Statistical Papers
https://doi.org/10.1007/s00362-024-01528-2

REGULAR ARTICLE

Is Fisher inference inferior to Neyman inference for policy analysis?

Rauf Ahmad[1] · Per Johansson[2] · Mårten Schultzberg[1]

Received: 5 May 2023 / Revised: 27 December 2023

Abstract
The increasing computational power has led to an increasing interest in Fisher's test in social science. As the Fisher and Neyman inference are based on different principles there is also an increasing interest in understanding the differential features of the two procedures. For example, Young (2018) found that the Fisher test has better size properties than the Neyman test in the situation with influential observations. Ding (2017), on the other hand, showed that the asymptotic variance of the mean-difference estimator (MDE) under Fisher inference is larger than that under Neyman inference, and that the asymptotic Fisher test is less powerful than the t-test even for the simplest case of homogeneous effect. Since MDE plays an important role for policy evaluation, these latter results are a concern for using Fisher's test as argued in Young (2018). With the aim of providing an understanding of the usefulness of the exact Fisher test for inference to the sample and to the population, this paper clarifies the results in Ding (2017). Using a novel Monte Carlo simulation following the same data generating processes as in Ding (2017), we demonstrate that the Fisher test has no worse power properties than the t-test even with heterogeneous effects.

Keywords t-statistic · Randomization test · Size · Power

1 Introduction

The testing problem of a population null hypothesis was of interest to Neyman and Pearson but not to Fisher. The differential approach to the philosophy of science between Neyman–Pearson and Fisher is quite discernible in Fisher's own polemic

The authors are grateful to Donald B. Rubin, Yannis Yatracos and seminar participants at Uppsala and Tsinghua Universities for their comments and suggestions.

✉ Per Johansson
per.johansson@statistics.uu.se

[1] Department of Statistics, Uppsala University, Uppsala, Sweden

[2] Department of Statistics, Uppsala University and YMSC, Tsinghua University, Uppsala, Sweden

Published online: 20 February 2024

article wherein he put it as *differences in logical points of view* (see Fisher 1955, p. 69). He further stated

> . . . we consider a continuum of hypotheses each eligible as null hypothesis, and it is the aggregate of frequencies calculated from each possibility in turns as true—including frequency of error, therefore only the "first kind", without any assumption of knowledge a priori–which supply [ ] the amounts of information available.

Basically, Fisher's argument concerned two issues with the Neyman–Pearson theory: (i) the assumption of repeated sampling from the same population, and (ii) the definition of errors of the second kind (type II error). For him, all one can do from a single experiment is to make inferences to the sample. Extrapolating the result by assuming repeated random sampling from a well-defined population with an aim of testing a population null made no sense for him.

Under the sharp null of no effect for any unit, the Fisher randomization test is a valid procedure for inference to the sample. The test has the correct level for an effect in the sample without needing any further assumption (Rubin 1980, 1986; Rosenbaum 2007). Under the additional stable unit treatment value assumption (Rubin 1980) the test has the correct level for, e.g. testing for an average treatment effect in the sample. Thus, with an interest of testing an average treatment effect in the sample we could either use the Fisher randomization test or the Neyman's test (a t-test) (cf. Ding 2017).

The motivation for this article stems from interesting results in Ding (2017) who compared the power of randomization test with the t-test for design based inference to the sample. Using finite population asymptotic theory, Ding (2017) showed that Fisher's test statistic, based on the mean-difference estimator (MDE), is approximately normal under the sharp null. He concludes that, if this normal approximation is used under the alternative, the Fisher test is less powerful than the t-test even for the simplest case of homogeneous effect. Moreover, the relative power of the t-test against Fisher's test is shown to be increasing with the size of the treatment effect.

The growing interest in the Fisher's test based on the MDE in social science (Athey and Imbens 2017; Young 2018) for inference to the superpopulation makes these results important and relevant to examine. One reason is that the results in Ding (2017) are only established under the null, so that any possibility of revealing the otherwise differences under the alternative remains unexplored. A second reason is that the results in Ding (2017) are in contrast to theoretical results based on repeated sampling, or superpopulation, assymptotics[1] and empirical results on Fisher inference to the superpopulation (Young 2018). The increasing popularity of computer based experimental designs[2] makes up yet a reason. The reason is that for some of these algorithms Neyman–Pearson inference is not an option, but the Fisher randomization test is. Thus, if this test is less efficient than the Neyman test, this suggests that the efficiency gains from some of the designs may well be lost when conducting the inference.

[1] For example, Lehmann (1959); Romano (1990); Chung and Romano (2013).

[2] For example, Morgan and Rubin (2012); Bertsimas et al. (2015); Kallus (2018); Lauretto et al. (2017); Krieger et al. (2019); Johansson and Schultzberg (2020); Kapelner et al. (2021); Johansson et al. (2021).

In line with Fisher we define the population null as the scientific null and the sample null as the statistical null. We discuss super population asymptotics and finite population asymptotics for both Neyman and Fisher randomization test based on the MDE for the two estimands; the population average treatment effect (PATE) and the sample average treatment effect (SATE). Following the same data generating processes as in Ding (2017), we then conduct a simulation study to assess the power properties of Fisher's test based on MDE for inference to both the PATE and SATE. To our knowledge, this simulation study is the first of its kind, expectedly due to the computational complexity of calculating power of the Fisher randomization tests.

Our simulation results show no overall superiority of Neyman's test over Fisher's test for any effect size. Instead, the property of a test being most powerful in this case seems to depend on the characteristics of the outcomes in the given sample. For most samples, and over most subsets of allocations, however, the tests show similar performance.

In addition, we find that with heterogeneous treatment effect, both tests have in general the wrong size when testing the population null in a single experiment. The results illustrates Fisher's concern about the "continuum of hypotheses each eligible as null hypothesis". The general attitude in the research community is that the t-test is preferable to the exact test as it is not restricted to the unrealistic assumption of homogeneous treatment effects. However, with heterogeneous effects the statistical null differ from scientific null. The implication is that size of test against a scientific null is in general only correct under repeated sampling from the same population and this holds for both tests.

The results also display that finite sample asymptotics is useful as it allows for Neyman–Pearson design based inference for a fixed, and quite small, sample size. However, it also shows that a comparison of repeated sampling inference with design based asymptotics is not meaningful as asymptotically the MDE has the same variance and SATE equal PATE.

It may also be added that our results are in line with those in Young (2018) who studied the differences in performance for the inference to the population using Monte Carlo simulation and reanalyzed 53 experimental papers culled from the journals of the American Economic Association. He found similar performance when there are no influential observations, but that the exact Fisher test based on the t-statistic (Chung and Romano 2013) has better size properties than the Neyman test. Note further that our focus is on the sample, or experiment, which clarifies the case for both tests with inference to the sample in case of heterogeneous effects. Since Young (2018)'s focus is on differences in inference to the population, the two studies are complementary to each other.

We may also refer to a few recent variants and extensions of the theory in different directions. Ding and Dasgupta (2018), by re-formulating the testing problem of an equal average treatment effect as a general linear hypothesis (GLH) test, consider Wald, ANOVA-type, and least-squares based test statistics, with emphasis on certain special cases. Wu and Ding (2021) discuss certain alternative strategies to Fisher's randomization test to test equality of average treatment effects. By re-formulating the testing problem as a general linear hypothesis (GLH), using an appropriate GLH matrix consisting of contrast comparisons, they consider Wald, ANOVA-type, and least-squares based test statistics, with focus on certain special cases. The same testing problem is also considered in Zhao and Ding (2021), additionally by adjusting the responses for the presence of covariates.

We begin, in the next section, with a brief orientation to the Neyman and Fisher lines of inference using Neyman (1923)'s potential outcome framework. The simulation scheme, comparing power of Neyman test and exact Fisher randomization test (FRT) with large n, is discussed in Sec. 3, where its results are presented and discussed in Sec. 4. The paper concludes with some general discussion in Sec. 5.

2 Neyman and Fisher inference

Let $Y_i(w) \in R$ denote the potential outcome for unit i with binary indicator w referring to the treatment group, i.e. $w = 1$ implies *treatment* and $w = 0$ implies *control*. In a completely randomized experiment with n units, n_1 units are assigned to the treatment group and n_0 units to the control. As a unit can only be assigned to one of the two groups, letting $W_i = 1$ or 0 if unit i is assigned to treatment or control group, respectively, the observed outcome can be either $Y_i(W_i = 1)$ or $Y_i(W_i = 0)$. In a more compact form, we can thus write an observed outcome for unit i as

$$Y_i = W_i Y_i(1) + (1 - W_i)Y_i(0). \tag{1}$$

Our inference for Y_i, or any linear combinations thereof, will be based on the so-called SUTVA assumption (Rubin 1980), which implies that there is no interference between individuals and the same treatment. The quantity of our main interest, the mean difference estimator (MDE), is defined, using (1), as

$$\widehat{\tau} = \overline{Y}_1 - \overline{Y}_0, \tag{2}$$

where

$$\overline{Y}_w = \frac{1}{n_w} \sum_{i:W_i=w}^{n_w} Y_i, \, w = 0, 1.$$

For the inference of $\widehat{\tau}$, we need additional assumptions to be stated later.
Throughout the paper we consider a sample of size n to be a random sample from a (potential) finite population of N units, $n \leq N$, and special cases thereof. For clarity, we will index $\widehat{\tau}$ to indicate over which distribution we randomize, i.e, what sampling design is being considered. For example, we let $\widehat{\tau}_{N,n}$ denote the estimator over random sampling from the population and over random treatment assignment within the random sample. When $n = N$, we for simplicity denote $\widehat{\tau}_{N,n} = \widehat{\tau}_n$ and when sampling from the super population we denote the estimator $\widehat{\tau}_{\infty,n}$.

2.1 Neyman inference

Given population and sample sizes, N and n, respectively, a total of $S = \binom{N}{n}$ random samples can be drawn in this set up. Let $\mathbf{u}_s^n$ be a vector containing the indices of the units in the sth sample, for $s = 1, ..., S$. The sample average treatment effect (SATE) for the sth sample then follows as

$$\text{SATE}(\mathbf{u}_s^n) = \frac{1}{n} \sum_{i \in \mathbf{u}_s^n} (Y_i(1) - Y_i(0)), \tag{3}$$

where the population average treatment effect (PATE) is

$$\tau = \mu_1 - \mu_0,$$

with $\mu_w = \frac{1}{N} \sum_{i=1}^{N} Y_i(w)$, $w = 0, 1$, denoting the population mean. Note that with homogeneous treatment effects $\text{SATE}(\mathbf{u}_s^n) = \tau$, $\forall s = 1, ..., S$.
Now, it can be shown that (see Aronow et al. 2014)

$$V(\widehat{\tau}_{N,n}) = \frac{1}{N-1} \left\{ \frac{N - n_1}{n_1} \sigma_{Y(1)}^2 + \frac{N - n_0}{n_0} \sigma_{Y(0)}^2 + 2\sigma_{Y(1),Y(0)} \right\}, \tag{4}$$

where

$$\begin{aligned} \sigma_{Y(w)}^2 &= \frac{1}{N} \sum_{i=1}^{N} (Y_i(w) - \mu_w)^2 \text{ and } \sigma_{Y(1)Y(0)} \\ &= \frac{1}{N} \sum_{i=1}^{N} (Y_i(1) - \mu_1)(Y_i(0) - \mu_0), \, w = 0, 1. \end{aligned}$$

Likewise, the variance of the heterogeneous treatment effect follows as

$$\begin{aligned} \sigma_\tau^2 &= \frac{1}{N} \sum_{i=1}^{N} (Y_i(1) - Y_i(0) - (\mu_1 - \mu_0))^2 \\ &= \sigma_{Y(1)}^2 + \sigma_{Y(0)}^2 - 2\sigma_{Y(1),Y(0)}, \end{aligned}$$

so that, we can re-write (4) as

$$\begin{aligned} V(\widehat{\tau}_{N,n}) &= \frac{1}{N-1} \left\{ \frac{N - n_1}{n_1} \sigma_{Y(1)}^2 + \frac{N - n_0}{n_0} \sigma_{Y(0)}^2 + \sigma_{Y(1)}^2 + \sigma_{Y(0)}^2 - \sigma_\tau^2 \right\} \\ &= \frac{N}{N-1} \left\{ \frac{1}{n_1} \sigma_{Y(1)}^2 + \frac{1}{n_0} \sigma_{Y(0)}^2 \right\} - \frac{1}{N-1} \sigma_\tau^2. \end{aligned} \tag{5}$$

With σ_τ^2 fixed, the last term in (5) vanishes and the multiplying factor $N/(N-1) \to 1$ for $N \to \infty$, reducing $V(\widehat{\tau}_{N,n})$ to the usual form of variance of two independent samples, i.e.,

$$V(\widehat{\tau}_{N,n}) = \left\{ \frac{1}{n_1}\sigma^2_{Y(1)} + \frac{1}{n_0}\sigma^2_{Y(0)} \right\} [1 + o(1)]. \tag{6}$$

Then, the standard central limit theorem (CLT) applies under random sampling mechanism. It is easy to show (see Theorem 5.1 in the Appendix) that, as $n, N \to \infty$,

$$\frac{\widehat{\tau}_{N,n} - \tau}{\sqrt{\mathrm{Var}(\widehat{\tau}_{N,n})}} \xrightarrow{\mathcal{D}} N(0, 1), \tag{7}$$

The proof follows from the discussion above, or as a special case of Example 6 in Li and Ding (2017).

It is evident that under the super population assumption

$$\frac{\widehat{\tau}_{\infty,n} - \tau}{\sqrt{\mathrm{Var}(\widehat{\tau}_{\infty,n})}} \xrightarrow{\mathcal{D}} N(0, 1),$$

but with $\sigma^2_{Y(w)} = \mathrm{E}(Y_i(w) - \mu_w)^2$, $\mu_w = \mathrm{E}(Y_i(w))$, $w = 0, 1$ in (6).

With regard to Neyman's within-sample inference, let

$$\overline{Y}(w) = \frac{1}{n}\sum_{i=1}^{n} Y_i(w) \text{ and } S^2_{Y(w)} = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i(w) - \overline{Y}(w))^2, \; w = 0, 1.$$

This helps us define

$$V(\widehat{\tau}_n) = \frac{S^2_{Y(1)}}{n_1} + \frac{S^2_{Y(0)}}{n_0} - \frac{S^2_{\tau}}{n}, \tag{8}$$

as originally given by Neyman (1923), where S^2_τ is the sample variance of the heterogeneous treatment, defined as

$$S^2_\tau = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i(1) - Y_i(0) - (\overline{Y}(1) - \overline{Y}(0))^2.$$

This structure can be used to test Neyman's SATE hypotheses

$$H_0^n(N): \; \mathrm{SATE}(\mathbf{u}_s^n) = 0 \; \text{ vs. } \; H_1^n(N): \; \mathrm{SATE}(\mathbf{u}_s^n) \neq 0. \tag{9}$$

Recall also that the corresponding PATE hypotheses are

$$H_0(N): \tau = 0 \; \text{ vs. } \; H_1(N): \tau \neq 0. \tag{10}$$

It follows (see Theorem 5.2 in the Appendix) from $\widehat{\tau}_n$ as test statistic that for SATE hypotheses,

$$\frac{\widehat{\tau}_n - \tau}{\sqrt{\mathrm{Var}(\widehat{\tau}_n)}} \xrightarrow{\mathcal{D}} N(0, 1), \tag{11}$$

where $\mathrm{Var}(\widehat{\tau}_n)$ is given in (8) and $\widehat{\tau}_{N,n} = \widehat{\tau}_n$ for $N = n$, so that $\mathrm{SATE}(\mathbf{u}_s^n) \to \tau$ when $n \to \infty$.

Neyman (1923) proposed

$$\widehat{\mathrm{Var}(\widehat{\tau}_n)} = \frac{s_{Y_1}^2}{n_1} + \frac{s_{Y_0}^2}{n_0}, \tag{12}$$

as a consistent estimator of the variances, simultaneously, for the inference to the sample (cf. Eqn. 11) and to the population for $N \to \infty$ (cf. Eqn. 7), where

$$s_{Y_w}^2 = \frac{1}{n_w - 1} \sum_{i:W_i=w}^{n_w} (Y_i - \overline{Y}_w)^2,$$

Note that, (12) is obtained by ignoring S_τ^2/n in $V(\widehat{\tau}_n)$ making $\widehat{\mathrm{Var}(\widehat{\tau}_n)}$ an upper-bound estimator of $\mathrm{Var}(\widehat{\tau}_n)$. Under rather weak assumptions (see assumptions (5.1) and (5.2) in the Appendix) we can replace $\mathrm{Var}(\widehat{\tau}_n)$ in equation (11) by the Neyman variance estimator $\widehat{\mathrm{Var}(\widehat{\tau}_n)}$.

Thus, for a realized experiment where the jth allocation is randomly selected, $j = 1, ..., \binom{n}{n_1}$, the asymptotic p-value of a two-sided test can be approximated by

$$\pi_N = 2\Phi(-\widehat{\tau}_n^j / \sqrt{\widehat{Var(\widehat{\tau}_n^j)}})), \tag{13}$$

where $\widehat{\tau}_n^j$ is the estimate of the jth allocation based on the sample $\mathbf{u}_s^n$ and $\Phi(.)$ is the distribution function of the normal distribution.

To summarize, the same test statistics are being used for inference to the PATE and SATE. In order to establish (11) we are implicitly assuming that the experiment is conducted on the whole populations.

2.2 Fisher's exact randomization test

Consider Fisher's null and alternative hypotheses, respectively, as

$$H_0^n(F) : Y_i(1) = Y_i(0)\ \forall\ i \in \mathbf{u}_s^n \ \text{ vs. } \ H_1^n(F) : Y_i(1) \neq Y_i(0),\ i \in \mathbf{u}_s^n,$$

where $H_0^n(F)$ coincides with Neyman's null, $H_0^n(N)$, in (10) under homogeneous treatment effect within the sample, i.e. we can write

$$H_0^n(F) : SATE(\mathbf{u}_s^n) = 0 \text{ against } SATE(\mathbf{u}_s^n) \neq 0.$$

The exact Fisher randomization test (FRT) is performed by estimating the treatment effect under all possible permutations of the 'potential' outcomes under $H_0^n(F)$. To see this, let $A = \binom{n}{n_1}$ and let the matrix $\mathbf{W} = (w_{ij}) \in R^{n \times A}$ arrange all possible random allocations in a complete randomized experiment such that $w_{ij} = 0$ if unit i is not treated and $w_{ij} = 1$ if treated. Denoting $\mathbf{Y}(w) = (Y_1(w), ..., Y_n(w))'$, $w = 0, 1$, the vector of observed outcomes is defined as

$$\mathbf{Y}= \mathbf{W}^j\mathbf{Y}(1) + (\mathbf{1} - \mathbf{W}^j)\mathbf{Y}(0),$$

where $\mathbf{W}^j = (W_1^j, ..., W_n^j)'$ is the specific allocation vector for jth experiment $j = 1, ...A$ and $\mathbf{1}$ is a vector of 1's. The exact p-value for a two-sided hypothesis can then be obtained as

$$\pi_F = \Pr\left\{|\widehat{\tau}(\mathbf{W}, \mathbf{Y})|) \geq |\widehat{\tau}_n^j||H_0^n(F)\right\} \tag{14}$$

where $\widehat{\tau}(\mathbf{W}, \mathbf{Y})$ is the symmetric distribution of estimates under the null over all A allocations in $\mathbf{W}$. As this test is derived solely from the actual randomization, the size of the test is always correct.

To consider the asymptotic case, a direct application of a central limit theorem in Ding (2017) ensures asymptotic normality of $\widehat{\tau}(\mathbf{W}, \mathbf{Y})$ under $H_0^n(F)$, with limiting variance

$$Var(\widehat{\tau}(\mathbf{W}, \mathbf{Y}|H_0^n(F)) = \frac{n}{n_1 n_0} s^2,$$

where $s^2 = (n-1)^{-1}\sum_{i=1}^n (Y_i - \overline{Y})^2$ and $\overline{Y} = n^{-1}\sum_{i=1}^n Y_i$. Comparing this variance of the normal approximation of Fisher's exact test with that of the Neyman's conservative test, the discrepancy comes out to be (see Ding 2017)

$$\begin{aligned}&\mathrm{Var}(\widehat{\tau}(\mathbf{W}, \mathbf{Y}|H_0^n(F)) - Var(N)\\ &= \left(\frac{1}{n_0} - \frac{1}{n_1}\right) \times (S_1^2 - S_0^2) + \frac{1}{n}(\overline{Y}(1) - \overline{Y}(0))^2 + o_p(n^{-1}),\end{aligned} \tag{15}$$

using $Var(N) = S^2_{Y(1)}/n_1 + S^2_{Y(0)}/n_0$. This implies that, Fisher's and Neyman's tests are asymptotically equivalent under the null if either $n_0 = n_1$ or if $S_1^2 = S_0^2$. Otherwise, the relative difference of variances grows with the size of the treatment effect.

2.2.1 Permutation testing

Let $F_{Y(w)}$ denote the distribution of the potential outcomes in the population, where $w = 0, 1$. For a random sample of observed outcomes $\{Y_i(W_i)\}_{i=1}^n$ that are exchangeable (e.g. under SATE$(\mathbf{u}_s^n) = \tau$ with homoscedasticity), Hoeffding (1951)'s well-know permutation CLT gives (see also Li and Ding 2017; Boos and Stefanski 2013), under the null,

$$\widehat{\tau}(\mathbf{W}, \mathbf{Y})/\sqrt{\mathrm{Var}(\widehat{\tau}_{\infty,n})} \xrightarrow{\mathcal{D}} N(0, 1),$$

as $n \to \infty$. Romano (1990) showed that if the distributions of $\{Y_i(1)\}_{i=1}^{n_1}$ and $\{Y_i(0)\}_{i=1}^{n_0}$ have common mean μ and finite variances σ_1^2 and σ_0^2, and if $n_1/n_0 \to 1$, then

$$\widehat{\tau}(\mathbf{W}, \mathbf{Y}) \xrightarrow{d} \widehat{\tau}_{\infty,n}/\sqrt{V(\widehat{\tau}_{\infty,n})} \text{ as } n \to \infty.$$

Note that, if the exact FRT is carried out using the statistic

$$\widehat{\tau}_{\infty,n}/\sqrt{\frac{s^2_{Y_1}}{n_1} + \frac{s^2_{Y_0}}{n_0}},$$

then its type I error asymptotically coincides with the nominal α under $H_0(N)$, and it also retains the exact error rate of α in finite samples under the sharp null (Chung and Romano 2013). Furthermore, under normality of the super populations, i.e. $Y(w) \sim N(\mu_w, \sigma^2)$, $w = 0, 1$, the exact FRT is the UMP test (Lehmann 1959, §5.8). Further, if n_1/n_0 is bounded, where $n \to \infty$, then the exact FRT can be approximated with the standard t-test.

3 Power of exact FRT for large *n*

Now, consider power. For exact FRT, power under an alternative is a fixed quantity for given set of $\mathbf{Y}(1)$ and $\mathbf{Y}(0)$ in the design space $\mathbf{W}$. For each allocation vector $\mathbf{W}^j$, $j = 1, ..., A$, there is a corresponding $\mathbf{Y}$ and an exact p-value, π_F, defined by Eqn. (14). The power of the exact FRT for inference to the units of the sample is defined as the proportion of the A allocations in $\mathbf{W}$ that achives π_F smaller than or equal to α. A Monte Carlo simulation of "exact power" would thus require A^2 calculations.

For large n, π_F can be approximated through simulations, and estimate π_F using $M < A$ randomly drawn allocations. For sufficiently large M, this approximation, say $\widehat{\pi}_F$, will be close to π_F. Achieving a similar accuracy for estimated power through simulation-based approximation obviously needs M^2 computations.

We dedicate this section to discuss an alternative approximation strategy, where we do exact FRT test and power computations by using independent subsets of allocations, and then averaging the results over the subsets. To motivate the case, we begin by briefly reviewing the small simulation study reported in Ding (2017).

3.1 Motivation

The following description is directly taken from Ding (2017). Let $Y(0) \sim N(0, 1/16)$ and $n = 100$. For $Y(1)$, the data generation process (DGP) is bifurcated as: (a) $Y(1) \sim N(\tau, 1/16)$, $n_1 = n_0 = 50$, and (b) $Y(1) \sim N(\tau, 1/4)$, $n_1 = 70$, $n_0 = 30$, where $\tau = 1/10$ in both (a) and (b). Since, $Y_i(0) \neq Y_i(1)$, $\forall i \in \mathbf{u}_s^{100}$ for all generated samples, $\tau = 0$ will differ from $H_0^n(F)$ or, generally, from $H_0^n(N)$ in both DGPs. To make it more precise, let $Y_i(0) = \varepsilon_{0i}$ and $Y_i(1) = \tau + \varepsilon_{1i}$, assuming $\mathrm{E}(\varepsilon_{0i}) =$

$\mathrm{E}(\varepsilon_{1i}) = 0$ in the super population. Then, for a sample of size $n = 100$,

$$\mathrm{SATE}(\mathbf{u}_s^{100}) = \tau + \overline{\varepsilon}_{1s} - \overline{\varepsilon}_{0s},$$

with $\overline{\varepsilon}_{ws} = \sum_{i=1}^{100} \varepsilon_{wi}/100$, $w = 0, 1$, where $\overline{\varepsilon}_{0\,s} \neq \overline{\varepsilon}_{1\,s}$ for most samples. Averaging over 1000 simulation runs, Ding (2017) computes the p-values of the Neyman's test as

$$2\Phi\left(-\widehat{\tau}^{j}_{100}/\sqrt{\widehat{\mathrm{Var}(\widehat{\tau}^{j}_{100})}}\right),\quad j = 1, ..., 1000,$$

and the p-values of the exact FRT are approximated as

$$\widehat{\pi}^{j}_{F} = \frac{1}{M}\sum_{m=1}^{M} \mathbf{1}\left(|\widehat{\tau}(\mathbf{W}^m, \mathbf{Y})| \geq |\widehat{\tau}^{j}_{100}|\right),\quad j = 1, ..., 1000, \tag{16}$$

where $M \equiv 10^5$ and $\mathbf{1}(\cdot)$ is an indicator function. This, however, implies that the simulation set up in Ding (2017) is not for testing $H_0^n(N)$. As alluded to above, this set up rather pertains to the power for testing $H_0(N)$, although using only a single sample from this super population.

From Table 1 in Ding (2017), where the results of the two DGPs are reported, we note an overall power, for case (a), as 0.512 for Neyman's test and 0.497 for Fisher's test. The results are not only similar, but also close to the expected power under repeated sampling, i.e. $= 1 - \Phi(-0.04) + \Phi(-3.96) = 0.516$. For case (b), the power of Neyman's test is 0.07 while that of Fisher's test is 0.008, which is even lower than the nominal level. Both are however far from the expected power under repeated sampling which, in this case, is $= 1 - \Phi(0.6302) + \Phi(-3.2898) = 0.265$.

The crux of the aforementioned comparison is that Ding (2017) does not consider all possible estimates $\widehat{\tau}^{j}_{100}$, , $j = 1, ..., A$, rather only the estimates in the set $\widehat{\tau}^{j}_{100}$, $j = 1, ..., 1000$, in a subset $\mathbf{W}_{B_1}$ of $\mathbf{W}$. The subset $\mathbf{W}_{B_1}$, with $\mathrm{card}(\mathbf{W}_{B_1}) = 1000$, is only one set out of $\binom{A}{1000}$ with $A = \binom{100}{50}$ and $\binom{100}{70}$, respectively. Another random subset, say $\mathbf{W}_{B_2}$, would most likely give other results. The problem is less pertinent for Neyman inference since the corresponding statistic does not depend on the empirical distribution over all allocations. However as seen for the DGP (b), the power can be quite far from expected even for Neyman's test.

3.2 Exact FRT in allocation subsets

The power of a size α FRT can be computed as

$$p_F = \frac{1}{A}\sum_{j=1}^{A} \mathbf{1}(\pi^{j}_{F} \leq \alpha),$$

where

$$\pi_F^j = \frac{1}{A}\sum_{m=1}^{A} \mathbf{1}\left(|\widehat{\tau}(\mathbf{W}^m, \mathbf{Y})| \geq |\widehat{\tau}_n^j|\right), \quad m = 1, ..., A. \tag{17}$$

Ding (2017) selected a subset of 1000 allocations and calculated the power within this set, i.e.

$$p_{F|1000} = \frac{1}{1000}\sum_{j=1}^{1000} \mathbf{1}(\widehat{\pi}_F^j \leq \alpha).$$

The power calculated over the subset $\mathbf{W}_{B_1}$ may not be a good approximation of the real power of the exact FRT, whether or not the complete set of allocations or the Monte Carlo approximation of the exact p-value is used. If 10^5 allocations provide good enough precision for the approximation of p-values, then the power should also be well approximated by the same set. This, however, requires 10^{10} iterations in one cell, which is simply not possible. We, instead, will make use of the algorithm developed in Johansson and Schultzberg (2020).

In the re-randomization context, Johansson and Schultzberg (2020) suggest an alternative to Monte Carlo approximation of the p-value with large n. For the following discussion of their approach, it is instructive to reformulate the definition of the SATE (cf. Eqn. 3) as the average of all potential estimates in a sample, i.e.

$$\text{SATE}(\mathbf{u}_s^n) = \frac{1}{A}\sum_{j \in \mathbf{W}} \widehat{\tau}_n^j. \tag{18}$$

In (18), it follows from the symmetry that unbiasedness of the MDE stems from that for any single allocation; e.g. $\mathbf{W}^j = (0, 1, 0, 1, ..., 0, 1)'$, there exist a mirror allocation with 1's and 0's exchanged.[3] This means, the unbiasedness can be preserved for any set of allocations $\mathbf{W}_{B_k}$, $k = 1, ..., K$, with cardinality larger than two, as long as the set includes only pairs of mirror allocations. To emphasize a set containing only mirror allocations, we add the superscript $*$. For example, $\mathbf{W}_{B_k}^*$ is a set of allocations of cardinality B_k, i.e. $\text{card}(\mathbf{W}_{B_k}^*) = B_k$, containing $B_k/2$ pairs of mirror allocations.

Here we simply take K random subsets, all of size B^*, where B^* is small enough to conduct the exact test. The exact p-value for a two sided hypothesis test for a given sample for each subset of allocations is thus defined as

$$\pi_{F|B_k} = \Pr(|\widehat{\tau}(\mathbf{W}_{B_k}^*, \mathbf{Y})|) \geq |\widehat{\tau}_n^j(k, s)|, k = 1, ..., K,$$

where $\widehat{\tau}_n^j(k, s)$ denotes an estimate for kth subset in sth sample. Even though the level of this test is always correct and the MDE is unbiased, it is important to note that the distributions of $\widehat{\tau}(\mathbf{W}_{B_k}^*, \mathbf{Y})$ and $\widehat{\tau}(\mathbf{W}, \mathbf{Y})$ generally differ, which explains why

[3] This always holds for balanced experiments. With n small and odd, the unbiasedness of the mean difference estimator may no longer hold; see e.g. (Morgan and Rubin 2012, p. 9) for an example with $n = 3$ and $n_1 = 2$.

the power with $\mathbf{W}^*_{B_k}$ generally differs from the *exact* power. Computing power by averaging over the K subsets reduces the approximation error.

4 Simulation study of exact FRT

To validate the discussion so far, we perform three simulation studies. The first re-examines the small simulation in Ding (2017). The second uses the same DGP but with $n = 12$. In the third, we study the performance of the two strategies in a pair wise stratified experiment. The two latter cases allow us to obtain the *exact* power of the exact FRT.

4.1 Case I: re-examining Ding (2017)'s study

We extend the DGPs in Ding (2017), keeping $Y(0) \sim N(0, 1/16)$ throughout where, for the alternative, we let (a) $Y(1) \sim N(\tau, 1/16)$ with $n_1 = n_0 = 50$, (b) $Y(1) \sim N(\tau, 1/4)$ with $n_1 = 70, n_0 = 30$, and (c) $Y(1) \sim N(\tau, 1/4)$ with $n_1 = 30, n_0 = 70$. Settings (a) and (b) with $\tau = 0.10$ are as in Ding (2017), whereas setting (c) is considered as an extension thereof.

For power, we consider an increasing alternative with $\tau = \{0, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3\}$. As another extension, the same simulations are also conducted under homogeneous effects, where the outcome under treatment is generated as $Y(1) = Y(0) + \tau$. As in Ding (2017), we test $H_0(N) : \tau = 0$ using the 5% level. But we repeat the test over $s = 1, ...100$. Likewise, we let $B^* = 1000$, but calculate the power of Neyman and the exact test for each random subset $\mathbf{W}^*_{B_k}$, $k = 1, ..., 50$ of $\mathbf{W}$. That is, for Neyman test we calculate the power as

$$p_N(k, s, \tau) = \frac{1}{1000} \sum_{j \in B_k} \mathbf{1}(\pi^j_N(s, \tau) \leq 0.05),$$

with $k = \{1, ..., 50\}$, $s = \{1, ..., 100\}$ and τ given above, where each p-value, $\pi^j_N(s, \tau)$, is the standard p-value from a two-sample t-test with the Satterthwaite approximation of the degrees of freedom (Welch 1947),[4] The power of Fisher's test is computed as

$$p_{F|B_k}(k, s, \tau) = \frac{1}{1000} \sum_{j \in B_k} \mathbf{1}(\pi^j_{F|B_k}(k, s, \tau) \leq 0.05), \tag{19}$$

with k, s, τ as above, where, for $j = 1, ..., 1000$,

$$\pi^j_{F|B_k}(k, s, \tau) = \frac{1}{1000} \sum_{m \in B^*_k} \mathbf{1}(|\widehat{\tau}(\mathbf{W}^{*m}_{B_k}, \mathbf{Y}))| \geq |\widehat{\tau}^j_n(k, s)|).$$

[4] In R it follows by default using the base function `t.test()`.

The number of replicates in a complete MC simulation should be $1.0089 \times 10^{29} (= \binom{100}{50})$ for DGPs (a) and (b), and 2.9372×10^{25} $(= \binom{100}{70} = \binom{100}{30})$ for (c). Here the number of replicates are $50 \times 1000 = 50,000$ for each cell when averaging over the subsets.

The results for all DGPs and effect types (homogeneous vs. heterogeneous) are summarized using a four-way analysis of variance (ANOVA), with factors (i) Inference (Fisher and Neyman), (ii) Effect size (seven levels), (iii) Subset (50 levels), and (iv) Sample (100 levels). We restrict the analysis to second order interactions. The upper panel of Table 1 reports for the homogeneous effects, with results corresponding to DGPs (a), (b) and (c) in columns 3–4, 5–6 and 7–8, respectively; the lower panel depicts the same set up for the heterogeneous case.

The strength of the coefficient of determination, R^2, exceeding 99% for all layouts, is an encouraging indicator for the summarization of results by ANOVA. The first take-home message from Table 1 is that the effect size is the most important factor, as expected. Secondly, the sample factor contributes quite substantially in explaining the variance, both as main effect but also in terms of interaction with the effect size. Again as expected, this particularly holds for heterogeneous effects case. Thirdly, differences in inference contributes to a small extent to the variation. However, as this factor has just one degrees of freedom (DF), the F-statistic is quite high, even though not close to the F-statistic for the sample (including interaction effects). The final conclusion pertains to the Subset factor, in that it does not add much to the explanation of the variance, which means the approximation errors incurred by using the subsets in the simulations is small.

Figure 1 depicts the simulation results across τ, with heterogeneous and homogeneous effects displayed in left and right panels, and DGPs (a), (b) and (c) displayed in upper, middle and lower panels, respectively. The box plots display inference to the experiment across 100 random samples, i.e. the fraction of rejected tests across τ in each of 100 experiments. With homogeneous effect, the fraction of rejected tests provide the size when $\tau = 0$ and power when $\tau > 0$. Inference to the population is obtained by averaging the fraction of the rejected over the 100 random samples, as displayed by the power curves.

For homogeneous effects, we observe similar results with respect to inference to the experiment and to the population. With respect to inference to the experiment, the Fisher test has by design the correct size, while there is a very small divergence for the Neyman test. There is a small divergence in power within samples. The maximum divergence in power between the two tests is 12%, for $\tau = 0.15$ in panel B. This suggests that, for homogeneous effects, the conclusion from a single experiment does not generally depend on the type of test conducted. As seen by the power curves, the two tests are indistinguishable for the inference to the population.

For heterogeneous effect, we observe substantial variation in fraction of rejected within samples. As expected, a substantial amount of experiments rejects the hypothesis of $\tau = 0$. The pattern is similar between the two tests in the balanced case and equal variance for case (a). The maximum divergence in power between the two tests is 12.5%, for $\tau = 0.15$. However with unbalanced designs and unequal variances substantial differences between the tests are seen. With $n_1 = 70$ (panel (b)) there are more cases rejecting $\tau = 0$ for the Neyman test, while it is the reverse when $n_1 = 30$

Table 1 ANOVA results for the Monte Carlo simulation with $n = 100$

	Df	Sum Sq	F value	Sum Sq	F value	Sum Sq	F value
Homogeneous							
		a		b		c	
Inference	1	0.75	4921.36	2.92	16401.53	2.92	16401.53
Effect	6	10013.54	10939388.72	9945.00	9311889.78	9945.00	9311889.78
Subset	49	0.01	1.86	0.02	2.55	0.02	2.55
Sample	99	30.42	2014.43	38.14	2164.63	38.14	2164.63
Inference× effect	6	0.88	956.43	2.77	2591.40	2.77	2591.40
Inference× subset	49	70.03	3.46	0.03	3.05	0.03	3.05
Inference× sample	99	0.24	16.12	0.74	41.90	0.74	41.90
Effect× subset	294	0.04	0.90	0.05	0.91	0.05	0.91
Effect× sample	594	36.47	402.48	37.49	354.56	37.49	354.56
Subset× sample	4851	1.99	2.69	2.41	2.79	2.41	2.79
Residuals	63951	9.76	($R^2 = 0.99$)	11.38	($R^2 = 0.99$)	11.38	($R^2 = 0.99$)
Heterogeneous							
		a		b		c	
Inference	1	0.53	5030.25	354.56	139538.48	354.56	139538.48
Effect	6	9743.08	15421849.24	7725.75	506754.31	7725.75	506754.31
Subset	49	0.01	1.19	0.01	0.08	0.01	0.08
Sample	99	542.94	52084.12	741.58	2948.02	741.58	2948.02
Inference× effect	6	0.81	1283.38	110.79	7266.91	110.79	7266.91
Inference× subset	49	0.00	0.70	0.01	0.07	0.01	0.07
Inference× sample	99	0.56	53.58	21.57	85.74	21.57	85.74
Effect× subset	294	0.03	0.90	0.05	0.07	0.05	0.07
Effect× sample	594	1419.12	22689.51	1752.97	1161.44	1752.97	1161.44
Subset× sample	4851	0.49	0.96	1.19	0.10	1.19	0.10
Residuals	63951	6.73	($R^2 = 0.99$)	162.49	($R^2 = 0.99$)	162.49	($R^2 = 0.99$)

(panel (c)). Thus with heterogeneous effects in unbalanced design, the conclusion for Fisher's test from a single experiment may likely differ from that using t-test. These differences in inference to the experiment is summarized in the power curves in the figure. For (a), the size and power over random sampling is the same for both strategies, which confirms the result in Lehmann (1959, §5.8). For (b), the average size of the t-test exceeds 5% and, neglecting the size distortion, the t-test is also more powerful than the exact Fisher test. For (c), the roles of the two tests are reversed.

The last thing to note from the figure is that there is no sign of a diverging difference in power with the effect size as suggested from Eqn. (15), neither for inference in the single experiment nor for inference over random sampling.

Note that, since we approach the numerical assessment by conducting exact Fisher inference within each subset and then averaging over all subsets, its usefulness can be gauged from the act that the results in Fig. 1 seem validating the theoretical results in

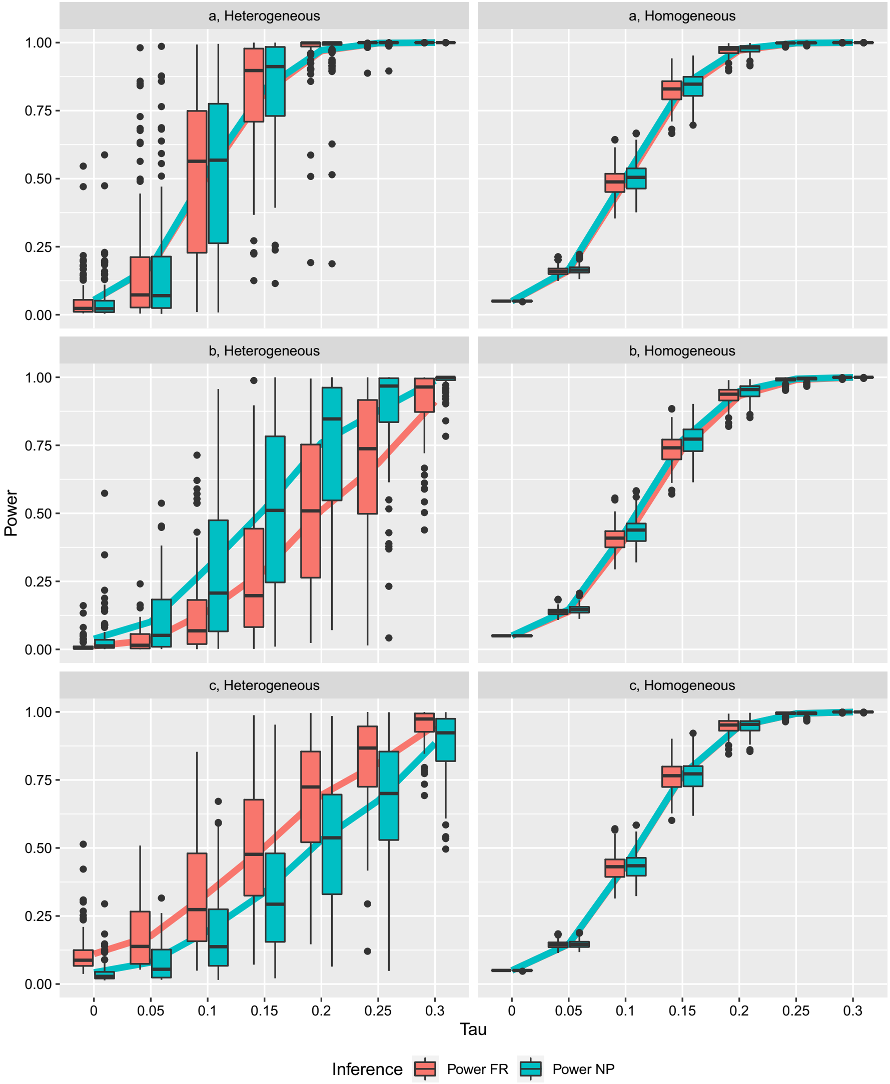

Fig. 1 Power comparison of Fisher and Neyman test for 100 independent samples in a complete randomized experiment with $n = 100$, where $n_1 = 50$ and $\sigma^2_{Y(0)} = \sigma^2_{Y(1)} = 1/16$ in the top panel, $n_1 = 70$ and $\sigma^2_{Y(1)} = 1/4$ in the middle panel, $n_1 = 30$ and $\sigma^2_{Y(1)} = 1/4$ in the bottom panel. Solid line is the power averaged over sets and samples

Lehmann (1959). In the next two sub-sections, we consider the performance in small experiment, where simulations are conducted within the complete set of allocations.

4.2 Small sample version

We use the same DGP as above, also keeping 100 random draws from the super population, now with $n = 12$ and $\tau = \{0, 0.25, 0.5, 0.75, 1, 1.25\}$, where n_1 is 6, 8 and 4 for DGP (a), (b) and (c), respectively. Figure 2 presents the results from the simulations across τ.

The box plots and lines display, respectively, average size and power over 100 samples. The results for heterogeneous and homogeneous effects are displayed in the left and right panels, respectively, where DGPs (a)–(c) are displayed in top, middle and bottom panels, respectively. The overall picture is the same as with $n = 100$. There is substantial variation in power for heterogeneous effects, but also the power curves for the two strategies overlap. There is also no sign of a diverging difference in power with the effect size as suggested by Eqn. (15). For homogeneous effects, there is some size distortion for t-test, and exact Fisher test is somewhat more powerful.

4.2.1 Pairwise stratification

Now we consider a pairwise stratified experiment. Let $Y_{ij}(w)$, $w = 0, 1,\ i = 1, ..., n$, $j = 1, 2$, be potential outcomes of units in a matched-pair experiment with $2n$ units (n pairs). The within-pair estimator

$$\widehat{\tau_i} = W_i(Y_{i1} - Y_{i2}) + (1 - W_i)(Y_{i2} - Y_{i1})$$

is unbiased for the within-pair average causal effect τ_i, where

$$\widehat{\tau_n}(s) = \frac{1}{n}\sum_{i=1}^{n}\widehat{\tau_i}$$

is an unbiased estimator of the sample treatment effect

$$\text{SATE}(\mathbf{u}_s^n) = \frac{1}{n}\sum_{i=1}^{n}\tau_i.$$

Following Ding (2017), the conservative variance estimator of $\widehat{\tau_n}(s)$ is

$$\text{Var}(N) = \frac{1}{n(n-1)}\sum_{i=1}^{n}(\widehat{\tau_i} - \widehat{\tau_n}(s))^2,$$

where Ding’s Theorem 4 show that

$$\text{Var}(\widehat{\tau}(\mathbf{W}, \mathbf{Y}|H_0^F)) = \frac{1}{n^2}\sum_{i=1}^{n}\widehat{\tau}_i^2,$$

so that

$$\text{Var}(\widehat{\tau}(\mathbf{W}, \mathbf{Y}|H_0^F)) - \text{Var}(N) = \frac{1}{n}\,\text{SATE}(\mathbf{u}_s^n)^2 + o_p(n^{-1}). \tag{20}$$

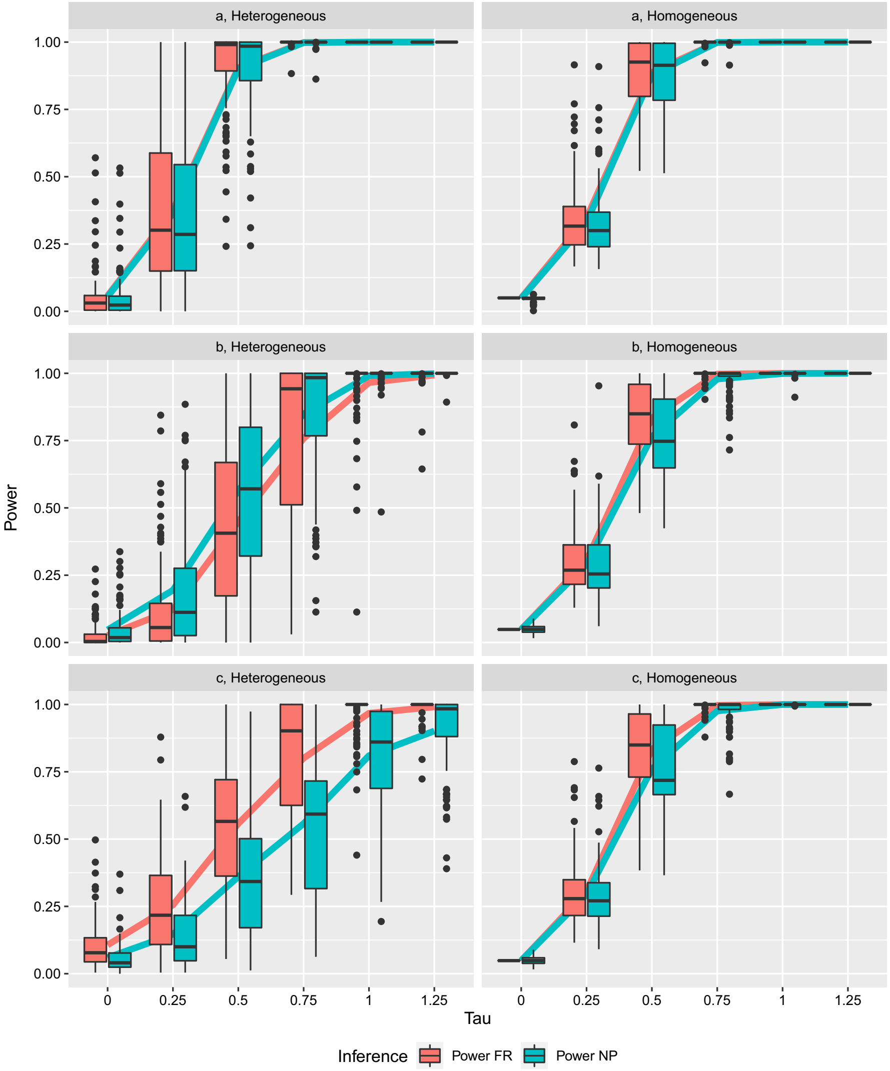

Fig. 2 Difference in power between Fisher and Neyman inference over $\binom{12}{6} = 924$, $\binom{12}{8} = 495$ and $\binom{12}{4} = 495$ possible allocations in 100 independent samples, respectively, in top, middle and lower panels. The solid line is the power averaged over the 100 samples

We draw 100 independent samples of size $n = 30$. For each sample, the p-values for all $2^{n/2} = 32{,}768$ possible allocations under the paired design are calculated. We again assume

$$Y(0) \sim N(0, 1/16), \tag{21}$$

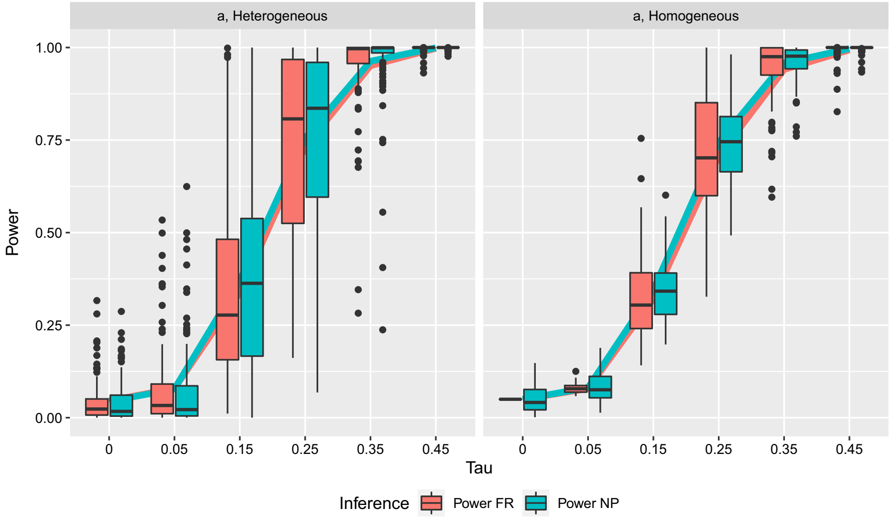

Fig. 3 Difference in power between Fisher and Neyman inference over all $2^{15} = 32,768$ possible allocations under pairwise stratified randomization in 100 independent samples with $n = 30$. The solid line is the power averaged over samples

and generate the counterfactual as

$$\text{(a)}: \ Y(1) \sim N(\tau, 1/16) \tag{22}$$

$$\text{(b)}: \ Y(1) = Y_i(0) + \tau, \tag{23}$$

where $\tau = \{0, 0.05, 0.15, 0.25, 0.35, 0.45\}$. The size and power for each sample is computed as the proportion of corresponding p-values below $\alpha = 0.05$. The results for heterogeneous effect under DGP (a) are shown in the left panel of Fig. 3 and those of homogeneous effect under DGP (b) are in the right panel. The most important finding is that the power curves of the two tests overlap so that we cannot discover any divergence in the power of t-test in comparison with the Fisher's test, as expected from Eqn. (20).

5 Discussion

Ding (2017) gives interesting theoretical results on the comparison of Neyman's and Fisher's two-sample inference based on the theory of potential outcomes. The present paper examines to what extent Ding (2017)'s results apply for the exact Fisher test for inference to the sample under the alternative. Based on the same data generating processes as in Ding (2017), we conduct a Monte Carlo study that captures the finite, but large, sample power properties of the exact Fisher randomization test. The results show no overall superiority of the Neyman test over the exact Fisher randomization test for any effect size. Instead, the property of a test being most powerful in this case seems to depend on the characteristics of the outcomes in the given sample.

For heterogeneous treatment effect, both tests have in general wrong size when testing the population (or scientific) null in a single experiment, which illustrates Fisher's concern (Fisher 1955). The crux in the single experiment case is that the sample average treatment effect (SATE) depends on the units sampled to the experiment and, with N fixed, SATE in general differs from zero. This fact perhaps pertains to what Fisher (1955, p. 69) meant with the statement "we consider a continuum of hypotheses, each eligible as null hypothesis". The within-sample asymptotic theory solves the problem by assuming the sample to be infinite, but most experiments are conduced on a finite sample whence the Neyman and Fisher tests have the same problem testing the scientific null.

It is however interesting to note that when testing for a sample (or statistic) null, the two tests may give different conclusions, at least so in unbalanced designs and with unequal variances. This gives some food for thought for theoretical research since at the end of the day, all we as statisticians have is the results from a single experiment.

Funding Open access funding provided by Uppsala University.

Appendix

Theorem 5.1 *Consider $\widehat{\tau}$ in (2) under a simple random sampling scheme, and let N, n_1, n_0, be as defined above. Given that n_1/n_0 is bounded, we have, as $n, N \to \infty$,*

$$\frac{\widehat{\tau}_{N,n} - \tau}{\sqrt{\mathrm{Var}(\widehat{\tau}_{N,n})}} \xrightarrow{\mathcal{D}} N(0,1),$$

with

$$\mathrm{Var}(\widehat{\tau}_{N,n}) = \left\{ \frac{1}{n_1}\sigma^2_{Y(1)} + \frac{1}{n_0}\sigma^2_{Y(0)} \right\} [1 + o(1)].$$

For such limit theorem, see Erdös and Renyi (1959) and Hajek (1960), with unified and generalized results presented in Li and Ding (2017).

Theorem 5.2 *For $\widehat{\tau}_n$ as a test statistic for $H_0^n(N)$, we have, as $n \to \infty$,*

$$\frac{\widehat{\tau}_n - \tau}{\sqrt{\mathrm{Var}(\widehat{\tau}_n)}} \xrightarrow{\mathcal{D}} N(0,1),$$

where

$$\mathrm{Var}(\widehat{\tau}_n) = \frac{S^2_{Y(1)}}{n_1} + \frac{S^2_{Y(0)}}{n_0} - \frac{S^2_{\tau}}{n},$$

Corollary 5.1 *Under Assumptions 5.1-5.2, the limit in Theorem 5.2 remains valid by replacing* $\mathrm{Var}(\widehat{\tau}_n)$ *with its consistent estimator* $\frac{s^2_{Y_1}}{n_1} + \frac{s^2_{Y_0}}{n_0}$.

Assumption 5.1 Let $n_k/n \to c_k \in (0, 1)$, as $n_k \to \infty$, $k = 0, 1$, where $n = n_1 + n_0$.

Assumption 5.2 $\max\limits_{1\leq i\leq n} d^2_{i,w}/\sum_i d^2_{i,w} \to 0$, as $n \to \infty$, where $d_{i,w} = Y_{iw} - \overline{Y}_w$, $w = 0, 1$.

References

Aronow PM, Green DP, Lee DK (2014) Sharp bounds on the variance in randomized experiments. Ann Stat 42(3):850–871

Athey S, Imbens G (2017) Chapter 3 - the econometrics of randomized experimentsa. In: Banerjee AV, Duflo E (eds) Handbook of Field Experiments, volume 1 of Handbook of Economic Field Experiments. North-Holland, Amsterdam, pp 73–140

Bertsimas D, Johnson M, Kallus N (2015) The power of optimization over randomization in designing experiments involving small samples. Oper Res 63(4):868–876

Boos DD, Stefanski LA (2013) Essential statistical inference: theory and methods. Springer, New York

Chung E, Romano JP (2013) Exact and asymptotically robust permutation tests. Ann Stat 41(2):484–507

Ding P (2017) A paradox from randomization-based causal inference. Stat Sci 32(3):331–345

Ding P, Dasgupta T (2018) A randomization-based perspective on analysis of variance: a test statistic robust to treatment effect heterogeneity. Biometrika 105(1):45–56

Erdös P, Renyi A (1959) On the central limit theorem for samples from a finite population. Publ Math Inst Hung Acad Sci 4:49–61

Fisher R (1955) Statistical methods and scientific induction. J Royal Stat Soc: Ser B 17(1):67–78

Hajek J (1960) Limiting distributions in simple random sampling from a finite population. Math Inst Hung Acad Sci 5:361–374

Hoeffding W (1951) A combinatorial central limit theorem. Ann Math Stat 22(4):558–566

Johansson P, Rubin DB, Schultzberg M (2021) On optimal rerandomization designs. J Royal Stat Soc: Ser B 83(2):395–403

Johansson P, Schultzberg M (2020) Rerandomization strategies for balancing covariates using pre-experimental longitudinal data. J Comput Graph Stat 29(4):798–813

Kallus N (2018) Optimal a priori balance in the design of controlled experiments. J Royal Stat Soc Ser B: Stat Methodol 80(1):85–112

Kapelner A, Krieger A, Sklar M, Shalit U, Azriel D (2021) Harmonizing optimized designs with classic randomization in experiments. Am Stat 75(2):195–206

Krieger A, Azriel D, Kapelner A (2019) Nearly random designs with greatly improved balance. Biometrika 106(3):695–701

Lauretto MS, Stern RB, Morgan KL, Clark MH, Stern JM (2017) Haphazard intentional allocation and rerandomization to improve covariate balance in experiments. AIP Conf Proc 1853:050003

Lehmann EL (1959) Testing statistical hypotheses, 1st edn. Springer, Cham

Li X, Ding P (2017) General forms of finite population central limit theorems with applications to causal inference. J Am Stat Assoc 112(520):1759–1769

Morgan KL, Rubin DB (2012) Rerandomization to improve covariate balance in experiments. Ann Stat 40(2):1263–1282

Neyman J (1923) On the application of probability theory to agricultural experiments. Essay on principles. Section 9. Transl Stat Sci (1990) 5(4):465–472

Romano JP (1990) On the behavior of randomization tests without a group invariance assumption. J Am Stat Assoc 85(411):686–692

Rosenbaum PR (2007) Interference between units in randomized experiments. J Am Stat Assoc 102(477):191–200

Rubin DB (1980) Randomization analysis of experimental data: the Fisher randomization test comment. J Am Stat Assoc 75(371):591–593

Rubin DB (1986) Comment. J Am Stat Assoc 81(396):961–962

Welch BL (1947) The generalization of 'student's' problem when several different population variances are involved. Biometrika 34(1–2):28–35

Wu J, Ding P (2021) Randomization tests for weak null hypotheses in randomized experiments. J Am Stat Assoc 116(536):1898–1913

Young A (2018) Channeling Fisher: randomization tests and the statistical insignificance of seemingly significant experimental results. Q J Econ 134(2):557–598

Zhao A, Ding P (2021) Covariate-adjusted Fisher randomization tests for the average treatment effect. J Econom 225:278–294