Statistical Science
2017, Vol. 32, No. 3, 331–345
DOI: 10.1214/16-STS571

A Paradox from Randomization-Based Causal Inference[1]

Peng Ding

Abstract. Under the potential outcomes framework, causal effects are defined as comparisons between potential outcomes under treatment and control. To infer causal effects from randomized experiments, Neyman proposed to test the null hypothesis of zero average causal effect (Neyman's null), and Fisher proposed to test the null hypothesis of zero individual causal effect (Fisher's null). Although the subtle difference between Neyman's null and Fisher's null has caused a lot of controversies and confusions for both theoretical and practical statisticians, a careful comparison between the two approaches has been lacking in the literature for more than eighty years. We fill this historical gap by making a theoretical comparison between them and highlighting an intriguing paradox that has not been recognized by previous researchers. Logically, Fisher's null implies Neyman's null. It is therefore surprising that, in actual completely randomized experiments, rejection of Neyman's null does not imply rejection of Fisher's null for many realistic situations, including the case with constant causal effect. Furthermore, we show that this paradox also exists in other commonly-used experiments, such as stratified experiments, matched-pair experiments and factorial experiments. Asymptotic analyses, numerical examples and real data examples all support this surprising phenomenon. Besides its historical and theoretical importance, this paradox also leads to useful practical implications for modern researchers.

Key words and phrases: Average null hypothesis, Fisher randomization test, potential outcome, randomized experiment, repeated sampling property, sharp null hypothesis.

1. INTRODUCTION

Ever since Neyman's seminal work, the potential outcomes framework (Neyman, 1923/1990; Rubin, 1974) has been widely used for causal inference in randomized experiments (e.g., Neyman, 1935; Hinkelmann and Kempthorne, 2008; Imbens and Rubin, 2015). The potential outcomes framework permits making inference about a finite population of interest, with all potential outcomes fixed and randomness coming solely from the physical randomization of the treatment assignments. Historically, Neyman (1923/1990) was interested in obtaining an unbiased estimator with a repeated sampling evaluation of the average causal effect, which corresponded to a test for the null hypothesis of zero average causal effect. On the other hand, Fisher (1935a) focused on testing the sharp null hypothesis of zero individual causal effect, and proposed the Fisher Randomization Test (FRT). Both Neymanian and Fisherian approaches are randomization-based inference, relying on the physical randomization of the experiments. Neyman's null and Fisher's null are closely related to each other: the latter implies the former, and they are equivalent under the constant causal effect assumption. Both approaches have existed for

Peng Ding is Assistant Professor, Department of Statistics, University of California, Berkeley, 425 Evans Hall, Berkeley, California 94720, USA (e-mail: pengdingpku@berkeley.edu).

[1]Discussed in 10.1214/16-STS582, 10.1214/16-STS590, 10.1214/16-STS600, 10.1214/17-STS610; rejoinder at 10.1214/17-STS571REJ.

many decades and are widely used in current statistical practice. They are now introduced at the beginning of many causal inference courses and textbooks (e.g., Rubin, 2004; Imbens and Rubin, 2015). Unfortunately, however, a detailed comparison between them has not been made in the literature.

In the past, several researchers (e.g., Rosenbaum, 2002, page 40) believed that "in most cases, their disagreement is entirely without technical consequence: the same procedures are used, and the same conclusions are reached." However, we show, via both numerical examples and theoretical investigations, that the rejection rate of Neyman's null is higher than that of Fisher's null in many realistic randomized experiments, using their own testing procedures. In fact, Neyman's method is always more powerful if there is a nonzero constant causal effect, the very alternative most often used for the Fisher-style inference. This finding immediately causes a seeming paradox: logically, Fisher's null implies Neyman's null, so how can we fail to reject the former while rejecting the latter?

We demonstrate that this surprising paradox is not unique to completely randomized experiments, because it also exists in other commonly-used experiments such as stratified experiments, matched-pair experiments and factorial experiments. The result for factorial experiments helps to explain the surprising empirical evidence in Dasgupta, Pillai and Rubin (2015) that interval estimators for factorial effects obtained by inverting a sequence of FRTs are often wider than Neymanian confidence intervals.

The paper proceeds as follows. We review Neymanian and Fisherian randomization-based causal inference in Section 2 under the potential outcomes framework. In Section 3, we use both numerical examples and asymptotic analyses to demonstrate the paradox from randomization-based causal inference in completely randomized experiments. Section 4 shows that a similar paradox also exists in other commonly-used experiments. Section 5 extends the scope of the paper to improved variance estimators and comments on the choices of test statistics. Section 6 illustrates the asymptotic theory of this paper with some finite sample real-life examples. We conclude with a discussion in Section 7, and relegate all the technical details to the Supplementary Material (Ding, 2017).

2. RANDOMIZED EXPERIMENTS AND RANDOMIZATION INFERENCE

We first introduce notation for causal inference in completely randomized experiments, and then review the Neymanian and Fisherian perspectives for causal inference.

2.1 Completely Randomized Experiments and Potential Outcomes

Consider N units in a completely randomized experiment. Throughout our discussion, we make the Stable Unit Treatment Value Assumption (SUTVA; Cox, 1992; Rubin, 1980), that is, there is only one version of the treatment, and interference between subjects is absent. The SUTVA allows us to define the potential outcome of unit i under treatment t as $Y_i(t)$, with $t = 1$ for treatment and $t = 0$ for control. The individual causal effect is defined as a comparison between two potential outcomes, for example, $\tau_i = Y_i(1) - Y_i(0)$. However, for each subject i, we can observe only one of $Y_i(1)$ and $Y_i(0)$ with the other one missing, and the individual causal effect τ_i is not observable. The observed outcome is a deterministic function of the treatment assignment T_i and the potential outcomes, namely, $Y_i^{\text{obs}} = T_iY_i(1) + (1 - T_i)Y_i(0)$. Let $\boldsymbol{Y}^{\text{obs}} = (Y_1^{\text{obs}}, \ldots, Y_N^{\text{obs}})'$ be the observed outcome vector. Let $\boldsymbol{T} = (T_1, \ldots, T_N)'$ be the treatment assignment vector, and $\boldsymbol{t} = (t_1, \ldots, t_N)' \in \{0, 1\}^N$ be its realization. Completely randomized experiments satisfy $\text{pr}(\boldsymbol{T} = \boldsymbol{t}) = N_1!N_0!/N!$, if $\sum_{i=1}^N t_i = N_1$ and $N_0 = N - N_1$. Note that in Neyman's (1923/1990) potential outcomes framework, all the potential outcomes are fixed numbers, and only the treatment assignment vector is random. In general, we can view this framework with fixed potential outcomes as conditional inference given the values of the potential outcomes. In the early literature, Neyman (1935) and Kempthorne (1955) are two research papers, and Kempthorne (1952), Hodges and Lehmann (1964), Chapter 9, and Scheffé (1959), Chapter 9, are three textbooks using potential outcomes for analyzing experiments.

2.2 Neymanian Inference for the Average Causal Effect

Neyman (1923/1990) was interested in estimating the finite population average causal effect:

$$\tau = \frac{1}{N}\sum_{i=1}^N \tau_i = \frac{1}{N}\sum_{i=1}^N \{Y_i(1) - Y_i(0)\} = \bar{Y}_1 - \bar{Y}_0,$$

where $\bar{Y}_t = \sum_{i=1}^N Y_i(t)/N$ is the finite population average of the potential outcomes $\{Y_i(t) : i = 1, \ldots, N\}$. He proposed an unbiased estimator

$$\widehat{\tau} = \bar{Y}_1^{\text{obs}} - \bar{Y}_0^{\text{obs}} \tag{1}$$

for τ, where $\bar{Y}_t^{\text{obs}} = \sum_{\{i:T_i=t\}} Y_i^{\text{obs}}/N_t$ is the sample mean of the observed outcomes under treatment t. The sampling variance of $\widehat{\tau}$ over all possible randomizations is

$$\text{var}(\widehat{\tau}) = \frac{S_1^2}{N_1} + \frac{S_0^2}{N_0} - \frac{S_\tau^2}{N}, \tag{2}$$

depending on $S_t^2 = \sum_{i=1}^N \{Y_i(t) - \bar{Y}_t\}^2/(N-1)$, the finite population variance of the potential outcomes $\{Y_i(t) : i = 1, \ldots, N\}$, and $S_\tau^2 = \sum_{i=1}^N (\tau_i - \tau)^2/(N-1)$, the finite population variance of the individual causal effects $\{\tau_i : i = 1, \ldots, N\}$. Note that previous literature sometimes used slightly different notation for S_τ^2, for example, S_{1-0}^2 (Rubin, 1990; Imbens and Rubin, 2015). Because we can never jointly observe the pair of potential outcomes for each unit, the variance of individual causal effects, S_τ^2, is not identifiable from the observed data. Recognizing this difficulty, Neyman (1923/1990) suggested using

$$\widehat{V}(\text{Neyman}) = \frac{s_1^2}{N_1} + \frac{s_0^2}{N_0} \tag{3}$$

as an estimator for $\text{var}(\widehat{\tau})$, where $s_t^2 = \sum_{\{i:T_i=t\}} (Y_i^{\text{obs}} - \bar{Y}_t^{\text{obs}})^2/(N_t - 1)$ is the sample variance of the observed outcomes under treatment t. However, Neyman's variance estimator overestimates the true variance, in the sense that $E\{\widehat{V}(\text{Neyman})\} \geq \text{var}(\widehat{\tau})$, with equality holding if and only if the individual causal effects are constant: $\tau_i = \tau$ or $S_\tau^2 = 0$. The randomization distribution of $\widehat{\tau}$ enables us to test the following Neyman's null hypothesis:

$$H_0(\text{Neyman}) : \tau = 0.$$

Under $H_0(\text{Neyman})$ and based on the Normal approximation in Section 3.3, the p-value from Neyman's approach can be approximated by

$$p(\text{Neyman}) \approx 2\Phi\left\{-\frac{|\widehat{\tau}^{\text{obs}}|}{\sqrt{\widehat{V}(\text{Neyman})}}\right\}, \tag{4}$$

where $\widehat{\tau}^{\text{obs}}$ is the realized value of $\widehat{\tau}$, and $\Phi(\cdot)$ is the cumulative distribution function of the standard Normal distribution. With nonconstant individual causal effects, Neyman's test for the null hypothesis of zero average causal effect tends to be "conservative," in the sense that it rejects less often than the nominal significance level when the null is true.

2.3 Fisherian Randomization Test for the Sharp Null

Fisher (1935a) was interested in testing the following sharp null hypothesis:

$$H_0(\text{Fisher}) : Y_i(1) = Y_i(0), \quad \forall i = 1, \ldots, N.$$

This null hypothesis is sharp because all missing potential outcomes can be uniquely imputed under $H_0(\text{Fisher})$. The sharp null hypothesis implies that $Y_i(1) = Y_i(0) = Y_i^{\text{obs}}$ are all fixed constants, so that the observed outcome for subject i is Y_i^{obs} under any treatment assignment. Although we can perform randomization tests using any test statistics capturing the deviation from the null, we will first focus on the randomization test using $\widehat{\tau}(\boldsymbol{T}, \boldsymbol{Y}^{\text{obs}}) = \widehat{\tau}$ as the test statistic, in order to make a direct comparison to Neyman's method. We will comment on other choices of test statistics in Section 5.1. Again, the randomness of $\widehat{\tau}(\boldsymbol{T}, \boldsymbol{Y}^{\text{obs}})$ comes solely from the randomization of the treatment assignment $\boldsymbol{T}$, because $\boldsymbol{Y}^{\text{obs}}$ is a set of constants under the sharp null. The p-value for the two-sided test under the sharp null is

$$p(\text{Fisher}) = \text{pr}\{|\widehat{\tau}(\boldsymbol{T}, \boldsymbol{Y}^{\text{obs}})| \geq |\widehat{\tau}^{\text{obs}}| | H_0(\text{Fisher})\},$$

measuring the extremeness of $\widehat{\tau}^{\text{obs}}$ with respect to the null distribution of $\widehat{\tau}(\boldsymbol{T}, \boldsymbol{Y}^{\text{obs}})$ over all possible randomizations. In practice, we can approximate the exact distribution of $\widehat{\tau}(\boldsymbol{T}, \boldsymbol{Y}^{\text{obs}})$ by Monte Carlo. We draw, repeatedly and independently, completely randomized treatment assignment vectors $\{\boldsymbol{T}^1, \ldots, \boldsymbol{T}^M\}$, and with large M the p-value can be well approximated by

$$p(\text{Fisher}) \approx \frac{1}{M} \sum_{m=1}^M I\{|\widehat{\tau}(\boldsymbol{T}^m, \boldsymbol{Y}^{\text{obs}})| \geq |\widehat{\tau}^{\text{obs}}|\}.$$

Eden and Yates (1933) performed the FRT empirically, and Welch (1937) and Pitman (1937, 1938) studied its theoretical properties. Rubin (1980) first used the name "sharp null," and Rubin (2004) viewed the FRT as a "stochastic proof by contradiction." For more discussion about randomization tests, please see Rosenbaum (2002) and Edgington and Onghena (2007).

3. A PARADOX FROM NEYMANIAN AND FISHERIAN INFERENCE

Neymanian and Fisherian approaches reviewed in Section 2 share some common properties but differ fundamentally. They both rely on the distribution induced by the physical randomization, but they test two

different null hypotheses and evolve from different statistical philosophies. In this section, we first compare Neymanian and Fisherian approaches using simple numerical examples, highlighting a surprising paradox. We then explain the paradox via an asymptotic analysis.

3.1 Initial Numerical Comparisons

We compare Neymanian and Fisherian approaches using numerical examples with both balanced and unbalanced experiments. In our simulations, the potential outcomes are fixed, and the simulations are carried out over randomization distributions induced by the treatment assignments. The significance level is 0.05, and M is 10^5 for the FRT.

EXAMPLE 1 (Balanced experiments with $N_1 = N_0$). The potential outcomes are independently generated from Normal distributions $Y_i(1) \sim N(1/10, 1/16)$ and $Y_i(0) \sim N(0, 1/16)$, for $i = 1, \ldots, 100$. The individual causal effects are not constant, with $S_\tau^2 = 0.125$. Further, once drawn from the Normal distributions above, the potential outcomes are fixed. We repeatedly generate 1000 completely randomized treatment assignments with $N = 100$ and $N_1 = N_0 = 50$. For each treatment assignment, we obtain the observed outcomes and implement two tests for Neyman's null and Fisher's null. As shown in Table 1(a), it never happens that we reject Fisher's null but fail to reject Neyman's null. However, we reject Neyman's null but fail to reject Fisher's null in 15 instances.

EXAMPLE 2 (Unbalanced experiments with $N_1 \neq N_0$). The potential outcomes are independently generated from Normal distributions $Y_i(1) \sim N(1/10, 1/4)$ and $Y_i(0) \sim N(0, 1/16)$, for $i = 1, \ldots, 100$. The individual causal effects are not constant, with $S_\tau^2 = 0.313$. They are kept as fixed throughout the simulations. The unequal variances are designed on purpose, and we will reveal the reason for choosing them later in Example 3 of Section 3.4. We repeatedly generate 1000 completely randomized treatment assignments with $N = 100$, $N_1 = 70$ and $N_0 = 30$. After obtaining each observed data set, we perform two hypothesis testing procedures, and summarize the results in Table 1(b). The pattern in Table 1(b) is more striking than in Table 1(a), because it happens 62 times in Table 1(b) that we reject Neyman's null but fail to reject Fisher's null. For this particular set of potential outcomes, Neyman's testing procedure has a power $62/1000 = 0.062$, slightly larger than 0.05, but Fisher's testing procedure has a power $8/1000 = 0.008$, much smaller than 0.05 even though the sharp null is not true. We will explain in Section 3.4 the reason why the FRT could have a power even smaller than the significance level under some alternative hypotheses.

3.2 Statistical Inference, Logic and Paradox

Logically, Fisher's null implies Neyman's null. Therefore, Fisher's null should be rejected if Neyman's null is rejected. However, this is not always true from the results of statistical inference in completely randomized experiments. We observed in the numerical examples above that it can be the case that

$$p(\text{Neyman}) < \alpha_0 < p(\text{Fisher}), \tag{5}$$

in which we reject Neyman's null but not Fisher's null, if we choose the significance level to be α_0 (e.g.,

TABLE 1
Numerical examples

(a) Balanced experiments with $N_1 = N_0 = 50$, corresponding to Example 1

	Not reject H_0(Fisher)	**Reject H_0(Fisher)**	
Not reject H_0(Neyman)	488	0	
Reject H_0(Neyman)	15	497	power(Neyman) = 0.512
		power(Fisher) = 0.497	

(b) Unbalanced experiments with $N_1 = 70$ and $N_0 = 30$, corresponding to Example 2

	Not reject H_0(Fisher)	**Reject H_0(Fisher)**	
Not reject H_0(Neyman)	930	0	
Reject H_0(Neyman)	62	8	power(Neyman) = 0.070
		power(Fisher) = 0.008	

$\alpha_0 = 0.05$). When (5) holds, an awkward logical problem appears. In the remaining part of this section, we will theoretically explain the empirical findings in Section 3.1 and the consequential logical problem.

3.3 Asymptotic Evaluations

While Neyman's testing procedure has an explicit form, the FRT is typically approximated by Monte Carlo. In order to compare them, we first discuss the asymptotic Normalities of $\widehat{\tau}$ and the randomization test statistic $\widehat{\tau}(\boldsymbol{T}, \boldsymbol{Y}^{\text{obs}})$. We provide a simplified way of doing variance calculation and a short proof for asymptotic Normalities of both $\widehat{\tau}$ and $\widehat{\tau}(\boldsymbol{T}, \boldsymbol{Y}^{\text{obs}})$, based on the finite population Central Limit Theorem (CLT; Hoeffding, 1952; Hájek, 1960; Lehmann, 1999; Freedman, 2008). Although recent work by Li and Ding (2017) for general cases can imply our result, we provide an elementary discussion of the problem. Before the formal asymptotic results, it is worth mentioning the exact meaning of "asymptotics" in the context of finite population causal inference. We need to embed the finite population of interest into a hypothetical infinite sequence of finite populations with increasing sizes, and also require the proportions of the treatment units to converge to a fixed value. Essentially, all the population quantities (e.g., τ, S_1^2, etc.) should have the index N, and all the sample quantities (e.g., $\widehat{\tau}$, s_1^2, etc.) should have double indices N and N_1. However, for the purpose of notational simplicity, we sacrifice a little bit of mathematical precision and drop all the indices in our discussion.

THEOREM 1. *As $N \to \infty$, the sampling distribution of $\widehat{\tau}$ satisfies*

$$\frac{\widehat{\tau} - \tau}{\sqrt{\text{var}(\widehat{\tau})}} \xrightarrow{d} \mathcal{N}(0, 1).$$

In practice, the true variance $\text{var}(\widehat{\tau})$ is replaced by its "conservative" estimator $\widehat{V}(\text{Neyman})$, and the resulting test rejects less often than the nominal significance level on average. While the asymptotics for the Neymanian unbiased estimator $\widehat{\tau}$ does not depend on the null hypothesis, the following asymptotic Normality for $\widehat{\tau}(\boldsymbol{T}, \boldsymbol{Y}^{\text{obs}})$ is true only under the sharp null hypothesis.

THEOREM 2. *Under $H_0(\text{Fisher})$ and as $N \to \infty$, the null distribution of $\widehat{\tau}(\boldsymbol{T}, \boldsymbol{Y}^{\text{obs}})$ satisfies*

$$\frac{\widehat{\tau}(\boldsymbol{T}, \boldsymbol{Y}^{\text{obs}})}{\sqrt{\widehat{V}(\text{Fisher})}} \xrightarrow{d} \mathcal{N}(0, 1),$$

where $\bar{Y}^{\text{obs}} = \sum_{i=1}^{N} Y_i^{\text{obs}}/N$, $s^2 = \sum_{i=1}^{N}(Y_i^{\text{obs}} - \bar{Y}^{\text{obs}})^2/(N-1)$, and $\widehat{V}(\text{Fisher}) = Ns^2/(N_1 N_0)$.

Therefore, the p-value under $H_0(\text{Fisher})$ can be approximated by

$$p(\text{Fisher}) \approx 2\Phi\left\{-\frac{|\widehat{\tau}^{\text{obs}}|}{\sqrt{\widehat{V}(\text{Fisher})}}\right\}. \tag{6}$$

From (4) and (6), the asymptotic p-values obtained from Neymanian and Fisherian approaches differ only due to the difference between the variance estimators $\widehat{V}(\text{Neyman})$ and $\widehat{V}(\text{Fisher})$. Therefore, a comparison of the variance estimators will explain the different behaviors of the corresponding approaches. In the following, we use the conventional notation $R_N = o_p(N^{-1})$ for a random quantity satisfying $N \cdot R_N \to 0$ in probability as $N \to \infty$ (Lehmann, 1999).

THEOREM 3. *Asymptotically, the difference between the two variance estimators is*

$$\begin{aligned}&\widehat{V}(\textit{Fisher}) - \widehat{V}(\textit{Neyman})\\ &= (N_0^{-1} - N_1^{-1})(S_1^2 - S_0^2) + N^{-1}(\bar{Y}_1 - \bar{Y}_0)^2\\ &\quad + o_p(N^{-1}).\end{aligned} \tag{7}$$

The difference between the variance estimators depends on the ratio of the treatment and control sample sizes, and differences between the means and variances of the treatment and control potential outcomes. The "conservativeness" of Neyman's test does not cause the paradox; if we use the true sampling variance rather than the estimated variance of $\widehat{\tau}$ for testing, then the paradox will happen even more often.

In order the verify the asymptotic theory above, we go back to compare the variances in the previous numerical examples.

EXAMPLE 3 (Continuations of Examples 1 and 2). We plot in Figure 1 the variances $\widehat{V}(\text{Neyman})$ and $\widehat{V}(\text{Fisher})$ obtained from the numerical examples in

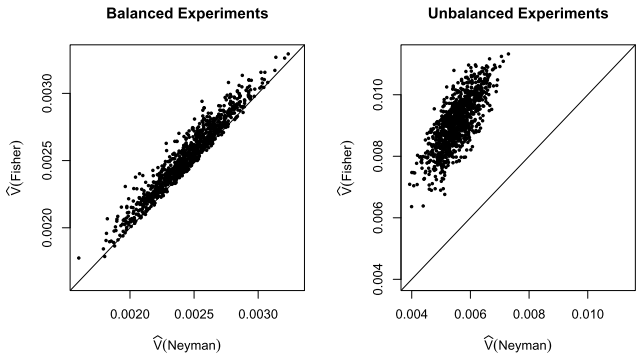

FIG. 1. *Variance estimators in balanced and unbalanced experiments.*

Section 3.1. In both the left and the right panels, $\widehat{V}(\text{Fisher})$ tends to be larger than $\widehat{V}(\text{Neyman})$. This pattern is more striking on the right panel with unbalanced experiments designed to satisfy $(N_0^{-1} - N_1^{-1})(S_1^2 - S_0^2) > 0$. It is thus not very surprising that the FRT is much less powerful than Neyman's test, and it rejects even less often than nominal 0.05 level as shown in Table 1(b).

3.4 Theoretical Comparison

Although quite straightforward, Theorem 3 has several helpful implications to explain the paradoxical results in Section 3.1.

Under $H_0(\text{Fisher})$, $\bar{Y}_1 = \bar{Y}_0, S_1^2 = S_0^2$, and the difference between the two variances is of higher order, namely, $\widehat{V}(\text{Fisher}) - \widehat{V}(\text{Neyman}) = o_p(N^{-1})$. Therefore, Neymanian and Fisherian methods coincide with each other asymptotically under the sharp null. This is the basic requirement, because both testing procedures should generate correct type one errors under this circumstance.

For the case with constant causal effect, we have $\tau_i = \tau$ and $S_1^2 = S_0^2$. The difference between the two variance estimators reduces to

$$\widehat{V}(\text{Fisher}) - \widehat{V}(\text{Neyman}) = \tau^2/N + o_p(N^{-1}). \tag{8}$$

Under $H_0(\text{Neyman})$, $\bar{Y}_1 = \bar{Y}_0$, and the difference between the two variances is of higher order, and two tests have the same asymptotic performance. However, under the alternative hypothesis, $\tau = \bar{Y}_1 - \bar{Y}_0 \neq 0$, and the difference above is positive and of order $1/N$, and Neyman's test will reject more often than Fisher's test. With larger effect size $|\tau|$, the powers differ more.

For balanced experiments with $N_1 = N_0$, the difference between the two variance estimators reduces to the same formula as (8), and the conclusions are the same as above.

For unbalanced experiments, the difference between two variances can be either positive or negative. In practice, if we have prior knowledge $S_1^2 > S_0^2$, unbalanced experiments with $N_1 > N_0$ are preferable to improve estimation precision. In this case, we have $(N_0^{-1} - N_1^{-1})(S_1^2 - S_0^2) > 0$ and $\widehat{V}(\text{Fisher}) > \widehat{V}(\text{Neyman})$ for large N. Surprisingly, we are more likely to reject Neyman's null than Fisher's null, although Neyman's test itself is conservative with nonconstant causal effect implied by $S_1^2 > S_0^2$.

From the above cases, we can see that Neymanian and Fisherian approaches generally have different performances, unless the sharp null hypothesis holds. Fisher's sharp null imposes more restrictions on the potential outcomes, and the variance of the randomization distribution of $\widehat{\tau}$ pools the within and between group variances across treatment and control arms. Consequently, the resulting randomization distribution of $\widehat{\tau}$ has larger variance than its repeated sampling variance in many realistic cases. Paradoxically, in many situations, we tend to reject Neyman's null more often than Fisher's null, which contradicts the logical fact that Fisher's null implies Neyman's null.

Finally, we consider the performance of the FRT under Neyman's null with $\bar{Y}_1 = \bar{Y}_0$, which is often of more interest in social sciences. If $S_1^2 > S_0^2$ and $N_1 > N_0$, the rejection rate of Fisher's test is smaller than Neyman's test, even though $H_0(\text{Neyman})$ holds but $H_0(\text{Fisher})$ does not. Consequently, the difference-in-means statistic $\widehat{\tau}(\boldsymbol{T}, \boldsymbol{Y}^{\text{obs}})$ has no power against the sharp null, and the resulting FRT rejects even less often than the nominal significance level. However, if $S_1^2 > S_0^2$ and $N_1 < N_0$, the FRT may not be more "conservative" than Neyman's test. Unfortunately, the FRT may reject more often than the nominal level, yielding an invalid test for Neyman's null. Gail et al. (1996), Lang (2015), and Lin et al. (2017) found this phenomenon in numerical examples, and we provide a theoretical explanation.

3.5 Binary Outcomes

We close this section by investigating the special case with binary outcomes, for which more explicit results are available. Let $p_t = \bar{Y}(t)$ be the potential proportion and $\widehat{p}_t = \bar{Y}_t^{\text{obs}}$ be the sample proportion of one under treatment t. Define $\widehat{p} = \bar{Y}^{\text{obs}}$ as the proportion of one in all the observed outcomes. The results in the following corollary are special cases of Theorems 1 to 3.

COROLLARY 1. *Neyman's test is asymptotically equivalent to the "unpooled" test*

$$\frac{\widehat{p}_1 - \widehat{p}_0}{\sqrt{\widehat{p}_1(1-\widehat{p}_1)/N_1 + \widehat{p}_0(1-\widehat{p}_0)/N_0}} \xrightarrow{d} \mathcal{N}(0,1) \tag{9}$$

under $H_0(\text{Neyman})$; and Fisher's test is asymptotically equivalent to the "pooled" test

$$\frac{\widehat{p}_1 - \widehat{p}_0}{\sqrt{\widehat{p}(1-\widehat{p})(N_1^{-1} + N_0^{-1})}} \xrightarrow{d} \mathcal{N}(0,1) \tag{10}$$

under $H_0(\text{Fisher})$. The asymptotic difference between the two tests is due to

$$\begin{aligned} &\widehat{V}(\text{Fisher}) - \widehat{V}(\text{Neyman}) \\ &= (N_0^{-1} - N_1^{-1})\{p_1(1-p_1) - p_0(1-p_0)\} \\ &\quad + N^{-1}(p_1 - p_0)^2 + o_p(N^{-1}). \end{aligned} \tag{11}$$

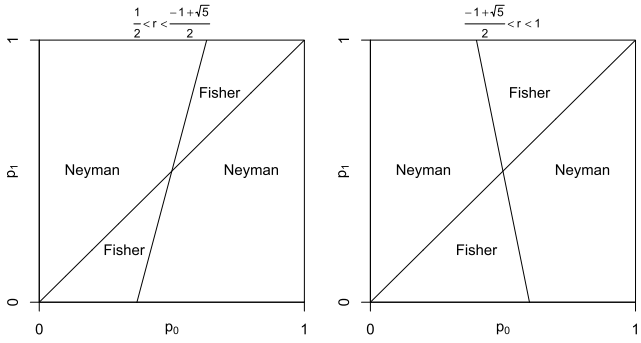

FIG. 2. *Binary outcome with different proportions $r = N_1/N$. Neyman's test is more powerful in the regions marked by "Neyman."*

For the case with binary outcomes, we can draw analogous but slightly different conclusions to the above. Under Neyman's null, $p_1 = p_0$ and the two tests are asymptotically equivalent. Therefore, the situation that the FRT is invalid under Neyman's null will never happen for binary outcomes. In balanced experiments, Neyman's test is always more powerful than Fisher's test under the alternative with $p_1 \neq p_0$. For unbalanced experiments, the answer is not definite, but equation (12) allows us to determine the region of (p_1, p_0) that favors Neyman's test for a given level of the ratio $r = N_1/N$. When $r > 1/2$, Figure 2 shows the regions in which Neyman's test is asymptotically more powerful than Fisher's test according to the value of r. When $r < 1/2$, the region has the same shape by symmetry. We provide more details about Figure 2 in the Supplementary Material (Ding, 2017).

Note that Fisher's test is equivalent to Fisher's exact test, and (10) is essentially the Normal approximation of the hypergeometric distribution (Barnard, 1947; Cox, 1970; Ding and Dasgupta, 2016). The two tests in (9) and (10) are based purely on randomization inference, which have the same mathematical forms as the classical "unpooled" and "pooled" tests for equal proportions under two independent Binomial models. Our conclusion is coherent with Robbins (1977) and Eberhardt and Fligner (1977) that the "unpooled" test is more powerful than the "pooled" one with equal sample size. For hypothesis testings in two by two tables, Greenland (1991) observed similar theoretical results as Corollary 1 but gave a different interpretation. Recently, Rigdon and Hudgens (2015) and Li and Ding (2016) constructed exact confidence intervals for τ by inverting a sequence of FRTs.

4. UBIQUITY OF THE PARADOX IN OTHER EXPERIMENTS

The paradox discussed in Section 3 is not unique to completely randomized experiments. As a direct generalization of the previous results, the paradox will appear in each stratum of stratified experiments. We will also show its existence in two other widely-used experiments: matched-pair designs and factorial designs. In order to minimize the confusion about the notation, each of the following two subsections is self-contained.

4.1 Matched-Pair Experiments

Consider a matched-pair experiment with $2N$ units and N pairs matched according to their observed characteristics. Within each matched pair, we randomly select one unit to receive treatment and the other to receive control. Let T_i be i.i.d. Bernoulli$(1/2)$ for $i = 1, \ldots, N$, indicating treatment assignments for the matched pairs. For pair i, the first unit receives treatment and the second unit receives control if $T_i = 1$; and otherwise if $T_i = 0$. Under the SUTVA, we define $(Y_{ij}(1), Y_{ij}(0))$ as the potential outcomes of the jth unit in the ith pair under treatment and control, and the observed outcomes within pair i are $Y_{i1}^{\text{obs}} = T_i Y_{i1}(1) + (1 - T_i) Y_{i1}(0)$ and $Y_{i2}^{\text{obs}} = T_i Y_{i2}(0) + (1 - T_i) Y_{i2}(1)$. Let $\boldsymbol{T} = (T_1, \ldots, T_N)'$ and $\boldsymbol{Y}^{\text{obs}} = \{Y_{ij}^{\text{obs}} : i = 1, \ldots, N; j = 1, 2\}$ denote the $N \times 1$ treatment assignment vector and the $N \times 2$ observed outcome matrix, respectively. Within pair i,

$$\widehat{\tau}_i = T_i\big(Y_{i1}^{\text{obs}} - Y_{i2}^{\text{obs}}\big) + (1 - T_i)\big(Y_{i2}^{\text{obs}} - Y_{i1}^{\text{obs}}\big)$$

is unbiased for the within-pair average causal effect

$$\tau_i = \big\{Y_{i1}(1) + Y_{i2}(1) - Y_{i1}(0) - Y_{i2}(0)\big\}/2.$$

Immediately, we can use

$$\widehat{\tau} = \frac{1}{N}\sum_{i=1}^{N} \widehat{\tau}_i$$

as an unbiased estimator for the finite population average causal effect

$$\tau = \frac{1}{N}\sum_{i=1}^{N} \tau_i = \frac{1}{2N}\sum_{i=1}^{N}\sum_{j=1}^{2}\big\{Y_{ij}(1) - Y_{ij}(0)\big\}.$$

Imai (2008) discussed Neymanian inference for τ and identified the variance of $\widehat{\tau}$ with the corresponding variance estimator. To be more specific, he calculated

$$\text{var}(\widehat{\tau}) = \frac{1}{4N^2}\sum_{i=1}^{N}\big\{Y_{i1}(1) + Y_{i1}(0) - Y_{i2}(1) - Y_{i2}(0)\big\}^2,$$

and proposed a variance estimator

$$\widehat{V}(\text{Neyman}) = \frac{1}{N(N-1)}\sum_{i=1}^{N}(\widehat{\tau}_i - \widehat{\tau})^2.$$

Again, the variance estimator is "conservative" for the true sampling variance because $E\{\widehat{V}(\text{Neyman})\} \geq \text{var}(\widehat{\tau})$ unless the within-pair average causal effects are constant. The repeated sampling evaluation above allows us to test Neyman's null hypothesis of zero average causal effect:

$$H_0(\text{Neyman}) : \tau = 0.$$

On the other hand, Rosenbaum (2002) discussed intensively the FRT in matched-pair experiments under the sharp null hypothesis:

$$H_0(\text{Fisher}) : Y_{ij}(1) = Y_{ij}(0),$$
$$\forall i = 1, \ldots, N; \forall j = 1, 2,$$

which is, again, much stronger than Neyman's null. For the purpose of comparison, we choose the test statistic with the same form as $\widehat{\tau}$, denoted as $\widehat{\tau}(\boldsymbol{T}, \boldsymbol{Y}^{\text{obs}})$. In fact, Fisher (1935a) used this test to analyze Charles Darwin's data on the relative growth rates of cross- and self-fertilized corns. In practice, the null distribution of this test statistic can be calculated exactly by enumerating all the 2^N randomizations or approximated by Monte Carlo. For our theoretical investigation, we have the following results.

THEOREM 4. *Under the sharp null hypothesis,* $E\{\widehat{\tau}(\boldsymbol{T}, \boldsymbol{Y}^{\text{obs}})|H_0(\textit{Fisher})\} = 0$, *and*

$$\widehat{V}(\textit{Fisher}) \equiv \text{var}\{\widehat{\tau}(\boldsymbol{T}, \boldsymbol{Y}^{\text{obs}})|H_0(\textit{Fisher})\}$$
$$= \frac{1}{N^2}\sum_{i=1}^{N}\widehat{\tau}_i^2.$$

Therefore, for matched-pair experiments, the difference in the variances is

$$\widehat{V}(\textit{Fisher}) - \widehat{V}(\textit{Neyman}) = \tau^2/N + o_p(N^{-1}).$$

The asymptotic Normality of the two test statistics holds because of the Lindberg–Feller CLT for independent random variables and, therefore, the different power behaviors of Neyman and Fisher's tests is again due to the above difference in the variances. Under H_0(Neyman), the difference is a higher order term, leading to asymptotically equivalent behaviors of Neymanian and Fisherian inferences. However, under the alternative hypothesis with nonzero τ, the same paradox appears again in matched-pair experiments: we tend to reject with Neyman's test more often than with Fisher's test.

For matched-pair experiments with binary outcomes, we let $m^{\text{obs}}_{y_1 y_0}$ be the number of pairs with treatment outcome y_1 and control outcome y_0, where $y_1, y_0 \in \{0, 1\}$. Consequently, we can summarize the observed data by a two by two table with cell counts $(m^{\text{obs}}_{11}, m^{\text{obs}}_{10}, m^{\text{obs}}_{01}, m^{\text{obs}}_{00})$. Theorem 4 can then be further simplified as follows.

COROLLARY 2. *In matched-pair experiments with binary outcomes, Neyman's test is asymptotically equivalent to*

$$\frac{m^{\text{obs}}_{10} - m^{\text{obs}}_{01}}{\sqrt{m^{\text{obs}}_{10} + m^{\text{obs}}_{01} - (m^{\text{obs}}_{10} - m^{\text{obs}}_{01})^2/N}} \xrightarrow{d} \mathcal{N}(0, 1) \tag{12}$$

under H_0(*Neyman*), *and Fisher's test is asymptotically equivalent to*

$$\frac{m^{\text{obs}}_{10} - m^{\text{obs}}_{01}}{\sqrt{m^{\text{obs}}_{10} + m^{\text{obs}}_{01}}} \xrightarrow{d} \mathcal{N}(0, 1) \tag{13}$$

under H_0(*Fisher*). *And the asymptotic difference between the two tests is due to*

$$\widehat{V}(\textit{Fisher}) - \widehat{V}(\textit{Neyman})$$
$$= (m^{\text{obs}}_{10} - m^{\text{obs}}_{01})^2/N^3 + o_p(N^{-1}).$$

Note that the number of discordant pairs, $m^{\text{obs}}_{10} + m^{\text{obs}}_{01}$, is fixed over all randomizations under the sharp null hypothesis and, therefore, Fisher's test is equivalent to the exact test based on $m^{\text{obs}}_{10} \sim \text{Binomial}(m^{\text{obs}}_{10} + m^{\text{obs}}_{01}, 1/2)$. Its asymptotic form (13) is the same as the McNemar test under a super population model (Agresti and Min, 2004).

4.2 Factorial Experiments

Fisher (1935a) and Yates (1937) developed the classical factorial experiments in the context of agricultural experiments, and Wu and Hamada (2009) provided a comprehensive modern discussion of design and analysis of factorial experiments. Although rooted in randomization theory (Kempthorne, 1955; Hinkelmann and Kempthorne, 2008), the analysis of factorial experiments is dominated by linear and generalized linear models, with factorial effects often defined as model parameters. Realizing the inherent drawbacks of the predominant approaches, Dasgupta, Pillai and Rubin (2015) discussed causal inference from 2^K factorial experiments using the potential outcomes framework, which allows for defining the causal estimands based on potential outcomes instead of model parameters.

We first briefly review the notation for factorial experiments adopted by Dasgupta, Pillai and Rubin (2015). Assume that we have K factors with levels $+1$ and -1. Let $\boldsymbol{z} = (z_1, \ldots, z_K)' \in \mathcal{F}_K = \{+1, -1\}^K$, a K-dimensional vector, denote a particular treatment combination. The number of possible values of $\boldsymbol{z}$ is $J = 2^K$, for each of which we define $Y_i(\boldsymbol{z})$ as the corresponding potential outcome for unit i under the SUTVA. We use a J-dimensional vector $\boldsymbol{Y}_i$ to denote all potential outcomes for unit i, where $i = 1, \ldots, N = r \times 2^K$ with an integer r representing the number of replications of each treatment combination. Without loss of generality, we will discuss the inference of the main factorial effect of factor 1, and analogous discussion also holds for general factorial effects due to symmetry. The main factorial effect of factor 1 can be characterized by a vector $\boldsymbol{g}_1$ of dimension J, with one half of its elements being $+1$ and the other half being -1. Specifically, the element of $\boldsymbol{g}_1$ is $+1$ if the corresponding z_1 is $+1$, and -1 otherwise. For example, in 2^2 experiments, we have $\boldsymbol{Y}_i = (Y_i(+1,+1),\ Y_i(+1,-1),\ Y_i(-1,+1),\ Y_i(-1,-1))'$ and $\boldsymbol{g}_1 = (+1,+1,-1,-1)'$. We define $\tau_{i1} = 2^{-(K-1)}\boldsymbol{g}_1'\boldsymbol{Y}_i$ as the main factorial effect of factor 1 for unit i, and

$$\tau_1 = \frac{1}{N}\sum_{i=1}^{N}\tau_{i1} = 2^{-(K-1)}\boldsymbol{g}_1'\bar{\boldsymbol{Y}}$$

as the average main factorial effect of the factor 1, where $\bar{\boldsymbol{Y}} = \sum_{i=1}^{N}\boldsymbol{Y}_i/N$.

For factorial experiments, we define the treatment assignment as $W_i(\boldsymbol{z})$, with $W_i(\boldsymbol{z}) = 1$ if the ith unit is assigned to $\boldsymbol{z}$, and 0 otherwise. Therefore, we use $\boldsymbol{W}_i = \{W_i(\boldsymbol{z}) : \boldsymbol{z} \in \mathcal{F}_K\}$ as the treatment assignment vector for unit i, and let $\boldsymbol{W}$ be the collection of all the unit-level treatment assignments. The observed outcomes are deterministic functions of the potential outcomes and the treatment assignment, namely, $Y_i^{\text{obs}} = \sum_{\boldsymbol{z}\in\mathcal{F}_K} W_i(\boldsymbol{z})Y_i(\boldsymbol{z})$ for unit i, and $\boldsymbol{Y}^{\text{obs}} = (Y_1^{\text{obs}}, \ldots, Y_N^{\text{obs}})'$ for all the observed outcomes. Because

$$\bar{Y}^{\text{obs}}(\boldsymbol{z}) = \frac{1}{r}\sum_{\{i:W_i(\boldsymbol{z})=1\}} Y_i^{\text{obs}} = \frac{1}{r}\sum_{i=1}^{N} W_i(\boldsymbol{z})Y_i(\boldsymbol{z})$$

is unbiased for $\bar{Y}(\boldsymbol{z})$, we can unbiasedly estimate τ_1 by

$$\widehat{\tau}_1 = 2^{-(K-1)}\boldsymbol{g}_1'\bar{\boldsymbol{Y}}^{\text{obs}},$$

where $\bar{\boldsymbol{Y}}^{\text{obs}}$ is the J-dimensional vector for the average observed outcomes. Dasgupta, Pillai and Rubin (2015) showed that the sampling variance of $\widehat{\tau}_1$ is

$$\text{var}(\widehat{\tau}_1) = \frac{1}{2^{2(K-1)}r}\sum_{\boldsymbol{z}\in\mathcal{F}_K} S^2(\boldsymbol{z}) - \frac{1}{N}S_1^2, \tag{14}$$

where $S^2(\boldsymbol{z}) = \sum_{i=1}^{N}\{Y_i(\boldsymbol{z}) - \bar{Y}(\boldsymbol{z})\}^2/(N-1)$ is the finite population variance of the potential outcomes under treatment combination $\boldsymbol{z}$, and $S_1^2 = \sum_{i=1}^{N}(\tau_{i1} - \tau_1)^2/(N-1)$ is the finite population variance of the unit level factorial effects $\{\tau_{i1} : i = 1, \ldots, N\}$. Similar to the discussion in completely randomized experiments, the last term S_1^2 in (14) cannot be identified, and consequently the variance in (14) can only be "conservatively" estimated by the following Neyman-style variance estimator:

$$\widehat{V}_1(\text{Neyman}) = \frac{1}{2^{2(K-1)}r}\sum_{\boldsymbol{z}\in\mathcal{F}_K} s^2(\boldsymbol{z}),$$

where the sample variance of outcomes $s^2(\boldsymbol{z}) = \sum_{\{i:W_i(\boldsymbol{z})=1\}}\{Y_i^{\text{obs}} - \bar{Y}^{\text{obs}}(\boldsymbol{z})\}^2/(r-1)$ under treatment combination $\boldsymbol{z}$ is unbiased for $S^2(\boldsymbol{z})$. The discussion above allows us to construct a Wald-type test for Neyman's null of zero average factorial effect for factor 1:

$$H_0^1(\text{Neyman}) : \tau_1 = 0.$$

On the other hand, based on the physical act of randomization in factorial experiments, the FRT allows us to test the following sharp null hypothesis:

$$H_0(\text{Fisher}) : Y_i(\boldsymbol{z}) = Y_i^{\text{obs}}, \quad \forall \boldsymbol{z} \in \mathcal{F}_K, \forall i = 1, \ldots, N. \tag{15}$$

This sharp null restricts all factorial effects for all the individuals to be zero, which is much stronger than H_0^1(Neyman). For a fair comparison, we use the same test statistic as $\widehat{\tau}_1$ in our randomization test, and denote $\widehat{\tau}_1(\boldsymbol{W}, \boldsymbol{Y}^{\text{obs}})$ as a function of the treatment assignment and observed outcomes. Under the sharp null (15), the randomness of $\widehat{\tau}_1(\boldsymbol{W}, \boldsymbol{Y}^{\text{obs}})$ is induced by randomization, and the following theorem gives us its mean and variance.

THEOREM 5. *Under the sharp null,* $E\{\widehat{\tau}_1(\boldsymbol{W}, \boldsymbol{Y}^{\text{obs}})|H_0(\textit{Fisher})\} = 0$, *and*

$$\widehat{V}_1(\textit{Fisher}) \equiv \text{var}\{\widehat{\tau}_1(\boldsymbol{W}, \boldsymbol{Y}^{\text{obs}})|H_0(\textit{Fisher})\} = \frac{1}{2^{2(K-1)}r}Js^2,$$

where $\bar{Y}^{\text{obs}} = \sum_{i=1}^{N} Y_i^{\text{obs}}/N$ *and* $s^2 = \sum_{i=1}^{N}(Y_i^{\text{obs}} - \bar{Y}^{\text{obs}})^2/(N-1)$ *are the sample mean and variance of all the observed outcomes.*

Based on Normal approximations, comparison of the p-values reduces to the difference between $\widehat{V}_1(\text{Neyman})$ and $\widehat{V}_1(\text{Fisher})$, as shown in the theorem below.

THEOREM 6. *With large r, the difference between $\widehat{V}_1(\text{Neyman})$ and $\widehat{V}_1(\text{Fisher})$ is*

$$\begin{aligned}&\widehat{V}_1(\text{Fisher}) - \widehat{V}_1(\text{Neyman})\\ &= \frac{1}{2^{3K-1}r}\sum_{z\in\mathcal{F}_K}\sum_{z'\in\mathcal{F}_K}\{\bar{Y}(z)-\bar{Y}(z')\}^2\\ &\quad + o_p(r^{-1}).\end{aligned} \tag{16}$$

Formula (8) is a special case of formula (17) with $K = 1$ and $r = N_1 = N_0 = N/2$, because complete randomized experiments are special cases of factorial experiments with a single factor. Therefore, in factorial experiments with the same replicates r at each level, the paradox always exists under alternative hypothesis with nonzero τ_1, just as in balanced completely randomized experiments.

5. IMPROVEMENTS AND EXTENSIONS

We have shown that the seemingly paradoxical phenomenon in Section 3 is due to the fact that Neyman's test is more powerful than Fisher's test in many realistic situations. The previous sections restrict the discussion on the difference-in-means statistic. We will further comment on the importance of this choice, and other possible alternative test statistics. Moreover, the original forms of Neyman's and Fisher's tests are both suboptimal. We will discuss improved Neymanian and Fisherian inference, and the corresponding paradox.

5.1 Choice of the Test Statistic

First, as hinted by Ding and Dasgupta (2016), for randomized experiments with binary outcomes, all test statistics are equivalent to the difference-in-means statistic. We formally state this conclusion in the following theorem.

THEOREM 7. *For completely randomized experiments, matched-pair experiments, and 2^K factorial experiments, if the outcomes are binary, then all test statistics are equivalent to the difference-in-means statistic.*

Therefore, for binary data, the choice of test statistic is not a problem.

Second, for continuous outcomes, the difference-in-means statistic is important, because it not only serves as a candidate test statistic for the sharp null hypothesis but also an unbiased estimator for the average causal effect. In the illustrating example in Section 6.3, practitioners are interested in finding the combination of several factors that achieves an optimal mean response.

For continuous outcomes, we have more options of test statistics. For instance, the Kolmogorov–Smirnov and Wilcoxon–Mann–Whitney statistics are also useful candidates for the FRT. However, the Neymanian analogues of these two statistics have not been established in the literature, and direct comparisons of the Fisherian and Neymanian using these two statistics are not obvious at this moment. In the Supplementary Material (Ding, 2017), we illustrate by numerical examples that the conservative nature of the FRT is likely to be true for these two statistics, because we find that the randomization distributions under the sharp null hypothesis is more disperse than those under weaker null hypotheses. Please see the Supplementary Materials (Ding, 2017) for more details, and it is our future research topic to pursue the theoretical results.

5.2 Improving the Neymanian Variance Estimators

For completely randomized experiments, Neyman (1923/1990) used $S_\tau^2 \geq 0$ as a lower bound, which is not the sharp bound. Recently, for general outcomes Aronow, Green and Lee (2014) derived the sharp bound of S_τ^2 based on the marginal distributions of the treatment and control potential outcomes using the Frechét–Hoeffding bounds (Nelsen, 2006); for binary outcomes Robins (1988) and Ding and Dasgupta (2016) gave simple forms. These improvements result in smaller variance estimators.

For matched-pair experiments, Imai (2008) improved the Neymanian variance estimator by using the Cauchy–Schwarz inquality. We are currently working on deriving sharp bounds for the variance of estimated factorial effects.

In summary, Neyman's test is even more powerful with improved variance estimators, which further bolsters the paradoxical situation wherein we reject Neyman's null but fail to reject Fisher's sharp null.

5.3 Improving the FRT and Connection with the Permutation Test

In the permutation test literature, some authors (e.g., Neuhaus, 1993; Janssen, 1997; Chung and Romano, 2013; Pauly, Brunner and Konietschke, 2015) suggested using the Studentized version of $\widehat{\tau}$, that is, $\widehat{\tau}/\sqrt{\widehat{V}(\text{Neyman})}$, as the test statistic. When the experiment is unbalanced, the FRT using this test statistic

has exact type one error under Fisher' null and correct asymptotic type one error under Neyman's null. However, this does not eliminate the paradox discussed in this paper. First, we have shown that this paradox arises even in balanced experiments, but this test statistic tries to correct the invalid asymptotic type one error under Neyman's null in unbalanced experiments. Second, Section 5.1 has shown that for binary outcomes any test statistic is equivalent to $\widehat{\tau}$ and, therefore, this Studentized test statistic will not change the paradox at least for binary outcomes. Third, the theories of permutation tests and randomization tests do not have a one-to-one mapping, although they often give the same numerical results. The theory of permutation tests assumes exchangeable units drawn from an infinite super-population, and the theory of randomization tests assumes fixed potential outcomes in a finite population and random treatment assignment. Consequently, the correlation between the potential outcomes never plays a role in the theory of permutation tests, but it plays a central role in the theory of randomization inference as indicated by Neyman's (1923/1990) seminal work and our discussion above.

6. ILLUSTRATIONS

In this section, we will use real-life examples to illustrate the theory in the previous sections. The first two examples have binary outcomes and, therefore, there is no concern about the choice of test statistic. The goal of the third example, a 2^4 full factorial experiment, is to find the optimal combination of the factors and, therefore, the difference-in-means statistic is again a natural choice for a test statistic.

6.1 A Completely Randomized Experiment

Consider a hypothetical completely randomized experiment with binary outcome (Rosenbaum, 2002, page 191). Among the 32 treated units, 18 of them have outcome being 1, and among the 21 control units, 5 of them have outcome being 1. The Neymanian p-value based on the improved variance estimator in Robins (1988) and Ding and Dasgupta (2016) is 0.004. The Fisherian p-value based on the FRT or equivalently Fisher's exact test is 0.026, and the Fisherian p-value based on Normal approximation in (10) is 0.020. The Neymanian p-value is smaller, and if we choose significance level at 0.01 then the paradox will appear in this example.

6.2 A Matched-Pair Experiment

The observed data of the matched-pair experiment in Agresti and Min (2004) can be summarized by the two by two table with cell counts $(m_{11}^{\text{obs}}, m_{10}^{\text{obs}}, m_{01}^{\text{obs}}, m_{00}^{\text{obs}}) = (53, 8, 16, 9)$. The Neymanian one-sided p-value based on (12) is 0.049. The Fisherian p-value based on the FRT is 0.076, and the Fisherian p-value based on Normal approximation in (13) is 0.051. Again, Neyman's test is more powerful than Fisher's test.

6.3 A 2^4 Full Factorial Experiment

In the "Design of Experiments" course in Fall 2014, a group of Harvard undergraduate students, Taylor Garden, Jessica Izhakoff and Zoe Rosenthal, followed Box's (1992) famous paper helicopter example for factorial experiments, and tried to identify the optimal combination of the four factors: paper type (construction paper, printer paper), paperclip type (small paperclip, large paperclip), wing length (2.5 inches, 2.25 inches) and fold length (0.5 inch, 1.0 inch), with the first level coded as -1 and the second level coded as $+1$. For more details, please see Box (1992). For each combination of the factors, they recorded two replicates of the flying times (in seconds) of the helicopters. We display the data in Table 2.

We show the Neymanian and Fisherian results in the upper and lower panel of Figure 3, respectively. Fig-

TABLE 2
A 2^4 factorial design and observed outcomes

F_1	F_2	F_3	F_4	Replicate 1	Replicate 2
−1	−1	−1	−1	1.60	1.55
−1	−1	−1	1	1.70	1.63
−1	−1	1	−1	1.44	1.38
−1	−1	1	1	1.56	1.61
−1	1	−1	−1	1.40	1.45
−1	1	−1	1	1.36	1.38
−1	1	1	−1	1.43	1.40
−1	1	1	1	1.32	1.27
1	−1	−1	−1	1.81	1.86
1	−1	−1	1	1.70	1.57
1	−1	1	−1	2.04	2.06
1	−1	1	1	1.68	1.61
1	1	−1	−1	1.58	1.28
1	1	−1	1	1.43	1.49
1	1	1	−1	1.51	1.54
1	1	1	1	1.53	1.38

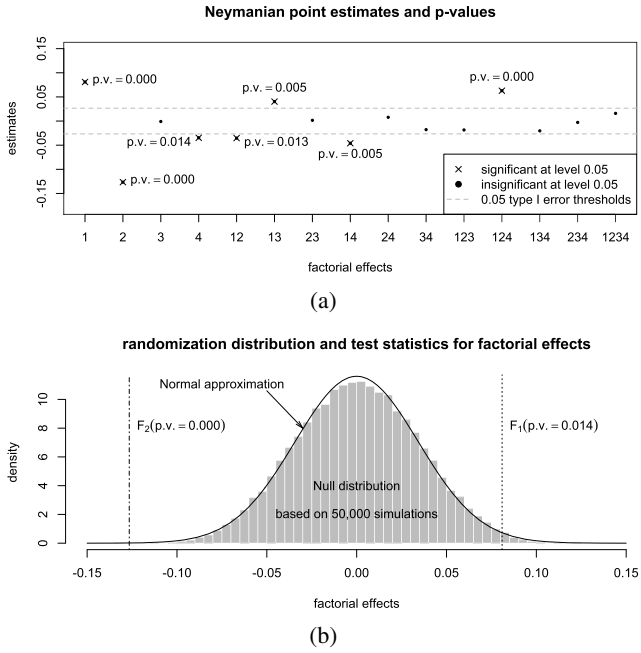

FIG. 3. *Randomization-based inference for a* 2^4 *full factorial experiment*. (a) *Neymanian inference. Factorial effects* F_1, F_2, F_4, F_1F_2, F_1F_3, F_1F_4 *and* $F_1F_2F_4$ *are significant at level* 0.05. (b) *Fisherian inference. Factorial effects* F_1 *and* F_2 *are significant*.

ure 3(a) shows both Neymanian point estimates and p-values for the 15 factorial effects. Seven of them, F_1, F_2, F_4, F_1F_2, F_1F_3, F_1F_4 and $F_1F_2F_4$, are significant at level 0.05, and after the Bonferroni correction, three of them, F_1, F_2, $F_1F_2F_4$, are still significant. Figure 3(b) shows the randomization distribution of the factorial effects under the sharp null hypothesis by a grey histogram. Note that all factorial effects have the same randomization distribution, because all of them are essentially a comparison of a random half versus the other half of the observed outcomes. Even though the sample size 32 is not huge, the randomization distribution is well approximated by the Normal distribution with mean zero and variance $\widehat{V}_1(\text{Fisher})$. Strikingly, only two factorial effects, F_1 and F_2, are significant, and after the Bonferroni correction only F_2 is significant. We further calculate the variance estimates: $\widehat{V}_1(\text{Neyman}) = 0.025$ and $\widehat{V}_1(\text{Fisher}) = 0.034$. The empirical findings in this particular example with finite sample are coherent with our asymptotic theory developed in Section 4.2. In this example, the Neymanian method can help detect more significant factors for achieving optimal flying time, while the more conservative Fisherian method may miss important factors.

7. DISCUSSION

7.1 Historical Controversy and Modern Discussion

Neyman (1923/1990) proposed to use potential outcomes for causal inference and derived mathematical properties of randomization; Fisher (1926) advocated using randomization in physical experiments, which was considered by Neyman "as one of the most valuable of Fisher's achievements" (Reid, 1982, page 44). Fisher (1935a), Section II, pointed out that "the actual and physical conduct of an experiment must govern the statistical procedure of its interpretation." Neyman and Fisher both proposed statistical procedures for analysis of randomized experiments, relying on the randomization distribution itself. However, whether Neyman's null or Fisher's null makes more sense in practice goes back to the famous Neyman–Fisher controversy in a meeting of the Royal Statistical Society (Neyman, 1935; Fisher, 1935b). After their 1935 controversy, Anscombe (1948), Kempthorne (1952) and Cox (1992) provided some further discussion on the usefulness and limitations of the two null hypotheses. For instance, the authors acknowledged that Neyman's null is mathematically weaker than Fisher's null, but both null hypotheses seem artificial requiring either individual causal effects or the average causal effect be exactly zero for finite experimental units. For Latin square designs, Wilk and Kempthorne (1957) developed theory under Neyman's view, and Cox (1958) argued that in most situations the Fisherian analysis was secure. Recently, Rosenbaum (2002), page 39, gave a very insightful philosophical discussion about the controversy, and Sabbaghi and Rubin (2014) revisited this controversy and its consequences. Fienberg and Tanur (1996) and Cox (2012) provided more historical aspects of causal inference and in particular the Neyman–Fisher controversy.

While the answer may depend on different perspectives of practical problems, we discussed only the consequent seeming paradox of Neymanian and Fisherian testing procedures for their own null hypotheses. Both our numerical examples and asymptotic theory showed that we encounter a serious logical problem in the analysis of randomized experiments, even though both Neyman's and Fisher's tests are valid Frequentists' tests, in the sense of controlling correct type one errors under their own null hypotheses. Our numerical examples and theoretical analysis reach a conclusion different from Rosenbaum (2002).

7.2 Randomization-Based and Regression-Based Inference

In the current statistical practice, it is also very popular among applied researchers to use regression-based methods to analyze experimental data (Angrist and Pischke, 2008). Assume the a linear model for the observed outcomes: $Y_i^{\text{obs}} = \alpha + \beta T_i + \varepsilon_i$, where $\varepsilon_i, \ldots, \varepsilon_N$ are independently and identically distributed (i.i.d.) as $\mathcal{N}(0, \sigma^2)$. The hypothesis of zero treatment effect is thus characterized by $H_0(LM) : \beta = 0$. The usual ordinary least squares variance estimator for the regression coefficient may not correctly reflect the true variance of $\widehat{\tau}$ under randomization. Schochet (2010), Samii and Aronow (2012), Lin (2013) and Imbens and Rubin (2015) pointed out that we can solve this problem by using the Huber–White heteroskedasticity-robust variance estimator (Huber, 1967; White, 1980), and the corresponding Wald test is asymptotically the same as Neyman's test. In Theorem A.1 of the Supplementary Material (Ding, 2017), we further build an equivalence relationship between Rao's score test and the FRT. For more technical details, please see the Supplementary Material (Ding, 2017). Previous results, as well as Theorem A.1, do justify the usage of linear models in analysis of experimental data.

7.3 Interval Estimation

Originally, Neyman (1923/1990) proposed an unbiased estimator for the average causal effect τ with a repeated sampling evaluation, which was later developed into the concept of the confidence interval (Neyman, 1937). In order to compare Neyman's approach with the FRT, we converted the interval estimator into a hypothesis testing procedure. As a dual, we can also invert the FRT for a sequence of null hypotheses to get an interval estimator for τ (Pitman, 1937, 1938; Rosenbaum, 2002). For example, we consider the sequence of sharp null hypotheses with constant causal effects:

$$H_0^{\delta}(\text{Fisher}) : Y_i(1) - Y_i(0) = \delta, \quad \forall i = 1, \ldots, N. \tag{17}$$

The interval estimator for τ with coverage rate $1 - \alpha$ is

$$\{\delta : \text{Fail to reject } H_0^{\delta}(\text{Fisher}) \text{ by the FRT at significant level } \alpha\}.$$

Dasgupta, Pillai and Rubin (2015) found some empirical evidence in factorial designs that the above interval is wider than the Neymanian confidence interval. Due to the duality between hypothesis testing and interval estimation, our results about hypothesis testing can partially explain the phenomenon about interval estimation in Dasgupta, Pillai and Rubin (2015). To avoid making assumptions such as constant causal effects in (17), we restricted the theoretic discussion to only hypothesis testings. It is our future work to extend the theory to interval estimations.

7.4 Practical Implications

We highlight some practical implications of our theory developed in the above sections.

First, the FRT is usually less powerful than Neyman's test, even for the simplest case with constant causal effect. Practitioners should keep in mind that the FRT may miss important treatment factors. Our examples in Section 6 and the empirical evidence in Dasgupta, Pillai and Rubin (2015) have confirmed our theoretical results.

Second, in the presence of treatment effect heterogeneity, the FRT may not be a valid test for the null hypothesis of zero average causal effect. Therefore, practitioners, especially those who are interested in social sciences, should always be aware of this potential danger of using the FRT, if the observed data show substantive heterogeneity in treatment and control groups. Furthermore, as Cox (1958) pointed out, in the presence of treatment effect heterogeneity, focusing only on the average causal effect is often not adequate, and detecting and explaining such heterogeneity may be more helpful. Treatment effect variation is another important issue beyond the current scope of our paper. Ding, Feller and Miratrix (2016) investigate this problem under the randomization framework.

Third, although we have shown that the FRT is less powerful in many realistic cases, we do not conclude that Neymanian inference trumps Fisherian inference. All our comparisons are based on asymptotics under regularity conditions, and the conclusion may not be true with small sample sizes or "irregular" potential outcomes. Therefore, Fisherian inference is still useful for small sample problems and exact inference. In practice, we should always check the discrepancy between the Normal approximation and the exact randomization distribution as in Figure 3(b) before applying our theoretical results to applied problems.

ACKNOWLEDGMENTS

I want to thank Professors Donald Rubin, Arthur Dempster, Tyler VanderWeele, James Robins, Alan Agresti, Fan Li, Peter Aronow, Sander Greenland and Judea Pearl for their comments. Dr. Avi Feller at

Berkeley, Dr. Arman Sabbaghi at Purdue and Misses Lo-Hua Yuan and Ruobin Gong at Harvard helped edit early versions of this paper. I am particularly grateful to Professors Tirthankar Dasgupta and Luke Miratrix for their continuous encouragement and help during my writing of this paper. A group of Harvard undergraduate students, Taylor Garden, Jessica Izhakoff and Zoe Rosenthal, collected the data from a 2^4 full factorial design for the final project of Professors Dasgupta and Rubin's course "Design of Experiments" in Fall 2014. They kindly shared their interesting data with me. Based on an early version of this paper, I received the 2014 Arthur P. Dempster Award from the Arthur P. Dempster Fund of the Harvard Statistics Department, generously established by Professor Stephen Blyth. I am also grateful to the detailed technical comments from one reviewer and many helpful historical comments from the other reviewer.

SUPPLEMENTARY MATERIAL

Supplementary Material (DOI: 10.1214/16-STS571SUPP; .pdf). Appendix A.1 gives two useful lemmas for randomized experiments. Appendix A.2 gives the proofs of all the theorems and corollaries in the main text. Appendix A.3 comments on the regression-based causal inference, and establishes a new connection between Rao's score test and the FRT. Appendix A.4 shows more details about generating Figure 2 in the main text. Appendix A.5 discusses the behaviors of the FRT using the Kolmogorov–Smirnov and Wilcoxon–Mann–Whitney statistics.

REFERENCES

AGRESTI, A. and MIN, Y. (2004). Effects and non-effects of paired identical observations in comparing proportions with binary matched-pairs data. *Stat. Med.* **23** 65–75.

ANGRIST, J. D. and PISCHKE, J. S. (2008). *Mostly Harmless Econometrics*: *An Empiricist's Companion*. Princeton Univ. Press, Princeton, NJ.

ANSCOMBE, F. J. (1948). The validity of comparative experiments. *J. Roy. Statist. Soc. Ser. A* **111** 181–200; discussion, 200–211. MR0030181

ARONOW, P. M., GREEN, D. P. and LEE, D. K. K. (2014). Sharp bounds on the variance in randomized experiments. *Ann. Statist.* **42** 850–871. MR3210989

BARNARD, G. A. (1947). Significance tests for 2×2 tables. *Biometrika* **34** 123–138. MR0019285

BOX, G. E. P. (1992). Teaching engineers experimental design with a paper helicopter. *Qual. Eng.* **4** 453–459.

CHUNG, E. and ROMANO, J. P. (2013). Exact and asymptotically robust permutation tests. *Ann. Statist.* **41** 484–507. MR3099111

COX, D. R. (1958). The interpretation of the effects of non-additivity in the Latin square. *Biometrika* **45** 69–73.

COX, D. R. (1970). *The Analysis of Binary Data*. Methuen & Co., Ltd., London. MR0282453

COX, D. R. (1992). *Planning of Experiments*. Wiley, New York. Reprint of the 1958 original. MR1175752

COX, D. R. (2012). Statistical causality: Some historical remarks. In *Causality*: *Statistical Perspectives and Applications* (C. Berzuini, P. Dawid and L. Bernardinelli, eds.) 1–5. Wiley, New York.

DASGUPTA, T., PILLAI, N. S. and RUBIN, D. B. (2015). Causal inference from 2^K factorial designs by using potential outcomes. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **77** 727–753. MR3382595

DING, P. (2017). Supplement to "A paradox from randomization-based causal inference." DOI:10.1214/16-STS571SUPP.

DING, P. and DASGUPTA, T. (2016). A potential tale of two-by-two tables from completely randomized experiments. *J. Amer. Statist. Assoc.* **111** 157–168. MR3494650

DING, P., FELLER, A. and MIRATRIX, L. W. (2016). Randomization inference for treatment effect variation. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **78** 655–671.

EBERHARDT, K. R. and FLIGNER, M. A. (1977). Comparison of two tests for equality of two proportions. *Amer. Statist.* **31** 151–155. MR0488444

EDEN, T. and YATES, F. (1933). On the validity of Fisher's z-test when applied to an actual example of non-normal data. *J. Agric. Sci.* **23** 6–17.

EDGINGTON, E. S. and ONGHENA, P. (2007). *Randomization Tests*, 4th ed. Chapman & Hall/CRC, Boca Raton, FL. With 1 CD-ROM (Windows). MR2291573

FIENBERG, S. E. and TANUR, J. M. (1996). Reconsidering the fundamental contributions of Fisher and Neyman on experimentation and sampling. *Int. Stat. Rev.* **64** 237–253.

FISHER, R. A. (1926). The arrangement of field experiments. *Journal of the Ministry of Agriculture of Great Britain* **33** 503–513.

FISHER, R. A. (1935a). *The Design of Experiments*, 1st ed. Oliver and Boyd, Edinburgh.

FISHER, R. A. (1935b). Comment on "Statistical problems in agricultural experimentation". *Suppl. J. R. Stat. Soc.* **2** 154–157. 173.

FREEDMAN, D. A. (2008). On regression adjustments to experimental data. *Adv. in Appl. Math.* **40** 180–193. MR2388610

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

GREENLAND, S. (1991). On the logical justification of conditional tests for two-by-two contingency tables. *Amer. Statist.* **45** 248–251.

HÁJEK, J. (1960). Limiting distributions in simple random sampling from a finite population. *Magyar Tud. Akad. Mat. Kutató Int. Közl.* **5** 361–374. MR0125612

HINKELMANN, K. and KEMPTHORNE, O. (2008). *Design and Analysis of Experiments*, *Vol.* 1: *Introduction to Experimental Design*, 2nd ed. Wiley, New York. MR2363107

HODGES, J. L. JR. and LEHMANN, E. L. (1964). *Basic Concepts of Probability and Statistics*. Holden-Day, Inc., San Francisco, CA–London–Amsterdam. MR0185709

HOEFFDING, W. (1952). The large-sample power of tests based on permutations of observations. *Ann. Math. Stat.* **23** 169–192. MR0057521

HUBER, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proc. Fifth Berkeley Sympos. Math. Statist. and Probability* (*Berkeley, Calif.*, 1965/66), *Vol. I: Statistics* 221–233. Univ. California Press, Berkeley, CA. MR0216620

IMAI, K. (2008). Variance identification and efficiency analysis in randomized experiments under the matched-pair design. *Stat. Med.* **27** 4857–4873. MR2528770

IMBENS, G. W. and RUBIN, D. B. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge Univ. Press, New York. MR3309951

JANSSEN, A. (1997). Studentized permutation tests for non-i.i.d. hypotheses and the generalized Behrens–Fisher problem. *Statist. Probab. Lett.* **36** 9–21. MR1491070

KEMPTHORNE, O. (1952). *The Design and Analysis of Experiments*. Wiley, New York; Chapman & Hall, London. MR0045368

KEMPTHORNE, O. (1955). The randomization theory of experimental inference. *J. Amer. Statist. Assoc.* **50** 946–967. MR0071696

LANG, J. B. (2015). A closer look at testing the "no-treatment-effect" hypothesis in a comparative experiment. *Statist. Sci.* **30** 352–371. MR3383885

LEHMANN, E. L. (1999). *Elements of Large-Sample Theory*. Springer, New York. MR1663158

LI, X. and DING, P. (2016). Exact confidence intervals for the average causal effect on a binary outcome. *Stat. Med.* **35** 957–960.

LI, X. and DING, P. (2017). General forms of finite population central limit theorems with applications to causal inference. *J. Amer. Statist. Assoc.* To appear.

LIN, W. (2013). Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique. *Ann. Appl. Stat.* **7** 295–318. MR3086420

LIN, W., HALPERN, S. D., PRASAD KERLIN, M. and SMALL, D. S. (2017). A "placement of death" approach for studies of treatment effects on ICU length of stay. *Stat. Methods Med. Res.* **26** 292–311. MR3592727

NELSEN, R. B. (2006). *An Introduction to Copulas*, 2nd ed. Springer, New York.

NEUHAUS, G. (1993). Conditional rank tests for the two-sample problem under random censorship. *Ann. Statist.* **21** 1760–1779. MR1245767

NEYMAN, J. (1935). Statistical problems in agricultural experimentation (with discussion). *Suppl. J. R. Stat. Soc.* **2** 107–180.

NEYMAN, J. (1937). Outline of a theory of statistical estimation based on the classical theory of probability. *Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Sci.* **236** 333–380.

NEYMAN, J. (1990). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statist. Sci.* **5** 465–472. Translated from the 1923 Polish original and edited by D. M. Dabrowska and T. P. Speed. MR1092986

PAULY, M., BRUNNER, E. and KONIETSCHKE, F. (2015). Asymptotic permutation tests in general factorial designs. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **77** 461–473. MR3310535

PITMAN, E. J. G. (1937). Significance tests which may be applied to samples from any populations. *Suppl. J. R. Stat. Soc.* **4** 119–130.

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

REID, C. (1982). *Neyman—From Life*. Springer, New York. MR0680939

RIGDON, J. and HUDGENS, M. G. (2015). Randomization inference for treatment effects on a binary outcome. *Stat. Med.* **34** 924–935. MR3310672

ROBBINS, H. (1977). A fundamental question of practical statistics. *Amer. Statist.* **31** 97.

ROBINS, J. M. (1988). Confidence intervals for causal parameters. *Stat. Med.* **7** 773–785.

ROSENBAUM, P. R. (2002). *Observational Studies*, 2nd ed. Springer, New York. MR1899138

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. Amer. Statist. Assoc.* **75** 591–593.

RUBIN, D. B. (1990). Comment on J. Neyman and causal inference in experiments and observational studies: "On the application of probability theory to agricultural experiments. Essay on principles. Section 9" [*Ann. Agric. Sci.* **10** (1923), 1–51]. *Statist. Sci.* **5** 472–480. MR1092987

RUBIN, D. B. (2004). Teaching statistical inference for causal effects in experiments and observational studies. *J. Educ. Behav. Stat.* **29** 343–367.

SABBAGHI, A. and RUBIN, D. B. (2014). Comments on the Neyman–Fisher controversy and its consequences. *Statist. Sci.* **29** 267–284. MR3264542

SAMII, C. and ARONOW, P. M. (2012). On equivalencies between design-based and regression-based variance estimators for randomized experiments. *Statist. Probab. Lett.* **82** 365–370. MR2875224

SCHEFFÉ, H. (1959). *The Analysis of Variance*. Wiley, New York; Chapman & Hall, London. MR0116429

SCHOCHET, P. Z. (2010). Is regression adjustment supported by the Neyman model for causal inference? *J. Statist. Plann. Inference* **140** 246–259. MR2568136

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

WHITE, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica* **48** 817–838. MR0575027

WILK, M. B. and KEMPTHORNE, O. (1957). Non-additivities in a Latin square design. *J. Amer. Statist. Assoc.* **52** 218–236. MR0088137

WU, C. F. J. and HAMADA, M. S. (2009). *Experiments: Planning, Analysis, and Optimization*, 2nd ed. Wiley, Hoboken, NJ. MR2583259

YATES, F. (1937). The design and analysis of factorial experiments. Technical communication 35, Imperial Bureau of Soil Sciences, Harpenden.