Statistical Science
2017, Vol. 32, No. 3, 346–348
DOI: 10.1214/16-STS582
Main article DOI: 10.1214/16-STS571

Understanding Ding's Apparent Paradox

Peter M. Aronow and Molly R. Offer-Westort

1. INTRODUCTION

We are grateful for the opportunity to comment on "A Paradox From Randomization-Based Causal Inference" (Ding, 2017), an interesting discussion of the properties of Fisher's randomization test (FRT) with a comparison to a Wald-type test based on a variance estimator originally proposed by Neyman (1990). The article illustrates that the use of a Wald-type test using Neyman's variance estimator and the FRT (with the difference in means as a test statistic) can lead to situations where the null of zero average causal effect (Neyman's null) is rejected, while the sharp null of zero individual causal effects (Fisher's null) is not. The article in large part attributes this apparent paradox, which persists asymptotically, to a difference in the implied variances of the reference distributions used to construct tests. In this comment, we seek to situate Ding's (2017) findings in established statistical results, and to explain how the apparent paradox is a direct consequence of two well-known results.

We summarize our contribution as follows: (i) Rao-type tests may be sub-optimal under nonlocal alternatives (Engle, 1984), and as Ding (2017) shows, the FRT is asymptotically equivalent to a Rao-type test assuming constant effects. Rather than the FRT, a Wald-type analogue to the FRT using the difference in means exists (considered by Freedman, 2008, Samii and Aronow, 2012, Gerber and Green, 2012, and Lin, 2013) is a valid test of Fisher's null, and does not suffer from the potential pathology of Rao-type tests. Furthermore, this Wald-type test—equivalent to a pooled-variance two-sample z-test—is asymptotically equivalent to the FRT under local alternatives. Reichardt and Gollob (1999) discuss this point, with reference to Mosteller and Rourke (1973). (ii) As highlighted by Pratt (1964), Romano (1990) and Freedman (2008), the behavior of tests under incorrect working assumptions may depend on the joint distribution of the data. Analogous to the Behrens–Fisher problem, a test of the null of no average effect that assumes there is no effect on the variance may be more or less powerful than a test of no average effect that makes no such additional assumption. We illustrate this by comparing the Wald-type analogue to the FRT to the standard Wald-type test of Neyman's null. Combining (i) and (ii), we see that Ding's (2017) apparent paradox follows from the use of a suboptimal test under nonlocal alternatives and the well-known behavior of tests of joint hypotheses when one of the constituent hypotheses is false. (iii) We also note an error in Ding's (2017) Theorem 7 and suggest a refinement that is correct at full generality.

Note, we consider only asymptotic results, and do not discuss the finite N differences between exact tests and tests that are only known to be asymptotically valid. Finite N differences may account in part for Ding's (2017) simulation results; while perhaps of practical importance, such differences are in our view theoretically uninteresting as the source of a potential paradox.

2. A WALD-TYPE ANALOGUE TO THE FISHER RANDOMIZATION TEST

Ding (2017) demonstrates the asymptotic equivalence of the FRT and Rao's score test as applied to a linear model assuming homoskedasticity. This result builds on prior findings from Romano (1990), Freedman (2008) and Samii and Aronow (2012). Freedman (2008) considers the operating characteristics of a Wald test assuming a homoskedastic linear model; Samii and Aronow (2012) give the implied variance of this test a randomization basis, by establishing that the implied variance is equivalent to the randomization distribution of the difference-in-means estimator if the treatment effect is assumed to be constant and equal to the observed estimate. Gerber and Green (2012) also advocate for this Wald-type approach in practice. Ding (2017) further reproduces Lin (2013) and Samii and Aronow's (2012) result, showing

Peter M. Aronow is Associate Professor, Departments of Political Science and Biostatistics, Yale University, 77 Prospect Street, New Haven, Connecticut 06520, USA (e-mail: peter.aronow@yale.edu). Molly R. Offer-Westort is a graduate student, Departments of Political Science and Statistics and Data Science, Yale University, 77 Prospect Street, New Haven, Connecticut 06520, USA (e-mail: molly.offer-westort@yale.edu).

that a Wald test from a linear model allowing for heteroskedasticity is equivalent to a Wald-type test using one of Neyman's variance estimators,

$$\widehat{V}(\text{Neyman}) = S_1^2/N_1 + S_0^2/N_0 + o_p(N^{-1}).$$

[We follow Ding's, 2017 notation throughout. We will also assume complete random assignment with the asymptotic scaling and regularity conditions of Freedman, 2008, primarily bounded fourth (cross) moments on the limit distribution of potential outcomes with $\lim_{N\to\infty} N_1/N = p$, where $0 < p < 1$. These regularity conditions appear to be acknowledged in the final paragraph of Ding, 2017.]

To summarize, Ding (2017) demonstrates that the FRT is asymptotically equivalent to a Rao score test under homoskedasticity, and reiterates that the Neyman test is equivalent to using a Wald test under heteroskedasticity. The asymptotic equivalence of the FRT and the Rao test illustrates why the FRT may have poor power relative to Wald-type analogues. Rao tests are known to have the potential to be suboptimal for nonlocal alternatives (Engle, 1984). In this setting, the reference distribution for the Rao test is constructed assuming the null hypothesis, $\tau = \overline{Y}_1 - \overline{Y}_0 = 0$. The Wald test statistic, however, is constructed in part from the empirical distribution, using as a working assumption $\tau = \widehat{\tau}$ in order to construct a reference distribution local to the observed data. Comparing the two tests based on their asymptotic approximations, the Wald test performs better than the Rao score test under settings where the true value of τ is not local to 0. We investigate this point below.

As implied by Ding [(2017), equation (7)], we have that the variance of the FRT reference distribution,

$$\widehat{V}(\text{Fisher}) = S_1^2/N_0 + S_0^2/N_1 + \tau^2/N + o_p(N^{-1}).$$

Suppose now that instead of using FRT, we were to use the Wald-type analogue considered by Freedman (2008), Gerber and Green (2012) and Samii and Aronow (2012). Ding (2017) notes the variance estimator associated with the Wald-type analogue in Section A.3.1, and refers to it as $\widehat{V}(\text{OLS})$, where

$$\widehat{V}(\text{OLS}) = S_1^2/N_0 + S_0^2/N_1 + o_p(N^{-1}).$$

Freedman's (2008) results straightforwardly imply that a Wald-type test using $\widehat{V}(\text{OLS})$ is an asymptotically valid test of Fisher's null, but that it may not be a valid test of the null that $\tau = 0$ when treatment effects are heterogeneous (as is possible under Neyman's null). It follows that

$$\widehat{V}(\text{Fisher}) - \widehat{V}(\text{OLS}) = \tau^2/N + o_p(N^{-1}),$$

and the FRT is suboptimal given nonlocal alternatives. The reason for this is intuitive: in constructing the implied variance estimate, $\widehat{V}(\text{Fisher})$ fails to account for any location shift across potential outcomes (akin to the combined variance), whereas $\widehat{V}(\text{OLS})$ does (akin to the pooled variance) (Reichardt and Gollob, 1999, pages 124–125). Via standard linearization arguments, $\widehat{V}(\text{OLS})$ will provide a better approximation to $V(\widehat{\tau})$ in large samples when effects are indeed nonlocal and constant. (Aronow, 2013, establishes the consistency of analogous linearized variance estimators for a broad class of nonstandard experimental designs under constant effects.)

Under Pitman-type local alternatives of the form $\tau_N = o(1)$, we have $\widehat{V}(\text{Fisher}) - \widehat{V}(\text{OLS}) = o_p(N^{-1})$. The only asymptotic distinction between $\widehat{V}(\text{Fisher})$ and $\widehat{V}(\text{OLS})$ emerges with nonlocal alternatives. But as Engle [(1984), page 786], notes, "If the alternative is not close to the null, then presumably both tests would reject with very high probability for large samples; the asymptotic behavior of tests for nonlocal alternatives is usually not of particular interest." Thus, as we proceed, we focus on the Wald-type test using $\widehat{V}(\text{OLS})$, which will allow us to highlight the remaining source of Ding's (2017) apparent paradox.

3. COMPARING WALD TO WALD, OR THE RETURN OF THE BEHRENS–FISHER PROBLEM

Ding [(2017), equation (7)], considers $\widehat{V}(\text{Fisher}) - \widehat{V}(\text{Neyman})$; let us consider an analogue, replacing $\widehat{V}(\text{Fisher})$ with its Wald-type analogue $\widehat{V}(\text{OLS})$:

$$\begin{aligned}\widehat{V}(\text{OLS}) - \widehat{V}(\text{Neyman}) &= (S_1^2/N_0 + S_0^2/N_1) \\ &\quad - (S_1^2/N_1 + S_0^2/N_0) \\ &\quad + o_p(N^{-1}).\end{aligned}$$

A necessary condition for $\operatorname{plim}_{N\to\infty} N\widehat{V}(\text{OLS}) \neq \operatorname{plim}_{N\to\infty} N\widehat{V}(\text{Neyman})$ is $S_1 \neq S_0$. [As noted above, this result also holds for $\widehat{V}(\text{Fisher}) - \widehat{V}(\text{Neyman})$ under any Pitman-type alternative such that $\tau_N = o(1)$.] We unpack this result.

Applying arguments from Neyman (1990), it is well known that $\lim_{N\to\infty} NV(\widehat{\tau}) \leq S_1^2/p + S_0^2/(1-p)$, where $p = \lim_{N\to\infty} N_1/N$. Ergo, given the finite population central limit theorem, Wald-type tests constructed using $\widehat{V}(\text{Neyman})$ will tend to be conservative. Under constant effects, Neyman (1990) further shows that this holds with equality: $\lim_{N\to\infty} NV(\widehat{\tau}) = S_1^2/p + S_0^2/(1-p)$. Furthermore, if the maintained hypothesis of constant effects is true, then it must be the case that $S_1 = S_0$ and, therefore, it would also be the case that $\lim_{N\to\infty} NV(\widehat{\tau}) = S_0^2/p + S_1^2/(1-p)$; in such a case, $\widehat{V}(\text{Neyman})$ and $\widehat{V}(\text{OLS})$ therefore asymptotically agree. If we are not in the world of constant effects, it is possible that $S_0 \neq S_1$, and there exist limit distributions such that $\lim_{N\to\infty} NV(\widehat{\tau}) > S_0^2/p + S_1^2/(1-p)$, implying that using $\widehat{V}(\text{OLS})$ may yield anticonservative inferences and invalid tests of Neyman's null.

Fisher's null can be viewed as a joint hypothesis that implies two features of the underlying distribution of potential outcomes: $\tau = 0$ and $S_1 = S_0$. Neyman's null implies no such assumption about $S_1 = S_0$. This setting may seem familiar, as it is analogous to the Behrens–Fisher problem. If all we are interested in is testing Neyman's null: $\tau = 0$ (cf., no location shift), an additional working assumption that $S_0 = S_1$ (cf., no scale shift) may lead to unusual behavior if indeed $S_0 \neq S_1$. If both $\tau \neq 0$ and $S_0 \neq S_1$, then Wald-type tests using $\widehat{V}(\text{OLS})$ may be more or less powerful than tests using $\widehat{V}(\text{Neyman})$. Or, as Lin (2013) illustrates, two wrongs can sometimes make a right, and tests of Fisher's null may be more powerful in testing the null that $\tau = 0$. But two wrongs can also sometimes make a wrong, as if $S_1^2/p + S_0^2/(1-p) < S_0^2/p + S_1^2/(1-p)$, then the converse will hold, and Ding's (2017) apparent paradox emerges. [Note that this condition is equivalent to ($p > 1/2$ and $S_1 > S_0$) or ($p < 1/2$ and $S_1 < S_0$).] Just as with the Behrens–Fisher problem, where an incorrect working assumption about the scale shift may lead to lower power in testing the null hypothesis of no location shift, there is no generic reason why, when $S_1 \neq S_0$, a test that assumes $S_1 = S_0$ would be more powerful in testing the hypothesis that $\tau = 0$.

4. A NOTE ON THEOREM 7

Ding [(2017), Theorem 7], states, "For completely randomized experiments, matched-pair experiments, and 2^K factorial experiments, if the outcomes are binary, then all test statistics are equivalent to the difference-in-means statistic." From this, Ding (2017) concludes, "Therefore, for binary data, the choice of test statistic is not a problem." While perhaps practically true for many test statistics, this claim is not correct when taken at full generality, for example, consider a test statistic that is not strictly increasing in the difference in means (e.g., $|\widehat{\tau}|$). For such test statistics, the resulting inferences may substantively differ. Here, we state a weaker claim verifiable at full generality, which has been known since Copas [(1973), equation (9)]: for completely randomized experiments, if outcomes are binary, then under the sharp null hypothesis of no treatment effect, all test statistics may be written as a function of the difference-in-means statistic. Stronger statements are likely available.

5. CONCLUSION

Our comment does not subtract from the paper as a useful exposition of tests commonly used to test the Fisher and Neyman nulls of no effect in randomized experiments. Ding (2017) has drawn attention to an important topic in the domain of randomization-based causal inference. We hope that by placing Ding's (2017) contribution in the context of a broader statistical literature, we have shed light on the reasons for this apparent paradox.

REFERENCES

ARONOW, P. M. (2013). Model assisted causal inference. Ph.D. thesis, Dept. Political Science, Yale Univ., New Haven, CT.

COPAS, J. B. (1973). Randomization models for the matched and unmatched 2 × 2 tables. *Biometrika* **60** 467–476. MR0448746

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

ENGLE, R. F. (1984). Wald, likelihood ratio, and Lagrange multiplier tests in econometrics. In *Handbook of Econometrics*, *Vol.* 2 775–826. Elsevier, Amsterdam.

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

GERBER, A. S. and GREEN, D. P. (2012). *Field Experiments: Design, Analysis, and Interpretation*. Norton, New York.

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

MOSTELLER, F. and ROURKE, R. E. (1973). *Sturdy Statistics: Nonparametrics and Order Statistics*. Addison-Wesley, Reading, MA.

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

PRATT, J. W. (1964). Robustness of some procedures for the two-sample location problem. *J. Amer. Statist. Assoc.* **59** 665–680. MR0166871

REICHARDT, C. S. and GOLLOB, H. F. (1999). Justifying the use and increasing the power of a *t* test for a randomized experiment with a convenience sample. *Psychol. Methods* **4** 117–128.

ROMANO, J. P. (1990). On the behavior of randomization tests without a group invariance assumption. *J. Amer. Statist. Assoc.* **85** 686–692. MR1138350

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