Statistical Science
2017, Vol. 32, No. 3, 356–361
DOI: 10.1214/17-STS610
Main article DOI: 10.1214/16-STS571

An Apparent Paradox Explained

Wen Wei Loh, Thomas S. Richardson and James M. Robins

We thank Peng Ding for bringing to light a paradox underlying the conventional conceptualization of Neymanian versus Fisherian inference for causal effects: although the Fisher null is a submodel of the Neyman null, Ding demonstrates in simulations that the Neyman test can reject the Neyman null without the Fisher test rejecting the Fisher null in two designs: balanced and unbalanced.

Ding restricts his analysis to asymptotic considerations. In particular, he explains the paradox by differences in large sample variances. We show that, for the balanced design, this explanation is incorrect empirically and also theoretically under Pitman asymptotics, as the asymptotic variances are equal; rather the paradox is wholly due to the Neyman test being anticonservative under the Fisher null in finite samples. Thus the paradox will disappear in large samples.

We conclude by addressing the implicit question raised by Ding's analysis: Are there better choices for test statistics and reference distributions for testing the Neyman and Fisher nulls that both avoid the small sample anticonservative behavior of the Neyman test against the Fisher null, and at the same time avoid the paradox at all sample sizes, while providing optimal test performance against (local) alternatives? We close by recommending a specific procedure.

1. FREQUENTIST p-VALUES: A REVIEW

Given an observation $\mathbf{x}^\circ$, suppose that we wish to test the simple null hypothesis that $\mathbf{x}^\circ$ arose from a particular density $f(\mathbf{x};\theta)$. A test is performed by comparing the observed value of a test statistic $r^\circ = r(\mathbf{x}^\circ)$ to a reference distribution $m(r)$, resulting in a candidate p-value:

$$\begin{aligned} &\mathrm{pv}(r,m,\theta;\mathbf{x}^\circ) \equiv \Pr_m\big[R \geq r^\circ\big] \quad \text{if } f(\mathbf{x}^\circ;\theta) > 0; \\ &\mathrm{pv}(r,m,\theta;\mathbf{x}^\circ) \equiv 0, \quad \text{otherwise,} \end{aligned} \tag{1}$$

where $R \sim m(\cdot)$; our notation for "pv" emphasizes that the candidate p-value depends on both the choice of test statistic and reference distribution. We use $\chi(r,m,\theta,\alpha;\mathbf{x}^\circ) \equiv I[\mathrm{pv}(r,m,\theta;\mathbf{x}^\circ) \leq \alpha]$ to be the corresponding α-level test. In a slight abuse of notation, we equivalently write $\mathrm{pv}(r,m,\theta;r^\circ)$ and $\chi(r,m,\theta,\alpha;r^\circ)$. We use $f_\theta(r) \equiv f(r;\theta)$ to be the marginal for $R = r(\mathbf{X})$, when $\mathbf{X} \sim f(\mathbf{x};\theta)$.

A candidate p-value $\mathrm{pv}(r,m,\theta;\mathbf{X})$ is said to be *conservative (at level α) for θ* if under $f(\mathbf{x};\theta)$, the probability $\Pr_\theta[\mathrm{pv}(r,m,\theta;\mathbf{X}) \leq \alpha]$ is $\leq \alpha$, *anticonservative* if $> \alpha$, *exact* if $= \alpha$. For $m(r) = f_\theta(r)$, $\mathrm{pv}(r, f_\theta, \theta; \mathbf{X})$ is exact at any level α^*, such that for some r^*, $f(r^*;\theta) > 0$ and $\Pr_\theta[r(\mathbf{X}) \geq r^*] = \alpha^*$; and is otherwise conservative. The following lemma demonstrates that $\chi(r, f_{\theta_0}, \theta_0, \alpha; \mathbf{X})$ is at least as powerful as any other conservative test $\chi(r,m,\theta_0,\alpha;\mathbf{X})$.

LEMMA 1. *If $\chi(r,m,\theta_0,\alpha;\mathbf{X})$ is a conservative α-level test for θ_0, then for any $\mathbf{x}^\circ$, if $\chi(r,m,\theta_0,\alpha;\mathbf{x}^\circ)$ rejects, so does $\chi(r, f_{\theta_0}, \theta_0, \alpha; \mathbf{x}^\circ)$.*

PROOF. By definition, $\chi(r,m,\theta_0,\alpha;\mathbf{X})$ is a conservative α-level test for θ_0 iff

$$\alpha \geq \Pr_{\theta_0}\big[r(\mathbf{X}) \geq c_\alpha\big] \equiv \mathrm{pv}(r, f_{\theta_0}, \theta_0; c_\alpha), \tag{2}$$

where c_α is the least c^* such that $\Pr_m(c^*) > 0$ and $\Pr_m[R \geq c^*] \leq \alpha$.

If $\chi(r,m,\theta_0,\alpha;\mathbf{x}^\circ) = 1$, then either $f(\mathbf{x}^\circ;\theta_0) = 0$, in which case the claim is trivial, or $r(\mathbf{x}^\circ) \geq c_\alpha$. In this case, $\mathrm{pv}(r, f_{\theta_0}, \theta_0; r(\mathbf{x}^\circ)) \leq \mathrm{pv}(r, f_{\theta_0}, \theta_0; c_\alpha) \leq \alpha$, so $\chi(r, f_{\theta_0}, \theta_0, \alpha; \mathbf{x}^\circ) = 1$. □

In what follows, in a minor abuse of notation, we will often write $\chi(r,m,\theta_0,\alpha;\mathbf{X})$ as $\chi(r,m,\theta_0,\alpha)$.

We use Θ_0 to denote a composite null hypothesis and define:

$$\mathrm{pv}(r,m_\theta,\Theta_0;\mathbf{x}^\circ) \equiv \sup_{\theta\in\Theta_0} \mathrm{pv}(r,m_\theta,\theta;\mathbf{x}^\circ) \quad \text{and}$$

$$\chi(r,m_\theta,\Theta_0,\alpha;\mathbf{x}^\circ) \equiv I\big[\mathrm{pv}(r,m_\theta,\Theta_0;\mathbf{x}^\circ) \leq \alpha\big],$$

Wen Wei Loh is Postdoctoral Research Associate, Department of Biostatistics, University of North Carolina, CB #7420, Chapel Hill, North Carolina 27599, USA (e-mail: wloh@u.washington.edu). Thomas S. Richardson is Professor and Chair, Department of Statistics, University of Washington, Box 354322, Seattle, Washington 98195, USA (e-mail: thomasr@u.washington.edu). James M. Robins is Mitchell L. and Robin LaFoley Dong Professor of Epidemiology, Department of Epidemiology, Harvard School of Public Health, 677 Huntington Avenue, Boston, Massachusetts 02115, USA (e-mail: robins@hsph.harvard.edu).

TABLE 1
Components of the two approaches described. To be precise, m_Z is the distribution of $|Z|$ where Z is standard normal, sometimes called the folded (standard) normal distribution

"Approach"	**Null hypothesis Θ**	**Test statistic $r(\cdot)$**	**Reference distribution $m(\cdot)$**
"Fisher"	$\Theta_{\mathcal{F}}$: for all i, $y_i(0) = y_i(1)$	$r_{\mathcal{F}}(\mathbf{x}^\circ) \equiv \lvert\widehat{\tau}\rvert \equiv \lvert\bar{y}_1^\circ - \bar{y}_0^\circ\rvert$	$m_{\mathcal{F}}(\cdot) \equiv$ randomization
"Neyman"	$\Theta_{\mathcal{N}}$: $\frac{1}{N}\sum_i y_i(0) = \frac{1}{N}\sum_i y_i(1)$	$r_{\mathcal{N}}(\mathbf{x}^\circ) \equiv \frac{\lvert\bar{y}_1^\circ - \bar{y}_0^\circ\rvert}{\sqrt{s_1^2/N_1 + s_0^2/N_0}}$	$m_{\mathcal{Z}}(\cdot) \equiv$ "std. normal"

to be the *supremum p-value and test for a composite null hypothesis* Θ_0 based on r and m_θ; note we allow m_θ to depend on θ.[1] By definition, if a supremum test $\chi(r, m_\theta, \Theta_0, \alpha; \mathbf{x}^\circ)$ rejects, then so do the supremum tests of subhypotheses $\chi(r, m_\theta, \Theta_0', \alpha; \mathbf{x}^\circ)$ for every $\Theta_0' \subset \Theta_0$. It follows from Lemma 1 that, for any conservative test $\chi(r, m_\theta, \Theta_0, \alpha; \mathbf{X})$, if $\chi(r, m_\theta, \Theta_0, \alpha; \mathbf{x}^\circ)$ rejects then so does $\chi(r, f_\theta, \Theta_0, \alpha; \mathbf{x}^\circ)$.

1.1 The Randomization Model

For a binary treatment, the population is defined by the (matrix) parameter $\theta = \{y_i(0), y_i(1)\}$ of $2N$ potential outcomes. The Fisher and Neyman nulls are

$$\Theta_{\mathcal{F}} \equiv \{\theta : \text{for all } i,\ y_i(0) = y_i(1)\} \quad \text{and}$$

$$\Theta_{\mathcal{N}} \equiv \left\{\theta : \sum_{i=1}^N y_i(0) = \sum_{i=1}^N y_i(1)\right\},$$

respectively. Define $\mathbf{y}(\mathbf{t}) \equiv \{t_i y_i(1) + (1 - t_i) y_i(0)\}$ to be the set of outcomes that would be observed under assignment $\mathbf{t}$. The observed data are $\mathbf{x}^\circ = (\mathbf{t}^\circ, \mathbf{y}^\circ)$, where $\mathbf{y}^\circ \equiv \mathbf{y}(\mathbf{t}^\circ)$. Similarly, given θ, the randomization distribution $f(\mathbf{t})$ of $\mathbf{T}$ induces a distribution $f(\mathbf{t}, \mathbf{y}; \theta)$ depending on θ over $\mathbf{X} = (\mathbf{T}, \mathbf{y}(\mathbf{T}))$. Let $m_{\mathcal{R},\theta}(\mathbf{x}) = f(\mathbf{t}, \mathbf{y}; \theta)$ denote this randomization distribution.

By consistency, $\mathbf{x}^\circ = (\mathbf{t}^\circ, \mathbf{y}^\circ)$ determines either $y_i(0)$ if $t_i = 0$ or $y_i(1)$ if $t_i = 1$. Consequently, there is only a single $\theta_{\mathcal{F}}^\circ \in \Theta_{\mathcal{F}}$ for which $\mathbf{x}^\circ$ is in the support of $f(\mathbf{x}; \theta)$, namely the one in which $y_i(0) = y_i(1) = y_i^\circ$. Thus for any r, m, $\text{pv}(r, m, \Theta_{\mathcal{F}}; \mathbf{x}^\circ) = \text{pv}(r, m, \theta_{\mathcal{F}}^\circ; \mathbf{x}^\circ)$. In this sense, having seen $\mathbf{y}^\circ$, the Fisher null reduces to the simple null, $\theta = \theta_{\mathcal{F}}^\circ$. Let $m_{\mathcal{F}}(\mathbf{x}) \equiv m_{\mathcal{R},\theta_{\mathcal{F}}^\circ}(\mathbf{x})$ be the *Fisher randomization distribution*; note that, although suppressed in this notation, by definition $m_{\mathcal{F}}(\cdot)$ depends on the observed data through $\mathbf{y}^\circ$. Given any statistic $r(\cdot)$, $\text{pv}(r, m_{\mathcal{F}}, \theta_{\mathcal{F}}^\circ; \mathbf{X})$ is an exact p-value for the null $\theta_{\mathcal{F}}^\circ$ associated with observed data $\mathbf{x}^\circ$.

In Section 3 below, we briefly describe methods for computing the supremum p-value for the Neyman null, $\text{pv}(r, m_{\mathcal{R},\theta}, \Theta_{\mathcal{N}}; \mathbf{x}^\circ)$, using the randomization distribution $m_{\mathcal{R},\theta}$ under every $\theta \in \Theta_{\mathcal{N}}$. This is harder because, in contrast to the Fisher null, the set of $\theta \in \Theta_{\mathcal{N}}$ for which $\mathbf{x}^\circ$ is in the support of $f(\mathbf{x}; \theta)$ is a $N - 1$ dimensional subspace. See also Rigdon and Hudgens (2015) and Chiba (2015) who invert tests based on $m_{\mathcal{R},\theta}$ to construct exact confidence intervals for the average causal effect τ.

2. EXPLANATION OF THE "PARADOX" EVENT

Ding writes that the "Fisher approach" differs from the "Neyman approach" in three ways as indicated in Table 1: the null hypothesis, the test statistic and the reference distribution. Ding equates the word "paradox" with the event that the Neyman test rejects $[\chi(r_{\mathcal{N}}, m_{\mathcal{Z}}) = 1]$, but the Fisher test fails to do so $[\chi(r_{\mathcal{F}}, m_{\mathcal{F}}) = 0]$;[2] we use quotation marks as we do not regard the occurrence of this event as particularly paradoxical. We now explain the source of the phenomena.

2.1 The Balanced Case and Small Samples

Observe that Ding changes both test statistic and reference distribution simultaneously. As a consequence, Ding fails to discover that in his balanced Example 1 ($N_1 = N_0 = 50$) the "paradox" does not arise, as he suggests, from the differences under the alternative between "variances," $\widehat{V}_{\mathcal{F}} \equiv Ns^2/N_0N_1$ and $\widehat{V}_{\mathcal{N}} \equiv s_1^2/N_1 + s_0^2/N_0$; rather, it is wholly due to the fact that p-values based on the reference distribution $m_{\mathcal{Z}}$ are anticonservative in finite samples.

[1] In this setting some authors also allow the reference distribution to depend on the observed data; see Bayarri and Berger (2000) and Robins, van der Vaart and Ventura (2000).

[2] Here $\chi(r_{\mathcal{N}}, m_{\mathcal{Z}}, \alpha; \mathbf{x}^\circ) \equiv \chi(r_{\mathcal{N}}, m_{\mathcal{Z}}, \Theta_{\mathcal{N}}, \alpha; \mathbf{x}^\circ) = \chi(r_{\mathcal{N}}, m_{\mathcal{Z}}, \theta^*, \alpha; \mathbf{x}^\circ)$, for all $\theta^* \in \Theta_{\mathcal{N}}$, such that $f(\mathbf{x}^\circ; \theta^* > 0)$. Likewise, $\chi(r_{\mathcal{F}}, m_{\mathcal{F}}, \alpha; \mathbf{x}^\circ) \equiv \chi(r_{\mathcal{F}}, m_{\mathcal{F}}, \Theta_{\mathcal{F}}, \alpha; \mathbf{x}^\circ) = \chi(r_{\mathcal{F}}, m_{\mathcal{F}}, \theta_{\mathcal{F}}^\circ, \alpha; \mathbf{x}^\circ)$. When, as here, we are comparing tests with the same α-level this is also omitted.

TABLE 2
Balanced experiments with $N_1 = N_0 = 50$, *and* $Y_i(1) \sim \mathcal{N}(1/10, 1/16)$ *as in Example* 1 *but with a zero individual treatment effect* $Y_i(0) = Y_i(1)$ *so that both the Fisher and Neyman nulls hold*; *the results were based on* 15,000 *randomizations*

(a) Comparing exact p-values resulting from Fisher's statistic $r_{\mathcal{F}}$ under $m_{\mathcal{F}}$ and approximate p-values from Neyman's statistic $r_{\mathcal{N}}$ under $m_{\mathcal{Z}}$

		$r_{\mathcal{F}}, m_{\mathcal{F}}$		
$N = 100, \alpha = 0.10$		Not reject	Reject	Power
$r_{\mathcal{N}}, m_{\mathcal{Z}}$	Not reject	0.8905	0.000	
	Reject	0.0029	0.1065	0.1095
	Power		0.1065	

(b) Comparing exact p-values resulting from Fisher's test statistic $r_{\mathcal{F}}$ under $m_{\mathcal{F}}$ and Neyman's test statistic $r_{\mathcal{N}}$ under $m_{\mathcal{F}}$

		$r_{\mathcal{F}}, m_{\mathcal{F}}$		
$N = 100, \alpha = 0.10$		Not reject	Reject	Power
$r_{\mathcal{N}}, m_{\mathcal{F}}$	Not reject	0.8935	0.0000	
	Reject	0.000	0.1065	0.1065
	Power		0.1065	

In fact, even in simulations under the Fisher Null, the "paradox" event occurs in spite of the fact that the probability limits of $N\widehat{V}_{\mathcal{N}}$ and $N\widehat{V}_{\mathcal{F}}$ are equal; see Table 2(a). For $\alpha = 0.1$, the test $\chi(r_{\mathcal{N}}, m_{\mathcal{F}})$ based on the Neyman statistic $r_{\mathcal{N}}$ under the Fisher randomization distribution $m_{\mathcal{F}}$ perfectly agrees with $\chi(r_{\mathcal{F}}, m_{\mathcal{F}})$; see Table 2(b).[3] This shows that it is the difference between reference distributions that is wholly responsible, not the choice of test statistic. In fact, this perfect agreement is true for all balanced designs (lines 1–4 and 9–12) in Table 3. Line 4 corresponds to Ding's balanced Example 1 (except for the random seed used).

Furthermore, since in each design in Table 3, including the unbalanced, the event $\{\chi(r_{\mathcal{N}}, m_{\mathcal{Z}}) = 1, \chi(r_{\mathcal{N}}, m_{\mathcal{F}}) = 0\}$ occurs, it follows from Lemma 1 (with $r = r_{\mathcal{N}}$, $m = m_{\mathcal{Z}}$, $\theta_0 = \theta^{\circ}_{\mathcal{F}}$, $f_{\theta_0} = m_{\mathcal{F}}$) that $\chi(r_{\mathcal{N}}, m_{\mathcal{Z}})$ is anticonservative under $\theta^{\circ}_{\mathcal{F}}$ at $\alpha = 0.1$ regardless of the design; see also Lang (2015), page 363.

[3]This test is described in Ding [(2017), Section 5.3]; see also references therein.

2.2 Asymptotic Considerations[4]

It is useful to define the standardized Fisher statistic $r_{s\mathcal{F}}(\mathbf{x}^{\circ}) \equiv |\overline{y}^{\circ}_1 - \overline{y}^{\circ}_0|/\{\widehat{V}_{\mathcal{F}}\}^{1/2}$ and recall $r_{\mathcal{N}}(\mathbf{x}^{\circ}) = |\overline{y}^{\circ}_1 - \overline{y}^{\circ}_0|/\{\widehat{V}_{\mathcal{N}}\}^{1/2}$. Given any observed data $\mathbf{x}^{\circ}$ with $\widehat{V}_{\mathcal{F}} > 0$, $\text{pv}(r_{s\mathcal{F}}, m_{\mathcal{F}}; \mathbf{x}^{\circ}) = \text{pv}(r_{\mathcal{F}}, m_{\mathcal{F}}; \mathbf{x}^{\circ})$ and $\chi(r_{s\mathcal{F}}, m_{\mathcal{F}}, \alpha; \mathbf{x}^{\circ}) = \chi(r_{\mathcal{F}}, m_{\mathcal{F}}, \alpha; \mathbf{x}^{\circ})$ since $\widehat{V}_{\mathcal{F}}$ is fixed given $\mathbf{x}^{\circ}$. Ding shows that for fixed α, under his asymptotics

$$\left|\chi(r_{s\mathcal{F}}, m_{\mathcal{F}}, \alpha; \mathbf{x}^{\circ}_N) - \chi(r_{s\mathcal{F}}, m_{\mathcal{Z}}, \alpha; \mathbf{x}^{\circ}_N)\right| \to 0$$

for any sequence $\mathbf{x}^{\circ}_N$ and compatible populations $\theta_N = \{y_i(0), y_i(1)\}$.[5] One may also show that $|\chi(r_{\mathcal{N}}, m_{\mathcal{F}}, \alpha; \mathbf{x}^{\circ}_N) - \chi(r_{\mathcal{N}}, m_{\mathcal{Z}}, \alpha; \mathbf{x}^{\circ}_N)| \to 0$ under the same conditions. Hence as $N \to \infty$, any "paradox" must be explained by the difference in the test statistics $r_{s\mathcal{F}}$ and $r_{\mathcal{N}}$ rather than by the difference in the reference distributions $m_{\mathcal{F}}$ and $m_{\mathcal{Z}}$.[6]

2.2.1 *Balanced design*. We now return to Ding's asymptotic analysis of the "paradox" in the balanced design. For this design, Ding correctly showed that $\widehat{V}_{\mathcal{F}} - \widehat{V}_{\mathcal{N}} = \tau^2/N + o_p(1/N)$, but then suggested this implied that the Neyman test $\chi(r_{\mathcal{N}}, m_{\mathcal{Z}})$ should be more powerful than the Fisher test $\chi(r_{s\mathcal{F}}, m_{\mathcal{F}})$ in large samples. This conclusion is incorrect under asymptotics with a fixed α-level. First, under a Pitman alternative of order $\tau = N^{-1/2}$, $\tau^2/N = o_p(1/N)$ so the probability the two tests disagree converges to 0 (i.e., they are asymptotically equivalent), and thus the tests have the same asymptotic variance and power. In fact, for the asymptotic variances to differ, τ would have to be order 1. But then the asymptotic power of both tests would be one. The finite sample exact concordance of the tests $\chi(r_{\mathcal{N}}, m_{\mathcal{F}})$ and $\chi(r_{\mathcal{F}}, m_{\mathcal{F}})$ in our simulations in the balanced case at sample size $N = 20$ and $\alpha = 0.1$ is already in line with these Pitman asymptotics.

2.2.2 *Unbalanced design*. In the unbalanced design with $(N_1/N_0 - 1) \to c > 0$, the two tests $\chi(r_{\mathcal{N}}, m_{\mathcal{Z}})$ and $\chi(r_{s\mathcal{F}}, m_{\mathcal{F}})$ are asymptotically equivalent (so the paradox occurs with limiting probability 0) unless we are under an alternative with $\tau = kN^{-1/2}$

[4]Our asymptotics assume the regularity conditions in Aronow, Green and Lee (2014). Further, when we write $\chi(r, m, \theta, \alpha)$, we view this as a sequence of tests $\{\chi(r_N, m_N, \theta_N, \alpha)\}$ where r, m and θ, but not α, change with the sample size N.

[5]Here we assume that $\lim_{N\to\infty} N\widehat{V}_{\mathcal{F}} > 0$.

[6]If one uses two different test statistics to perform two different tests of even a single simple null hypothesis, the tests, of course, may lead to different conclusions. This well-known phenomena is not normally viewed as a paradox.

TABLE 3

Simulation results based on 15,000 *simulations. All tests were performed with* $\alpha = 0.1$. *For settings where* $N_1 = N_0$, *the distribution for* $Y_i(1)$ *was* $Y_i(1) \sim \mathcal{N}(1/10, 1/16)$; *for settings where* $N_1 \neq N_0$, *the distribution for* $Y_i(1)$ *was* $Y_i(1) \sim \mathcal{N}(1/10, 1/4)$. *For the constant Individual Causal Effect,* ICE $= d$, *and* $Y_i(0) = Y_i(1) - d$; *for the settings with non-constant* ICE $\neq d$, *the Average Causal Effect* ACE $= a$, *and* $Y_i(0) \sim \mathcal{N}(1/10 - a, 1/16)$. *Rows* 4 *and* 8 *correspond to Ding's Examples* 1 *and* 2.
[*A realization in which* $\{\chi(r_{\mathcal{N}}, m_{\mathcal{F}}) = 1, \chi(r_{\mathcal{N}}, m_{\mathcal{Z}}) = 0\}$ *was never observed so this column is omitted*]

Design						**Monte Carlo average (empirical rejection rate)**			**Ding's "Paradox"**				
N_0	N_0	S_0	S_1	ACE	ICE	$\chi(r_{\mathcal{F}}, m_{\mathcal{F}})$	$\chi(r_{\mathcal{N}}, m_{\mathcal{Z}})$	$\chi(r_{\mathcal{N}}, m_{\mathcal{F}})$	$\mathbf{0} = \chi(r_{\mathcal{F}}, m_{\mathcal{F}})$ $\mathbf{1} = \chi(r_{\mathcal{N}}, m_{\mathcal{Z}})$	$\chi(r_{\mathcal{F}}, m_{\mathcal{F}})$ $\chi(r_{\mathcal{N}}, m_{\mathcal{F}})$	$\chi(r_{\mathcal{N}}, m_{\mathcal{F}})$ $\chi(r_{\mathcal{N}}, m_{\mathcal{Z}})$	$\chi(r_{\mathcal{N}}, m_{\mathcal{F}})$ $\chi(r_{\mathcal{F}}, m_{\mathcal{F}})$	$\chi(r_{\mathcal{N}}, m_{\mathcal{Z}})$ $\chi(r_{\mathcal{F}}, m_{\mathcal{F}})$
50	50	0.25	0.25	0	0	0.1065	0.1095	0.1065	0.0029	0	0.0029	0	0
		0.25	0.25	0	$\neq 0$	0.0101	0.0106	0.0101	0.0005	0	0.0005	0	0
		0.25	0.25	0.1	0.1	0.6318	0.6357	0.6318	0.0039	0	0.0039	0	0
		0.25	0.25	0.1	$\neq 0.1$	0.7202	0.7289	0.7202	0.0087	0	0.0087	0	0
30	70	0.5	0.5	0	0	0.1012	0.1050	0.1008	0.0111	0.0083	0.0042	0.0087	0.0073
		0.25	0.5	0	$\neq 0$	0.0012	0.0075	0.0053	0.0063	0.0041	0.0021	0	0
		0.5	0.5	0.1	0.1	0.2526	0.2530	0.2415	0.0136	0.0084	0.0115	0.0195	0.0132
		0.25	0.5	0.1	$\neq 0.1$	0.0873	0.3071	0.2928	0.2198	0.2055	0.0143	0	0
10	10	0.25	0.25	0	0	0.1041	0.1221	0.1041	0.0180	0	0.0180	0	0
		0.25	0.25	0	$\neq 0$	0.0299	0.0389	0.0299	0.0090	0	0.0090	0	0
		0.25	0.25	0.1	0.1	0.2121	0.2407	0.2121	0.0286	0	0.0286	0	0
		0.25	0.25	0.1	$\neq 0.1$	0.1133	0.1376	0.1133	0.0243	0	0.0243	0	0
6	14	0.5	0.5	0	0	0.1047	0.1340	0.1040	0.0333	0.0116	0.0300	0.0123	0.0039
		0.25	0.5	0	$\neq 0$	0.0033	0.0291	0.0166	0.0258	0.0133	0.0125	0	0
		0.5	0.5	0.1	0.1	0.1229	0.1616	0.1231	0.0453	0.0164	0.0385	0.0161	0.0066
		0.25	0.5	0.1	$\neq 0.1$	0.0059	0.0655	0.0408	0.0597	0.0349	0.0247	0	0

and $(S_{1N}/S_{0N})^2 - 1 \to b \neq 0$.[7] Under this alternative, the "paradox" will occur with positive limiting probability if and only if $b > 0$. Furthermore, a calculation shows that $\chi(r_{s\mathcal{F}}, m_{\mathcal{F}}, \alpha)$ has asymptotic power less than 1, in spite of the fact that, since b is nonzero, this alternative differs from the Fisher null (for which $b = 0$) by order 1. It follows that against this alternative the "paradox" could be prevented asymptotically by replacing $r_{s\mathcal{F}}$ with a different statistic r^* for which the associated test $\chi(r^*, m_{\mathcal{F}}, \alpha)$ has asymptotic power 1. However, this would introduce other trade-offs that we discuss below.

3. TOWARD A METHODOLOGICAL SOLUTION TO THE "PARADOX"

Exact p-values for the Neyman null. A simple solution to avoiding the "paradox" and obtaining tests that are not anticonservative, is to compute exact p-values $\text{pv}(r, m_{\mathcal{R},\theta}, \theta, \alpha; \mathbf{x}^\circ)$ for all $\theta \in \Theta_{\mathcal{N}}$, and then find the supremum $\text{pv}(r, m_{\mathcal{R},\theta}, \Theta_{\mathcal{N}}, \alpha; \mathbf{x}^\circ)$. If the response $y_i \in \{0, 1\}$, then $\Theta_{\mathcal{N}}$ is the finite set given by the intersection of the 2-d integer lattice $(\{0\} \cup \mathbb{Z}^+)^2$, and the convex polyhedron given by

$$
\begin{aligned}
&0 \le \theta_{00} \le n_{0+},\\
&0 \le \theta_{01} \le \min(n_{11} + n_{00}, n_{01} + n_{10}),\\
&0 \le \theta_{11} \le n_{1+},\\
&\max(n_{11}, n_{10}) \le \theta_{01} + \theta_{11} \le \min(N - n_{01}, N - n_{00}),
\end{aligned} \tag{3}
$$

where n_{yt} denotes the number of units i with observed outcomes $y_i = y$ and $t_i = t$, while θ_{ab} is the number with potential outcomes $y_i(0) = a$, $y_i(1) = b$. Loh and Richardson (2017) describe algorithms for computing exact p-values for all $\theta \in \Theta_{\mathcal{N}}$ for $\widehat{\tau}$, $r_{\mathcal{N}}$ and the likelihood ratio; the latter is a nonmonotonic function of $\widehat{\tau}$.[8]

An alternative to $r_{\mathcal{N}}$. For a continuous outcome, there will typically be too many populations in $\Theta_{\mathcal{N}}$ to compute exact p-values for them all, necessitating the use of asymptotics. To obtain maximum power, while asymptotically protecting the Neyman null at level α, one can replace $r_{\mathcal{N}}$ by $r_{\mathcal{A}}(\mathbf{x}) \equiv (\overline{y}_1 - \overline{y}_0)/\{\widehat{V}^H_{\mathcal{N}}\}^{0.5}$, where $\widehat{V}^H_{\mathcal{N}}$ is the variance estimator of Aronow, Green and Lee (2014); $N\widehat{V}^H_{\mathcal{N}}$ converges in probability to a limit $NV^H_{\mathcal{N}}$, which never exceeds $NV_{\mathcal{N}}$ and is a sharp upper bound for $NV_{\widehat{\tau}}$, where $V_{\widehat{\tau}} = S_1^2/N_1 + S_0^2/N_0 - S_\tau^2/N$ is the unidentified variance of $\widehat{\tau}$. $\chi(r_{\mathcal{A}}, m_{\mathcal{Z}}, \alpha)$ will have strictly greater power asymptotically than $\chi(r_{\mathcal{N}}, m_{\mathcal{Z}}, \alpha)$ against Pitman alternatives to the Neyman null,[9] except for the subset of those Pitman alternatives that are local to the hypothesis that the marginal limit distributions of $y(0)$ and $y(1)$ are the same. For binary Y, $\widehat{V}^H_{\mathcal{N}}$ is asymptotically equivalent to the variance estimators of Robins (1988) and Ding and Dasgupta (2016).[10] Finally, like $r_{\mathcal{N}}$ and $r_{s\mathcal{F}}$, $|\chi(r_{\mathcal{A}}, m_{\mathcal{F}}, \alpha; \mathbf{x}_N{}^\circ) - \chi(r_{\mathcal{A}}, m_{\mathcal{Z}}, \alpha; \mathbf{x}_N{}^\circ)| \to 0$. Hence $\chi(r_{\mathcal{A}}, m_{\mathcal{F}})$ and $\chi(r_{\mathcal{A}}, m_{\mathcal{Z}})$ are asymptotically equivalent, but, unlike $\chi(r_{\mathcal{A}}, m_{\mathcal{Z}})$, $\chi(r_{\mathcal{A}}, m_{\mathcal{F}})$ is guaranteed not to be anticonservative under the Fisher null in small samples.[11]

Alternatives to $r_{\mathcal{F}}$ and $r_{s\mathcal{F}}$. We have shown that with Ding's choice of test statistics (i) the probability the "paradox" occurs converges to zero under local alternatives to the Fisher null but (ii) the paradox can happen with limiting positive probability under certain identifiable nonlocal alternatives[12] to the Fisher null, since Ding's Fisher test does not have limiting power 1 against these alternatives. It is natural to ask whether there exist test statistics whose associated Fisher test might have asymptotic power 1 against the Fisher null for all identifiable nonlocal alternatives, thereby avoiding the paradox for such alternatives. The test $\chi(r^*, m_{\mathcal{F}})$ with r^* a member of the class of generalized Kolmogorov–Smirnov [KS] test statistics (Præstgaard, 1995) should have this property under weak regularity conditions. However, use of $\chi(r^*, m_{\mathcal{F}})$ would reintroduce the "paradox" under local alternatives when $\chi(r_{\mathcal{N}}, m_{\mathcal{Z}})$, $\chi(r_{\mathcal{N}}, m_{\mathcal{F}})$, $\chi(r_{\mathcal{A}}, m_{\mathcal{Z}})$ or $\chi(r_{\mathcal{A}}, m_{\mathcal{F}})$ is used as a test of the Neyman null because (a) even under the Fisher null the asymptotic rejection region of $\chi(r^*, m_{\mathcal{F}})$ is not a subset of the others and (b) the limiting power of $\chi(r^*, m_{\mathcal{F}})$ will be less than

[7] Such alternatives are not possible for Y binary since $(S_{1N}/S_{0N})^2 - 1$ will be order N^{-1} if τ is order $N^{-1/2}$.

[8] Ding's Theorem 7 is not correct as stated, since for binary outcomes not all test statistics are equivalent to $\widehat{\tau}$. Ding's proof establishes the weaker claim that *under the Fisher null* every statistic is a *(possibly nonmonotonic, noninjective) function of* $\widehat{\tau}$.

[9] See, for example, Table 3 row 6 of Aronow, Green and Lee (2014), where $V^H_{\mathcal{N}}/V_{\mathcal{N}}$ is significantly less than 1. The calculations are relevant to a Pitman alternative to the Neyman null as the means of the Beta distributions generating $y(1)$ and $y(0)$ are the same, but as the variances are different, it is not local to the Fisher null.

[10] The claim made by Ding and Dasgupta that their estimator "improved on" that of Robins (1988) and Aronow, Green and Lee (2014) is incorrect. Ding has submitted a correction (Ding, 2016).

[11] See also Chung and Romano [(2013), Example 2.1], and references therein.

[12] We call a sequence of populations $\{\theta_N\}$ an identifiable nonlocal alternative to the Fisher null if there exists a consistent test of that alternative under the Fisher null.

that of the other four statistics under, for example, a constant individual treatment effect alternative with τ of order $N^{-1/2}$, because $\chi(r^*, m_{\mathcal{F}})$ spreads its power over many directions.

Summary and Recommendations

Taken together our points above provide answers to the following:

Q: Can one find a test of the Fisher null and a test of the Neyman null such that: (i) regardless of sample size or the true $\theta = \{y_i(0), y_i(1)\}$, the "paradox" of the Neyman test accepting and the Fisher test rejecting can never occur; (ii) the Fisher test is never anticonservative under the Fisher null regardless of sample size; (iii) asymptotically, the Neyman test is a conservative test of the Neyman null with power at least as great as that of any other conservative test against any local[13] alternative to the Neyman null?

A: Yes. Select, as both the Fisher and Neyman test, the *single test* $\chi(r_{\mathcal{A}}, m_{\mathcal{F}})$ that uses $r_{\mathcal{A}}$ as test statistic and the Fisher randomization distribution as reference distribution.

Q: Can one find a test of the Fisher null and a test of the Neyman null satisfying (i)–(iii) plus (iv): the Fisher test has limiting power 1 against all (identifiable) nonlocal alternatives to the Fisher null?

A: No. As argued above, to satisfy (iv), any Fisher test (e.g., a generalized KS test) must have power in many directions, while to satisfy (ii) and (iii) the Neyman test must be $\chi(r_{\mathcal{A}}, m_{\mathcal{F}})$ which concentrates most of its power in a single direction. Thus this combination would reintroduce the "paradox," violating (i).

Conclusion. In the context of a finite population causal model, if one is somewhat more interested in testing for the presence of an average causal effect, than in testing the Fisher null, and cannot compute exact p-values for all laws in the Neyman null, then we recommend using $\chi(r_{\mathcal{A}}, m_{\mathcal{F}})$, which satisfies (i) to (iii) above.

[13]We restrict to local alternatives because if τ is order 1 then all test statistics that we consider have limiting power one against the Neyman null, and hence also the Fisher null.

ACKNOWLEDGMENTS

This research was supported by US National Institutes of Health Grant R01 AI032475 and US Office of Naval Research Grant N00014-15-1-2672.

REFERENCES

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

BAYARRI, M. J. and BERGER, J. O. (2000). *P* values for composite null models. *J. Amer. Statist. Assoc.* **95** 1127–1142, 1157–1170. With comments and a rejoinder by the authors. MR1804239

CHIBA, Y. (2015). Exact tests for the weak causal null hypothesis on a binary outcome in randomized trials. *J. Biom. Biostat.* **6** 244.

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

DING, P. (2016). Personal communication.

DING, P. (2017). A paradox from randomization-based causal inference. *Statist. Sci.* **32** 331–345.

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

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

LOH, W. W. and RICHARDSON, T. S. (2017). Likelihood analysis for the finite population Neyman–Rubin binary causal model. In preparation.

PRÆSTGAARD, J. T. (1995). Permutation and bootstrap Kolmogorov–Smirnov tests for the equality of two distributions. *Scand. J. Stat.* **22** 305–322. MR1363215

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

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

ROBINS, J. M., VAN DER VAART, A. and VENTURA, V. (2000). Asymptotic distribution of p values in composite null models. *J. Amer. Statist. Assoc.* **95** 1143–1167, 1171–1172. MR1804240