectively. The Fisher randomization tests using F and X^2 give different conclusions at the commonly used significance level of 0.05. In this unbalanced experiment, the Fisher randomization test using F is less powerful.

S10. MORE SIMULATION WITH NONNORMAL OUTCOMES

S10·1. *Type I error of the Fisher randomization test using F*

In this subsection, we use simulation to evaluate the finite sample performance of the Fisher randomization test using F under H_0(Neyman). We consider the following three cases, where $\mathcal{E}$ denotes an exponential distribution with mean 1.

Case S1. For balanced experiments with sample sizes $N = 45$ and $N = 120$, we generate potential outcomes under two cases: (S1.1) $Y_i(1) \sim \mathcal{E}$, $Y_i(2) \sim \mathcal{E}/0.7$, $Y_i(3) \sim \mathcal{E}/0.5$; and (S1.2) $Y_i(1) \sim \mathcal{E}$, $Y_i(2) \sim \mathcal{E}/0.5$, $Y_i(3) \sim \mathcal{E}/0.3$. These potential outcomes are independently generated, and standardized to have zero means.

Case S2. For unbalanced experiments with sample sizes $(N_1, N_2, N_3) = (10, 20, 30)$ and $(N_1, N_2, N_3) = (20, 30, 50)$, we generate potential outcomes under two cases: (S2.1) $Y_i(1) \sim \mathcal{E}$, $Y_i(2) = 2Y_i(1)$, $Y_i(3) = 3Y_i(1)$; and (S2.2) $Y_i(1) \sim \mathcal{E}$, $Y_i(2) = 3Y_i(1)$, $Y_i(3) = 5Y_i(1)$. These potential outcomes are standardized to have zero means. In this case, $p_1 < p_2 < p_3$ and $S^2_{\cdot}(1) < S^2_{\cdot}(2) < S^2_{\cdot}(3)$.

Case S3. For unbalanced experiments with sample sizes $(N_1, N_2, N_3) = (30, 20, 10)$ and $(N_1, N_2, N_3) = (50, 30, 20)$, we generate potential outcomes under two cases: (S3.1) $Y_i(1) \sim \mathcal{E}$, $Y_i(2) = 1.2Y_i(1)$, $Y_i(3) = 1.5Y_i(1)$; and (S3.2) $Y_i(1) \sim \mathcal{E}$, $Y_i(2) = 1.5Y_i(1)$, $Y_i(3) = 2Y_i(1)$. These potential outcomes are standardized to have zero means. In this case, $p_1 > p_2 > p_3$ and $S^2_{\cdot}(1) < S^2_{\cdot}(2) < S^2_{\cdot}(3)$.

We follow §6·1 and obtain the same conclusions about the Fisher randomization test using F, because Figures 1 and S5 exhibit the same pattern.

In Figure 5(a), for case (S1.1), the rejection rates are 0.022 and 0.014, and for case (S1.2), the rejection rates are 0.030 and 0.030, for sample sizes $N = 45$ and $N = 120$ respectively. In Figure 5(b), for case (S2.1), the rejection rates are 0.018 and 0.024, and for case (2.2), the rejection rates are 0.026 and 0.018, for sample sizes $N = 45$ and $N = 120$ respectively. The Monte Carlo standard errors are all close to but no larger than 0.003.

In Figure 5(c), for case (S3.1), the rejection rates are 0.076 and 0.086, and for case (S3.2), the rejection rates are 0.108 and 0.109, for sample sizes $N = 45$ and $N = 120$ respectively, with all Monte Carlo standard errors no larger than 0.008. In these two cases, the Fisher randomization test using F does not preserve correct type I error.

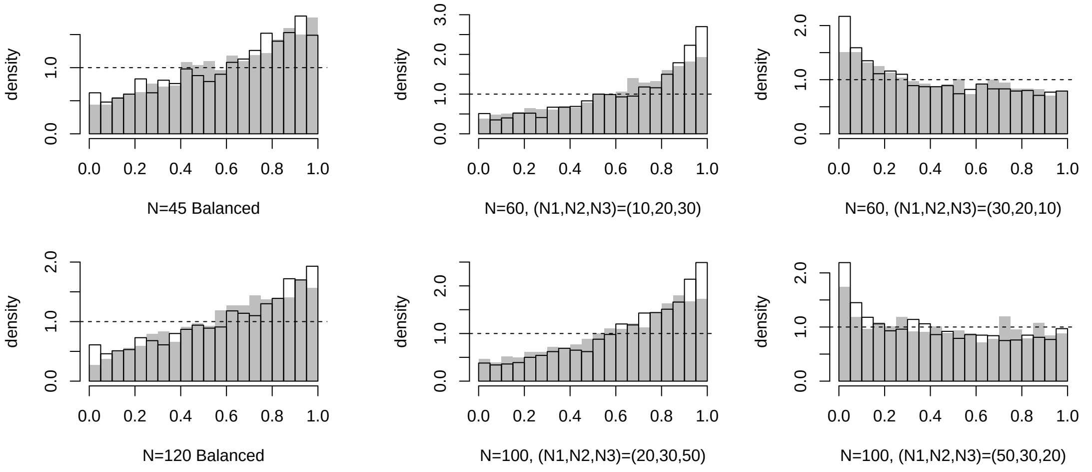
S10·2. *Type I error of the Fisher randomization test using X^2*

We follow §6·2, generate the same data as §S10·1, and obtain the same conclusions about the Fisher randomization test using X^2, because Figures 2 and S6 exhibit the same pattern. All the Monte Carlo standard errors of the rejection rates below are close but no larger than 0.005.

In Figure 6(a), for case (S1.1), the rejection rates are 0.034 and 0.018, and for case (S1.2), the rejection rates are 0.048 and 0.029, for sample sizes $N = 45$ and $N = 120$ respectively. In Figure 6(b), for case (S2.1), the rejection rates are 0.032 and 0.035, and for case (S2.2), the rejection rates are 0.025 and 0.036, for sample sizes $N = 45$ and $N = 120$ respectively. In Figure 6(c), for case (S3.1), the rejection rates are 0.060 and 0.062, and for case (S3.2), the rejection rates are 0.054 and 0.044, for sample sizes $N = 45$ and $N = 120$ respectively. This, coupled with Figure S5, agrees with our theory that the Fisher randomization test using X^2 can control type I error under H_0(Neyman) better than using F.

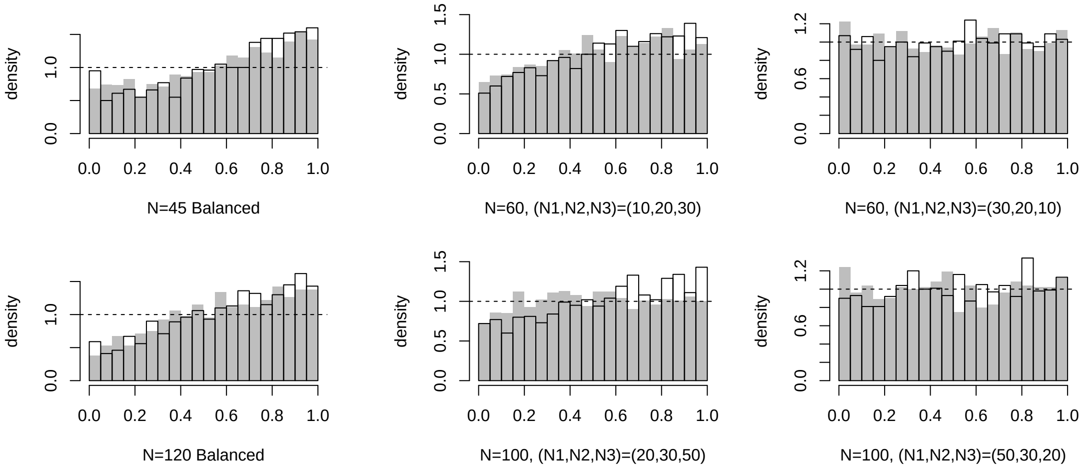
S10·3. *Power comparison of the Fisher randomization tests using F and X^2*

We follow §6·3 to compare the powers of the Fisher randomization tests using F and X^2. We consider the following cases and summarize the results in Figure S7.



(a) Balanced experiments, case S1 (b) Unbalanced experiments, case S2 (c) Unbalanced experiments, case S3

Fig. S5: Histograms of the p-values under $H_0(\text{Neyman})$ based on the Fisher randomization tests using X^2, with grey histogram and white histograms for the first and second sub-cases.

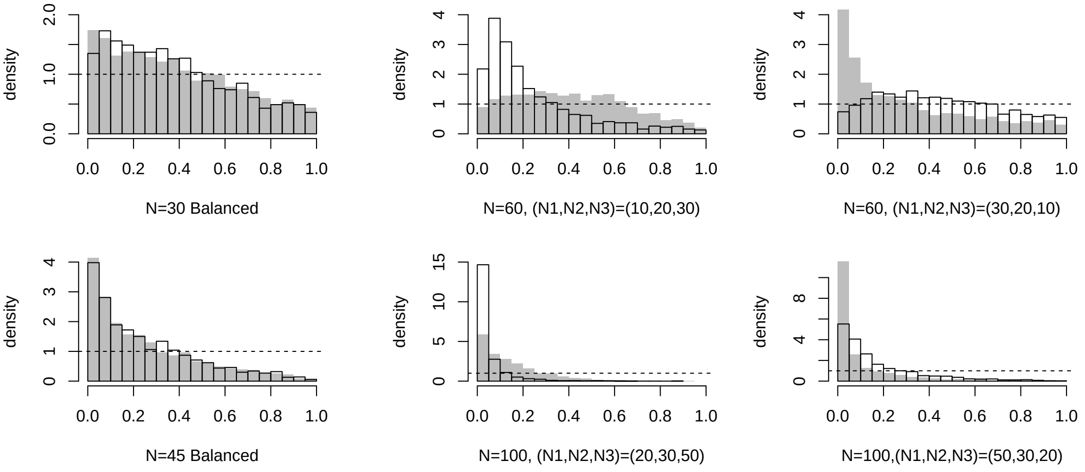


(a) Balanced experiments, case S1 (b) Unbalanced experiments, case S2. (c) Unbalanced experiments, case S3

Fig. S6: Histograms of the p-values under $H_0(\text{Neyman})$ based on the Fisher randomization tests using X^2, with grey histogram and white histograms for the first and second sub-cases.

Case S4. For balanced experiments with sample sizes $N = 30$ and $N = 45$, we generate potential outcomes from $Y_i(1) \sim \mathcal{E}$, $Y_i(2) \sim \mathcal{E}/0.7$, $Y_i(3) \sim \mathcal{E}/0.5$. These potential outcomes are independently generated, and shifted to have means $(0, 0.5, 1)$.

Case S5. For unbalanced experiments with sample sizes $(N_1, N_2, N_3) = (10, 20, 30)$ and $(N_1, N_2, N_3) = (20, 30, 50)$, we first generate $Y_i(1) \sim \mathcal{E}$ and standardize them to have mean zero, and we then generate $Y_i(2) = 3Y_i(1) + 1$ and $Y_i(3) = 5Y_i(1) + 2$. In this case, $p_1 < p_2 < p_3$ and $S_\cdot^2(1) < S_\cdot^2(2) < S_\cdot^2(3)$.



(a) Balanced experiments, case S4 (b) Unbalanced experiments, case S5 (c) Unbalanced experiments, case S6

Fig. S7: Histograms of the p-values under alternative hypotheses based on the Fisher randomization tests using F and X^2, with grey histograms for X^2 and white histograms for F.

Case S6. For unbalanced experiments with sample sizes $(N_1, N_2, N_3) = (30, 20, 10)$ and $(N_1, N_2, N_3) = (50, 30, 20)$, we generate potential outcomes the same as the above case S5. In this case, $p_1 > p_2 > p_3$ and $S_{\cdot}^2(1) < S_{\cdot}^2(2) < S_{\cdot}^2(3)$.

When the sample sizes are positively associated with the variances of the potential outcomes, the Fisher randomization test using F has larger power than that using X^2. However, when the treatment groups are balanced or when the sample sizes are negatively associated with the variances of the potential outcomes, the Fisher randomization test using F has smaller power than that using X^2. We report the rejection rates below with all the Monte Carlo standard errors no larger than 0.01.

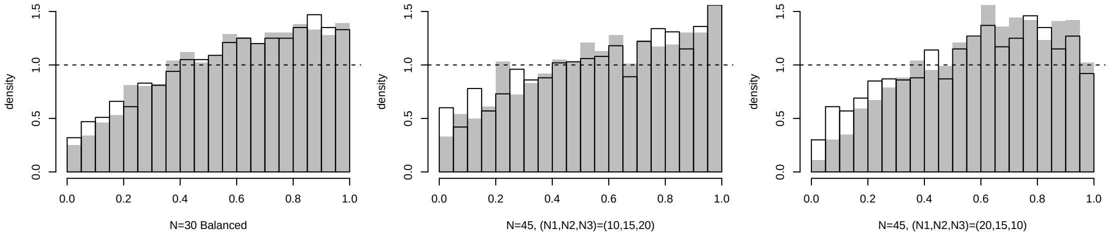
For case S4, the rejection rates using X^2 and F are 0.087 and 0.066 with sample size $N = 30$, and 0.207 and 0.198 with sample size $N = 45$. For case S5, the powers using X^2 and F are 0.044 and 0.106 with sample size $N = 60$, and 0.293 and 0.729 with sample size $N = 100$. For case S6, the rejection rates using X^2 and F are 0.211 and 0.037 with sample size $N = 60$, and 0.578 and 0.274 with sample size $N = 100$.

S10·4. *Finite sample evaluation of Corollary* 4 *with skewed outcomes*

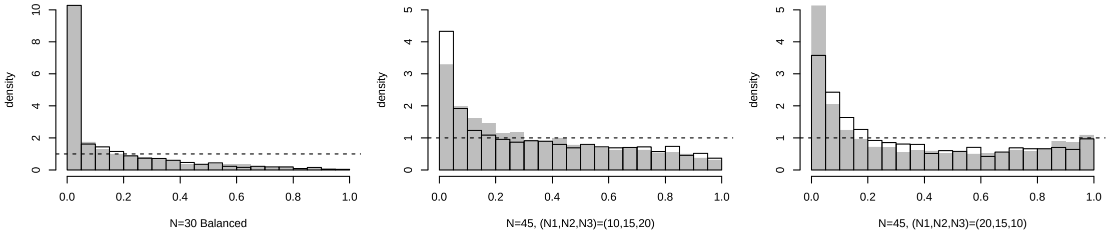
We first generate log-normal potential outcomes $Y_i(1) \sim \exp\{\mathcal{N}(0,1)\}$, $Y_i(2) \sim \exp\{\mathcal{N}(1,1)\}$, and $Y_i(3) \sim \exp\{\mathcal{N}(2,1)\}$, and then standard them to have equal finite population means 0 and variances 1.

Under H_0(Neyman), the p-values of the Fisher randomization test using F and X^2 are shown in Figure S8(a). With sample size $(N_1, N_2, N_3) = (10, 10, 10)$, the rejection rates using X^2 and F are 0.012 and 0.016; with sample size $(10, 15, 20)$, the rejection rates are 0.016 and 0.028; with sample size $(20, 15, 10)$, the rejection rates are 0.006 and 0.015. The Monte Carlo standard errors are all close to but no larger than 0.004.

Under alternative hypotheses, the p-values of the Fisher randomization test using F and X^2 are shown in Figure S8(b). With sample size $(N_1, N_2, N_3) = (10, 10, 10)$, we shift the potential outcomes by constants $(0, 0.5, 1)$, and the rejection rates using X^2 and F are 0.514 and 0.512;



(a) $H_0(\text{Neyman})$ holds



(b) $H_0(\text{Neyman})$ does not hold

Fig. S8: Histograms of the p-values under equal finite population variances based on the Fisher randomization tests using F and X^2, with grey histograms for X^2 and white histograms for F.

with sample size $(10, 15, 20)$, we shift the potential outcomes by constants $(0, 0.2, 0.5)$, and the rejection rates are 0.164 and 0.215; with sample size $(20, 15, 10)$, we shift the potential outcomes by constants $(0, 0.2, 0.5)$, and the rejection rates are 0.256 and 0.179. The Monte Carlo standard errors are all close but no larger than 0.011.

In finite samples, we observe moderate difference between the Fisher randomization tests using X^2 and F even with homoskedastic potential outcomes, although Corollary 4 ensures their asymptotic equivalence.