Statistical Science
2017, Vol. 32, No. 3, 362–366
DOI: 10.1214/17-STS571REJ
Main article DOI: 10.1214/16-STS571
© Institute of Mathematical Statistics, 2017

Rejoinder: A Paradox from Randomization-Based Causal Inference

Peng Ding

I enjoyed reading the critical but insightful comments from Aronow and Offer-Westort, Chung, Bailey, Loh, Richardson and Robins. In my original paper, I mainly discussed a "paradox" that arose in the Neymanian and Fisherian randomization-based inferences using the difference-in-means statistic. The theory was essentially an asymptotic analysis of the powers of the Neymanian and Fisherian tests, which depended on the choice of the test statistic.

Motivated by the comments, I will extend the scope of the original paper by discussing three other issues. First, if we use the Fisher randomization tests (FRTs) for both the Neymanian and Fisherian null hypotheses, then the paradox disappears (Section 1). Second, I will review some classical results and recent progress about the FRTs with treatment effect heterogeneity (Section 2). Third, I will discuss general test statistics and the remaining difficulties in evaluating their properties (Section 3). For simplicity, I will focus on treatment-control experiments.

1. EXACT INFERENCE WITHOUT PARADOXES

The paradox in Ding (2017) results from different inferential frameworks, one using the Wald test based on the repeated sampling distribution, and the other using an exact randomization test under the sharp null hypothesis. Frequentists' properties guarantee that both methods are valid for their own purposes and criteria, but do not guarantee that the final results regarding "reject" or "not-reject" are compatible with logic. This paradox, however, will go away if we consider only exact inference based on the FRTs. I will elaborate on this perspective below, reviewing some progress and highlighting the remaining difficulties.

Modifying Loh, Richardson and Robins (LRR)'s notation, I use $\theta = \{Y_i(1), Y_i(0)\}_{i=1}^N$ for the potential outcomes of the finite population,

$$\Theta(\text{Fisher}) = \{\theta : Y_i(1) = Y_i(0) \text{ for all } i\}$$

for the potential outcomes satisfying Fisher's null hypothesis, and

$$\Theta(\text{Neyman}) = \left\{\theta : \sum_{i=1}^N Y_i(1) = \sum_{i=1}^N Y_i(0)\right\}$$

for the potential outcomes satisfying Neyman's null hypothesis. Given all the potential outcomes for the finite units, θ, choosing a test statistic, we can compute the p-value associated with this test statistic by using the FRT, denoted by $p(\theta)$ where we highlight its dependence on θ. Because there is only a single set of potential outcomes in $\Theta(\text{Fisher})$, the p-value for Fisher's null hypothesis is simply $p(\theta_0)$ where $\theta_0 \in \Theta(\text{Fisher})$. The p-value for Neyman's null hypothesis is $\max_{\theta\in\Theta(\text{Neyman})} p(\theta)$, which by definition is larger than $p(\theta_0)$ because $\Theta(\text{Fisher}) \subset \Theta(\text{Neyman})$. Following this testing strategy for Fisher's and Neyman's null hypotheses, no logical incoherence will appear.

The above strategy sounds general, but it is rarely used in practice except for special cases with binary outcomes. For binary outcomes, Rigdon and Hudgens (2015) first obtained the p-values for all possible sets of potential outcomes, and then inverted tests to construct finite sample exact confidence intervals for the average causal effect. This is feasible because for binary outcomes the number of all possible potential outcomes is finite. By utilizing stochastic ordering information, Li and Ding (2016) further reduced the computational cost of Rigdon and Hudgens (2015). Lu, Ding and Dasgupta (2015a) made an attempt to construct potential outcomes for ordinal data, which turned out to be much more complicated than binary data.

Beyond binary outcomes, computing $p(\theta)$ for all possible sets of potential outcomes is almost intractable unless we are willing to impose some "model" assumptions in the sense of Rosenbaum [(2002), Chapter 5]. We illustrate this point using three examples below.

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

EXAMPLE 1. The classical constant causal effect model (Rosenbaum, 2002) implies $\tau_i = c$, and consequently Fisher's and Neyman's null hypotheses are both equivalent to $c = 0$ or $\tau_i = 0$ for all units i.

EXAMPLE 2. A linear causal effect model satisfies $Y_i(1) = aY_i(0) + b$, and consequently Fisher's null hypothesis implies $a = 1$ and $b = 0$, and Neyman's null hypothesis implies $b = (1-a)\bar{Y}_0$ and $Y_i(1) = aY_i(0) + (1-a)\bar{Y}_0$. Fisher's null hypothesis is sharp, but Neyman's null hypothesis is not because a is unknown. In order to compute the p-value under Neyman's null hypothesis, we need to conduct a grid search over the domain of a. In particular, for a fixed a, Neyman's null hypothesis is sharp in the sense that we can determine all the missing potential outcomes based on the observed ones and, therefore, we can compute the p-value using the FRT, denoted by $p(a)$. The p-value for Neyman's null hypothesis is then $\sup_a p(a)$. Without any constraints, we can search over $a \in (-\infty, +\infty)$. However, the parameter a controls the variance ratio between the treatment and control potential outcomes because $S_1^2 = a^2 S_0^2$. In practice, if we have prior knowledge about this variance ratio, then we can search over a subset of $(-\infty, +\infty)$.

EXAMPLE 3. A general model for individual causal effect could be

$$\tau_i = Y_i(1) - Y_i(0) = f\big(Y_i(0), \gamma\big) \quad (i = 1, \ldots, N)$$

parametrized by a K dimensional parameter γ, where $\gamma \in \Gamma \subset \mathcal{R}^K$. Fisher's null hypothesis implies that $f(Y_i(0), \gamma) = 0$ for all i; Neyman's null hypothesis implies $\sum_{i=1}^N f(Y_i(0), \gamma) = 0$, which further restricts γ to be within a smaller subset $\Gamma(\text{Neyman}) \subset \Gamma$. To compute the p-value for Neyman's null hypothesis, we need to conduct a grid search over the domain $\Gamma(\text{Neyman})$ of the parameter γ.

However, exact inference often relies on unverifiable model assumptions as shown above. Otherwise, we have to rely on the finite population central limit theorems to construct asymptotic tests or confidence sets. Models and asymptotics are different strategies for dealing with unknown nuisance parameters, which in our case, include all unknown potential outcomes. In practice, we are facing a tradeoff.

2. FRTS THAT ARE ROBUST TO TREATMENT EFFECT HETEROGENEITY

In the original paper, I mainly focused on comparing the powers of the Neymanian and Fisherian tests, mentioning only in passing the possible anticonservative behavior of the FRT using $\hat{\tau}$ as the test statistic for Neyman's null hypothesis. See the last paragraph of Section 3.4 of Ding (2017). Chung highlighted this point and stated that "[w]ithout controlling the level of tests, comparing the power of the tests has less credibility" under Neyman's null hypothesis, with which I totally agree and want to discuss more.

Under the randomization inference framework, some issues have been tackled by Ding and Dasgupta (2016) in the analysis of variance, with the treatment-control experiment being a special case. In the example below, I will give a more detailed discussion of the treatment-control experiment because it is a direct consequence of Theorem 3 of Ding (2017).

EXAMPLE 4. If $Y_i(1) = aY_i(0) + (1-a)\bar{Y}_0$ for all i as in Example 2, then $\tau = 0$ and $\tau_i = (a-1) \times Y_i(0) + (1-a)\bar{Y}_0$. Asymptotically, the FRT using $\hat{\tau}$ is invalid under Neyman's null hypothesis if and only if $\widehat{V}(\text{Fisher}) = Ns^2/(N_1 N_0)$ is asymptotically smaller than the true sampling variance of $\hat{\tau}$, $\text{var}(\hat{\tau}) = S_1^2/N_1 + S_0^2/N_0 - S_\tau^2/N$. Based on Theorem 3 of Ding (2017), some simple algebra shows that asymptotically $\widehat{V}(\text{Fisher}) < \text{var}(\hat{\tau})$ reduces to

$$\big(N_0^{-1} - N_1^{-1}\big)\big(S_1^2 - S_0^2\big) + N^{-1} S_\tau^2 < 0.$$

Let $r = N_1/N$ be the proportion of the treated units. Because $S_1^2 = a^2 S_0^2$ and $S_\tau^2 = (a-1)^2 S_0^2$, the above inequality reduces to

$$\left(\frac{1}{1-r} - \frac{1}{r}\right)(a^2 - 1) + (a-1)^2 < 0$$

$$\iff \quad (a-1)\big\{\big(r^2 - 3r + 1\big)a - \big(r^2 + r - 1\big)\big\} > 0.$$

Figure 1 shows the regions that the FRT using $\hat{\tau}$ yields correct and incorrect asymptotic type one errors under Neyman's null hypothesis.

In Example 4 with treatment effect heterogeneity, the FRT using $\hat{\tau}$ will yield an incorrect type one error for Neyman's null hypothesis for a wide range of (r, a). Motivated by previous discussions of permutation tests in a super population, Ding and Dasgupta's (2016) recent result implies that the FRT using the studentized statistic $t_{\text{DD}} = \hat{\tau}/\sqrt{\widehat{V}(\text{Neyman})}$ will yield an exact test for Fisher's null hypothesis and asymptotically conservative test for Neyman's null hypothesis. Therefore, this FRT is robust to treatment effect heterogeneity.

Interestingly, LRR also proposed to use the FRT using a studentized statistic $t_{\text{LRR}} = \hat{\tau}/\sqrt{\widehat{V}'(\text{Neyman})}$ based on the improved variance estimator in Aronow,

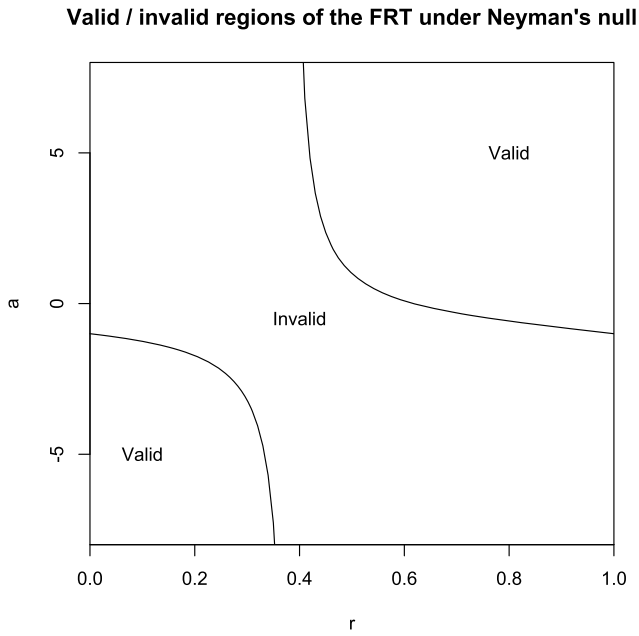

FIG. 1. *The FRT using $\widehat{\tau}$ under Neyman's null hypothesis under the linear treatment effect model in Examples 2 and 4.*

Green and Lee (2014). LRR's purpose of using t_{LRR} is to avoid the paradox of Ding (2017). Moreover, like the FRT using t_{DD}, the FRT using t_{LRR} will also be exact under Fisher's sharp null hypothesis and asymptotically valid under Neyman's null hypothesis.

To summarize, although the FRT is exact under Fisher's sharp null hypothesis, we need to choose the test statistic carefully to avoid the pathological phenomenon discussed in Ding (2017) and to make the FRT more robust to treatment effect heterogeneity. Overall, studentizing the difference-in-means statistic seems an attractive choice.

3. GENERAL TEST STATISTICS AND THEIR NEYMANIAN ANALYSES

The theory of the original paper focused on using $\widehat{\tau}$ in the Fisherian and Neymanian tests. Sections 5.1 and A.5 of Ding (2017) conjectured that similar results might hold for the Kolmogorov–Smirnov (KS) and the Wilcoxon–Mann–Whitney (WMW) statistics. LRR considered the KS statistic for the consistency of the FRT, and Chung cautioned that the FRT using the WMW statistic was often invalid for Neyman's null hypothesis from a super population perspective (Chung and Romano, 2016).

Currently, the real challenge is due to the lack of Neymanian analyses of the KS statistic and the WMW statistic, although these statistics have been widely used in the FRT under Fisher's sharp null hypothesis. Current literature does study the asymptotic properties of general rank statistics and U-statistics under super population models, but these results are not directly applicable under the Neyman-type finite population inference.

For instance, Chung argued that the WMW statistic

$$\begin{aligned}\widehat{\tau}_{\mathrm{WMW}} &= \frac{1}{N_0 N_1}\sum_{T_i=0}\sum_{T_j=1} I\big(Y_i^{\mathrm{obs}} \le Y_j^{\mathrm{obs}}\big) \\ &= \frac{1}{N_0 N_1}\sum_{i=1}^{N}\sum_{j=1}^{N} T_j(1-T_i) I\big\{Y_i(0) \le Y_j(1)\big\}\end{aligned}$$

could be viewed as an estimator of $\tau_{\mathrm{WMW}}^{\mathrm{SP}} = \mathrm{pr}\{Y_i(0) \le Y_j(1)\}$ for $i \neq j$. However, as my notation indicates, $\tau_{\mathrm{WMW}}^{\mathrm{SP}}$ is a super population causal parameter assuming that $\{Y_i(1), Y_i(0)\}_{i=1}^{N}$ are i.i.d. draws from a random vector $\{Y(1), Y(0)\}$. Moreover, current super population analysis assumed that the treatment and control potential outcomes were independent of each other (Chung and Romano, 2016). In fact, under the randomization inference framework, because

$$\begin{aligned}E\big\{T_j(1-T_i)\big\} &= \mathrm{pr}(T_i = 0, T_j = 1) \\ &= \binom{N-2}{N_1-1}\Big/\binom{N}{N_1} \\ &= \frac{N_1 N_0}{N(N-1)} \quad (i \neq j)\end{aligned}$$

we can show that

$$\begin{aligned}E(\widehat{\tau}_{\mathrm{WMW}}) &= \frac{1}{N(N-1)}\sum\sum_{i\neq j} I\big\{Y_i(0) \le Y_j(1)\big\} \\ &\equiv \tau_{\mathrm{WMW}}^{\mathrm{FP}},\end{aligned}$$

where $E(\cdot)$ denotes expectations over all possible randomizations, and $\tau_{\mathrm{WMW}}^{\mathrm{FP}}$ is the finite population causal parameter corresponding to the WMW statistic. However, obtaining the randomization distribution of $\widehat{\tau}_{\mathrm{WMW}}$ is much more challenging than obtaining that of $\widehat{\tau}$. Nevertheless, it seems apparent that this randomization distribution will depend on the association between the individual potential outcomes, which did not appear in the super population calculation (Chung and Romano, 2016).

More importantly, does the finite population causal parameter $\tau_{\mathrm{WMW}}^{\mathrm{FP}}$ make sense in practice? It involves comparison between the potential outcomes of different units, which seems inferior to a direct comparison

of the potential outcomes of the same unit, for example,

$$\frac{1}{N}\sum_{i=1}^{N} I\{Y_i(0) \le Y_i(1)\}.$$

Unfortunately, these types of causal parameters cannot be identified even with large samples and only bounds of them can be obtained (Fan and Park, 2010, Lu, Ding and Dasgupta, 2015b); randomization-based inference for them are even more challenging.

The comparison based only on $\widehat{\tau}$ seems limited, but it is not obvious how to go beyond it. More research should be done.

4. OTHER TWO ISSUES RAISED BY THE DISCUSSANTS

4.1 Pitman Alternative and Local Power

In Section 4.3 of the first arXiv version of this paper (Ding, 2014), I discussed "locally most powerful test" and considered the Pitman-type local alternative. I showed that choosing a special sequence of local alternative hypotheses with constant causal effects $\tau_i = c/\sqrt{N}$ for some fixed c,

$$\widehat{V}(\text{Fisher}) - \widehat{V}(\text{Neyman}) = o_p(N^{-1}),$$

and, therefore, Neymanian and Fisherian inferences have the same asymptotic behavior under this sequence of local alternatives. Unfortunately, this discussion was dropped during the review and revision of Ding (2017), but was brought up again by Aronow and Offer-Westort and LRR.

LRR thought my explanation was flawed under the Pitman asymptotics. However, the Pitman asymptotics is just one choice of asymptotic analysis, and it is not the gold standard even in the classical hypothesis testing literature (Serfling, 1980, Chapter 10). The Pitman asymptotics is useful for explaining other phenomena, but I found it limited at least for the purpose of explaining the numerical examples in Table 1 of Ding (2017).

LRR argued that the "paradox" was merely due to finite sample problems based on their Table 2(a). However, the relative magnitude of the $(2, 1)$th element of their Table 2(a) is too small compared to those of the $(2, 1)$th elements of Tables 1(a) and 1(b) of Ding (2017).

4.2 Potential Outcomes Beyond Treatment-Unit Additivity

Professor Bailey claimed that she had "never used the potential outcomes framework." This is puzzling to me, because her equation (1) "$Y_i(t) = \tau_t + Z_i$" uses the *potential outcomes* $Y_i(t)$ and assumes additivity of the treatment effect τ_t and the unit characteristic Z_i. However, the advantage of the Neyman–Rubin potential outcomes framework is to allow for treatment effect heterogeneity, under which asymptotically conservative repeated sampling evaluation is often possible.

Professor Bailey pointed out that Fisher used neither the notation of potential outcomes nor the terminologies "sharp null hypothesis" nor "randomization test," with which I totally agree. However, from my own experience of studying the current literature of causal inference, the potential outcomes framework is transparent about the inferential goal and flexible enough to be used in different inferential frameworks. Even though Fisher did not use the potential outcomes framework, as an assistant professor in the department founded by Neyman, I feel obligated to use it to continue the Neyman tradition.

ACKNOWLEDGMENTS

I want to thank the Editor (Professor Peter Green) and the anonymous Associate Editor for inviting discussions on my paper and for giving me the opportunity to write the rejoinder. Drs. Arman Sabbaghi at Purdue, Avi Feller at Berkeley and Anqi Zhao at Harvard gave constructive comments on early versions of this rejoinder, and Miss Lo-Hua Yuan edited my draft carefully.

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

CHUNG, E. and ROMANO, J. P. (2016). Asymptotically valid and exact permutation tests based on two-sample U-statistics. *J. Statist. Plann. Inference* **168** 97–105. MR3412224

DING, P. (2014). A paradox from randomization-based causal inference. Technical report. Available at arXiv:1402.0142v1.

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

DING, P. and DASGUPTA, T. (2016). A randomization-based perspective of analysis of variance: A test statistic robust to treatment effect heterogeneity. Available at arXiv:1608.01787.

FAN, Y. and PARK, S. S. (2010). Sharp bounds on the distribution of treatment effects and their statistical inference. *Econometric Theory* **26** 931–951. MR2646485

LI, X. and DING, P. (2016). Exact confidence intervals for the average causal effect on a binary outcome. *Stat. Med.* **35** 957–960. MR3457618

LU, J., DING, P. and DASGUPTA, T. (2015a). Construction of alternative hypotheses for randomization tests with ordinal outcomes. *Statist. Probab. Lett.* **107** 348–355. MR3412795

LU, J., DING, P. and DASGUPTA, T. (2015b). Treatment effects on ordinal outcomes: Causal estimands and sharp bounds. Available at arXiv:1507.01542.

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

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

SERFLING, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. Wiley, New York. MR0595165