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).