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; Average null hypothesis; Fisher randomization test; One-way layout; Repeated sampling property; Sharp null hypothesis

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 of 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 two 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 ex-
experiment, Neyman’s null hypothesis allows for treatment effect heterogeneity, which is weaker than Fisher’s null 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, 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 and controls asymptotic type I error under Neyman’s null.

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. Under the stable unit treatment value assumption (Rubin, 1980), let \( Y_i(j) \) denote the potential outcome (Neyman, 1923; Rubin, 1974) of unit \( i \) when assigned to treatment level \( j \), where \( i = 1, \ldots, N \) and \( j = 1, \ldots, J \). 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

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

where \( \bar{Y}(j) = N^{-1} \sum_{i=1}^{N} Y_i(j) \) 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 \( i \)th unit is assigned to treatment \( j \), and 0 otherwise, where \( i = 1, \ldots, N \) and \( j = 1, \ldots, J \). Equivalently, it can be represented by the discrete random variable \( W_i = \sum_{j=1}^{J} jW_i(j) \in \{1, \ldots, J\} \), the treatment received by unit \( i \). Let \( (W_1, \ldots, W_N) \) be the treatment assignment vector, and \( (w_1, \ldots, w_N) \) denote its realization. If \( (N_1, \ldots, N_J) \) units are assigned at random to treatments \( (1, \ldots, J) \) respectively, the treatment assignment mechanism satisfies \( \Pr((W_1, \ldots, W_N) = (w_1, \ldots, w_N)) = \prod_{j=1}^{J} N_j!/N! \) 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 \( Y_i^{\text{obs}} = \sum_{j=1}^{J} W_i(j)Y_i(j) \) (\( i = 1, \ldots, N \)).

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:

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

Under (1), all the \( J \) potential outcomes \( Y_i(1), \ldots, Y_i(J) \) are equal to the observed outcome \( Y_i^{\text{obs}} \), 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 (1) 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 de-
parture from (1). Define \( \overline{Y}_i^{\text{obs}}(j) = N_j^{-1} \sum_{i=1}^{N_j} W_i(j)Y_i^{\text{obs}} \) as the sample average of the observed outcomes within treatment level \( j \), and \( \overline{Y}_i^{\text{obs}} = N^{-1} \sum_{i=1}^{N} Y_i^{\text{obs}} \) as the sample average of all the observed outcomes. Define \( s_{\text{obs}}^2(j) = (N_j - 1)^{-1} \sum_{i=1}^{N_j} W_i(j)(Y_i^{\text{obs}} - \overline{Y}_i^{\text{obs}}(j))^2 \) and \( s_{\text{obs}}^2 = (N - 1)^{-1} \sum_{i=1}^{N} (Y_i^{\text{obs}} - \overline{Y}_i^{\text{obs}}(j))^2 \) as the corresponding sample variances with divisors \( N_j - 1 \) and \( N - 1 \), respectively. Let

\[
\text{SSTre} = \sum_{j=1}^{J} N_j \{ \overline{Y}_i^{\text{obs}}(j) - \overline{Y}_i^{\text{obs}} \}^2
\]

be the treatment sum of squares, and

\[
\text{SSRes} = \sum_{j=1}^{J} \sum_{i:W_i(j)=1} \{ Y_i^{\text{obs}} - \overline{Y}_i^{\text{obs}}(j) \}^2 = \sum_{j=1}^{J} (N_j - 1)s_{\text{obs}}^2(j)
\]

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

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

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

It is believed that the distribution of (4) under Fisher’s null 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 (4) 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 quite non-trivial. Below, we state a known result on the approximate randomization distribution of (4), in which we use the notation \( A_N \sim B_N \) to represent two sequences of random variables \( \{A_N\}_{N=1}^{\infty} \) and \( \{B_N\}_{N=1}^{\infty} \) that have the same asymptotic distribution.

**Theorem 1.** Assume Fisher’s null. Over repeated sampling of \( (W_1, \ldots, W_N) \), the expectations of the residual and treatment sums of squares are \( E(\text{SSRes}) = (N - J)s_{\text{obs}}^2 \) and \( E(\text{SSTre}) = (J-1)s_{\text{obs}}^2 \), and the asymptotic distribution of (4) is

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

**Remark 1.** As \( N \to \infty \), both the statistic \( F \) and random variable \( F_{J-1,N-J} \) are asymptotically \( \chi^2_{J-1}/(J-1) \). The original \( F \) approximation for randomization inference for a finite population was derived by cumbersome moment matching between statistic (4) and the corresponding \( F_{J-1,N-J} \) distribution (Welch, 1937; Pitman, 1938; Kempthorne, 1952). In the Supplementary Material, we provide a simpler proof based on the finite population central limit theorem (Hájek, 1960), and throughout the paper we assume the regularity conditions required by this theorem.

**Remark 2.** Under Fisher’s null, 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 in Theorem 1 are calculated with respect to the distribution of the treatment assignment. Also, the ratio of expectations of the numerator MSTre and denominator MSRes of (4) is 1 under Fisher’s null.
4. Sampling properties of the F statistic under Neyman’s null hypothesis

In Section 3, we discussed the randomization distribution, i.e., sampling distribution under Fisher’s null, of the $F$ statistic in (4). However, the sampling distribution of the $F$ statistic under Neyman’s null hypothesis of zero average treatment effect (Neyman, 1923, 1935), i.e.,

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

(5)
is often of major interest but is under-investigated. Neyman’s null (5) imposes weaker restrictions on the potential outcomes than Fisher’s null (1), making it impossible to compute the exact, or even approximate distribution of the $F$ statistic under such hypothesis. However, analytical expressions for $E(\text{SSTre})$ and $E(\text{SSRes})$ can still be derived under Neyman’s null along the lines of Theorem 1, and can be used to gain insights about the consequences of testing Neyman’s null using the Fisher randomization test with the $F$ statistic in (4).

For treatment level $j = 1, \ldots, J$, define $p_j = N_j/N$ as the proportion of the units, and $S^2(j) = (N - 1)^{-1} \sum_{i=1}^{N} \{Y_i(j) - \bar{Y}(j)\}^2$ as the finite population variances of potential outcomes. Let $\bar{Y}(\cdot) = \sum_{j=1}^{J} p_j \bar{Y}(j)$ and $S^2 = \sum_{j=1}^{J} p_j S^2(j)$ be the weighted averages of the finite population means and variances. The sampling distribution of the $F$ statistic in (4) depends crucially on the finite population variance of the unit-level treatment effects

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

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

Kempthorne (1955) obtained the following result on the sampling expectations of $\text{SSRes}$ and $\text{SSTre}$ for balanced designs under the assumption of strict additivity:

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

(6)

This result implies that under strict additivity and balanced treatment assignment, $E(\text{MStr} - \text{MStre}) = 0$ under Neyman’s null, and provides a heuristic justification for testing Neyman’s null using the Fisher randomization test with the $F$ statistic. However, strict additivity combined with Neyman’s null implies Fisher’s null, 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

$$\Delta = \sum_{j<j'} \sum_{j} p_j p_{j'} S^2(j-j')$$

(7)

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 $\Delta$ 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 given by (3) is $E(\text{SSRes}) = \sum_{j=1}^{J} (N_j - 1)S^2(j)$, and the expectation of the treat-
A randomization-based perspective of analysis of variance

The sum of squares given by (2) is

\[ E(\text{SSTre}) = \sum_{j=1}^{J} N_j \{ \bar{Y}_j(j) - \bar{Y}(\cdot) \}^2 + \sum_{j=1}^{J} (1 - p_j) S^2(j) - \Delta, \]

which reduces to \( E(\text{SSTre}) = \sum_{j=1}^{J} (1 - p_j) S^2(j) - \Delta \) under Neyman’s null.

**Corollary 1.** Under Neyman’s null with strict additivity in Definition 1, or, equivalently, under Fisher’s null, 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 \( p_j = 1/J \),

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

Furthermore, under Neyman’s null, \( E(\text{SSRes}) = (N - J) S^2 \) 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 (6) 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 Neyman’s null 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 Neyman’s null, the difference between the mean squares of the residual and the treatment is

\[ E(\text{MSRes} - \text{MSTre}) = \frac{(N - 1) J}{(J - 1)(N - J)} \sum_{j=1}^{J} (p_j - J^{-1}) S^2(j) + \frac{\Delta}{J - 1}. \]

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. When the \( p_j \)’s and \( S^2(j) \)’s are positively associated, the Fisher randomization test using \( F \) tends to be conservative under Neyman’s null; when the \( p_j \)’s and \( S^2(j) \)’s are negatively associated, the Fisher randomization test using \( F \) may not control correct type I error under Neyman’s null. The numerical examples in Section 6.1 will make this clearer.

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 Neyman’s null in unbalanced experiments, we propose to use the following \( X^2 \) test statistic for the Fisher randomization test. Define the weighted average of the sample means as

\[ \bar{Y}_{\text{w}} = \frac{\sum_{j=1}^{J} N_j \bar{Y}_j(j)}{\sum_{j=1}^{J} N_j s^2_{\text{obs}}(j)}. \]
Define the $X^2$ test statistic as

$$X^2 = \sum_{j=1}^{J} \frac{N_j}{\hat{s}_{\text{obs}}^2(j)} \left( \bar{Y}^\text{obs}(j) - \bar{Y}_w \right)^2,$$

which is the weighted average of the $J$ treatment squares. This test statistic has been exploited in the classical analysis of variance literature (e.g., James, 1951; Welch, 1951; 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; 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 Fisher’s null, because we always have an exact test for Fisher’s null 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 under both Fisher’s and Neyman’s null hypotheses 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 Fisher’s null, the asymptotic distribution of $X^2$ is $\chi^2_{J-1}$. Under Neyman’s null, the asymptotic distribution of $X^2$ is stochastically dominated by $\chi^2_{J-1}$, i.e., for any constant $a > 0$, $\text{pr}(X^2 \geq a) \leq \text{pr}(\chi^2_{J-1} \geq a)$.

**Remark 3.** Under Fisher’s null, the randomization distribution of $\text{SSTre}/\hat{s}_{\text{obs}}^2$ follows $\chi^2_{J-1}$ asymptotically as shown in the Supplementary Material. Under Neyman’s null, however, the asymptotic distribution of $\text{SSTre}/\hat{s}_{\text{obs}}^2$ is not $\chi^2_{J-1}$, 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 $\chi^2_{J-1}$ randomization distribution under Fisher’s null, and has an asymptotic distribution that is stochastically dominated by $\chi^2_{J-1}$ under Neyman’s null.

Therefore, under Neyman’s null, 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 Fisher’s or Neyman’s null is $X^2 > x_{1-\alpha}$, the $1 - \alpha$ quantile of the $\chi^2_{J-1}$ distribution. Because the asymptotic distribution of $X^2$ under Neyman’s null is stochastically dominated by $\chi^2_{J-1}$, 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). This feature 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 (8) 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.
A randomization-based perspective of analysis of variance

\textbf{Corollary 4.} If $S_1^2(1) = \cdots = S_J^2(J)$, then $(J-1)F \sim X^2$, which are both asymptotically $\chi^2_{J-1}$ under Fisher’s null or Neyman’s null.

The condition $S_1^2(1) = \cdots = S_J^2(J)$ holds under treatment effect additivity in Definition 1. Under this condition, 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.

\section{Simulation Studies}

6.1. Type I error of the Fisher randomization test using $F$

In this subsection, we use simulation studies to evaluate the finite sample performance of the Fisher randomization test using $F$ under Neyman’s null. We consider the following three cases, where $N(\mu, \sigma^2)$ denotes a normal distribution with mean $\mu$ and variance $\sigma^2$.

Case 1. For balanced experiments with sample sizes $N = 45$ and $N = 120$, we generate potential outcomes under two cases: (1.1) $Y_i(1) \sim N(0, 1)$, $Y_i(2) \sim N(0, 1.2)$, $Y_i(3) \sim N(0, 1.5)$; and (1.2) $Y_i(1) \sim N(0, 1)$, $Y_i(2) \sim N(0, 2)$, $Y_i(3) \sim N(0, 3)$. 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) $Y_i(1) \sim N(0, 1)$, $Y_i(2) = 2Y_i(1)$, $Y_i(3) = 3Y_i(1)$; and (2.2) $Y_i(1) \sim N(0, 1)$, $Y_i(2) = 3Y_i(1)$, $Y_i(3) = 5Y_i(1)$. These potential outcomes are standardized to have zero means. In this case, $p_1 < p_2 < p_3$ and $S_1^2(1) < S_2^2(2) < S_3^2(3)$.

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 cases: (3.1) $Y_i(1) \sim N(0, 1)$, $Y_i(2) = 2Y_i(1)$, $Y_i(3) = 3Y_i(1)$; and (3.2) $Y_i(1) \sim N(0, 1)$, $Y_i(2) = 3Y_i(1)$, $Y_i(3) = 5Y_i(1)$. These potential outcomes are standardized to have zero means. In this case, $p_1 > p_2 > p_3$ and $S_1^2(1) < S_2^2(2) < S_3^2(3)$.

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 Figures 1(a)–1(c) corresponding to cases 1–3 above.

In Figure 1(a) for balanced experiments, the Fisher randomization testing using $F$ is always conservative with $p$-values distributed towards 1. With larger heterogeneity in the potential outcomes, the histograms of the $p$-values, i.e., the white histograms with black borders, have larger masses near 1.

In Figure 1(b) for unbalanced experiments, the sample sizes under each treatment level are increasing in the variances of the potential outcomes. The Fisher randomization test using $F$ is always 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.

In Figure 1(c) for unbalanced experiments, the sample sizes under different treatment levels are decreasing in the variances of the potential outcomes. 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.

These empirical findings are coherent with our theory in Section 4. In practice, if the sample sizes under different treatment levels are decreasing in the sample variances of the observed outcomes, then the Fisher randomization testing using $F$ may not yield correct type I error under Neyman’s null.
(a) Balanced experiments: grey histogram for $Y_i(1) \sim \mathcal{N}(0, 1)$, $Y_i(2) \sim \mathcal{N}(0, 1.2^2)$, $Y_i(3) \sim \mathcal{N}(0, 1.5^2)$, and white histogram with borders for $Y_i(1) \sim \mathcal{N}(0, 1)$, $Y_i(2) \sim \mathcal{N}(0, 2^2)$, $Y_i(3) \sim \mathcal{N}(0, 5^2)$.

(b) Unbalanced experiments: grey histogram for $Y_i(1) \sim \mathcal{N}(0, 1)$, $Y_i(2) = 2Y_i(1)$, $Y_i(3) = 3Y_i(1)$, and white histogram with borders for $Y_i(1) \sim \mathcal{N}(0, 1)$, $Y_i(2) = 3Y_i(1)$, $Y_i(3) = 5Y_i(1)$.

(c) Unbalanced experiments: grey histogram for $Y_i(1) \sim \mathcal{N}(0, 1)$, $Y_i(2) = 2Y_i(1)$, $Y_i(3) = 3Y_i(1)$, and white histogram with borders for $Y_i(1) \sim \mathcal{N}(0, 1)$, $Y_i(2) = 3Y_i(1)$, $Y_i(3) = 5Y_i(1)$.

Fig. 1. Histograms of the $p$-values over repeated sampling, based on the Fisher randomization test using $F$, under Neyman’s null.

6.2. Type I error of the Fisher randomization test using $X^2$

Figure 2 shows the same simulation studies as Figure 1, but with test statistic $X^2$. Figure 2(a) shows a similar pattern as Figure 1(a), indicating conservativeness of the Fisher randomization test using $X^2$. However, Figure 2(b) and Figure 2(c) show much better performance of the Fisher randomization test using $X^2$ than Figure 1(b) and Figure 1(c), because the $p$-values are much closer to uniform. This verifies our theory that the Fisher randomization test using $X^2$ can control the asymptotic type I error under Neyman’s null.

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 of nonzero average treatment effects. 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 $(0, 1, 2)$. 
Fisher randomization test using $F$ has smaller power than that using $X^2$ when the sample sizes are negatively associated with the variances of the potential outcomes, and when the sample sizes are positively associated with the variances of the potential outcomes, in Figures 3(a)–3(c), correspond to cases 4–6 above. When the experiments are balanced and obtain the $p$-value zero, and we then generate $(N_1, N_2, N_3) = (50, 30, 20)$. In this case, $p_1 < p_2 < p_3$ and $S_1^2(1) < S_2^2(2) < S_3^2(3)$.

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 $S_1^2(1) < S_2^2(2) < S_3^2(3)$.

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. When the experiments are balanced and 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 sample sizes are negatively associated with the variances of the potential outcomes, the Fisher randomization test using $F$ has smaller power than that using $X^2$. 

Fig. 2. Histograms of the $p$-values over repeated sampling, based on the Fisher randomization test using $X^2$, under Neyman's null.
(a) Balanced experiments, $Y_i(1) \sim \mathcal{N}(0, 1)$, $Y_i(2) \sim \mathcal{N}(0, 2^2)$, $Y_i(3) \sim \mathcal{N}(0, 3^2)$, and these potential outcomes are independently generated, and shifted to have means $(0, 1, 2)$: grey histogram for the FRT using $X^2$, and white histogram with borders for the FRT using $F$.

(b) Unbalanced experiments, $Y_i(1) \sim \mathcal{N}(0, 1)$ and standardized to have mean zero, $Y_i(2) = 3Y_i(1) + 1$ and $Y_i(3) = 5Y_i(1) + 2$: grey histogram for the Fisher randomization test using $X^2$, and white histogram with borders for the Fisher randomization test using $F$.

(c) Unbalanced experiments, the same potential outcomes as in Figure 3(b): grey histogram for the Fisher randomization test using $X^2$, and white histogram with borders for the Fisher randomization test using $F$.

Fig. 3. Histograms of the $p$-values over repeated sampling, based on the Fisher randomization tests using $F$ and $X^2$, under alternative hypotheses

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 (Hájek, 1960). Therefore, we recommend using $X^2$ as the test statistic for unbalanced completely randomized experiments.

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 apply Schur’s theorem (Schur, 1911; Stryan, 1973) to analyze the eigenvalues of the covariance matrix of \( \{Y^{\text{obs}}(1), \ldots, Y^{\text{obs}}(J)\} \) to obtain the properties of \( F \) and \( X^2 \) for general \( J \geq 2 \). 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 \( \hat{\tau}(j, j') \) is \( \hat{Y}^{\text{obs}}(j) - \hat{Y}^{\text{obs}}(j') \), which has sampling variance \( \text{var}\{\hat{\tau}(j, j')\} = S^2(j)/N_j + S^2(j')/N_j - S^2(j-j')/(N_j + N_j') \) and an conservative variance estimator \( s^2_{\text{obs}}(j)/N_j + s^2_{\text{obs}}(j')/N_j' \) (Neyman, 1923).

\textbf{Corollary 5.} When \( J = 2 \), the \( F \) and \( X^2 \) statistics reduce to

\[
F \approx \frac{\hat{\tau}^2(1, 2)}{s^2_{\text{obs}}(1)/N_2 + s^2_{\text{obs}}(2)/N_1}, \quad X^2 = \frac{\hat{\tau}^2(1, 2)}{s^2_{\text{obs}}(1)/N_1 + s^2_{\text{obs}}(2)/N_2},
\]

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 Fisher’s null, \( F \sim \chi^2_1 \) and \( X^2 \sim \chi^2_1 \). Under Neyman’s null, \( F \sim C_1 \chi^2_1 \) and \( X^2 \sim C_2 \chi^2_1 \), where

\[
C_1 = \lim_{N \to +\infty} \frac{\text{var}\{\hat{\tau}(1, 2)\}}{S^2(1)/N_2 + S^2(2)/N_1}, \quad C_2 = \lim_{N \to +\infty} \frac{\text{var}\{\hat{\tau}(1, 2)\}}{S^2(1)/N_1 + S^2(2)/N_2} \leq 1. \tag{9}
\]

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 Neyman’s null. In contrast, \( C_2 \) is always no larger than 1, and therefore using \( X^2 \) in the Fisher randomization test is conservative for testing Neyman’s null. Neyman (1923) proposed to use the square root of \( X^2 \) to test Neyman’s null 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.

Under a superpopulation model, Chung & Romano (2013) showed that using \( \hat{\tau}(1, 2) \) in the Fisher randomization test can be conservative or anti-conservative for testing the hypothesis of equal means of two samples, and suggested using the studentized statistic, or equivalently \( X^2 \), to remedy the problem of possibly inflated type I error. Chung & Romano (2013)’s test is asymptotically exact under the super population model.

After rejecting either Fisher’s or Neyman’s null hypothesis in (1) or (5), 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 \( \bar{Y}_i = \bar{Y}_i(j') \). According to Corollary 5, 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 other experimental designs and test statistics (Pauly et al., 2015; Chung & Romano, 2016a,b). We leave this to future work.

\textbf{Acknowledgement}

The research of Peng Ding was partially funded by Institute of Education Sciences, U.S.A. The authors thank Mr. Xinran Li, Dr. Zhichao Jiang, Miss Lo-Hua Yuan and Miss Robin Gong for helpful suggestions for early 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, more simulations and examples.

REFERENCES

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 the Royal Statistical Society, Series B (Statistical Methodology)* 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., 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.

HÁJEK, J. (1960). Limiting distributions in simple random sampling from a finite population. *Publications of the Mathematics Institute of the Hungarian Academy of Science* 5, 361–74.

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.

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

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

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

KEMPPTHORNE, 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.

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.

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. *Journal of Educational Psychology* 66, 688–701.

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.

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

WEEAHANDI, 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 material for
“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

Appendix A presents the proofs, Appendix B contains numerical examples, and Appendix C gives additional simulation studies.

APPENDIX A. PROOFS

To prove the theorems, we need the following basic 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

\[
\begin{align*}
\text{cov}\{W_i(j), W_i(j')\} &= -p_j(1 - p_j)/(N - 1), \quad (i \neq i') \\
\text{cov}\{W_i(j), W_i'(j')\} &= -p_j p_{j'}, \quad (j \neq j') \\
\text{cov}\{W_i(j), W_i'(j')\} &= p_j p_{j'}/(N - 1), \quad (i \neq i', j \neq j').
\end{align*}
\]

**Proof of Lemma S1.** The proof is straightforward. $\square$

**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 $S_{cd} = (S_c^2 + S_d^2 - S_{c-d}^2)/2$, where $S_{c-d}^2$ is the finite population variance of $(c_1 - d_1, \ldots, c_N - d_N)$. For $j \neq j'$,

\[
\begin{align*}
\text{var}\left\{ \frac{1}{N_j} \sum_{i=1}^{N} W_i(j) c_i \right\} &= \frac{1 - p_j}{N_j} S_c^2, \quad \text{cov}\left\{ \frac{1}{N_j} \sum_{i=1}^{N} W_i(j) c_i, \frac{1}{N_j} \sum_{i=1}^{N} W_i(j') d_i \right\} = -\frac{S_{cd}}{N}.
\end{align*}
\]
Proof of Lemma S2. Lemma S2 is known, and special forms appeared in Kempthorne (1955). We give an elementary proof for completeness. Applying Lemma S1, we have

$$\text{var}\left\{ \frac{1}{N_j} \sum_{i=1}^{N} Y_i(j)c_i \right\}$$

$$= \frac{1}{N_j^2} \text{var}\left\{ \sum_{i=1}^{N} W_i(j)(c_i - \bar{c}) \right\}$$

$$= \frac{1}{N_j^2} \left( \sum_{i=1}^{N} \text{var}\{ W_i(j)\}(c_i - \bar{c})^2 - \sum_{i \neq i'} \text{cov}\{ W_i(j), W_{i'}(j)\}(c_i - \bar{c})(c_{i'} - \bar{c}) \right)$$

$$= \frac{1}{N_j^2} \left( \sum_{i=1}^{N} p_j(1 - p_j)(c_i - \bar{c})^2 - \sum_{i \neq i'} \frac{p_j(1 - p_j)}{N - 1} (c_i - \bar{c})(c_{i'} - \bar{c}) \right)$$

$$= \frac{1}{N_j^2} \left( p_j(1 - p_j) \sum_{i=1}^{N} (c_i - \bar{c})^2 + \frac{p_j(1 - p_j)}{N - 1} \sum_{i=1}^{N} (c_i - \bar{c})^2 \right)$$

$$= \frac{1 - p_j}{N_j} S^2_c.$$  

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

$$\text{cov}\left\{ \frac{1}{N_j} \sum_{i=1}^{N} W_i(j)c_i, \frac{1}{N_{j'}} \sum_{i=1}^{N} W_{i'}(j)d_i \right\}$$

$$= \frac{1}{N_j N_{j'}} \text{cov}\left\{ \sum_{i=1}^{N} W_i(j)(c_i - \bar{c}), \sum_{i=1}^{N} W_{i'}(j')(d_i - \bar{d}) \right\}$$

$$= \frac{1}{N_j N_{j'}} \left( \sum_{i=1}^{N} \text{cov}\{ W_i(j), W_{i'}(j')\}(c_i - \bar{c})(d_i - \bar{d}) \right. \left. + \sum_{i \neq i'} \sum_{i' \neq i} \text{cov}\{ W_i(j), W_{i'}(j')\}(c_i - \bar{c})(d_{i'} - \bar{d}) \right)$$

$$= \frac{1}{N_j N_{j'}} \left( - \sum_{i=1}^{N} p_j p_{j'}(c_i - \bar{c})(d_i - \bar{d}) + \sum_{i \neq i'} \frac{p_j p_{j'}}{N - 1} (c_i - \bar{c})(d_{i'} - \bar{d}) \right)$$

$$= \frac{1}{N_j N_{j'}} \left( p_j p_{j'} \sum_{i=1}^{N} (c_i - \bar{c})(d_i - \bar{d}) + \frac{p_j p_{j'}}{N - 1} \sum_{i=1}^{N} (c_i - \bar{c})(d_i - \bar{d}) \right)$$

$$= -S_{cd}/N.$$  

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

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

Therefore, the finite population central limit theorem (Hájek, 1960), coupled with the variance and covariance formulas in (S1), implies

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

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

\[
F = \frac{\chi^2_{J-1}/(J-1)}{(N - 1 - \chi^2_{J-1})/(N - J)} \sim \frac{\chi^2_{J-1}/(J - 1)}{(N - 1) s^2_{\text{obs}} - \chi^2_{J-1} s^2_{\text{obs}}}/(N - J)
\]

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

\[
\text{cov}\{\bar{Y}_{\text{obs}}(j), \bar{Y}_{\text{obs}}(j')\} = \text{cov}\left\{ \frac{1}{N_j} \sum_{i=1}^{N} W_i(j)Y_i(j), \frac{1}{N_j} \sum_{i=1}^{N} W_i(j')Y_i(j') \right\}
\]

\[
= -\frac{1}{2N} \{S^2(j) + S^2(j') - S^2(j-j')\}.
\]
Therefore, the variance of $\bar{Y}_{\text{obs}} = \sum_{j=1}^{J} p_j \bar{Y}_{\text{obs}}(j)$ is

$$\text{var}(\bar{Y}_{\text{obs}}) = \sum_{j=1}^{J} p_j^2 \text{var}(\bar{Y}_{\text{obs}}(j)) + \sum_{j \neq j'} p_j p_{j'} \text{cov}(\bar{Y}_{\text{obs}}(j), \bar{Y}_{\text{obs}}(j'))$$

$$= \sum_{j=1}^{J} p_j^2 \frac{1 - p_j}{N_j} S^2(j) - \sum_{j \neq j'} p_j p_{j'} \frac{1}{2N} \{S^2(j) + S^2(j') - S^2(j-j')\}$$

$$= \frac{1}{N} \left\{ \sum_{j=1}^{J} p_j (1 - p_j) S^2(j) - \frac{1}{2} \sum_{j \neq j'} p_j p_{j'} S^2(j) - \frac{1}{2} \sum_{j \neq j'} p_j p_{j'} S^2(j-j') \right\}.$$ 

Because

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

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

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

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

Second, the covariance between $\bar{Y}_{\text{obs}}(j)$ and $\bar{Y}_{\text{obs}}(j')$ is

$$\text{cov}(\bar{Y}_{\text{obs}}(j), \bar{Y}_{\text{obs}}(j')) = p_j \text{var}(\bar{Y}_{\text{obs}}(j)) + \sum_{j' \neq j} p_{j'} \text{cov}(\bar{Y}_{\text{obs}}(j), \bar{Y}_{\text{obs}}(j'))$$

$$= \frac{1}{N} (1 - p_j) S^2(j) - \frac{1}{2N} \sum_{j' \neq j} p_{j'} \{S^2(j) + S^2(j') - S^2(j-j')\}.$$ 

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

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

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

$$\text{cov}(\bar{Y}_{\text{obs}}(j), \bar{Y}_{\text{obs}}(j')) = (2N)^{-1} \left\{ 2(1 - p_j) S^2(j) - (1 - p_j) S^2(j) - S^2 + p_j S^2(j) + \Delta_j \right\}$$

$$= (2N)^{-1} \left\{ S^2(j) - S^2 + \Delta_j \right\}.$$ 

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

$$\text{var}(\bar{Y}_{\text{obs}}(j) - \bar{Y}_{\text{obs}}) = \text{var}(\bar{Y}_{\text{obs}}(j)) + \text{var}(\bar{Y}_{\text{obs}}) - 2\text{cov}(\bar{Y}_{\text{obs}}(j), \bar{Y}_{\text{obs}})$$

$$= \frac{1}{N} \left\{ \frac{1 - p_j}{p_j} S^2(j) + \Delta - S^2(j) + S^2 - \Delta_j \right\}.$$
Finally, the expectation of the treatment sum of squares is

\[ E(\text{SSTre}) = E \left( \sum_{j=1}^{J} N_j \{ \bar{Y}^{\text{obs}}_j - \bar{Y}^{\text{obs}} \}^2 \right) \]

\[ = \sum_{j=1}^{J} N_j \left\{ \bar{Y}_j - \sum_{j=1}^{J} p_j \bar{Y}_j \right\}^2 + \sum_{j=1}^{J} p_j \left\{ \frac{1-p_j}{p_j} S^2(j) + \Delta - S^2(j) + S^2 - \Delta \right\}, \]

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

\[ E(\text{SSTre}) = \sum_{j=1}^{J} N_j \left\{ \bar{Y}_j - \sum_{j=1}^{J} p_j \bar{Y}_j \right\}^2 + \sum_{j=1}^{J} (1-p_j) S^2(j) + \Delta - S^2 + S^2 - 2\Delta \]

\[ = \sum_{j=1}^{J} N_j \left\{ \bar{Y}_j - \sum_{j=1}^{J} p_j \bar{Y}_j \right\}^2 + \sum_{j=1}^{J} (1-p_j) S^2(j) - \Delta. \quad \square \]

Under Neyman’s null, i.e., \( \bar{Y}_1 = \cdots = \bar{Y}_J \), or, equivalently, \( \bar{Y}_j - \sum_{j=1}^{J} p_j \bar{Y}_j = 0 \) for all \( j \), the expectation of the treatment sum of squares further reduces to

\[ E(\text{SSTre}) = \sum_{j=1}^{J} (1-p_j) S^2(j) - \Delta. \]

Because \( \{Y^{\text{obs}}_i : W_i(j) = 1\} \) 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., \( E\{s^2_{\text{obs}}(j)\} = S^2(j) \). Therefore, the mean of the residual sum of squares is

\[ E(\text{SSRes}) = E \left\{ (N_j - 1)s^2_{\text{obs}}(j) \right\} = \sum_{j=1}^{J} (N_j - 1)S^2(j). \]

**Proof of Corollary 1.** Additivity implies \( S^2 = S^2(j) \) for all \( j \) and \( \Delta = 0 \), and the conclusions follow. \( \square \)

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

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

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

Moreover, under Neyman’s null, \( E(\text{SSRes}) \) 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 \( E(\text{MSRes}) - E(\text{MSTre}) = \Delta/(J-1) \geq 0. \) \( \square \)
Proof of Corollary 3. Under Neyman’s null,

\[ E(\text{MSRes}) - E(\text{MSTre}) = \sum_{j=1}^{J} \left( \frac{N_j - 1}{N - J} - \frac{1 - p_j}{J - 1} \right) S^2(j) + \frac{\Delta}{J - 1} \]

\[ = \frac{(N - 1)J}{(J - 1)(N - J)} \sum_{j=1}^{J} (p_j - J^{-1})S^2(j) + \frac{\Delta}{J - 1}. \]

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 \( A \ast B \) to denote the element-wise product of \( A \) and \( B \), i.e., the \((i, j)\)-th element of \( A \ast B \) is the product of the \((i, j)\)-th elements of \( A \) and \( B \), \( A_{ij}B_{ij} \).

**Lemma S3.** If \( X \sim N_{J}(0, A) \), then \( X^T B X \sim \sum_{j=1}^{J} \lambda_j \xi_j \), where the \( \xi_j \)'s are iid \( \chi^2_1 \), and the \( \lambda_j \)'s are eigenvalues of \( BA \).

**Lemma S4.** If \( A \) is positive semidefinite and \( B \) is a correlation matrix, then the maximum eigenvalue of \( A \ast B \) does not exceed the maximum eigenvalue of \( A \).

**Proof of Theorem 3.** We first prove the result under Neyman’s null, and then view the result under Fisher’s null as a special case. Let \( Q_j = N_j/S^2(j) \) for \( j = 1, \ldots, J \), and \( Q = \sum_{j=1}^{J} Q_j \).

Define \( q^T_w = (Q_1^{1/2}, \ldots, Q_J^{1/2})/Q^{1/2} \), and \( P_w = I_J - q_wq^T_w \) is a projection matrix of rank \( J - 1 \). Let \( \bar{Y}_{\text{obs}} = Q^{-1} \sum_{j=1}^{J} Q_j \bar{Y}_{\text{obs}}(j) \) be a weighted average of the means of the observed outcomes. By Slutsky’s Theorem, \( X^2 \) has the same asymptotical distribution as

\[ X^2_0 = \sum_{j=1}^{J} Q_j \left\{ \bar{Y}_{\text{obs}}(j) - \bar{Y}_{\text{obs}} \right\}^2. \]

Define \( \rho_{jk} \) 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 \( \rho_{jk} \). The finite population central limit theorem (Hájek, 1960) implies

\[ V_0 \equiv \begin{bmatrix} Q_1^{1/2} & \{\bar{Y}_{\text{obs}}(1) - \bar{Y}(1)\} \\
Q_2^{1/2} & \{\bar{Y}_{\text{obs}}(2) - \bar{Y}(2)\} \\
\vdots \\
Q_J^{1/2} & \{\bar{Y}_{\text{obs}}(J) - \bar{Y}(J)\} \\ 
\end{bmatrix} \sim N_J \begin{bmatrix} 0, \\
\begin{pmatrix} 1 - p_{11} & -p_{12}/2 & \cdots & -p_{1J}/2 \\
-p_{21}/2 & 1 - p_{22} & \cdots & -p_{2J}/2 \\
\vdots & \vdots & \ddots & \vdots \\
-p_{J1}/2 & -p_{J2}/2 & \cdots & 1 - p_{JJ} \end{pmatrix} = P \ast R \end{bmatrix}, \]

where \( P = I_J - qq^T \) 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.
The treatment sum of squares reduces to
\[ X_0^2 = \sum_{j=1}^{J} Q_j \{ \bar{Y}_{\text{obs}}^{(j)} - \bar{Y}^{(j)} \}^2 - \frac{1}{Q} \left[ \sum_{j=1}^{J} Q_j \{ \bar{Y}_{\text{obs}}^{(j)} - \bar{Y}^{(j)} \} \right]^2, \]
which can be further rewritten as a quadratic form (cf. Chung & Romano, 2013)
\[ X_0^2 = V_0^T (I_J - q_w q_w^T) V_0 \equiv V_0^T P_w V_0. \]

According to Lemma S3, \( X_0^2 \) has asymptotic distribution \( \sum_{j=1}^{J-1} \lambda_j \xi_j \), where the \( \lambda_j \)'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, \( X_0^2 \sim \sum_{j=1}^{J-1} \lambda_j \xi_j \), where \( \lambda_j \leq 1 \) for all \( j \). Because the \( \chi_j^2 \) 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 \( \chi_{J-1}^2 \).

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 \( Y_i(1) = \cdots = Y_i(J) = Y_{\text{obs}} \). This restricts \( S^2(j) \) to be constant, and the correlation coefficients between potential outcomes to be 1. Correspondingly, \( P_w = P, R = 1, J \), 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 \( \chi_{J-1}^2 \), because \( PP = P \) has \( J - 1 \) nonzero eigenvalues and all of them are 1.

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

**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 \( S^2(1) = \cdots = S^2(J) = S^2 \), then \( \bar{Y}_{\text{obs}}^{(1)} = \bar{Y}_{\text{obs}}^{(2)} \) and
\[ X_0^2 = \sum_{j=1}^{J} \{ \bar{Y}_{\text{obs}}^{(j)} - \bar{Y}_{\text{obs}} \}^2 \quad = \quad \frac{\text{SSTre}}{S^2}. \]
Because \( \text{MSRes} = \sum_{j=1}^{J} (N_j - 1) s_{\text{obs}}^2(j) / (N - J) \) converges to \( S^2 \) in probability, Slutsky’s Theorem implies
\[ (J - 1)F = \frac{\text{SSTre}}{\text{MSRes}} \sim \frac{\text{SSTre}}{S^2}. \]
Therefore, \( (J - 1)F \sim X^2 \). Their asymptotic distributions under Fisher’s null and Neyman’s null follow from Theorems 1 and 2.

**Proof of Corollary 5.** First, we discuss \( F \). Because \( \bar{Y}_{\text{obs}}^{(1)} = p_1 \bar{Y}_{\text{obs}}^{(1)} + p_2 \bar{Y}_{\text{obs}}^{(2)} \), we have
\[ \bar{Y}_{\text{obs}}^{(1)} - \bar{Y}_{\text{obs}} = p_2 \bar{\tau}(1, 2), \quad \bar{Y}_{\text{obs}}^{(2)} - \bar{Y}_{\text{obs}} = -p_1 \bar{\tau}(1, 2). \]

The treatment sum of squares reduces to
\[ \text{SSTre} = N_1 \{ \bar{Y}_{\text{obs}}^{(1)} - \bar{Y}_{\text{obs}} \}^2 + N_2 \{ \bar{Y}_{\text{obs}}^{(2)} - \bar{Y}_{\text{obs}} \}^2 = N p_1 p_2 \bar{\tau}^2(1, 2), \]
and the residual sum of squares reduces to $SS_{Res} = (N_1 - 1)s_{obs}^2(1) + (N_2 - 1)s_{obs}^2(2)$. Therefore, the $F$ statistic reduces to

$$F = \frac{SS_{Treat}/(N-2)}{SS_{Res}/(N-2)} = \frac{\hat{\tau}^2(1,2)}{\frac{N(N_1-1)}{(N-2)N_1N_2} s_{obs}^2(1) + \frac{N(N_2-1)}{(N-2)N_1N_2} s_{obs}^2(2)} \approx \frac{\hat{\tau}^2(1,2)}{s_{obs}^2(1)/N_2 + s_{obs}^2(2)/N_1},$$

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 $F \sim F_{1,N-2} \sim \chi^2_1$ under Fisher’s null. However, under Neyman’s null, Neyman (1923), coupled with the finite population central limit theorem (Hájek, 1960), imply

$$\left\{ \frac{s^2(1)}{N_1} + \frac{s^2(2)}{N_2} - \frac{s^2(1-2)}{N} \right\}^{1/2} \sim \mathcal{N}(0, 1),$$

and $s_{obs}^2(j) \rightarrow S^2(j)$ in probability ($j = 1, 2$). Therefore, the asymptotic distribution of $F$ under Neyman’s null is $F \sim C_1 \chi^2_1$, where

$$C_1 = \lim_{N \rightarrow +\infty} \frac{S^2(1)/N_1 + S^2(2)/N_2 - S^2(1-2)/N}{S^2(1)/N_2 + S^2(2)/N}.$$

Second, we discuss $X^2$. Because

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

we have

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

$$\bar{Y}_{\text{obs}}(2) - \bar{Y}_{\text{obs}} = -\frac{N_1}{s_{obs}^2(1)} \hat{\tau}^2(1,2) / \left\{ \frac{N_1}{s_{obs}^2(1)} + \frac{N_2}{s_{obs}^2(2)} \right\}.$$

Therefore, the $X^2$ statistic reduces to

$$X^2 = \left\{ \frac{N_1}{s_{obs}^2(1)} \frac{N_2^2}{s_{obs}^2(2)} \hat{\tau}^2(1,2) + \frac{N_2}{s_{obs}^2(2)} \frac{N_1^2}{s_{obs}^2(1)} \hat{\tau}^2(1,2) \right\} / \left\{ \frac{N_1}{s_{obs}^2(1)} + \frac{N_2}{s_{obs}^2(2)} \right\}^2$$

$$= \frac{\hat{\tau}^2(1,2)}{s_{obs}^2(1)/N_1 + s_{obs}^2(2)/N_2}.$$

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

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

The constant $C_2$ is smaller than or equal to 1 with equality holding if the limit of $S^2(1-2)$ is zero, i.e., the unit-level treatment effects are constant asymptotically. \qed
A randomization-based perspective of analysis of variance

Fig. S1. Distributions of $\chi^2$. The histograms are the sampling distributions, the dotted lines are the asymptotic distributions, and the solid lines are the $\chi^2$ distribution.

APPENDIX B. NUMERICAL EXAMPLES

Example S1. 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

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

(S2)

and the second set of potential outcomes from

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

(S3)

After generating the potential outcomes, we center the $Y_i(j)$’s by subtracting the mean to make $\bar{Y}_i = 0$ for all $j$ so that Neyman’s null holds. Figure S1 shows the distributions of $\chi^2$ 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 $\chi^2$. 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 $\chi^2$ becomes larger.

Example S2. 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 Neyman’s null, 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 S3. 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 treat-
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 |    |
| **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 |

ment 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 of 0.05. In this unbalanced experiment, the Fisher randomization test using $F$ is less powerful.

**APPENDIX C. MORE SIMULATION STUDIES WITH NONNORMAL OUTCOMES**

**C.1. Type I error of the Fisher randomization test using $F$**

In this subsection, we use simulation studies to evaluate the finite sample performance of the Fisher randomization test using $F$ under Neyman’s null. We consider the following three cases, where $E$ 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) $Y_i(1) \sim E$, $Y_i(2) \sim E/0.7$, $Y_i(3) \sim E/0.5$; and (S1.2) $Y_i(1) \sim E$, $Y_i(2) \sim E/0.5$, $Y_i(3) \sim E/0.3$. 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) $Y_i(1) \sim E$, $Y_i(2) = 2Y_i(1)$, $Y_i(3) = 3Y_i(1)$; and (S2.2) $Y_i(1) \sim E$, $Y_i(2) = 3Y_i(1)$, $Y_i(3) = 5Y_i(1)$. These potential outcomes are standardized to have zero means. In this case, $p_1 < p_2 < p_3$ and $S^2_i(1) < S^2_i(2) < S^2_i(3)$.

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) $Y_i(1) \sim E$, $Y_i(2) = 1.2Y_i(1)$, $Y_i(3) = 1.5Y_i(1)$; and (S3.2) $Y_i(1) \sim E$, $Y_i(2) = 1.5Y_i(1)$, $Y_i(3) = 2Y_i(1)$. These potential outcomes are standardized to have zero means. In this case, $p_1 > p_2 > p_3$ and $S^2_i(1) < S^2_i(2) < S^2_i(3)$.

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.
A randomization-based perspective of analysis of variance

(a) Balanced experiments: grey histogram for $Y_i(1) \sim \mathcal{E}$, $Y_i(2) \sim \mathcal{E}/0.7$, $Y_i(3) \sim \mathcal{E}/0.5$, and white histogram with borders for $Y_i(1) \sim \mathcal{E}$, $Y_i(2) \sim \mathcal{E}/0.3$.

(b) Unbalanced experiments: grey histogram for $Y_i(1) \sim \mathcal{E}$, $Y_i(2) = 2Y_i(1)$, $Y_i(3) = 3Y_i(1)$, and white histogram with borders for $Y_i(1) \sim \mathcal{E}$, $Y_i(2) = 3Y_i(1), Y_i(3) = 5Y_i(1)$.

(c) Unbalanced experiments: grey histogram for $Y_i(1) \sim \mathcal{E}$, $Y_i(2) = 1.2Y_i(1), Y_i(3) = 1.5Y_i(1)$, and white histogram with borders for $Y_i(1) \sim \mathcal{E}$, $Y_i(2) = 1.5Y_i(1), Y_i(3) = 2Y_i(1)$.

Fig. S2. Histograms of the $p$-values over repeated sampling, based on the Fisher randomization test using $F$, under Neyman’s null

C.2. Type I error of the Fisher randomization test using $X^2$

We follow §6-2, generate the same data as Appendix C.1, and obtain the same conclusions about the Fisher randomization test using $X^2$, because Figures 2 and S3 exhibit the same pattern.

C.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 $(0, 0.5, 1)$.

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 $S^2(1) < S^2(2) < S^2(3)$.
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 \(S_1^2(1) < S_2^2(2) < S_3^2(3)\).

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 and 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 \(X^2\).

Fig. S3. Histograms of the \(p\)-values over repeated sampling, based on the Fisher randomization test using \(X^2\), under Neyman’s null

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.

HÁJEK, J. (1960). Limiting distributions in simple random sampling from a finite population. Publications of the Mathematics Institute of the Hungarian Academy of Science 5, 361–74.
A randomization-based perspective of analysis of variance

(a) Balanced experiments, $Y_i(1) \sim E$, $Y_i(2) \sim E/0.7$, $Y_i(3) \sim E/0.5$, and these potential outcomes are independently generated, and shifted to have means $(0, 0.5, 1)$: grey histogram for the FRT using $X^2$, and white histogram with borders for the FRT using $F$.

(b) Unbalanced experiments, $Y_i(1) \sim E$ and standardized to have mean zero, $Y_i(2) = 3Y_i(1) + 1$ and $Y_i(3) = 5Y_i(1) + 2$: grey histogram for the Fisher randomization test using $X^2$, and white histogram with borders for the Fisher randomization test using $F$.

(c) Unbalanced experiments, the same potential outcomes as in Figure 4(b): grey histogram for the Fisher randomization test using $X^2$, and white histogram with borders for the Fisher randomization test using $F$.

Fig. S4. Histograms of the $p$-values over repeated sampling, based on the Fisher randomization tests using $F$ and $X^2$, under alternative hypotheses

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

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.