Biometrika (2018), **105**, 1, *pp.* 45–56
Printed in Great Britain
doi: 10.1093/biomet/asx059
Advance Access publication 31 October 2017

A randomization-based perspective on 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 effect are considered in a finite-population setting associated with completely randomized experiments involving more than two treatments. The consequences of using the F statistic to conduct such a test are examined, and we argue 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. We propose to use an alternative test statistic, derive its asymptotic distributions under Fisher's and Neyman's null hypotheses, and demonstrate its advantages through simulations.

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

1. INTRODUCTION

One-way analysis of variance (Fisher, 1925; Scheffe, 1959) is perhaps the most commonly used tool to analyse 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 superpopulation model with identically and independently distributed normal error terms, or by using the asymptotic randomization distribution of the F statistic. Units in real-life experiments are rarely random samples from a superpopulation, making a finite-population randomization-based perspective on inference important (e.g., Rosenbaum, 2010; Dasgupta et al., 2015; Imbens & Rubin, 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 any assumptions about the underlying outcome distribution.

In causal inference from a finite population, two 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 when there is no 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 sometimes 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 that is robust with respect 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; Rubin, 1974) 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. For treatment level $j = 1,\ldots,J$, define $p_j = N_j/N$ as the proportion of the units and $S^2_\cdot(j) = (N-1)^{-1}\sum_{i=1}^{N}\{Y_i(j) - \bar{Y}_\cdot(j)\}^2$ as the finite-population variance of the potential outcomes.

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 $W_i = \sum_{j=1}^{J} jW_i(j)$, 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, and 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 outcome of unit i is a deterministic function of the treatment it has received and the potential outcomes, given by $Y_i^{\mathrm{obs}} = \sum_{j=1}^{J} W_i(j)Y_i(j)$.

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:

$$H_{0\mathrm{F}} : Y_i(1) = \cdots = Y_i(J) \quad (i = 1,\ldots,N).$$

Under $H_{0\mathrm{F}}$, all J potential outcomes $Y_i(1),\ldots,Y_i(J)$ equal 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 that under $H_{0\mathrm{F}}$ 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 the 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 the 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 $H_{0\mathrm{F}}$. 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 define $\bar{Y}_{\cdot}^{\mathrm{obs}} = N^{-1}\sum_{i=1}^{N} Y_i^{\mathrm{obs}}$ as the sample average of all the observed outcomes. Let $s^2_{\mathrm{obs}}(j) = (N_j-1)^{-1}\sum_{i=1}^{N} W_i(j)\{Y_i^{\mathrm{obs}} - \bar{Y}_{\cdot}^{\mathrm{obs}}(j)\}^2$ and $s^2_{\mathrm{obs}} = (N-1)^{-1}\sum_{i=1}^{N}(Y_i^{\mathrm{obs}} - \bar{Y}_{\cdot}^{\mathrm{obs}})^2$ be the corresponding sample variances with divisors $N_j - 1$ and $N-1$, respectively. Let

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

be the treatment sum of squares, and let

$$\mathrm{SS_R} = \sum_{j=1}^{J}\sum_{i:\,W_i(j)=1}\{Y_i^{\mathrm{obs}} - \bar{Y}_{\cdot}^{\mathrm{obs}}(j)\}^2 = \sum_{j=1}^{J}(N_j-1)s^2_{\mathrm{obs}}(j)$$

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

$$F = \frac{\mathrm{SS_T}/(J-1)}{\mathrm{SS_R}/(N-J)} \equiv \frac{\mathrm{MS_T}}{\mathrm{MS_R}} \tag{1}$$

is defined as the ratio of the treatment mean square $\mathrm{MS_T} = \mathrm{SS_T}/(J-1)$ to the residual mean square $\mathrm{MS_R} = \mathrm{SS_R}/(N-J)$.

The distribution of (1) under $H_{0\mathrm{F}}$ 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. Although it is relatively easy to show that (1) follows $F_{J-1,N-J}$ if the observed outcomes follow a normal linear model drawn from a superpopulation, arriving at such a result via a purely randomization-based argument is nontrivial. Below, we state a known result on the approximate randomization distribution of (1), and 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, $\bar{Y}_{\cdot}(j)$ has a finite limit, $S^2_{\cdot}(j)$ has a finite and positive limit, and $N^{-1}\max_{1\leqslant i\leqslant N}|Y_i(j) - \bar{Y}_{\cdot}(j)|^2 \to 0$.

THEOREM 1. *Assume $H_{0\mathrm{F}}$. Over repeated sampling of $(W_1,\ldots,W_N)$, the expectations of the residual and treatment sums of squares are $E(\mathrm{SS_T}) = (J-1)s^2_{\mathrm{obs}}$ and $E(\mathrm{SS_R}) = (N-J)s^2_{\mathrm{obs}}$, and as $N\to\infty$, the asymptotic distribution of* (1) *is*

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

Remark 1. In Theorem 1 and the following discussion, 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 $N\to\infty$. The original F approximation for randomization inference for a finite population was derived by cumbersome moment matching between the statistic (1) and the corresponding $F_{J-1,N-J}$ distribution (Welch, 1937; Pitman, 1938; Kempthorne, 1952). In

the Supplementary Material, we give a simpler proof based on the finite-population central limit theorem, similar to Silvey (1954).

Remark 2. Under H_{0F}, the total sum of squares is fixed, but its components SS_T and SS_R 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 the expectations of the numerator MS_T and the denominator MS_R of (1) is 1 under H_{0F}.

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

In § 3 we discussed the randomization distribution, i.e., the sampling distribution under H_{0F}, of the F statistic in (1). However, the sampling distribution of the F statistic under Neyman's null hypothesis of no average treatment effect (Neyman, 1923, 1935),

$$H_{0N} : \bar{Y}_\cdot(1) = \cdots = \bar{Y}_\cdot(J),$$

is often of interest but has received limited attention (Imbens & Rubin, 2015). This hypothesis imposes weaker restrictions on the potential outcomes than H_{0F}, making it impossible to compute the corresponding exact, or even approximate, distribution of F. However, analytical expressions for $E(\text{SS}_\text{T})$ and $E(\text{SS}_\text{R})$ can be derived under H_{0N} along the lines of Theorem 1, and can be used to gain insights into the consequences of testing H_{0N} using the Fisher randomization test with F.

Let $\bar{Y}_\cdot(\cdot) = \sum_{j=1}^J p_j \bar{Y}_\cdot(j)$ and $S^2 = \sum_{j=1}^J p_j S^2_\cdot(j)$ be the weighted averages of the finite-population means and variances. The sampling distribution of F depends crucially on the finite-population variance of the unit-level treatment effects,

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

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, if* $S^2_\tau(j,j') = 0$ *for all* $j \neq j'$.

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

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

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

$$\Delta = \sum_{j<j'} \sum p_j p_{j'} S^2_\tau(j,j')$$

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., if there is treatment effect heterogeneity, then $\Delta \neq 0$. Thus Δ is a measure of the 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* $E(\mathrm{SS_R}) = \sum_{j=1}^{J}(N_j - 1)S^2_{\cdot}(j)$*, and the expectation of the treatment sum of squares is*

$$E(\mathrm{SS_T}) = \sum_{j=1}^{J} N_j \{\bar{Y}_{\cdot}(j) - \bar{Y}_{\cdot}(\cdot)\}^2 + \sum_{j=1}^{J}(1 - p_j)S^2_{\cdot}(j) - \Delta,$$

which reduces to $E(\mathrm{SS_T}) = \sum_{j=1}^{J}(1 - p_j)S^2_{\cdot}(j) - \Delta$ *under* $H_{0\mathrm{N}}$.

COROLLARY 1. *Under* $H_{0\mathrm{N}}$ *with strict additivity in Definition 1 or, equivalently, under* $H_{0\mathrm{F}}$*, the results in Theorem 2 reduce to* $E(\mathrm{SS_R}) = (N - J)S^2$ *and* $E(\mathrm{SS_T}) = (J - 1)S^2$*, which coincide with the statements in Theorem 1.*

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

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

Furthermore, under $H_{0\mathrm{N}}$*,* $E(\mathrm{SS_R}) = (N - J)S^2$ *and* $E(\mathrm{SS_T}) = (J - 1)S^2 - \Delta$*, implying that the difference between the residual mean square and treatment mean square is* $E(\mathrm{MS_R} - \mathrm{MS_T}) = \Delta/(J - 1) \geqslant 0$.

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 $H_{0\mathrm{N}}$ using the Fisher randomization test with the F statistic may be conservative, in the sense that it may reject a null hypothesis less often than the nominal level. However, for unbalanced designs, the conclusion is not definite, as can be seen from the following corollary.

COROLLARY 3. *Under* $H_{0\mathrm{N}}$*, the difference between the residual and treatment mean square is*

$$E(\mathrm{MS_R} - \mathrm{MS_T}) = \frac{(N - 1)J}{(J - 1)(N - J)}\sum_{j=1}^{J}(p_j - J^{-1})S^2_{\cdot}(j) + \frac{\Delta}{J - 1}.$$

Corollary 3 shows that the residual mean square may be larger or smaller than that of the treatment, depending on the balance or lack thereof in the experiment and the variances of the potential outcomes. Under $H_{0\mathrm{N}}$, when the p_j and $S^2_{\cdot}(j)$ are positively associated, the Fisher randomization test using F tends to be conservative; when the p_j and $S^2_{\cdot}(j)$ 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 $H_{0\mathrm{N}}$ in unbalanced experiments, we propose to use (3) for the Fisher randomization test. Let $\hat{Q}_j = N_j/s^2_{\mathrm{obs}}(j)$, 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

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

This test statistic has been exploited in classical analysis of variance (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 permutation tests (e.g., Neuhaus, 1993; Janssen, 1997, 1999; Janssen & Pauls, 2003; Chung & Romano, 2013; Pauly et al., 2015).

Replacing F with (3) does not affect the validity of the Fisher randomization test for testing $H_{0\text{F}}$, because we always have an exact test for $H_{0\text{F}}$ no matter which test statistic we use. We show below that the Fisher randomization test using X^2 can also control the asymptotic Type I error for testing $H_{0\text{N}}$, so the Fisher randomization test using X^2 can control the Type I error under both $H_{0\text{F}}$ and $H_{0\text{N}}$ asymptotically, making X^2 a more attractive choice than the classical F statistic for the Fisher randomization test.

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

Remark 3. Under $H_{0\text{F}}$, the randomization distribution of $\text{SS}_\text{T}/s^2_{\text{obs}}$ follows χ^2_{J-1} asymptotically, as shown in the Supplementary Material. Under $H_{0\text{N}}$, however, the asymptotic distribution of $\text{SS}_\text{T}/s^2_{\text{obs}}$ 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 $H_{0\text{F}}$ and has an asymptotic distribution that is stochastically dominated by χ^2_{J-1} under $H_{0\text{N}}$.

Therefore, under $H_{0\text{N}}$, 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 $H_{0\text{F}}$ or $H_{0\text{N}}$ 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 $H_{0\text{N}}$ 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. The asymptotic conservativeness described above is not particular to our test statistic, but rather a feature of the finite-population inference (Neyman, 1923; Aronow et al., 2014; Imbens & Rubin, 2015). It distinguishes Theorem 3 from previous results on permutation tests (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 (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 strict additivity in Definition 1, the condition $S_{\cdot}^2(1) = \cdots = S_{\cdot}^2(J)$ 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; 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 $H_{0\mathrm{N}}$. 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 settings: (1A) $Y_i(1) \sim \mathcal{N}(0,1)$, $Y_i(2) \sim \mathcal{N}(0, 1{\cdot}2^2)$ and $Y_i(3) \sim \mathcal{N}(0, 1{\cdot}5^2)$; (1B) $Y_i(1) \sim \mathcal{N}(0,1)$, $Y_i(2) \sim \mathcal{N}(0, 2^2)$ and $Y_i(3) \sim \mathcal{N}(0, 3^2)$. These potential outcomes are independently generated and standardized to have zero mean.

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 settings: (2A) $Y_i(1) \sim \mathcal{N}(0,1)$, $Y_i(2) = 2Y_i(1)$ and $Y_i(3) = 3Y_i(1)$; (2B) $Y_i(1) \sim \mathcal{N}(0,1)$, $Y_i(2) = 3Y_i(1)$ and $Y_i(3) = 5Y_i(1)$. These potential outcomes are standardized to have zero mean. In this case, $p_1 < p_2 < p_3$ and $S_{\cdot}^2(1) < S_{\cdot}^2(2) < S_{\cdot}^2(3)$.

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 settings: (3A) $Y_i(1) \sim \mathcal{N}(0,1)$, $Y_i(2) = 2Y_i(1)$ and $Y_i(3) = 3Y_i(1)$; (3B) $Y_i(1) \sim \mathcal{N}(0,1)$, $Y_i(2) = 3Y_i(1)$ and $Y_i(3) = 5Y_i(1)$. These potential outcomes are standardized to have zero mean. In this case, $p_1 > p_2 > p_3$ and $S_{\cdot}^2(1) < S_{\cdot}^2(2) < S_{\cdot}^2(3)$.

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 Figs. 1(a)–(c) corresponding to cases 1–3 above. In the next few paragraphs we report the rejection rates associated with these cases along with their standard errors.

In Fig. 1(a), the Fisher randomization test using F is conservative with p-values distributed towards 1. With greater heterogeneity in the potential outcomes, the histograms of the p-values have larger masses near 1. For case (1A) the rejection rates are 0·010 and 0·018, and for case (1B) 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 greater than 0·003.

In Fig. 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 Fig. 1(a), when there is greater heterogeneity in the potential outcomes, the p-values have larger masses near 1. For case (2A) the rejection rates are 0·016 and 0·014, and for case (2B) 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 greater than 0·003.

In Fig. 1(c), the sample sizes under different treatment levels are decreasing in the variances of the potential outcomes. For case (3A) the rejection rates are 0·133 and 0·126, and for case (3B) 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 greater than 0·009. The Fisher randomization test using F does

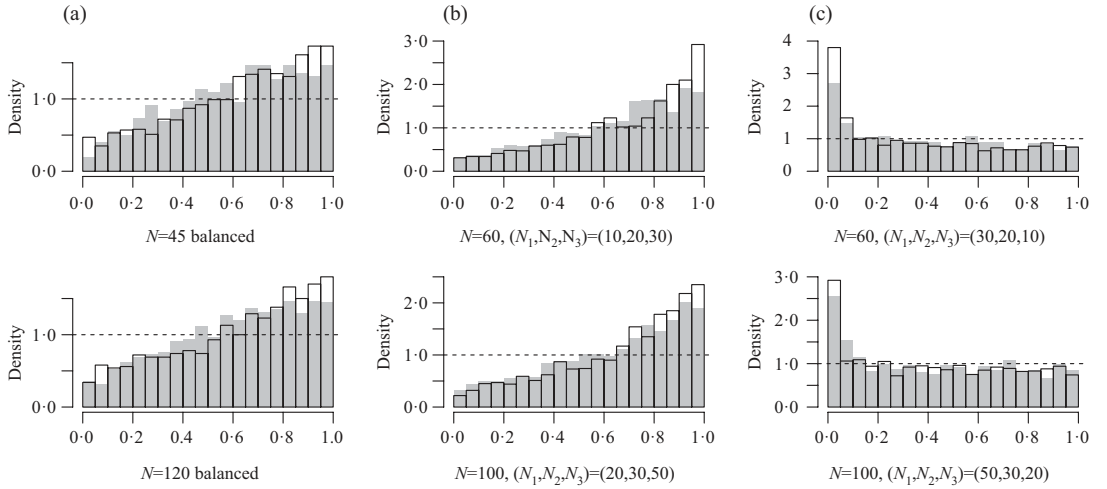


Fig. 1. Histograms of the p-values under H_{0N} based on the Fisher randomization tests using F: (a) balanced experiments, case 1; (b) unbalanced experiments, case 2; (c) unbalanced experiments, case 3. Grey and white histograms correspond to the subcases A and B, respectively.

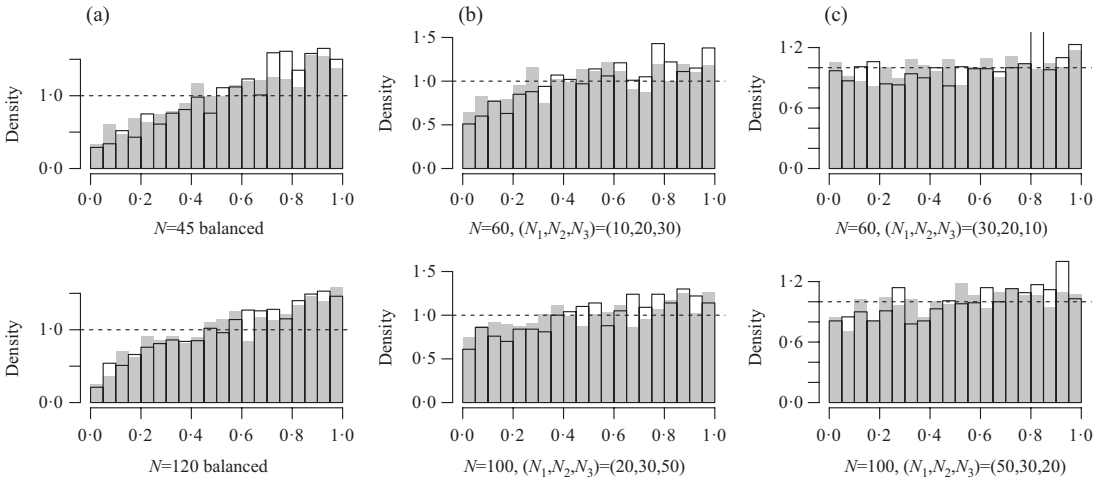


Fig. 2. Histograms of the p-values under H_{0N} based on the Fisher randomization tests using X^2: (a) balanced experiments, case 1; (b) unbalanced experiments, case 2; (c) unbalanced experiments, case 3. Grey and white histograms correspond to the subcases A and B, respectively.

not preserve correct Type I error, with p-values distributed towards 0. With greater heterogeneity in the potential outcomes, the p-values have larger masses near 0.

These empirical findings agree with our theory in § 4; that is, 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_{0N}.

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

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

Figure 2(a) is similar to Fig. 1(a). For case (1A) the rejection rates are 0·016 and 0·012, and for case (1B) 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 greater than 0·003.

Figure 2(b) shows better performance of the Fisher randomization test using X^2 than in Fig. 1(b), with p-values closer to uniform. For case (2A) the rejection rates are 0·032 and 0·038, and for case (2B) 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 greater than 0·004.

Figure 2(c) shows much better performance of the Fisher randomization test using X^2 than in Fig. 1(c), because the p-values are much closer to uniform. For case (3A) the rejection rates are 0·052 and 0·042, and for case (3B) 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 greater 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_{0N}.

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 $Y_i(1) \sim \mathcal{N}(0, 1)$, $Y_i(2) \sim \mathcal{N}(0, 2^2)$ and $Y_i(3) \sim \mathcal{N}(0, 3^2)$. These potential outcomes are independently generated and transformed 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 mean zero; 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 in the same way as in case 5 above. 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 datasets, 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, shown in Figs. 3(a)–(c), correspond to cases 4–6 above. The Monte Carlo standard errors for the rejection rates are all close to but no greater than 0·011.

For case 4, the rejection rates using X^2 and F are respectively 0·290 and 0·376 with sample size $N = 30$, and 0·576 and 0·692 with sample size $N = 45$. For case 5, the powers using X^2 and F are respectively 0·178 and 0·634 with sample size $N = 60$, and 0·288 and 0·794 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 higher power than that using X^2.

For case 6, the rejection rates using X^2 and F are respectively 0·494 and 0·355 with sample size $N = 60$, and 0·642 and 0·576 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 lower power than that using X^2.

6·4. *Simulation studies under other distributions and applications*

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

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

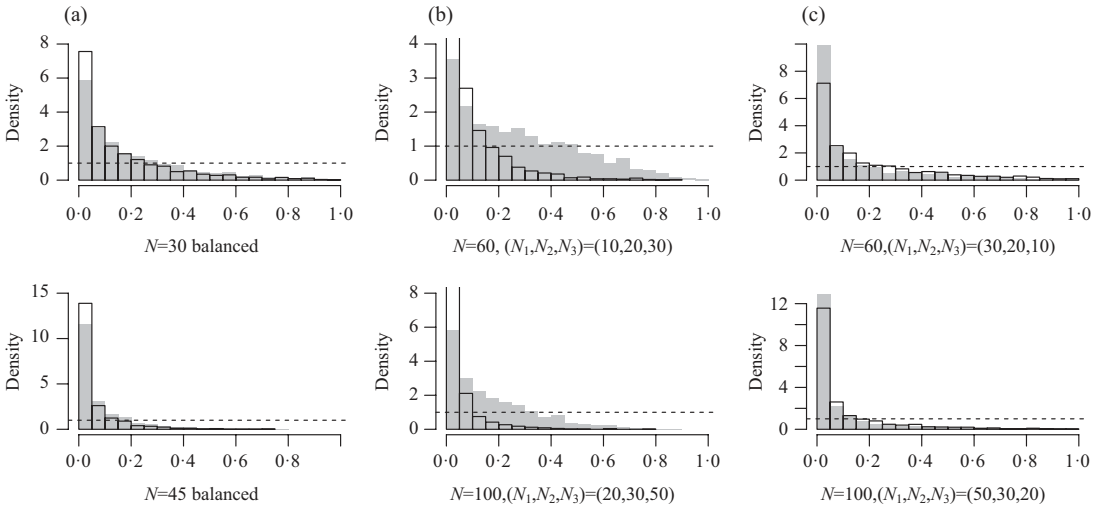


Fig. 3. Histograms of the p-values under alternative hypotheses based on the Fisher randomization tests using F and X^2: (a) balanced experiments, case 4; (b) unbalanced experiments, case 5; (c) unbalanced experiments, case 6. Grey histograms correspond to X^2 and white histograms to F.

7. DISCUSSION

As shown in the proofs of Theorems 1 and 3 in the Supplementary Material, we need to analyse 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, by considering the case of $J = 2$ we can gain more insight and 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 $\text{var}\{\hat{\tau}(j,j')\} = S_\cdot^2(j)/N_j + S_\cdot^2(j')/N_{j'} - S_\tau^2(j,j')/N$ and a 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 for 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 $H_{0\text{F}}$, $F \dot{\sim} \chi_1^2$ and $X^2 \dot{\sim} \chi_1^2$. Under $H_{0\text{N}}$, $F \dot{\sim} C_1\chi_1^2$ and $X^2 \dot{\sim} C_2\chi_1^2$, where

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

Depending on the sample sizes and the finite-population variances, C_1 can be either larger or smaller than 1. Consequently, using F in the Fisher randomization test can be conservative or anticonservative for testing $H_{0\text{N}}$. In contrast, C_2 is always no larger than 1, and therefore using X^2 in the Fisher randomization test is conservative for testing $H_{0\text{N}}$. Neyman (1923) proposed using the square root of X^2 to test $H_{0\text{N}}$ 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 $\hat{\tau}(1,2)$ has been widely used in the Fisher randomization test (Imbens & Rubin, 2015); it, however, can

yield either conservative or anticonservative tests for $H_{0\mathrm{N}}$, as shown by 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*

$$T^2 = \frac{\hat{\tau}^2(1,2)}{Ns^2_{\mathrm{obs}}/(N_1N_2)} \approx \frac{\hat{\tau}^2(1,2)}{s^2_{\mathrm{obs}}(1)/N_2 + s^2_{\mathrm{obs}}(2)/N_1 + \hat{\tau}^2(1,2)/N},$$

where the approximation is due to ignoring the difference between $(N, N_1 - 1, N_2 - 1)$ and (N, N_1, N_2). Under $H_{0\mathrm{F}}$, $T^2 \dot{\sim} F \dot{\sim} \chi^2_1$, and under $H_{0\mathrm{N}}$, $T^2 \dot{\sim} F \dot{\sim} C_1\chi^2_1$ with C_1 defined in (4).

Remark 5. Analogously, under the superpopulation model, Romano (1990) showed that the Fisher randomization test using $\hat{\tau}(1,2)$ can be conservative or anticonservative 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 superpopulation model.

After rejecting either $H_{0\mathrm{F}}$ or $H_{0\mathrm{N}}$, it is often of interest to test pairwise hypotheses; that is, for $j \neq j'$, $Y_i(j) = Y_i(j')$ for all i, or $\bar{Y}(j) = \bar{Y}(j')$. According to Corollaries 5 and 6, we recommend using the Fisher randomization test with test statistic $\hat{\tau}^2(j,j')/\{s^2_{\mathrm{obs}}(j)/N_j + s^2_{\mathrm{obs}}(j')/N_{j'}\}$, 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 analogy between our finite-population theory and the superpopulation theory of Chung & Romano (2013) 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, giving rise to the problem of multiple testings (Chung & Romano, 2016b). We leave these questions to future work.

ACKNOWLEDGEMENT

Ding was partially funded by the Institute of Education Sciences and the National Science Foundation, U.S.A. Dasgupta was partially funded by the National Science Foundation, U.S.A. The authors thank Xinran Li, Zhichao Jiang, Lo-Hua Yuan and Robin Gong for comments on earlier versions of the paper. We thank a reviewer and the associate editor for helpful comments.

SUPPLEMENTARY MATERIAL

Supplementary material available at *Biometrika* online includes proofs and more examples.

REFERENCES

ARONOW, P. M., GREEN, D. P. & LEE, D. K. (2014). Sharp bounds on the variance in randomized experiments. *Ann. Statist.* **42**, 850–71.

CHUNG, E. & ROMANO, J. P. (2013). Exact and asymptotically robust permutation tests. *Ann. Statist.* **41**, 484–507.

CHUNG, E. & ROMANO, J. P. (2016a). Asymptotically valid and exact permutation tests based on two-sample U-statistics. *J. Statist. Plan. Infer.* **168**, 97–105.

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

DASGUPTA, T., PILLAI, N. S. & RUBIN, D. B. (2015). Causal inference from 2^K factorial designs using the potential outcomes model. *J. R. Statist. Soc.* B **74**, 727–53.
DING, P. (2017). A paradox from randomization-based causal inference (with Discussion). *Statist. Sci.* **32**, 331–45.
DING, P., FELLER, A. & MIRATRIX, L. (2016). Randomization inference for treatment effect variation. *J. R. Statist. Soc.* B **78**, 655–71.
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. *J. Mult. Anal.* **153**, 255–65.
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. *Statist. Med.* **15**, 1069–92.
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**, 324–9.
JANSSEN, A. (1997). Studentized permutation tests for non-i.i.d. hypotheses and the generalized Behrens-Fisher problem. *Statist. Prob. Lett.* **36**, 9–21.
JANSSEN, A. (1999). Testing nonparametric statistical functionals with applications to rank tests. *J. Statist. Plan. Infer.* **81**, 71–93.
JANSSEN, A. & PAULS, T. (2003). How do bootstrap and permutation tests work? *Ann. Statist.* **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. *J. Am. Statist. Assoc.* **50**, 946–67.
KRISHNAMOORTHY, K., LU, F. & MATHEW, T. (2007). A parametric bootstrap approach for ANOVA with unequal variances: Fixed and random models. *Comp. Statist. Data Anal.* **51**, 5731–42.
LI, X. & DING, P. (2017). General forms of finite population central limit theorems with applications to causal inference. *J. Am. Statist. Assoc.* **112**, 1759–69.
LIN, W., HALPERN, S. D., PRASAD, K. M. & SMALL, D. S. (2017). A "placement of death" approach for studies of treatment effects on ICU length of stay. *Statist. Meth. Med. Res.* **26**, 292–311.
NEUHAUS, G. (1993). Conditional rank tests for the two-sample problem under random censorship. *Ann. Statist.* **21**, 1760–79.
NEYMAN, J. (1923). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statist. Sci.* **5**, 465–72.
NEYMAN, J. (1935). Statistical problems in agricultural experimentation (with Discussion). *Suppl. J. R. Statist. Soc.* **2**, 107–80.
PAULY, M., BRUNNER, E. & KONIETSCHKE, F. (2015). Asymptotic permutation tests in general factorial designs. *J. R. Statist. Soc.* B **77**, 461–73.
PITMAN, E. J. (1938). Significance tests which may be applied to samples from any populations: III. The analysis of variance test. *Biometrika* **29**, 322–35.
RICE, W. R. & GAINES, S. D. (1989). One-way analysis of variance with unequal variances. *Proc. Nat. Acad. Sci.* **86**, 8183–4.
ROMANO, J. P. (1990). On the behavior of randomization tests without a group invariance assumption. *J. Am. Statist. Assoc.* **85**, 686–92.
ROSENBAUM, P. R. (2010). *Design of Observational Studies*. New York: Springer.
RUBIN, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol*. 66, 688–701.
RUBIN, D. B. (1980). Comment on "Randomization analysis of experimental data: The Fisher randomization test" by D. Basu. *J. Am. Statist. Assoc.* **75**, 591–3.
SCHEFFE, H. (1959). *The Analysis of Variance*. New York: John Wiley & Sons.
SILVEY, S. D. (1954). The asymptotic distributions of statistics arising in certain non-parametric tests. *Glasgow Math. J.* **2**, 47–51.
WEERAHANDI, S. (1995). ANOVA under unequal error variances. *Biometrics* **51**, 589–99.
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**, 330–6.

[*Received on* 22 *July* 2016. *Editorial decision on* 20 *August* 2017]

Biometrika (2017), **xx**, x, *pp*. 1–14

Printed in Great Britain

Supplementary material for "A randomization-based perspective on 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

§S1 presents the proofs, §S2 contains examples, and §S3 gives additional simulation.

S1. 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*

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

*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 $j \neq j'$,*

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

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

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

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

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

□

Proof of Theorem 1. Under $H_{0\text{F}}$, $\{Y_i^{\text{obs}}: i=1,\ldots,N\}$ and $\text{SS}=(N-1)s_{\text{obs}}^2$ 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,

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

which further implies that

$$E(\text{SS}_\text{T}) = \text{SS} - E(\text{SS}_\text{R}) = (N - 1)s^2_{\text{obs}} - (N - J)s^2_{\text{obs}} = (J - 1)s^2_{\text{obs}}.$$

Applying Lemma S2, we have

$$\text{var}\{\bar{Y}_{\cdot}^{\text{obs}}(j)\} = \frac{1 - p_j}{N_j}s^2_{\text{obs}}, \quad \text{cov}\{\bar{Y}_{\cdot}^{\text{obs}}(j), \bar{Y}_{\cdot}^{\text{obs}}(j')\} = -\frac{s^2_{\text{obs}}}{N}. \tag{S1}$$

Therefore, the finite-population central limit theorem (Li & Ding, 2017, Theorem 5), coupled with the variance and covariance formulae in (S1), implies

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

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

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

□

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

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

Therefore,

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

Because

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

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

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

Second,

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

We further define $\sum_{j'\neq j} p_{j'} S_\tau^2(j,j') = \Delta_j$. 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

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

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

$$E(\mathrm{SS_T}) = \sum_{j=1}^{J} N_j\left\{\bar{Y}_\cdot(j) - \sum_{j=1}^{J} p_j\bar{Y}_\cdot(j)\right\}^2 + \sum_{j=1}^{J}(1-p_j)S_\cdot^2(j) + \Delta - S^2 + S^2 - 2\Delta$$
$$= \sum_{j=1}^{J} N_j\left\{\bar{Y}_\cdot(j) - \sum_{j=1}^{J} p_j\bar{Y}_\cdot(j)\right\}^2 + \sum_{j=1}^{J}(1-p_j)S_\cdot^2(j) - \Delta.$$

Under $H_{0\mathrm{N}}$, 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

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

Because $\{Y_i^{\mathrm{obs}} : W_i(j) = 1\}$ 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., $E\{s_{\mathrm{obs}}^2(j)\} = S_\cdot^2(j)$. Therefore, the mean of the residual sum of squares is

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

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 $S^2 = \sum_{j=1}^{J} S_\cdot^2(j)/J$, and therefore Theorem 2 implies

$$E(\mathrm{SS_R}) = \frac{N-J}{J}\sum_{j=1}^{J} S_\cdot^2(j) = (N-J)S^2,$$
$$E(\mathrm{SS_T}) = \frac{N}{J}\sum_{j=1}^{J}\{\bar{Y}_\cdot(j) - \bar{Y}_\cdot(\cdot)\}^2 + (J-1)S^2 - \Delta.$$

Moreover, under $H_{0\mathrm{N}}$, $E(\mathrm{SS_R})$ is unchanged, and $E(\mathrm{SS_T}) = (J-1)S^2 - \Delta$. Therefore, the expectation of the mean treatment squares is no larger than the expectation of the mean residual squares, because $E(\mathrm{MS_R}) - E(\mathrm{MS_T}) = \Delta/(J-1) \geq 0$. □

Proof of Corollary 3. Under $H_{0\mathrm{N}}$,

$$E(\mathrm{MS_R}) - E(\mathrm{MS_T}) = \sum_{j=1}^{J} \left(\frac{N_j - 1}{N - J} - \frac{1 - p_j}{J - 1} \right) S_{\cdot}^2(j) + \frac{\Delta}{J - 1}$$
$$= \frac{(N-1)J}{(J-1)(N-J)} \sum_{j=1}^{J} (p_j - J^{-1}) S_{\cdot}^2(j) + \frac{\Delta}{J-1}. \qquad \square$$

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_1^2, 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 $H_{0\mathrm{N}}$, and then view the result under $H_{0\mathrm{F}}$ as a special case.

Let $Q_j = N_j / S_{\cdot}^2(j)$ 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 $\bar{Y}_{w0}^{\mathrm{obs}} = Q^{-1} \sum_{j=1}^{J} Q_j \bar{Y}_{\cdot}^{\mathrm{obs}}(j)$ be a weighted average of the means of the observed outcomes. According to Li & Ding (2017, Proposition 3), $s_{\mathrm{obs}}^2(j) - S_{\cdot}^2(j) \to 0$ in probability $(j = 1, \ldots, J)$. By Slutsky's Theorem, X^2 has the same asymptotic distribution as

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

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

$$V_0 \equiv \begin{bmatrix} Q_1^{1/2}\{\bar{Y}_{\cdot}^{\mathrm{obs}}(1) - \bar{Y}_{\cdot}(1)\} \\ Q_2^{1/2}\{\bar{Y}_{\cdot}^{\mathrm{obs}}(2) - \bar{Y}_{\cdot}(2)\} \\ \vdots \\ Q_J^{1/2}\{\bar{Y}_{\cdot}^{\mathrm{obs}}(J) - \bar{Y}_{\cdot}(J)\} \end{bmatrix}$$
$$\dot{\sim} \mathcal{N}_J \left[0, \begin{pmatrix} 1 - p_1 & -p_1^{1/2} p_2^{1/2} \rho_{12} & \cdots & -p_1^{1/2} p_J^{1/2} \rho_{1J} \\ -p_2^{1/2} p_1^{1/2} \rho_{21} & 1 - p_2 & \cdots & -p_2^{1/2} p_J^{1/2} \rho_{2J} \\ \vdots & & & \vdots \\ -p_J^{1/2} p_1^{1/2} \rho_{J1} & -p_J^{1/2} p_2^{1/2} \rho_{J2} & \cdots & 1 - p_J \end{pmatrix} = P * R \right],$$

recalling $P = I_J - qq^{\mathrm{T}}$ and the element-wise product operator $*$. In the above, the mean and covariance matrix of the random vector V_0 follow directly from Lemmas S1 and S2.

Under $H_{0\mathrm{N}}$ with $\bar{Y}_{\cdot}(1) = \cdots = \bar{Y}_{\cdot}(J)$, we can verify that

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

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 $S_{\cdot}^2(j)$ 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 $H_{0\mathrm{F}}$ is the same as the permutation distribution. Therefore, applying Chung & Romano (2013) yields the same result for X^2 under $H_{0\mathrm{F}}$. □

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 $S_{\cdot}^2(1) = \cdots = S_{\cdot}^2(J) = S^2$, then $\bar{Y}_{w0}^{\mathrm{obs}} = \bar{Y}_{\cdot}^{\mathrm{obs}}$, and

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

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

$$(J-1)F = \frac{\mathrm{SS_T}}{\mathrm{MS_R}} \dot{\sim} \frac{\mathrm{SS_T}}{S^2}.$$

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

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

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

The treatment sum of squares reduces to

$$\mathrm{SS_T} = N_1 \left\{\bar{Y}_{\cdot}^{\mathrm{obs}}(1) - \bar{Y}_{\cdot}^{\mathrm{obs}}\right\}^2 + N_2 \left\{\bar{Y}_{\cdot}^{\mathrm{obs}}(2) - \bar{Y}_{\cdot}^{\mathrm{obs}}\right\}^2 = N p_1 p_2 \hat{\tau}^2(1,2),$$

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

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

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 $H_{0\text{F}}$. However, under $H_{0\text{N}}$, Neyman (1923), coupled with the finite-population central limit theorem (Li & Ding, 2017, Theorem 5), imply

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

and $s^2_{\text{obs}}(j) \to S^2_\cdot(j)$ in probability $(j = 1, 2)$. Therefore, the asymptotic distribution of F under $H_{0\text{N}}$ is $F \dot\sim C_1\chi^2_1$, where

$$C_1 = \lim_{N\to+\infty} \frac{S^2_\cdot(1)/N_1 + S^2_\cdot(2)/N_2 - S^2_\tau(1,2)/N}{S^2_\cdot(1)/N_2 + S^2_\cdot(2)/N_1}.$$

Second, we discuss X^2. Because

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

we have

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

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 $H_{0\text{F}}$. However, under $H_{0\text{N}}$, we can use an argument similar to that for F and obtain $X^2 \dot\sim C_2\chi^2_1$, where

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

The constant C_2 is smaller than or equal to 1 with equality holding if the limit of $S^2_\tau(1,2)$ 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

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

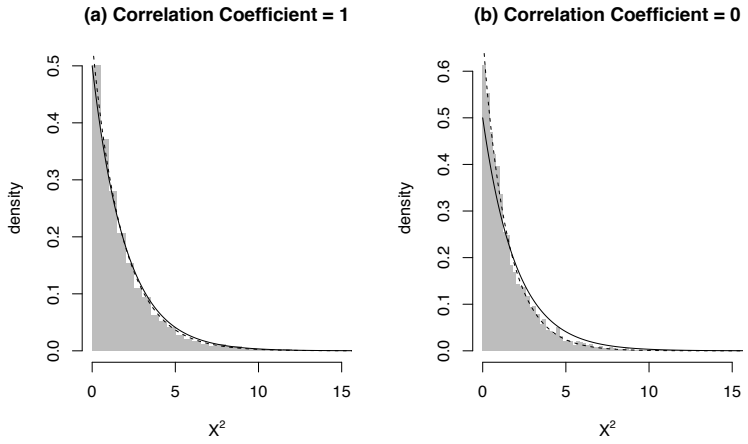


Fig. S1: Distributions of X^2. The histograms are the sampling distributions, the dotted lines are the asymptotic distributions, and the solid lines are the χ^2_2 distribution.

which implies the equivalent formula of T^2 in Corollary 6. Under $H_{0\mathrm{F}}$ or $H_{0\mathrm{N}}$, $\hat{\tau}(1,2) \to 0$ in probability, which coupled with Slutsky's Theorem, implies the asymptotic equivalence $T^2 \stackrel{\cdot}{\sim} F$. □

S2. 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

$$Y_i(1) \sim \mathcal{N}(0,1), \quad Y_i(2) = 3Y_i(1), \quad Y_i(3) = 5Y_i(1), \tag{S2}$$

and the second set of potential outcomes from

$$Y_i(1) \sim \mathcal{N}(0,1), \quad Y_i(2) \sim \mathcal{N}(0,3^2), \quad Y_i(3) \sim \mathcal{N}(0,5^2). \tag{S3}$$

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 $H_{0\mathrm{N}}$ holds. Figure S1 shows the distributions of X^2 over repeated sampling of the treatment assignment vector $(W_1, \ldots, W_N)$ for potential outcomes generated from (S2) and (S3). The true sampling distributions under both cases are stochastically dominated by χ^2_2. Under (S2), the correlation coefficients between the potential outcomes are 1; whereas under (S3), 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 $H_{0\mathrm{N}}$, 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.

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

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

S3. MORE SIMULATION WITH NONNORMAL OUTCOMES

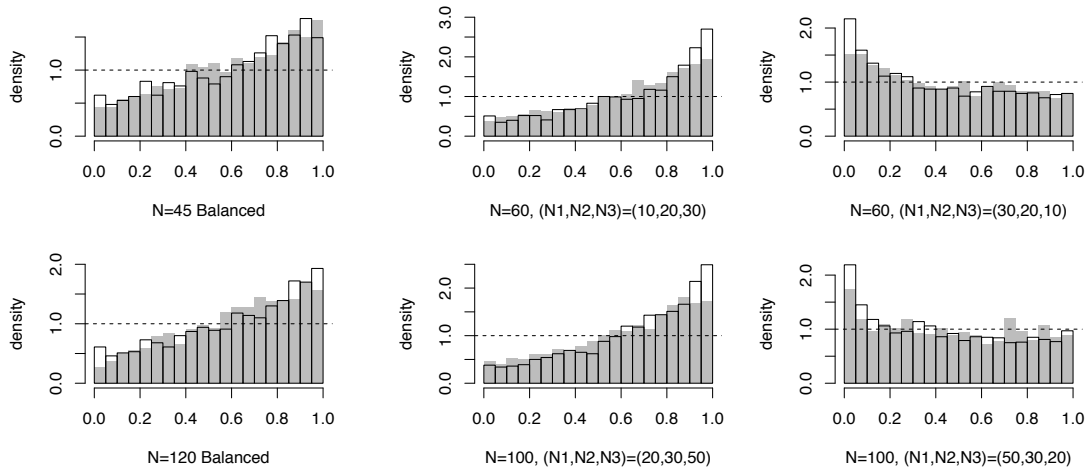
S3·1. *Type I error of the Fisher randomization test using F*

In this subsection, we use simulation to evaluate the finite sample performance of the Fisher randomization test using F under $H_{0\mathrm{N}}$. 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) $Y_i(1) \sim \mathcal{E}$, $Y_i(2) = 1.5Y_i(1)$, $Y_i(3) = 2Y_i(1)$.



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

Fig. S2: Histograms of the p-values under $H_{0\mathrm{N}}$ based on the Fisher randomization tests using X^2, with grey histogram and white histograms for the first and second sub-cases.

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 S2 exhibit the same pattern.

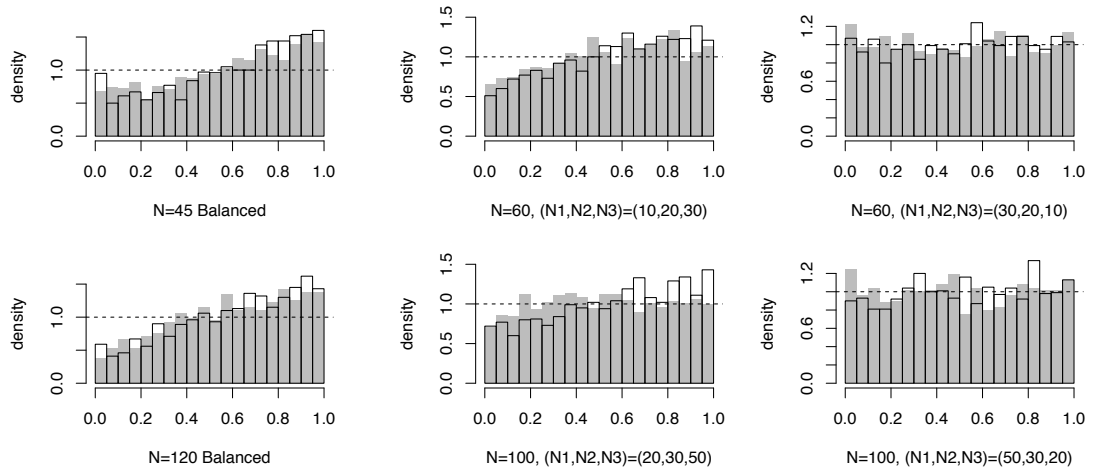
In Figure 2(a), for case (S1.1), the rejection rates are 0.022 and 0.014, and for case (S1.2), the rejection rates are 0.030 and 0.030, for sample sizes $N = 45$ and $N = 120$ respectively. In Figure 2(b), for case (S2.1), the rejection rates are 0.018 and 0.024, and for case (2.2), the rejection rates 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 2(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.

S3·2. *Type I error of the Fisher randomization test using X^2*

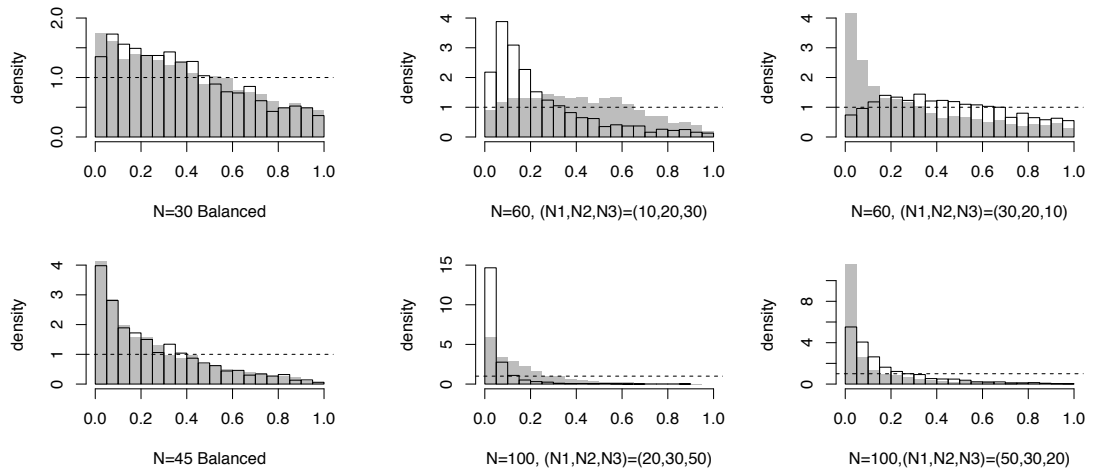
We follow §6·2, generate the same data as §S3·1, and obtain the same conclusions about the Fisher randomization test using X^2, because Figures 2 and S3 exhibit the same pattern. All the Monte Carlo standard errors of the rejection rates below are close but no larger than 0.005.

In Figure 3(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 3(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 3(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 S2, agrees with our theory that the Fisher randomization test using X^2 can control type I error under $H_{0\mathrm{N}}$ better than using F.



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

Fig. S3: Histograms of the p-values under $H_{0\mathrm{N}}$ based on the Fisher randomization tests using X^2, with grey histogram and white histograms for the first and second sub-cases.



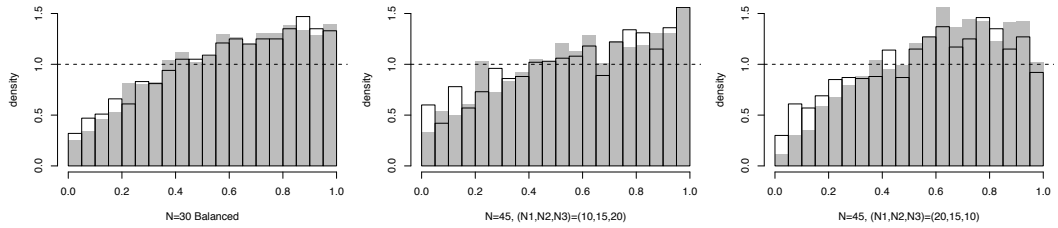
(a) Balanced experiments, case S4 (b) Unbalanced experiments, case S5 (c) Unbalanced experiments, case S6

Fig. S4: 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.

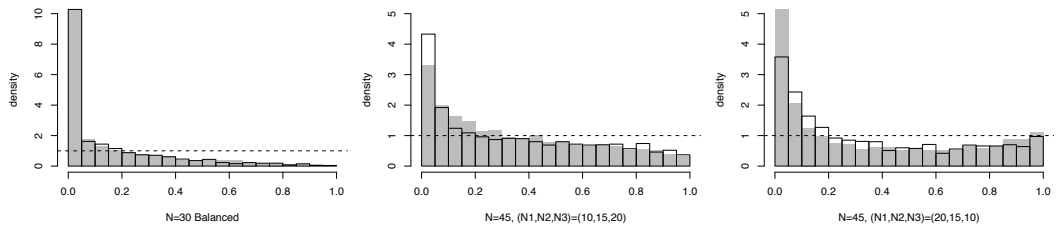
S3·3. *Power comparison of the Fisher randomization tests using F and X^2*

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 S4.

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 $(0, 0.5, 1)$.



(a) $H_{0\mathrm{N}}$ holds



(b) $H_{0\mathrm{N}}$ does not hold

Fig. S5: 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.

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 $S^2_\cdot(1) < S^2_\cdot(2) < S^2_\cdot(3)$.

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 $S^2_\cdot(1) < S^2_\cdot(2) < S^2_\cdot(3)$.

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 $N = 100$.

S3·4. *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 $H_{0\mathrm{N}}$, the p-values of the Fisher randomization test using F and X^2 are shown in Figure S5(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 S5(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; 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.

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.

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

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

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

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.

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

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

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

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