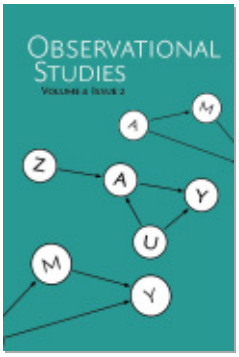
Beyond Statistical Criticism

Paul R. Rosenbaum, Dylan S. Small



Observational Studies, Volume 4, Issue 2, 2018, pp. 65-70 (Article)

Published by University of Pennsylvania Press
DOI: *https://doi.org/10.1353/obs.2018.0008*

Observational Studies 1 (2018) 65-70 Submitted 12/17; Published 1/18

Beyond Statistical Criticism

Paul R. Rosenbaum **rosenbaum@wharton.upenn.edu**
Dylan S. Small **dsmall@wharton.upenn.edu**
Department of Statistics
The Wharton School, University of Pennsylvania Philadelphia, PA 19104 U.S.A.

Abstract

In an admirable essay, Bross makes many useful observations. The goal, however, should be to take a step beyond statistical criticism, arriving at an objective statement about what the (research design + data) say and fail to say. Often this entails saying a bit less than one might like in exchange for saying something definite and objective.

1. What Bross says

Statistical criticism is a remarkably important topic about which remarkably little has been written. Bross is certainly right in saying: "[T]he quality of statistical criticism [... is often...] rather poor." He is also right in trying to "put the statistical critic on his mettle — not to muzzle him." He right again in suggesting as the standard that the critic of an empirical study "operates under the same ground rules as a proponent." Bross goes on to classify forms of statistical criticism that fall short of this standard. For instance, he writes:

> The bulk of statistical criticism is of the hit-and-run variety — the critic points out some real or fancied flaw and supposes that his job is done. Indeed, some critics appear to labor under the misconception that if some flaw can be found in a study, this automatically invalidates the author's conclusions. ... [I]t is not enough to spot flaws in a study: a responsible critic would go on to show how these flaws lead to a counterhypothesis that can explain the observations.

Continuing, he writes:

> Proponents of scientific hypotheses are often justly criticized for their "tubular vision" — a remarkable inability to "see" the evidence unfavorable to their hypothesis. Critics are equally subject to this type of defective vision.

Concerning "dogmatic criticism," Bross writes:

> Consider the following quotation from Sir R. A. Fisher, which has been echoed by other eminent critics: 'The evidence linking cigarette smoking with lung cancer, standing by itself, is inconclusive, as it is apparently impossible to carry out properly controlled experiments with human materials.' ... Because of the lack of randomization, there is a potential 'self-selection' bias (which suggests

> a counterhypothesis). If this counterhypothesis can be rendered tenable, then, indeed, the proponent's evidence is 'inconclusive.' Instead of attempting to make the self-selection hypothesis tenable, Fisher simply dismissed the entire body of epidemiological data.

In his Presidential Address to the American Statistical Association, Jerome Cornfield endorsed Bross' position:

> In statistical applications one can detect this same search for purity.... It ... shows up in a certain type of statistical criticism of scientific results, in which pointing to a potential weakness is considered equivalent to demolition. Bross' proposed ground rule for statistical criticism – that some effort to demonstrate the reality as well as the potentiality of the weakness be required – does not seem to have dimmed this quest for purity, at least as manifested in some recent statistical criticisms.

Bross concludes: "[M]y theme has been: we should not have a 'double standard' in science and statistics, one standard for proponents and another for critics."

There is much to admire in Bross' essay, and little with which we disagree.

2. Agree on less, rather than agree to disagree

2.1 The standard case: Is it chance?

Too often, statistical criticism ends where art criticism or culinary criticism or wine criticism ends, with an agreement to disagree. Is Picasso a greater painter than Matisse? You can point to paintings and offer commentary, but there is no knock-down argument that compels agreement. Is statistical criticism like that?

When statistical criticism achieves agreement, a specific pattern of argument commonly occurs. We end up agreeing about something objective that falls a bit short of absolutely settling the original disagreement. Take the simplest and most familiar case, but look at it from the perspective of a proponent and a critic. A proponent plots I independent and identically distributed observations of a Y_i against an X_i, notes that higher Y_i's tend to occur with higher X_i's, and suggests that Y_i exhibits some form of monotone association with X_i. A critic looks at the plot, and claims that a pattern like that is not indicative of a genuine association, and could be produced by bad luck when X and Y are independent. The proponent then uses some conventional statistical test, perhaps Kendall's correlation, testing the hypothesis of independence against alternatives of monotone association, obtaining a two-sided P-value. The P-value does not establish who is correct, the proponent or the critic, but it does clarify what each is saying in light of the data. Perhaps one of them is saying something outlandish, perhaps not. The P-value quantifies the amount of bad luck that would be needed to produce the observed pattern were X and Y independent; however, it does not logically prove that bad luck is or is not the explanation. If the P-value were $2/3$, then it would not take much bad luck to produce the association, whereas if the P-value were 10^{-10} then it would take quite a bit of bad luck. And, of course, there are intermediate situations. The P-value does not adjudicate the claims of the proponent

and the critic, but it does clarify what each is saying. One can recognize the P-value as an objective interpretation of what the proponent and critic are each saying, while leaving open the question of who is correct. If the P-value were 0.07, the study's audience might objectively see that neither the proponent nor the critic is saying something outlandish, yet the audience might divide, some siding with the proponent, others with the critic, on the basis of other evidence or considerations.

The P-value discussion just given exhibits a specific pattern. It steps back from one question to answer another question instead. The original question — who is correct, proponent or critic — is replaced by another question that can be answered objectively, namely how much bad luck would be needed to produce the ostensible pattern were X and Y independent. The objective answer describes what the research-design-plus-data say, and quite possibly they may say less than we might like. Nonetheless, the objective answer is a fact of the matter — Kendall's correlation did yield a particular P-value — even if this fact of the matter falls short of an absolute adjudication of the positions of proponent and critic. We sacrificed the pure and absolute, gaining in its place the objective, and we are better off for this exchange.

Statistical evidence invariably involves both a research design and data derived from that design, and in randomized trials or sample surveys, it may involve nothing else (Fisher 1935, Chapter 2). More often, statistical evidence involves a research design, data derived from that design, plus assumptions, sometimes quite fanciful assumptions, such as an infinite population of people from which an independent and identically distributed sample has been drawn. Some fanciful assumptions are inconsequential, in the sense that replacing them by more realistic assumptions does not materially alter conclusions; however, other assumptions play a crucial role in conclusions, so changing the assumptions changes the conclusions.

2.2 A familiar case: Is it selection bias?

In observational studies of treatment effects, one fanciful but consequential assumption is that treatments were randomly assigned with probabilities that are a function of observed covariates but not of potential outcomes given covariates, so-called ignorable treatment assignment. It is common to raise doubts about adjustments for observed covariates by calling into question this assumption of ignorable assignment, often postulating an unmeasured covariate for which adjustments are also required.

A proponent claims that there is strong evidence of causality in the observed association between treatment and outcome after adjustment for observed covariates. A critic denies this, saying instead that the association was produced by an unmeasured covariate, that treatment assignment is not ignorable. Who is correct? As in the case of P-values and chance, the matter will not be settled by a proof. However, we may step back from adjudicating the conflicting claims of proponent and critic and objectively clarify what each is saying. Perhaps one or the other is saying something outlandish, perhaps not. A sensitivity analysis does this; see Cornfield et al. (1959). It asks about the magnitude of bias from an unobserved covariate that would need to be present to alter the study's conclusion. It says: To explain the observed association between treatment and outcome as a selection bias due to nonrandomized treatment assignment, the bias would need to be

of such and such a magnitude, say Γ. True, this does not absolutely settle the disagreement between proponent and critic, but it does clarify what each is saying. It is quite a different thing to say that a tiny, barely perceptible bias in treatment assignment could explain an association, as opposed to saying that only an enormous bias could do so. We have taken a step back from whether bias produced the association. Instead, we have objectively clarified what is being said by a proponent who denies it is bias or a critic who asserts it is bias. It is a fact in the data that to explain the association between heavy smoking and lung cancer as a bias, that bias would have to be enormous. This fact is less than we might like — it is a step back to what the data say — but it is an important fact nonetheless.

2.3 Statistical criticism can undermine itself

A proof by contradiction assumes, for the sake of argument, that a claim is true, en route to showing that the claim is false. In a parallel way, assuming a critic is correct for the sake of argument may provide the means for showing he is incorrect. The critic says the treatment is without effect, that the association between treatment and outcome is entirely the product of selection bias, the product of who gets treated not of effects caused by treatment. The critic's claim has consequences, and those consequences may undermine the critic's claim.

Of course, a responsible proponent has done a sensitivity analysis, acknowledging that selection biases in excess of Γ could explain away the association between treatment and outcome. The proponent accepts the critic's claim momentarily for the sake of argument. What would follow from accepting the critic's claim as true? What if all the associations were the product of selection biases, with no causal effect anywhere? The proponent then shows that were the critic's claim true — were it all selection bias with no treatment effect — then a certain statistical analysis would be justified that would have been unjustified without the critic's claim. The proponent then does this added analysis, finding that the association between treatment and outcome would then be insensitive to a larger bias, $\Gamma' > \Gamma$. In this sense, the critic's claim undermines itself: Were it true, it would only make the study insensitive to larger biases. It is not that the critic's claim is false — we do not know that. We do know, objectively, that the critic's claim fails in its role as a criticism of the original study. Supposing the critic's claim to be true would only strengthen the proponent's position, so it fails as a criticism of the proponent's position. An example of this kind of reasoning is given in Rosenbaum (2015).

2.4 Aporia

A statistical critic may point to a logical inconsistency among data, assumptions, and scientific knowledge from other sources. If several propositions are logically inconsistent, then they cannot all be true; yet at a certain moment, we may not be in a position to identify the one or several false propositions that create logical inconsistency. Such a situation is said to be dissonant or said to be an aporia; see, for instance, Rescher (2009). An aporia is not a state of total ignorance, but rather an uncomfortable state of knowledge: logical inconsistency among propositions that I have good reason to believe is strong evidence that some of the propositions I have good reason to believe are, in fact, false. The acknowledgement of an aporia is an uncomfortable advance in understanding, a step beyond believing logi-

cally incompatible propositions while failing to recognize their incompatibility. Statistical criticism may bring an aporia to light without resolving it. The objective step back from settling the dispute between a proponent and a critic may be to acknowledge the existence of an aporia.

For instance, Yang et al. (2014) considered a plausible instrument or instrumental variable (IV) and used it to estimate a plausible beneficial treatment effect in one population in which the true treatment effect is unknown. They then applied the same instrument to a second population in which current medical opinion holds that this same treatment confers no benefit, finding that this IV suggests a benefit in this second population also. Specifically, there is debate about whether delivery by caesarean section improves the survival of extremely premature infants, but current medical opinion holds that it is of no benefit for otherwise healthy but slightly premature infants. In contrast, the IV analysis suggested a substantial benefit for both types of infants. In light of this, there is logical incompatibility between four items: (i) the data, (ii) the claim that extremely premature infants benefit from delivery by caesarean section, (iii) the claim that otherwise healthy, slight premature infants do not benefit, (iv) the claim that the IV is valid in both groups of babies. Removal of any one of (i), (ii), (iii) or (iv) would remove the inconsistency, but there is no basis for removing one and accepting the others. This aporia is a fact about the data, and it is good to acknowledge facts about the data, even though, in this case, it leaves us uncomfortably without resolution of the source of the inconsistency.

An aporia, though uncomfortable, is an advance in understanding: it can spur further investigation and further advances in understanding. Socrates, in Plato's *Meno*, thought that demonstrating aporia in a curious person's thinking would spur discovery. Socrates said of a befuddled young interlocutor who he put in an aporia:

> At first he did not know what [he thought he knew], and he does not know even now: but at any rate he thought he knew then, and confidently answered as though he knew, and was aware of no difficulty; whereas now he feels the difficulty he is in, and besides not knowing does not think he knows...[W]e have certainly given him some assistance, it would seem, towards finding out the truth of the matter: for now he will push on in the search gladly, as lacking knowledge; whereas then he would have been only too ready to suppose he was right...[Having] been reduced to the perplexity of realizing that he did not know...he will go on and discover something.

References

Cornfield, J. (1975). A statistician's apology. *Journal of the American Statistical Association*, 70, 7-14.

Cornfield, J., Haenszel, W., Hammond, E. C., Lilienfeld, A. M., Shimkin, M. B. and Wynder, E. L. (1959). Smoking and lung cancer: recent evidence and a discussion of some questions. *Journal of the National Cancer Institute*, 22, 173-203.

Fisher, R. A. (1935). *Design of Experiments*. Edinburgh: Oliver and Boyd.

Plato. *Plato in Twelve Volumes*, Vol. 3 translated by W.R.M. Lamb. Cambridge, MA, Harvard University Press; London, William Heinemann Ltd. 1967, Section 84a-c.

Rescher, N. (2009). *Aporetics: Rational Reliberation in the Race of Inconsistency*. Pittsburgh: University of Pittsburgh Press.

Rosenbaum, P. R. (2015). Some counterclaims undermine themselves in observational studies. *Journal of the American Statistical Association*, 110, 1389-1398.

Yang, F., Zubizarreta, J.R., Small, D.S., Lorch, S. and Rosenbaum, P.R. (2014). Dissonant conclusions when testing the validity of an instrumental variable. *The American Statistician*, 68, 253-263.