Observational Studies 4 (2018) 1-10 Submitted 1960; Published reprinted, 1/18

Statistical Criticism

Irwin D.J. Bross

Editor's Note: Irwin Bross (1921-2004) was Director of Biostatistics, Roswell Park Institute, Buffalo, NY from 1959-1983. This article was originally published in *Cancer*, 1960, 13, 394-400. The paper is reprinted with permission of the copyright holder, John Wiley & Sons. Comments by current researchers in observational studies follow.

IN THE GREAT debate over smoking and lung cancer the quality of statistical criticism was, I think, rather poor (despite the eminence of the critics). The bitter lesson to be learned was this: *The "rules of the game" for statistical criticism need to be spelled out more clearly and completely.* These remarks represent a first step in this direction. While I will draw my examples from the lung cancer debate, similar instances can be found in most of the scientific areas in which statistical methods are employed. Much that nowadays passes as statistical criticism is superficial and sophomoric in character and serves to obscure a scientific discussion rather than to clarify it.

Let me emphasize at the start that the purpose of ground rules is to put the statistical critic on his mettle – not to muzzle him. The rules of the game for the proponent of scientific hypothesis, discussed in various texts on statistics and logic, can help him to make the statements that are warranted by the data. If the *proponent* flagrantly violates these rules, they provide a basis for calling him to account (Berkson, 1958). In the same way, ground rules for a critic will help him to distinguish valid objections. Of course, they also provide a basis for calling a critic to account for irresponsible attacks on scientific study. If both proponents and critics have to watch their P's and Q's, we might hope that it would be easier to achieve broad agreement on scientific issues.

1. The Role of a Critic

As a first step toward the ground rules of statistical criticism, let us examine the roles of the critic and the proponent. In what follows, the critic will be considered as opposed to the proponent in the sense that he denies the proposed scientific hypothesis or at any rate denies that it has been demonstrated.

Although the critic's role appears purely negative, it has a positive side to it. Implicitly (and sometimes explicitly) he puts forth a counterhypothesis. This point may be clarified by a simple example. Let us say that a critic objects to the conclusions of a scientific study because the proponent has not used significance tests. This objection would be trivial if, for example, the value of the chi square was actually enormous. However, it would be a strong objection if a difference between two series (which was essential to the proponent's argument) was not significant when the test was performed. But why is this objection strong? Because the critic can now frame a tenable counterhypothesis that explains the

©2018 John Wiley & Sons.

results in terms of sampling variation alone. Since the proponent cannot rule out this counterhypothesis, he cannot establish his own hypothesis.

In much the same way a critic who objects to a bias in the design or a failure to control some established factor is, in fact, raising a counterhypothesis (even though he may not state it). Since the counterhypothesis is essential in the logical structure of criticism, it facilitates debate when it is explicitly stated (Brownlee, 1957). When the hypothesis is so stated, the basic question suggests itself: What is the responsibility of a critic with respect to his counterhypothesis?

2. A Criterion for Criticism

Consider the following tentative rule: The critic has the responsibility for showing that his counterhypothesis is tenable. In so doing, he operates under the *same* ground rules as a proponent.

This rule may appear to conflict with the principle that the *burden of proof* rests on a proponent, but this is not the case. Although both critic and proponent may operate under the same rules in establishing their respective hypotheses, there is a great difference in what happens next. When a critic has shown that his counterhypothesis is tenable, his job is done (while at this point the proponent's job is just beginning). A proponent's job is not finished as long as there is a tenable hypothesis that rivals the one he asserts.

Many critics seem to employ a rule that is much weaker than the tentative rule stated. They feel that a critic's responsibility ends when he merely *presents* a counterhypothesis without showing it to be tenable. This I regard as unrealistic because it imposes an impossible task on a proponent. He would be required to rule out every *conceivable* hypothesis. Since there are an unlimited number of such hypotheses, there would be no end to the proponent's labors. By restricting consideration to *tenable* hypotheses, the proponent's task becomes feasible (although onerous).

Tentative rule does not impose any impossible task upon the critic, since he can employ the usual scientific procedures to show that his counterhypothesis is tenable. For example, a minimal requirement would be that the effects predicted from the critic's hypothesis should be in line with the actual data, at least in direction and order of magnitude. The additional arguments needed to establish tenability depend on the nature of the hypothesis. For instance, if the hypothesis involves sampling variation alone, it would be tenable in any study employing samples. For hypotheses involving an artefact, the experience with this artefact in previous studies can be used to establish the direction and magnitude of the effect. It may even be possible to show the effect operating in the proponent's data. When a counterhypothesis involves a well known real factor, e.g., age or sex in an epidemiological study, it would be sufficient to mention the relationship, e.g., death rates from cancer tend to increase with age. However, when a counterhypothesis is novel or controversial, a critic (like a proponent in the same circumstances) will have to develop a strong argument.

For these reasons, the suggested rule for criticism seems to be both fair and feasible (and I will employ it in criticizing the critics).

3. Hit-and-Run Criticism

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 in validates the author's conclusions. Since the critic makes no attempt to develop a tenable counterhypothesis, his performance is on a par with that of a proponent who glances at his data and then jumps to his conclusion. Two examples should suffice to make this point plain. Quite a number of the critics of the Hammond and Horn (1954) study (along with other prospective studies) have called attention to the possibility of misclassification of the cause of death on death certificates. Most of these critics dropped the matter at that point (apparently under the impression that they had scored a hit). However, if they had followed the usual scientific procedures in developing a tenable hypothesis – if they had looked at other studies or existing theory (Bross, 1954) – they would have found that misclassification tends to *diminish* observed differentials. If they had compared the Hammond and Horn tabulations of "purified" data, i.e., cases with confirmed diagnosis, with the "unpurified" data, they would have *seen* how misclassification operated to reduce differentials.

Another example is Berkson's (1955) model for a selection bias based on an item in the protocol of the Hammond and Horn study (initially sick individuals were excluded). The model itself is a good example of constructive criticism, since it formulated the objection in a precise fashion that facilitated both theoretical and empirical investigation. However, I was amazed when, in talking with several statisticians, I encountered the opinion that this *model* seriously jeopardized the Hammond and Horn conclusions. Of course a model carries no weight in a scientific argument until it has been shown to be tenable. This particular model was untenable because it predicted that the differentials found in the study would *shrink* as time went by, whereas, if anything, the change with time was the *opposite* direction. What is more, even if the model had been tenable it would have been of little value in a counterhypothesis, since it could be shown mathematically that the bias could produce only slight differentials (the observed differentials had a different *order of magnitude*) (Brownlee, 1957).

We see, then, that it 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. If a critic fails to build a tenable hypothesis, he clearly fails in his duty.

4. Dogmatic Criticism

To show that his counterhypothesis is tenable, a critic may use arguments based on current statistical principles and practices. However, a critic has no license to make exaggerated claims, unfounded assumptions, or dogmatic assertions (even if the statements are quoted from statistical textbooks).

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 material.*" (Laurence, 1957) (The italics are mine.)

This blanket condemnation rests largely on one defect of the prospective studies as com pared to controlled experiments. The exposure to cigarette smoke, i.e., smoking habits, is determined by the personal choice of each individual, whereas ideally the exposure would be set by the experimenter (using a randomized allocation). 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 (involving carefully collected information on hundreds of thousands of individuals). He did so on the basis that the data do not meet certain theoretical standards for "properly controlled experimentation." This seems to me a gross violation of the empirical spirit of modern science *and* of modern statistics. It raises the theory of statistics, e.g., randomization, to the level of dogma.

5. Speculative Criticism

While I do not agree with those who say that there is no place for speculation in a scientific article, I do feel that there are definite restrictions on hypothetical excursions. For one thing, speculation should he clearly labeled as such; for another, speculations should not enter the conclusions. These restrictions apply equally to proponent and critic.

There is one type of counterhypothesis in which the temptation to speculate is very strong – chypotheses based on a new "real world" factor. A statistician should be especially careful with this type of substantive hypothesis because he is now in the domain of the subject matter field – he is functioning as an epidemiologist or sociologist or psychiatrist (depending on the nature of the new factor) rather than as a statistician.

The task of establishing the tenability of a substantive counterhypothesis is more difficult than that for a methodological counterhypothesis, since "local" ground rules, i.e., those of the particular scientific field, come into play. For example, in epidemiology a proposed new factor has to be consistent with the broad incidence patterns of the disease, e.g., geographic distribution, time trends, and sex ratios.

While numerous substantive counterhypotheses have been introduced in the lung cancer controversy, there has been practically no attempt to render such hypotheses tenable. Thus Berkson brought up Pearl's (1928) "rate of living" hypothesis (Berkson, 1958) but frankly admitted that: "Actually I do not know of any independent evidence for such an effect of smoking." He also cited the "constitution" hypothesis, noted one of its shortcomings, and remarked: "I do not profess to be able to track out the implications of the constitutional theory or to defend it..." While it is to Berkson's credit that he clearly labeled these two counterhypotheses as speculative, they appear to play an important role in his subsequent rejection of the "carcinogen" hypothesis, i.e., speculations enter his conclusion.

It may be argued that it is too stringent to require a critic to show that his substantive counterhypothesis is tenable because he is not actually *asserting* it but merely *suggesting* it as a possible line for future research. However, I fail to see how a critic contributes to the scientific process if the suggested avenue for research is, in fact, a dead end road. Nor can I see how a critic can expect to point out a sensible direction for research unless he explores the tenability of his counterhypothesis – for example, by checking whether his notion jibes with the incidence pattern for lung cancer.

The most striking feature of lung cancer incidence is the drastic increase in the age specific male death rates over the past generation. This rapid increase is virtually *unique* – the female death rate shows much less change, other cancer rates are fairly stable, and the rates for other causes of death either show relatively minor changes or else are rapidly *decreasing*. The peculiar behavior of the male lung cancer rates poses some difficult questions for any substantive hypothesis. Why is lung cancer thus singled out? Why are male death rates affected and not female death rates? Why should this have happened in the last generation? I leave it to the reader to put these questions to some of the counterhypotheses raised, e.g., those based on "stress," "genetic factors," and "constitution."

In my opinion, even a cursory exploration would have shown most of the critics that their substantive counterhypotheses were untenable. Had this been done, much of the confusion in the lung cancer debate would have been avoided.

6. Tubular Criticism

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. For example, Berkson (1958) complained that "virtually all of the evidence" that cigarette smoke is carcinogenic comes from epidemiological-statistical studies. He was unable to "see" the evidence from vital statistics, combustion chemistry, animal experiments, lung tissue pathology, etc.

Tubular vision also occurs in the examination of actual data. Since Berkson is one of the few critics who (1) dealt with data, (2) stated his counterhypotheses, and (3) made a serious effort to establish their tenability, I will draw my examples from his work (Berkson, 1958). However, judging *over-all* performance, I would say Berkson far excels the other critics.

To appreciate the illustrations, we first must understand the purpose of Berkson's analysis of the Doll and Hill data (Doll and Hill, 1956). His counterhypothesis was: "The observed associations are 'spurious'...the result of the interplay of various subtle and complicated 'biases.' " To establish tenability, Berkson first undertook to show that"... there can hardly be any doubt that association is shown for 'all or nearly all' causes of death... "in prospective studies. Before examining Berkson's arguments for this crucial point, let us see how it is used to establish the counterhypothesis. Berkson said: "For myself, I find it quite incredible that smoking should cause all these diseases...And if we are not crassly to violate the principle of Occam's razor, we should not attribute to each separate association a radically different explanation."

I would not interpret Berkson's remarks as a denial that an environmental factor, e.g., polluted milk, can be responsible for more than one disease. Hence tobacco smoke, which is chemically quite complex (containing nicotine, polycyclic hydrocarbons, etc.), might induce or aggravate several different diseases (e.g., lung cancer, coronary thrombosis, or chronic bronchitis) by radically different "specific" etiological mechanisms. We can, however, make a distinction between those diseases in which an etiological hypothesis based on chemical components in tobacco smoke can be supported by independent evidence (call these "specific" diseases) and the many other causes of death in which a corresponding hypothesis would be highly speculative (call these "nonspecific" diseases). Now if we find that many of the "nonspecific" diseases are associated with smoking, then I quite agree with Berkson that the "simple" hypothesis of a general bias running through the data is clearly preferable to the "complex" hypothesis requiring a large number of speculative hypotheses to account for the associations. Moreover, if we also find that the bias effect is similar in direction and magnitude to the effects found for the "specific" diseases, then we have a tenable counter-hypothesis for the whole of the data and the proponents of "specific" hypotheses are in a hopeless position.

Of course, this argument hinges on a demonstration that there is "generalized association" in the "nonspecific" diseases. For this purpose, Berkson started with a Doll and Hill tabulation (Table 29; Doll and Hill, 1956) that gave the death rates in 4 tobacco consumption categories for "lung cancer," "coronary thrombosis," "other respiratory diseases," "other cancers," and "other diseases."

For a significance test of "generalized association," Berkson suggested that: "Appropriate here is some form of permutation test..." He went on to say that: "However it is figured, the probability of getting by chance...consistently higher death rates among the heavy smokers than among any of the 3 categories of less than heavy smokers, in each of 5 predesignated categories of cause of death, and in agreement with the independently obtained similar finding in the prospective study of Hammond and Horn (1954), must be considered negligible."

Here, I think, is an instance of "tubular vision." Two of the 5 categories represent "specific" diseases while a third, "other respiratory diseases," largely reflects the influence of chronic bronchitis – another "specific" disease. In other words, most of the evidence that Berkson used to *deny* "specific" effects came from these very effects! Indeed, *unless* these "specific" effects are included, there is little evidence for a "generalized association" in Table 29. Thus, while a permutation type test is significant at the 5% level for the 2 "specific" diseases, the corresponding test for the 2 "nonspecific" causes is definitely not significant.

7. Tubular Criticism and Data

Berkson himself did not seem satisfied with his inferences from Table 29 for he proceeded to construct (from Doll and Hill tabulations), Table 34, which listed 15 causes of death (and hence permitted segregation of "specific" causes). This table is reproduced as Table 1 in this article.

Berkson clearly "sees" his "generalized association" operating in Table 1, but the only analytic evidence offered is: "The death rate for heavy smokers is higher than that for nonsmokers in 12 of the 15 categories, although in several instances the number of deaths, the differences of rates, or both, are small."

A "permutation" test that could be used on this evidence is the sign test. [Strictly speaking, the death rates in the different causes may not be independent because overenumeration in one cause might lead to underenumeration in a related cause.] Let I be the total number of "inversions" (i.e., cases in which the death rate was lower for the heavy smokers than for the nonsmokers). Let NI be the total number of "noninversions." Then,

Table 1: Death Rates for Various Smoking Classes for Individualized Categories of Disease (Report of Doll and Hill, 1956) (Data from Table 34 of Berkson, 1958)

	Death rate, standard.,/1000				
			Men smoking a daily average of:		
Category	No. deaths	Non- smok.	1-14 gm.	15-24 gm.	25+ gm.
Cancer					
Lung	84	0.07	0.47	0.86	1.66
Up. respir. & digest. tract	13	0.00	0.13	0.09	0.21
Stomach	32	0.41	0.36	0.10	0.31
Colon & rectum	57	0.44	0.54	0.37	0.74
Prostate	30	0.55	0.26	0.22	0.34
Other sites	88	0.64	0.72	0.76	1.02
Respir. dis					
Pulm. tuberculosis	19	0.00	0.16	0.18	0.29
Chron. bronchitis	42	0.12	0.29	0.39	0.72
Other respir. dis	65	0.69	0.55	0.54	0.40
Coronary thrombosis	508	4.22	4.64	4.60	5.99
Other cardiovas. dis.	279	2.23	2.15	2.47	2.25
Cereb. hemorrhage	227	2.01	1.94	1.86	2.33
Peptic ulcer	18	0.00	0.14	0.16	0.22
Violence	77	0.42	0.82	0.45	0.90
Other dis.	183	1.45	1.81	1.47	1.57

using the sign test:

$$\frac{[|NI - I| - 1]^2}{NI + I} = \frac{[|12 - 3| - 1]^2}{12 + 3} = \frac{64}{15} = 4.27.$$

Since the 5% level critical value is 3.84, the sign test is significant, and we would reject the null hypothesis that sampling variation alone can account for the result cited by Berkson. Unfortunately, a departure in the observe direction might be due to either "specific" or "generalized" association, and if we we try to subdivide the causes the numbers will be so small that the sign test will have little power.

However, we can call on the big brother of the sign test, the sequence sign test, to do the job. As before, we count "inversions" but this time we consider all 6 of the pairwise comparisons that can be made between the 4 death rates for each cause. In Table 2 we count a pairwise comparison as an "inversion" if the death rate to the right of the other member of the pair is the smaller one. Thus, for cerebral hemorrhage the sequence of rates is 2.01, 1.94, 1.86, and 2.33. Starting with 2.01 as the "left hand" member of the pair, we have "inversions" for 1.94 and 1.86 and a "noninversion" for 2.33. Moving on to 1.94 as the "left hand" member of the pair, we have an "inversion" for 1.86 and a "noninversion" for 2.33. Finally for the pair 1.86, 2.33 is a "noninversion." So for this cause we have 3 "inversions" and 3 "noninversions." Table 2 lists the results for the causes in Table 1.

Table 2: Inversion Count for Table 1 (By Cause)

Group	Inversions	Noninversions
"Specific" causes		
Lung cancer	0	6
Chronic bronchitis	0	6
Coronary thrombosis	1	5
SUBTOTAL A	1	17
Questionable causes		
Pulm. tuberculosis	0	6
Upper respiratory cancer	1	5
Peptic ulcer	0	6
SUBTOTAL B	1	17
SUMMARY TOTALS		
All "nonspecific"	25	29
All causes	27	63
"Nonspecific" cancers		
Stomach	5	1
Colon & rectum	2	4
Prostate	4	2
Other sites	0	6
SUBTOTAL C	11	13
"Nonspecific" major causes		
Cerebral hemorrhage	3	3
Other cardiovascular diseases	2	4
Other diseases	2	4
SUBTOTAL D	7	11
Other "nonspecific" causes		
Violence	1	5
Other respiratory dis.	6	0
SUBTOTAL E	7	5

For all causes, there are 27 inversions and 63 noninversions and the sequence sign test is:

$$K\frac{(|NI-I|-1)^2}{NI+I} = \frac{9}{13}\frac{(|63-27|-1)^2}{90} = 9.42$$

where

$$K = \frac{9}{2(\text{no. factor categories})+5} = \frac{9}{2(4)+5} = \frac{9}{13}$$

and represents an adjustment for the fact that the 6 pairwise comparisons in a cause are not independent. This test is significant at the 1% level.

Since the sequence sign test is more powerful than the sign test, we are now able to segregate the "specific" and "nonspecific" causes (Table 2). I have also separated off 3 causes (subtotal B in Table 2) that are of questionable value for our purposes. There are less than 20 deaths in each of these series, and they were tabulated separately only because "specific" effects were suspected.

The "specific" causes (subtotal A in Table 2) show up as highly significant (8.65), although there are only 3 of them. The "nonspecific" causes in toto show 25 inversions and 29 noninversions, which is close to the expected values under the null hypothesis, i.e., 27 and 27, and is, of course, not significant (0.12). We might expect a "generalized association" to show most clearly in the "non-specific" major causes since these 3 causes account for 40% of all the deaths, but this is not the case (subtotal D in Table 2). Again in view of Berkson's demand for an "explanation for the association shown with cancers...of such sites as the colon, stomach, and pancreas..." (Berkson, 1958) we might expect a striking result for "nonspecific" cancers, but this was not found (subtotal C in Table 2). In short, a permutation type analysis fails to detect Berkson's "generalized associations" in the Doll and Hill data (although it picks up associations for the "specific" diseases easily enough).

No permutation analysis is presented in Berkson's (1958) article and instead he simply cited 2 examples of nonspecific causes that he believed show evidence of association. Berkson had reservation about 1 of these causes, "violence," but the other ("cancer: other sites") is "notable": "...this group shows a graded increase of death rate with increased amount of smoking, from a rate of 0.64 for nonsmokers, to a rate of 1.02 for heavy smokers." However, as can be seen from Table 2, the cited example is the *only* "nonspecific" cause whose rates show a graded increase (I equals 0 and NI equals 6) with increased smoking. If this one cause is to be regarded as a strong argument for generalized association, what are we to make of the category "other respiratory diseases" that shows a very similar pattern but in the opposite direction (I equals 6 and NI equals 0)?

8. Summing Up

The lengthy illustration of "tubular vision" in the examination of data contains several important lessons for statisticians. First, it shows how dangerous it is – even for an experienced and competent statistician – to draw inferences by scanning data and picking out favorable cases. Second, it indicates how analytic tools can help to safeguard against the "tubular vision" to which we are all liable. Third, Berkson's (1958) approach – while not successful for the Doll and Hill data – illustrates how an argument for the tenability of a counterhypothesis can be developed from a proponent's own data. Fourth, the example shows that the task of a responsible critic can be as difficult and exacting as that of a responsible proponent (whereas "hit-and-run" criticism is child's play).

In my discussion of the role and responsibility of the statistical critic, my theme has been: we should not have a "double standard" in science and statistics, one standard for proponents and another for critics. The same ground rules should apply to both. If a proponent should not jump to his conclusions or base them on dogma or speculation, neither should a critic. If a proponent should be wary of "tubular vision," so should a critic. In short, we might frame the following "golden rule" for critics: Do unto a proponent as you would have *him* do.

Technical Appendix

Derivation of the sequence sign test follows directly from a result given on page 241 of Feller's *An Introduction to Probability Theory and Its Applications* (1957). Feller proves that for a single sequence (under the null hypothesis) the number of inversions is asymptotically normally distributed with a mean

$$(E_1) = \frac{n(n-1)}{4}$$

and variance

$$(V_1) = \frac{(2n+5)(n)(n-1)}{72} = \frac{(2n+5)(E_1)}{18}.$$

Assuming independence for M causes, we find at once that the total number of inversions (I) is asymptotically normally distributed with

$$E = M(E_1), V = M(V_1).$$

Since $\frac{(I-3)^2}{V}$ is asymptotically distributed as chi square with 1 degree of freedom, the sequence sign test (uncorrected) follows when the substitution $I + NI = 2E$ is made. I have included a correction for continuity analogous to the one used in the sign test (Mosteller, 1952).

References

Berkson, J. (1955). Statistical study of association between smoking and lung cancer. *Proc. Staff Meet. Mayo Clin.*, 30, 319-348.

Berkson, J. (1958). Smoking and lung cancer; some observations on 2 recent reports. *J. Am. Statist. A.*, 53: 28-38.

Bross, I.J. (1954). Misclassification in 2 × 2 tables. *Biometrics*, 10, 478-486.

Brownlee, K.A. (1957). Note on effects of nonresponse on surveys. *J. Am. Statist. A.*, 52, 29-32.

Doll, R. and Hill, A.B. (1956). Lung cancer and other causes of death in relation to smoking; second report on mortality of British doctors. *Brit. M. J.*, 2, 1071-1081.

Feller, W. (1957). *An Introduction to Probability Theory and Its Applications*, 2nd ed., Vol. I. New York, N.Y. John Wiley & Sons, Inc.

Fisher, R.A. (1957). Cancer and smoking. [Letter to the Editor.] *Nature*, 182, 596.

Fisher, R.A. (1958). Lung cancer and cigarettes? [Letter to the Editor.] *Nature*, 182, 108.

Hammond, E.C. and Horn, D. (1954). Relationship between human smoking habits and death rates: follow-up study of 187,766 men. *J.A.M.A.*, 155, 1316-1328.

Laurence, W.L. (1957). Cigarette-cancer links disputed. *New York Times* 107: (sect. 4) 7, Dec. 29, 1957.

Mosteller, F. (1952). Some statistical problems in measuring subjective response to drugs. *Biometrics*, 8, 220-226.

Pearl, R. (1928). *The Rate of Living: Being an Account of Some Experimental Studies on the Biology of Life Duration.* New York, N.Y. Albert A. Knopf, Inc.

Wynder, E.L., Bross, I.J., Cornfield, J. and O'Donnell, W.E. (1956). Lung cancer in women: study of environmental factors. *New England J. Med.*, 255, 1111-1121.

Observational Studies 4 (2018) 11-16 Submitted 11/17; Published 1/18

Statistical Criticism and Causality in *Prima Facie* Proof of Disparate Impact Discrimination

William B. Fairley wfairley@analysisandinference.com
William A. Huber whuber@analysisandinference.com
Analysis & Inference, Inc.
Springfield, PA 19064, U.S.A.

Statistical criticism in a legal case need not follow the same "rules of the game" as those set out by Bross (1960) for criticism in a scientific context. The law has its own stylized requirements for criticism engaged in by experts – statistical or otherwise. For example, the law formally assigns a "burden of proof" depending on the legal theories advanced by the parties. When that burden of proof is not laid on a statistical expert he or she may make a "hit and run criticism" (see Bross (1960, p. 395)). However, insofar as the expert is in the business of being persuasive, then hit and run in many contexts will still not suffice regardless of the formal assignment of burdens.

Law has another device for handling criticism. The decision maker, such as a judge or jury, is understood to be giving different "weights of evidence" to different witnesses depending on their persuasiveness. A hit and run criticism or a counter-hypothesis that is not tenable may be given little or no weight.

Writing for judges about how scientific evidence can be best used in legal proceedings (as applied particularly for econometric regression models), Finkelstein (1973, p. 146) has proposed that judges adopt certain protocols, one of which echoes Bross's proposal that a critic should offer a tenable counter-hypothesis. Finkelstein paints a picture of arguments converging towards "tacit agreement among the experts":

> When parties express criticisms by introducing alternative models, the process need not necessarily be a seesaw battle of conflicting econometric demonstrations. [T]here may instead be a progression towards greater refinement and correctness in statistical methodology which will not only be apparent to the decision maker, but which may also achieve results meriting at least tacit agreement among the experts.

Finkelstein echoes Bross in requiring a "superior alternative analysis" from an "objecting party":

> This experience suggests as the second protocol that (i) a party objecting to an econometric model introduced by another party should demonstrate the numerical significance [size of effect] of his objections whenever possible, and (ii) a party objecting to an econometric model of data designated by the decision maker for econometric analysis should produce a superior alternative analysis of that data.

©2018 William Fairley and William Huber.

A quantitative evaluation of effect size ("numerical significance") will be more persuasive (although it is not always available). An example is a suit brought by high ranking officers in the New York City Police Department who had been passed over for promotion to top posts by a new Commissioner. *Courtney v. City of N.Y.*, 20 F. Supp. 2d 655 (S.D.N.Y. 1998).[1] They alleged age discrimination. While these officers were in their fifties or older, many of the top posts had been given to younger officers; it was alleged that the new Commissioner had said he wanted to sweep out "old dead wood." On behalf of the plaintiffs, a city university professor testified that age and promotion were related because a chi-square test of the data showed a statistically significant difference in the rates of promotion between officers over 53 and officers under 53. The expert also provided data showing the older officers also had more experience (years of service), with the implication being that they were more qualified for that reason. In response to questioning by counsel for the Police Department, the expert declined to comment on whether the younger officers might have had more leadership ability, noting that while "experience" was "objective," no objective evidence on leadership ability was at hand.[2] The jury (perhaps impressed by this testimony of "numerical significance," or perhaps influenced by other facets of the evidence not presented here) found for the plaintiffs.

The expert seemed to infer age discrimination from an association between age and promotion. An obvious alternative explanation [3] was that the younger officers promoted did indeed have stronger leadership abilities – at least as judged by the new Commissioner regarding his agenda of taking radical departures in policing from his predecessor. Here was an instance of the numerical data at hand appearing, at least in the mind of the expert, to drive out the arguably more relevant but non-numerical data.[4]

In employment discrimination law, and in some other contexts, there is a doctrine of disparate impact in which a plaintiff may advance a prima facie case of discrimination

[1] One of the authors of this article testified on behalf of the Police Department.

[2] In other contexts, such as disparate impact in employment discrimination, courts have accepted subjective criteria. For example, in *Watson v. Ft. Worth Bank & Tr.*, 487 U.S. 977 (1988), the court wrote:

> In the context of subjective or discretionary employment decisions, the employer will often find it easier than in the case of standardized tests to produce evidence of a "manifest relationship to the employment in question." It is self-evident that many jobs, for example those involving managerial responsibilities, require personal qualities that have never been considered amenable to standardized testing. In evaluating claims that discretionary employment practices are insufficiently related to legitimate business purposes, it must be borne in mind that "[c]ourts are generally less competent than employers to restructure business practices, and unless mandated to do so by Congress they should not attempt it." ("[The] criteria [used by a university to award tenure], however difficult to apply and however much disagreement they generate in particular cases, are job related....It would be a most radical interpretation of Title VII for a court to enjoin use of an historically settled process and plainly relevant criteria largely because they lead to decisions which are difficult for a court to review"). (citations omitted)

[3] This explanation is supported by the *Peter Principle* (Peter and Hull, 1969), which posits that "every new member in a hierarchical organization climbs the hierarchy until he/she reaches his/her level of maximum incompetence." Accordingly, older members will tend to be less competent and less qualified for promotion than younger members, especially in the higher ranks of the hierarchy. Simulations bear out this conclusion: see, *inter alia*, Pluchino et al. (2009).

[4] Tribe (1971), in an influential law review article, complained that mathematical or statistical data and models could over-awe the legal decision maker, especially juries, and unfairly take the place of more subjective but more relevant evidence.

by demonstrating an imbalance in the proportion of a protected class (distinguished, for example, by age, race, or sex) among those gaining some advantage such as promotion or not being laid off.[5] Usually such an imbalance must be shown to be statistically significant and not negligible.[6] Thus, a numerical difference between proportions of minority and majority class members being laid off may in some instances be taken as sufficient for the case to proceed to a further stage.

Most of the time, stronger evidence is needed. As the Supreme Court noted in [*Watson v. Fort Worth Bank & Tr.*, 487 U.S. 994 (1988), the Supreme Court's "formulations...have consistently stressed that statistical disparities must be sufficiently substantial "that they raise...an inference of *causation.*" In other words, the statistical disparities must "show that the practice in question has *caused* the exclusion of applicants for jobs or promotions because of their membership in a protected group" *Id.* (emphases added).[7] See also *Stagi v. National R.R. Passenger Corp.*, 391 F. App'x 133, 136 n.3 (3d Cir. 2010).

A statistical difference is not always enough to establish even a *prima facie* case. Other factors than age could explain an association between age and being laid off. *Sheehan v. Daily Racing Form* was a case in which a layout editor at a newspaper in Chicago was laid off after the introduction of computer methods replaced his work.[8] 104 F.3d 940 (7th Cir. 1997). A greater proportion of older than younger employees were laid off. Table 1 gives the data ($\chi^2 = 9$, $p = 0.009$ (Fisher's Exact Test); no one was between 42 and 48).

In his opinion in *Sheehan* Judge Richard A. Posner wrote:

> More important is the expert's failure to correct for any potential explanatory variables other than age. Completely ignored was the more than remote possibility that age was correlated with a legitimate job related qualification, such as familiarity with computers. Everyone knows that younger people are on average more comfortable with computers than older people are, just as older people are

[5] The plaintiff must also identify an action or policy that (allegedly) caused the imbalance. See *Wards Cove Packing Co. v. Atonio*, 490 U.S. 657 (1989).

> [A]...plaintiff does not make out a case of disparate impact simply by showing that, "at the bottom line," there is racial imbalance in the work force. As a general matter, a plaintiff must demonstrate that it is the application of a specific or particular employment practice that has created the disparate impact under attack. Such a showing is an integral part of the plaintiff's prima facie case in a disparate-impact suit under Title VII [of the Civil Rights Act of 1964].

[6] Courts have placed an emphasis in finding discrimination on establishing the statistical significance of a difference. That emphasis is traced to the Supreme Court's decision in *Castaneda v. Partida*, 430 U.S. 482, 496 n.17(1977): "As a general rule for such large samples, if the difference between the expected value and the observed number is greater than two or three standard deviations, then the hypothesis that the jury drawing was random would be suspect to a social scientist." The court did not comment on the importance of the model or assumptions underlying a test of statistical significance. In practice, these are often ignored. For example, see cases discussed in Sugrue and Fairley (1983, pp. 953-955), especially *Bryan v. Koch*, 627 F.2d 612, 616-618 (2d Cir. 1980).

[7] Showing of causation is the key requirement. There is, however, the statement that causation can be shown by "sufficiently substantial" disparities. We have already stated above that disparities must be substantial in certain ways to demonstrate a *prima facie* case, There is the further suggestion here that the more substantial the disparities the better the demonstration of causation. This may be, but is not always, the case.

[8] The plaintiff in this case advanced not only a legal claim of disparate impact but disparate treatment (intentional discrimination).

Table 1: Laid Off Versus Age

	Laid Off?		
Age	Yes	No	Total
≤ 42	0	6	6
≥ 48	9	3	12
Total	9	9	18

Source: *Sheehan vs. Daily Racing Form*

> on average more comfortable with manual shift cars than younger people are. 104 F.3d 940, 942 (7th Cir. 1997).

Bross (1959, p. 395 left hand side) wrote:

> When a counterhypothesis involves a well-known real factor, e.g., age or sex in an epidemiological study, it would be sufficient to mention the relationship, e.g., death rates from cancer tend to increase with age.

Posner might say that his counter-hypothesis (computer facility was a factor correlated with age) is obviously superior to the hypothesis of age discrimination, which was supported by nothing more than a recital of who was laid off and their ages.

Also, see *Diehl v. Xerox Corp.*, 933 F. Supp. 1157, 1167-68 (W.D.N.Y. 1996).

> Although both experts concluded that the redeployment and dismissal rates among certain portions of the [employer's] workforce who were 40 years of age or older and [were] male employees were greater than those of other employee groups which disparities could not be caused by chance factors, Dr. Honig failed to conduct further statistical tests to determine what other factors *could* have accounted for those disparities. Instead, she concluded that since the disparities were not caused by chance, the high probabilities (1 in 100) demonstrated that they were caused by age or gender. Without conducting any other statistical tests to rule out factors other than age or gender and by relying solely on age and gender ratios, Dr. Honig's testimony is fatally flawed pursuant to *Wards Cove.*

Nevertheless, it is not uncommon to see experts rely solely on a statistical difference to support a *prima facie* case of discrimination. This is in marked contrast to the view in science or social science that correlation, whether statistically significant or not, does not (by itself) establish causality.[9] From a scientific perspective, statistical significance is not

[9] To cite just one example, Madden (1985, p. 80) writes of how finding sex discrimination in employment requires by definition adjusting for other factors that could account for pay differentials:

> Empirical work by economists has concentrated on measuring and accounting for sex differentials in productivity. The basic procedure used in empirical investigations of sex discrimination is to determine the magnitude of the pay differential *after adjusting for quantifiable sex differences in productivity-enhancing characteristics.* (emphasis supplied)

as important a desideratum in finding a difference – that can at least in part be assigned to being a member of a protected class – as "causal significance," if by this phrase is meant: "a causal relation between two variables is supported."[10] Establishing causal significance requires a very different and more extended inquiry than showing statistical significance.

In discrimination cases, we see a marked tendency for testifying experts to draw causal conclusions from tests of association, which of course goes beyond what such tests can support. For example, in cases of age discrimination in promotions, it is common to see experts opine based only on statistical differences that the effects are "age-based" or that age must be a "factor" in creating the differences. Along the same lines, it is common to see experts interpret a null hypothesis of independence or no association as a causal hypothesis. For example, in a sex discrimination case the null hypothesis might be described as a "gender-neutral" policy, suggesting that the alternative hypothesis is the policy is "gender-biased." Often such characterizations are unaccompanied by any evidence suggesting actual bias, and in many cases the absence of such evidence is actually stipulated. Similarly, the term "effect" is used, as in "the effect of race," to suggest to a lay audience a causal effect of a variable (like age, race, or gender) without having established causal significance and without explicitly claiming the effect is causal.

A good causal argument examines more than one variable: in Bross's terms, any variables suggested by plausible counterfactual hypotheses could be incorporated in the analysis. Courts have not, however, required that all possible explanatory variables be included to support a causal conclusion. *Diehl v. Xerox Corp.*, 933 F. Supp. 1167 (W.D.N.Y. 1996) is a case alleging age discrimination in employment following a reduction in force. The court wrote (933 F. Supp. 1157, 1169 (W.D.N.Y. 1996)):

> [Dr. Bloom testified that]...the probative value of a regression analysis is not eliminated simply because the analysis uses fewer than all critical variables. In Bazemore v. Friday, the Supreme Court reached the same conclusion:
>
> Importantly, it is clear that a regression analysis that includes less than all measurable variables may serve to prove a plaintiff's case. A plaintiff in a Title VII suit need not prove discrimination with scientific certainty; rather, his or her burden is to prove discrimination by a preponderance of the evidence.

From our point of view as sometime expert witnesses (and critics thereof) it often is unclear whether a demonstration of causality is required to support a given case. We see instances where experts equate association with causation. In some cases, but not all, courts criticize (or even exclude) experts for such interpretation. However, our experience is consistent with Bross's thesis insofar as judges and juries are more likely to be persuaded when a tenable counter-hypothesis can be supplied than when the only critical rejoinder is the mere suggestion that another factor might explain an outcome. However, the situation is not symmetrical between plaintiffs and defendants. A plaintiff always has the burden of showing causation; a defendant does not have this burden and only needs to rebut the plaintiff's case – without necessarily even suggesting other factors than those advanced by plaintiff.

[10] Tinkham (2010), a legal commenter, suggests that a *prima facie* case need only be supported by some relevant evidence and need not require a demonstration of causality. In practice, as in *Sheehan* and *Diehl*, we see cases where a failure to demonstrate causality called down unfavorable opinions from judges.

References

Bross, I. D. J. (1960). Statistical criticism. *Cancer*, 13, 394-400.

Finkelstein, M.O. (1973). Regression Models in Administrative Proceedings. *Harvard Law Review*, 86, No. 8.

Madden, J.F. (1985) The Persistence of Pay Differentials: The Economics of Sex Discrimination. In Laurie Larwood et al, Editors, *Women and Work: An Annual Review*, Volume 1, Sage.

Peter, L. and Hull, R. (1969). *The Peter Principle. Why Things Always Go Wrong.* William Morrow.

Pluchino, A., Rapisarda, A. and Garofalo, C. The Peter Principle Revisited: A Computational Study. arXiv:0907.0455 [physics.soc-ph].

Sugrue, T.J. and Fairley, W.B. (1983). A Case of Unexamined Assumptions: The Use and Misuse of the Statistical Analysis of Casteneda/Hazelwood in Discrimination Litigation. *Boston College Law Review*, July 1983.

Tinkham, T. (2010). The Uses and Misuses of Statistical Proof in Age Discrimination Claims. *Hofstra Labor and Employment Law Journal*, 27, Issue 2.

Tribe, L.H. (1971). "Trial by Mathematics: Precision and Ritual in the Trial Process." *Harvard Law Review*, 84, No. 6.

Cases Cited

Bazemore v. Friday, 478 U.S. 385 (1986).

Castaneda v. Partida, 430 U.S. 482 (1977).

Courtney v. City of N.Y., 20 F. Supp. 2d 655 (S.D.N.Y. 1998).

Diehl v. Xerox Corp., 933 F.Supp. 1157 (W.D.N.Y.1996)

Sheehan v. Daily Racing Form, Inc., 104 F.3d 940 (7th Cir. 1997).

Stagi v.National R.R. Passenger Corp., 391 F. (3d Cir. 2010).

Wards Cove Packing Co. v. Atonio, 490 U.S. 657 (1989).

Watson v. Ft. Worth Bank & Tr., 487 U.S. 977 (1988).

Observational Studies 4 (2018) 17-31 Submitted 12/17; Published 1/18

The Potential Usefulness of Bross's Principles of Statistical Criticism for the Evaluation of Statistical Evidence in Law and Public Policy

Joseph L. Gastwirth **jlgast@gwu.edu**
Department of Statistics
George Washington University
Washington, DC 20052, USA

1. Introduction

The classic paper by Bross (1960) should be read in conjunction with Cornfield's inequality, in the appendix of Cornfield, Haenszel et al. (1959) and described in Gastwirth (1988), Greenhouse (1982), Greenhouse (2009), Rosenbaum and Krieger (1990) and Rosenbaum (2002). The inequality states conditions that a suggested omitted variable needs to satisfy in order to "explain" a difference in the proportions of successes (or failures) between two groups. Many extensions of the original result allowing for sampling error or matched pairs and other designs have been developed over the years (Rosenbaum, 2002; Guo at al. 2013) and suggested for use in legal cases (Gastwirth, 1992). Although one may question the thoroughness of the analysis of the data in Table I of the article, the criteria Bross gives for statistical criticism remain relevant today.[1]

This commentary will focus on the applicability of the framework suggested by Bross for the evaluation of criticisms of statistical evidence in law and public policy as other commentators will discuss the advances in statistical methodology that provide a more comprehensive analysis of epidemiologic and related data sets. Section 2 reviews the role of statistical evidence in discrimination cases. These cases are brought under civil, rather than criminal law, so the trier of fact (jury or judge) decides the case based on the preponderance of the evidence or "more likely than not" standard, rather than the "beyond a reasonable doubt" standard used in criminal cases. Section 3 discusses how courts have considered criticisms of statistical evidence. Because our focus is on the usefulness of Bross's principles, some important legal aspects, such as whether the cases discussed in Section 3 concerned an appeal of a summary judgment or were a class action will not be emphasized.[2] Section

[1] Table I examines the death rates of many diseases, however, most toxic agents only affect one or a few diseases, e.g. workers exposed to benzene have an increased risk of leukemia. The sign test gives equal weight to each of the types of mortality, although epidemiologic studies had shown that smoking had a strong association with lung cancer. Statistical methods designed to detect trends in dose-response data, e.g. the Cochran-Armitage (1954, 1955) test or its extension (Mantel, 1963) to stratified data or combining the results of several studies would be more powerful.

[2] When a party moves for summary judgment, it is claiming that the opposing party has no case and the entire case should end. Summary judgment is warranted when "the pleadings, depositions, answers to interrogatories and admissions on file, together with the affidavits, if any, show that there is no genuine issue

©2018 Joseph Gastwirth.

4 describes the main studies that led to warning the public about the association between the use of aspirin to treat children with colds or chicken pox and their risk of subsequently developing a rare but serious disease, Reye Syndrome. Because the industry was able to raise questions about the early studies, without being held to the criteria stated by Bross (1960), slightly over three years elapsed from the time the FDA (November, 1982) felt the public should be notified and the start of the warning campaign in the United States in early 1985.

2. The Role of Statistical Evidence in Disparate Impact and Disparate Treatment Discrimination Cases

There are two categories of EEO cases, disparate impact and disparate treatment. Disparate impact cases concern the legitimacy of a job requirement, e.g. passing a written or physical test or possessing a certain level of education. When the proportion of applicants from a legally protected group, typically a race-ethnic minority or females, satisfying the requirement is significantly less than the corresponding proportion of majority applicants, an employer has the opportunity to discredit the plaintiffs' analysis by showing that the data contain serious errors or omit a major relevant variable. If the employer cannot show that the plaintiffs' statistics are defective, then they need to demonstrate that the requirement is necessary for the job. If the employer can demonstrate that the requirement is job-related, the plaintiff is given the opportunity to suggest an alternative criterion that achieves the objective of the requirement at issue but has less of a disparate impact. This approach was established by the Supreme Court in *Griggs v. Duke Power*, 401 U.S. 424, 91 S.Ct. 849 (1971).

In *International Brotherhood of Teamsters v. United States*, 431 U.S. 324 (1977), the U.S. Supreme Court defined disparate treatment as discriminatory acts in which "[t]he employer simply treats some people less favorably than others because of their race, color, religion, sex, or national origin." The plaintiff needs to show that the employer's decision was motivated by the employee's membership in a protected group and statistical evidence is relevant to the process of shifting burdens of production during the proceedings:

1. The plaintiff has a burden of establishing a *prima facie* case that, left unrebutted, raises an inference of discrimination.

2. If the plaintiff establishes a *prima facie* case, the defendant has the burden of producing a legitimate and non-discriminatory reason for its action.

3. Then the plaintiff has the burden of showing that the non-discriminatory justification given by the defendant is a mere pretext.

as to any material fact and that the moving party is entitled to judgment as a matter of law." Fed. R. Civ. R. Civ. P. Rule 56(c). Courts evaluate the evidence giving the opposing party, who would lose the case at that point, the benefit of the doubt. For example, even if the trial judge thinks the evidence favors the moving party, if a reasonable jury could view it in favor of the opposing party, summary judgment should not be granted. The seminal case in the U.S. is *Anderson v. Liberty Lobby, Inc.*, 477 U.S. 242, 242-243, 106 S. Ct. 2505, 2511 (1986). Thus, a party can prevail at summary judgment but still lose the case. Because class actions concern a large number of plaintiffs and usually involve a large amount of money, plaintiffs need to satisfy special rules, see Fed. R. Civ. Proc. 23. Statistical evidence is often used to show there is a common question, e.g. underpayment of minority or female employees.

In both types of cases, statistical evidence may be used by a plaintiff at the first or third stage and may be used by the defendant at the second stage to support its claim that its job requirement is related to successful performance of the job or its employment decisions arose from legitimate considerations. After the plaintiff submits statistical evidence at the first stage, the defendant may rebut the analysis by showing that it is flawed. Similarly, at the second stage of either type of case the plaintiff may point out defects in the defendant's analysis.

Using the framework suggested by Bross, at the first stage the plaintiff has the role of the *proponent* of a scientific hypothesis. Typically, their evidence will be a comparison of the success rates or average wages of minority and majority success rates, often adjusting for the effect of other appropriate factors, such as seniority or relevant educational background, with a regression or stratified analysis. Then the defendant has the role of the critic and the protocol suggested by Bross is especially relevant as it implies that courts should require the defendant to

1. Specify the deficiencies in the plaintiff's statistical presentation.

2. Show that they are sufficiently severe that they raise serious doubts about the inference made by the plaintiff's expert from the statistical analysis.

If, after taking into account the defendant's criticism of the plaintiff's analysis and the other evidence, the judge decides that the plaintiff's evidence is still sufficient to establish a *prima facie* case of disparate treatment, the defendant becomes the *proponent* of the hypothesis that any disparity is due to legitimate factors, rather than discrimination. Often they will submit a more detailed analysis, e.g. by incorporating more legitimate factors or stratifying the data into more appropriate subgroups.[3] At the similar stage of a disparate impact case, the defendant needs to submit a validation study showing the job requirement at issue is correlated with performance of the job. After the defendant submits its evidence, both statistical and non-statistical, the plaintiff has the role of the *critic* and courts should require them to offer criticisms meeting the above two criteria.

In disparate impact cases, if the defendant demonstrates that the job requirement at issue is job related, the plaintiff is given an opportunity to propose an alternative requirement that has less of a disparate impact but achieves the purpose of the original one, in selecting successful employees. Now the plaintiff is the *proponent* of an alternative requirement and the defendant may *criticize* the proposal by showing that it is not have the same predictive ability of on the job success as the original requirement.

[3] For example, in a promotion case, if an employer has several locations, say k, and typically promotes from within each of them, the defendant may rebut a plaintiffs' analysis of the data from all locations aggregated into a single 2x2 table, by stratifying the data into k sub-tables and showing that the Cochran-Mantel-Haenszel test is not significant.

3. The Applicability of these Principles of Statistical Criticism to Equal Employment Cases

The Supreme Court's discussion of the regression analyses[4] submitted by both parties in *Bazemore v. Friday*[5] illustrates how the principles of statistical criticism could assist the courts when they assess statistical evidence. The lower courts had rejected plaintiffs' regression analysis submitted to support their request that a class action be certified for the claim of the plaintiffs, black employees of the North Carolina Agricultural Extension Service, of discrimination in salaries and in the performance evaluation system. Prior to the Civil Rights Act of 1965, the service paid black employees less than whites and a secondary issue was whether pre-Act disparities in pay that continued after the Civil Rights Act was passed violated the new law.[6]

The expert for the United States analyzed salary data for 1974, 1975 and 1981 by regressing pay on race, education, tenure and job title because an official at the Service stated that salaries were determined from four factors, education, tenure, job title and job performance. He found that black employees earned $331 less than whites did in 1974, the disparity was $395 in 1975 and these were statistically significant at the .05 level. While there was disparity disadvantaging blacks it 1981, it was not statistically significant.

The defendant submitted a regression analysis for 1975 using similar predictors and found a significant disparity of $384, quite similar to those found by the plaintiffs. Furthermore, when they included quartile rank for that year, the disparity increased to $475.[7]

Apparently, the defendant offered a number of criticisms of the plaintiffs' regression, which the lower courts accepted. In particular: the County Chairman who were mainly white might skew the data to show whites earning more than blacks, several variables relating to job performance and variation in pay between counties. The Court noted that job title was included in plaintiffs' regressions, so the inclusion of County Chairman would be accounted for and that defendant's own regression showed that including job performance as reflected in the quartile rankings, did *not* explain the disparity.

The defendant did *not* submit evidence to support its claim that the disparities could be explained by county variations. The government, however, did present evidence that black employees were not located disproportionately in counties that contributed only a small amount to employee salaries.[8]

[4] There is a large literature concerning the use of statistics and regression in legal cases. See Gastwirth (1988), Fienberg (1989), Zeisel and Kaye (1997) and Finkelstein, and Levin (2000). Gray (2009) and Hersch and Bullock (2014) discuss how courts have evaluated data from many cases and suggest that they may under-weight plaintiffs regressions by giving too much credence to defendants' criticism. Greiner (2009) reviews the applicability of causal inference methods in civil rights cases. Sinclair and Pan (2009) and Graubard (2009) discuss the use of Peters-Belson regression, which fits a model to the majority group and compares the actual outcomes (e.g. pay or promotion) of the protected group to its predicted value from the majority equation

[5] 478 U.S. 385 (1986)

[6] The decision, Id. at n.8 noted that if the pre-1965 pay disparities continued, the employer violated the law. For employees hired afterwards, plaintiffs would need to provide evidence that new disparities were created. Prior to 1965, the Service maintained two racially segregated branches and paid black employees less than white employees.

[7] The defendant's expert said he was unable to explain why adding those rankings increased the disparity but mentioned that the ranking data was missing for 20% of the employees.

[8] The Extension Service was funded by money from the national, state and local governments.

The Court stated that an analysis, which accounts for the *major* factors, is acceptable and that failure to include additional predictors would affect its probativeness, rather than its admissibility as evidence. In a footnote, the opinion stated that a regression analysis could be so incomplete that the results would not be admissible, but this concern did not apply to plaintiffs' evidence in *Bazemore*. The Court rejected the standard adopted by the appellate court that the plaintiffs' regression should include *all* measurable factors. The Court reminded the lower courts that plaintiffs are not required to prove discrimination with scientific certainty; their burden is to prove discrimination by the preponderance of the evidence.

Had the courts known of the principles of statistical criticism in Bross (1960), they would have required the defendant to do more than suggest some factors, especially pay differentials among the state's counties; they would expect the defendant to offer evidence in support of that claim. Thus, the plaintiffs would not have needed to show that black employees were not concentrated in counties that contributed a small amount of funds.

Comments: (1) The evidence in the case is somewhat unusual in that the Service submitted a regression that included a possible explanatory variable reflecting job performance that actually increased the disparity. Normally, a defendant would not suggest a factor that increased a disparity as an explanation for it. (2) The Supreme Court did not discuss the data on quartile rankings as the appellate opinion examined the data for 1981 in each of the five (of six) districts with agents of both races, separately.[9] According to the appellate court, 751 F. 2d 662 (1984) at 673-74, the disparity in the proportions of black and white employees receiving evaluations in the lowest quartile, who would not receive a merit raise, were *not* statistically significant in *any* of the five districts.[10] Even though the proportion of blacks being ranked in the lowest quartile exceeded the proportion of whites in the five districts, the court apparently required that statistical significance be observed in at least one-half of the districts. The proper analysis uses the Cochran-Mantel-Haenszel method for combining 2x2 tables and finds a statistically significant difference at the .01 level (Gastwirth, 1988, p. 267-78), and the odds a black employee ranked in the lowest quartile was 2.17 or twice those of a white.

3.1 Examples of cases where judges adopted an approach similar to Bross's principles

The plaintiffs in *Randall vs. Rolls Royce*[11] were women who complained that they received less pay than similar male employees did. After the district court rejected their request that the case be certified as a class action, their appeal was considered by the 7th Circuit. The company established broad pay ranges for compensation categories for employees in jobs that were of equal value to it. In order to meet competition for job types that were in

[9] There were no black employees in one district.

[10] The court used a t-test to compare the proportions. Had it used Fisher's exact test, it would have observed a statistically significant result in the Northwest district as the test yields a two-tailed p-value of .04. More importantly, one should examine stratified data by an appropriate combination test, e.g. the Cochran-Mantel-Haenszel test after checking that the odds ratios are similar. Gastwirth, Miao and Pan (2017) reanalyze stratified data from an actual case illustrating the methodology and providing references to the literature.

[11] 637 F 3d 818 (7th Cir. 2011).

demand, the company created an additional narrow range for each category to allow it to match the "prevailing market wage", i.e. wages offered by other employers in the area.

In 2003, the year just before the complaint period, the plaintiffs noted that the average base pay of male employees in the five compensation categories relevant to the case was 5% higher than that of females. This differential persisted throughout the complaint period and if the difference were attributable to sex, the firm's failure to eliminate it would perpetuate discrimination and violate the law. The plaintiffs apparently submitted a regression analysis. By including the type of job in a more comprehensive regression, the defendant showed that gender was no longer a significant predictor of salary.

The opinion mentions that the plaintiff's expert made other errors, in addition to failing to adjust for differences in the jobs occupied by male and female employees. He included employees hired after the beginning of the complaint period, which did not make sense since the claim was that females were discriminated against because the company failed to erase a disparity that existed at the outset of the period. Moreover, he did not make a study of the reasons for differences in the starting salaries of these male and female hires.

The opinion followed the precepts of Dr. Bross, as it did not simply accept a suggestion that the type of job could possibly explain the 5% disparity in the salaries of male and female employees; rather it required the defendant to submit an appropriate regression. Had the plaintiffs found a significant gender difference in an analysis that included job type, then the issue of the coverage of the database used by the plaintiffs should be explored, i.e., the analysis should be redone by examining the appropriate employees. Irwin Bross' article, "Statistical Criticism," gives advice that is surprisingly current, given that it appeared in the journal *Cancer* nearly sixty years ago. Indeed, the only obviously dated aspects of this paper are the use of the generic male pronoun and the sense that it was still an open question whether cigarette smoking caused lung cancer.

The *Allen v. Seidman*[12] case concerned the disparate impact on African-American employees of a written exam used to determine whether to promote bank examiners employed at the FDIC. Fourteen of the 36 or 38.9% of the African-Americans passed, while 329 of 391 or 84.1% of Whites did. The pass rate of African-Americans was less than one-half that of Whites and the difference is highly significant.[13] The defendant suggested that the minority candidates might have had lesser qualifications, e.g. education. The opinion noted that all of the examinees had worked for five to fifteen years and at least one year at their current grade level at the agency and had obtained a recommendation from their regional director. Thus, the minority and White candidates appeared to be reasonably homogeneous. The defendant did not submit any evidence showing an imbalance between the two groups with respect to a job-related factor, e.g. seniority or education. The opinion continues with "since the defendant, while taking pot shots – none fatal at the plaintiffs' statistical comparison, did not bother to conduct its own regression analysis, which for all we know would have confirmed and strengthened the plaintiffs' simpler study."

The *Allen* opinion is a good example of a court following sound principles of statistical criticism. Even its description of the defendant's suggestions of possible explanatory factors,

[12] 881 F.2d 375 (7th Cir. 1989) upholding the district's finding of disparate impact in Allen v. Isaac. 39 FEP Cases, 1142 (N.D. Ill. 1986).

[13] Fisher's exact test yields a p-value of about 10^{-8} or less than one in a million.

without submitting evidence that they could explain the disparity, as taking "pot-shots" is similar to the words "hit and run" used by Bross.

In support of a claim of sex discrimination *EEOC v. General Telephone Co. of Northwest*, 829 F.2d 885 (9th Cir. 1989), the Commission's expert introduced a regression including the legitimate factors, showing that gender was statistically significantly negatively related to pay. The defendant did not submit a regression including additional variables or submit a different statistical analysis. Rather, it argued that differences in job interest among men and women would explain the disparity. The District court accepted this explanation and found for the defendant. Consistent with Bross's criteria for a critic, the Ninth Circuit reversed the trial court and remanded the case for reconsideration. The opinion notes "GenTel had to produce credible evidence that curing the alleged flaws would also cure the statistical disparity – proof which GenTel did not offer. Thus, we hold that the district court erred in uncritically accepting GenTel's assertion that the plaintiff's analyses were flawed where GenTel had failed to show that if the EEOC had "adequately" accounted for the alleged flaws, the disparities in its analyses would have been eliminated."

The statistical evidence in *Sheehan v. Purolator*, 839 F.2d 99 (2nd Cir. 1988) concerned whether female exempt employees, as a class, were discriminated against in pay, job assignment and promotion. The plaintiffs submitted a regression analysis, which showed disparities in pay between male and female employees.[14] The opinion notes that the defendant supported its claim that plaintiffs' regression was flawed because it did not include education and prior experience by *introducing* evidence that these factors indeed could explain the disparities.

3.2 A case where courts accepted "explanations" without requiring evidence they could explain the disparity

The statistical evidence in the *Equal Employment Opportunity Commission v. Sears, Roebuck & Co.*, 839 F.2d 302 (7th Cir. 1988) sex discrimination case focused on whether females had the same opportunity as males to be hired or promoted to commissioned sales jobs, which had more risk but generally had higher pay than non-commissioned sales positions. Plaintiffs' expert selected six factors that he thought might affect an applicant's chance of selection: (1) job applied for; (2) age; (3) education; (4) job type experience; (5) product line experience; and (6) commission product sales experience. A logistic regression including gender and these variables yielded a statistically significant negative coefficient for gender when the model was fit to national data, but not for all regions or territories

In addition to questioning the coding of some of the predictors, Sears argued that the predictor variables chosen by the expert did not include major variables such as interest in a commission sales position and females had less relevant experience and less education. Furthermore, some characteristics such as assertiveness, friendliness and motivation are better evaluated in an interview and cannot be obtained from a written application form. The appellate court thought that the experience variables used by the plaintiff were adequate but that Sears provided sufficient evidence from national surveys and some limited studies

[14] The appellate opinion does not report the independent variables used to predict an employee's salary.

by the company, demonstrating that women had noticeably less interest in the type of position than men did.[15]

Apparently, the courts did *not* require Sears to demonstrate that the general difference in interest in commissioned sales jobs, which was also true for individuals who actually applied, was of sufficient magnitude that it and incorporating years of relevant experience, could explain the disparities. The dissenting opinion of Judge Cudahy states, "Perhaps the most questionable aspect of the majority opinion is its acceptance of women's alleged low interest and qualifications for commission selling as a complete explanation for the huge statistical disparities favoring men."

This case has important implications for statisticians, who might become involved as an expert witness. The courts gave less credibility to the statistician's choice of predictors because he was not a labor economist and they were not necessarily the major predictors considered by Sears. Thus, it is important for a statistician to be involved in discovery in order to learn the criteria actually considered by an employer and obtain the information on those factors.

3.3 Two cases illustrating the need for principles of criticism to consider the availability of data on the major variables and for statistical experts to be given all the relevant data

In a very useful review of the need for statistical evidence to meet the standards of reliability set out by the Court in Daubert and the Federal Rules of Evidence, Rosenblum (2015) summarizes several cases where the plaintiff's evidence was insufficient.[16] Rosenblum (2015) discusses how courts have implemented these standards in equal employment cases. Typically, this occurs when the statistical presentation does not account or adjust for known job-related factors. Sometimes plaintiffs' expert has pooled data from a wide variety of positions or locations into one large sample, e.g., ignoring the fact that most hires for the job at issue come from the local area or a difference in the type of product in each location.[17] The *Bickerstaff v. Vassar College*, 196 F. 3d 435 (1999) case, however, is interesting for our purposes as it is less clear that the plaintiff's regression was so flawed that it should have been disregarded.

The African-American plaintiff was an Associate Professor, who in 1994 claimed that she was discriminated against because she had not been promoted to Full Professor. Her expert submitted a regression addressing whether salaries at Vassar varied due to race or

[15] The coding of some of the experience variables as yes (1) or no (0), instead of using the amount of relevant experience was noted as a limitation by the trial and appellate courts.

[16] In *Daubert v. Merrill Dow* 509 U.S. 579, 584-587 (1993), the Court established guidelines for the admission of scientific testimony to ensure it was relevant to the issues involved in the case and reliable. Two of the factors were whether the method had been peer reviewed and published and whether it had a known and acceptable error rate.

[17] The *Penk v. Oregon State Board of Education*, 816 F.2d 458, case illustrates this issue. Plaintiffs claimed that women were paid less than similarly qualified men throughout the college and university system. Thus, the two major research universities were included with colleges with a different purpose. Again, the courts rejected a regression analysis submitted by plaintiffs that did not include several major variables. The opinion states "Missing parts of the plaintiffs' interpretation of the board's decision-making equation included such highly determinative quality and productivity factors as teaching quality, community and institutional service, and quality of research and scholarship."

sex. Salaries were regressed on experience, rank, productivity and discipline. The opinion reported that salaries were determined based on scholarship (0-3 points), teaching (0-3 points) and service (0-2 points). The court criticized the analysis because it omitted two of the key variables, teaching and service.[18] Furthermore, the court observed that in light of the point system that the College employed, these variables were quantifiable and could be controlled for in a statistical analysis.

Because the quality of teaching and research should be reflected, in part, in their experience and rank, it is reasonable to ask whether the points awarded each year in the three categories were available. Employers are required to keep personnel records for a period, so they might well have been. If the plaintiff had or could have had access to this data[19], then the court would be justified in concluding the regression omitted too many key factors. On the other hand, if the College did not keep that information, it is unclear that the point system was actually the main determinant of faculty salaries or pay raises.

In contrast to *Bickerstaff*, plaintiff's expert in *Diehl v. Xerox Corp.* 933. F. Supp. 1157 (W.D.N.Y. 1996), that concerned whether older male workers were unfairly laid off, did not include performance ratings of employees. She claimed that the managers would be predisposed to give older employees lower evaluations given the general level of discrimination in society. The court noted that the expert had not conducted a statistical analysis to check whether there was evidence of a pattern of lower performance ratings at the firm and deemed defendant's regression analysis that included these factors more relevant.

It is interesting to contrast the *Diehl* and *General Telephone* cases with the opinion in *Sears Roebuck*. The *Sears* court accepted the defendant's evidence that women, in general, had a lower level of interest in certain types of jobs. The other two decisions said that the party asserting that a general societal pattern would explain a disparity or justify its not being included, as a factor in its analysis should demonstrate that the general pattern applies to the applicants or employees in the particular case.

The issue of whether performance ratings need to be included in statistical analyses submitted in promotion or layoff cases occurs frequently.[20] When possible, statistical analyses submitted by a plaintiff should demonstrate that members of the protected class received lower performance ratings or that, the ratings in the review used to determine the promotions or layoffs were significantly lower than previous ones.[21] Courts also consider noticeably

[18] The opinion 196 F. 3d 435 at 449-450 raised questions about how well the productivity variable controlled for scholarship.

[19] In *Carpenter v. Boeing Co.*, 456 F.3d 1183 (10th Cir. 2006) the plaintiffs did not request information on some of the factors listed in the Collective Bargaining Agreement that specified how opportunities for overtime were to be allocated and the court discounted their regression. The *Bickerstaff* opinion does not discuss the issue of data availability and it is possible that plaintiffs only requested data in electronic form, such as payroll data that included job-title and date of hire.

[20] For example, *Nitshke v. McDonnell Douglas Corp.*, 68 F. 3d 249 (8th Cir. 1995) and *Hutson v. McDonnell Douglas Corp.* 63 F. 3d 771 (8th Cir. 1995). The decision in *Smith v. Virginia Commonwealth University*, 84 F. 3d (4th Cir. 1996) states that whether performance ratings should be included is a question of fact. Thus, the facts to the specific case will determine the appropriateness of using performance ratings.

[21] Gastwirth (1997) illustrates an analysis showing that older workers received worse ratings than younger employees did from a case that settled just prior to trial. Yu (2009) presents an alternative analysis of the data, which yielded a p-value of .07 for a two-tailed test, just above the usual .05 level. In age discrimination cases, however, one-sided tests are appropriate as only employees over the age of 40 are covered by the law. Hence, both procedures would conclude that age affected the performance reviews. The fact that the

more negative reviews that occur after an employee has filed a charge of discrimination as evidence of retaliation.[22]

3.4 Implications for statistical experts

While the Court in *Bazemore* set out useful guidelines on the factors a statistical analysis needs to satisfy in order to be admitted in evidence, it did not discuss the criteria courts should use in evaluating criticisms. An important issue that was not addressed in Bazemore is how to determine the legitimate major factors. In cases involving universities or colleges, teaching, research and service are well known determinants of pay and promotion.[23] Legitimate factors in other cases might be seniority, education and prior experience in hiring cases, along with performance ratings or reviews in promotion and equal pay cases. Information about these factors, however, should be available in order for them to qualify as a major factor. Otherwise, there is no way for courts to oversee that they were applied fairly to all applicants or employees or even used in the decisions that are being scrutinized.

These considerations and the cases discussed indicate that statistical experts request information about all the factors used in making the employment decisions. If one does not utilize information about a particular factor, one should be prepared to justify doing so.[24] For equal employment and other civil cases, such as product liability, it is important that an expert be involved at the time of discovery, so they can ensure that the major factors and information about how they were used in the decision process can be studied.[25]

evaluations of several older employees were lower than the previous ones the received sufficed to defeat the defendant's summary judgment motion in *Woods v. The Boeing Company*, 2009 WL 4609678 (10th Cir. 2009).

[22] See *Wyatt v. City of Boston*, 35 F.3d 13, 15-16, (1st Cir. 1994) for a list of actions indicating retaliation and *Kim v. Nash Finch*, 123 F. 3d 1046 (8th Cir. 1997) for an example where many more negative comments were put in an Asian employee's file after a complaint.

[23] In *Fisher v. Vassar College*, 70 F.3d 1420 (1995), the court noted that service is amorphous. In part, this is because the decision makers, Chairs and Deans, typically appoint faculty members to Committees etc. Thus, information about the willingness of faculty members to participate as a member of a committee, rather than the actual appointments may be more relevant. In addition, the relative importance of community service may vary with the type of institution.

[24] The *Wado v. Xerox Corp.*, 991 F. Supp. 174, 184 (1998), affirmed *Smith v. Xerox*, 196 F. 3d 358 (2d Cir. 1999) discussed by Rosenblum (2015) is an example. In an age discrimination case, the expert did not include performance ratings because they assumed they were biased. This case is similar to *Diehl* discussed in the text.

[25] After a civil case is initiated, there is a period set aside for discovery. The parties ask questions to the other designed to ascertain relevant information. During this period, other employees and the testifying expert may be deposed so both sides are aware of the evidence that will be presented. In product liability cases concerning a new drug, the plaintiff will ask for studies the manufacturer made as well as other complaints from individuals who may have been harmed from using the drug. The defendant, will ask about the health of the plaintiff to see whether the illness may have arisen from another health problem, rather than the drug in question.

4. The Reye syndrome story: How the use of the principles of statistical criticism might have saved lives in both the U.S. and U.K.[26]

For a number of years, pediatric specialists suspected that the risk of children contracting a rare but very serious disease, Reye syndrome, increased after they received aspirin to alleviate symptoms of a cold or similar childhood disease.[27] After the case control study by Halpin et al. (1982), which confirmed a statistically significant association found in two earlier studies, the FDA (1982) initiated the process of warning the public. The proposed warning and background studies was submitted to the Office of Information and Regulatory Analysis at the Office of Management Budget for review.[28]

During the reviewing process, the Aspirin Institute, which represented the interests of the industry, raised several questions about the 1982 study, the FDA had relied on. The Halpin et al. (1982) study matched each case to one or two (when available) controls on their age, race, geographical, time and type of illness.[29] A logistic model controlled for the presence of fever, headache and sore throat yielded an estimated relative risk of 11.5 (p-value ¡ .001). In addition to submitting a detailed reanalysis of the data, which questioned the coding of some of the answers the respondents gave, the Institute argued that the association could be due to "recall" bias and the possibility that parents of cases, who knew about the association, might say they administered aspirin because the child developed Reye syndrome.

The principles of statistical criticism would have required the Institute to submit evidence that a much higher proportion of parents or guardians of cases knew about the association than the parents of the controls. Furthermore, it should have demonstrated a meaningful difference in the ability of parents of cases and controls to recall what they gave their child, even though Reye syndrome tends to occur within a couple of weeks after the prodromal illness.

Two criticisms seemingly had some validity. First, a higher proportion of cases reported a fever than controls and the maximum fever reached by cases generally exceeded that of the controls. The logistic equation had just used the presence (1) or absence (0) of fever. Second, the parents of cases, who were interviewed while their child was very ill, were under much more stress than the parents of controls were and this could have affected their response. The industry suggested the controls formed from children hospitalized for other diseases or visited the emergency room would have been more appropriate as the parents who responded to the questionnaire had been under stress.

Table 3 in Halpin et al. (1982) addressed the issue of the effect of fever by stratifying the data into four levels of the highest fever (none, low, middle, high) and observed that while the prevalence of fever was higher in the cases, that for each level of fever, a higher fraction of cases had taken aspirin than controls. The data in their Table 3 ignored the matching,

[26] The author served as a statistical consultant to the Office of Statistical Policy at OMB and assisted in the review of the epidemiologic studies the FDA relied on.

[27] See Trauner (1984) for a medical perspective and FDA (1982) for a chronology of the early studies.

[28] Section 6(a)(3)(c) of Executive Order12866, and the Regulatory Right-to-Know Act, require that proposed regulations be reviewed by that office. Circular A-4, issued in September 2003, describes the review process, which considers the costs and benefits of a proposed regulation.

[29] Most controls were obtained by locating classmates of a case who were absent from class, presumably with the prodromal illness, at the same time as the case.

i.e., it is reported as several 2x2 tables. Analyzing it with the Cochran-Mantel-Haenszel test yields an estimated odds ratio of 14.7 (p-value $< 10^{-5}$ and 3.5 was the lower end of a 95% confidence interval. Furthermore, at the highest level of maximum fever, all 41 cases used aspirin and 33 of 44 (75%) of the controls did. Thus, the prevalence of high maximum fever in the cases could not have met Cornfield's conditions for an omitted factor.

By the time the decision was made, little information was submitted concerning the potential effect of stress on the accuracy of recall. The government decided that another study should be conducted.[30] A Public Health Task Force was formed and planned the study was during 1983 and a pilot study was undertaken during mid-February through May 1984.[31] The data analysis was reviewed and made available in December 1984. The logistic regression model that controlled for fever and other symptoms yielded an estimated relative risk of 19.0 and the Task Force recommended that the public be warned.[32] When the cases were compared to each of the control groups, the two controls suggested by the industry had the highest estimated odds ratios (28.5 for emergency room controls and 70.2 for inpatient controls). This highlights the importance of the requirement stated by Bross that a critic do more than suggest a possible explanation; they need to submit evidence supporting a conjecture, here a care-giver being under stress, could create the relative risks around 10, found in the earlier studies.

The industry raised doubts about the study and the association to the Regulatory Office at OMB. In particular, it argued that the decision should be based on one study. Based on all the studies, in January 1985, the Office said it would approve the request to warn the public and the industry voluntarily conducted a warning campaign. The CDC reported that the number of cases of Reye syndrome reported to it dropped from 204 in 1984 to 98 in 1985 (see Table 4 in Gastwirth, 2013) for a longer time series and references).

The United Kingdom did not institute a public education campaign until a further study in 1986 confirmed the association and the number of cases declined soon afterwards (Hardie et al. 1996; Porter et al. 1990). Porter et al. (1990) studied the effect of the warning campaign. They report that children with febrile illnesses were 17 times more likely to have taken aspirin before hospitalization in 1985-6 than in 1998-9. They also indicate that about 15 to 25% of the caregivers in Belfast and London were aware of the association between aspirin use in children and Reye syndrome and less than 50% had heard of Reye syndrome.

5. Conclusion and Implications

The events leading up to warning the public about how to avoid a major cause of Reye syndrome illustrate the importance of the statistics profession developing and endorsing sound principles of statistical criticism. The principles suggested by Bross remain a sound guide, however, the context of the particular application should also be considered. In

[30] The review of the regulation needed to be expedited because of an unusual circumstance, totally unrelated to the scientific issues. President Reagan had given speech in Iowa and answered questions from the audience. One person asked about Reye syndrome and the proposed regulation, and he responded that the review would be completed within a month.

[31] Putting aside the few questionnaires, the industry raised doubts about, there still was a statistically increased relative risk, although lower than 10. A non-statistical aspect of the review was whether the increased risk was sufficiently high to justify the proposed wording of the warning.

[32] See Public Health Service (1985) for further details.

situations like the Reye syndrome one, where alternative medications, without such serious side effects, are available, the critic should be required to demonstrate that their criticisms or suggested alternative explanations really explain away the association. There may be other situations in public health where an effective alternative treatment does not exist. Then the proponent of the studies showing an increased risk may also need to demonstrate that the risk of harm out-weighs the benefit of the medicine.

In the context of discrimination cases under the disparate treatment approach, the cost to a plaintiff of a court's total rejection of their statistical evidence is likely to be that they will lose the case at that point. In contrast, if a court admits the plaintiff's evidence and allows the case to go forward, the defendant only has the burden of explaining how the disparity between the success rates of minority and majority employees or applicants arose from legitimate considerations. Thus, at the first stage of the proceedings, the plaintiff, who is the proponent in the framework of Bross, should be required to use the information about the major factors considered by the employer, provided the employer has preserved this information. If the plaintiff succeeds in establishing a statistically significant disparity between similarly qualified (with respect to these major factors) minority and majority employees, the defendant should be required to demonstrate that, the flaws or omitted factors in the plaintiff's evidence are *sufficiently severe* that the ultimate inference could be changed.

In the context of a criminal case, where the prosecution must prove the defendant "beyond a reasonable doubt", the scientific support underlying some types of evidence, e.g. bullet residue and even fingerprint evidence has been questioned. Bolck and Stamouli (2017) discuss some of the statistical issues and refer to many useful articles. The principles of statistical criticism will be helpful to courts in assessing whether a minor violation of an assumption in a statistical calculation or small amount of missing data in study showing a forensic technique has a certain degree of accuracy are severe enough to alter the ultimate impact of the evidence on a jury.

References

Armitage, P. (1955). Tests for linear trends in proportions and frequencies. *Biometrics*, 11, 375-386.

Bolck, A. and Stamouli, A. (2017). Likelihood Ratios for categorical evidence; Comparison of LR models applied to gunshot residue data. *Law, Probability and Risk*, 16, 71-90.

Bross, I. D. J. (1960). Statistical criticism. *Cancer*, 13, 394-400.

Cochran, W.G. (1954). Some methods for strengthening the common χ^2 tests. *Biometrics*, 10, 417-451.

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

FDA (1982). Labeling for Salicylate-Containing Drug Products. Federal Register, 47, Dec. 28, 1982, 57886-57901.

Fienberg, S.E. (Ed.) (1989). *The Evolving Role of Statistical Assessments as Evidence in the Courts.* New York, NY: Springer-Verlag.

Finkelstein, M.O. and Levin, B. (2001). *Statistics for Lawyers*, 2nd Ed. New York, NY: Springer-Verlag

Gastwirth, J.L. (1988). *Statistical Reasoning in Law and Public Policy* Vol. 1 Statistical Concepts and Issues of Fairness. Orlando, FL: Academic Press.

Gastwirth, J.L. (1992). Methods for assessing the sensitivity of statistical comparisons Used in Title VII cases to omitted variables. *Jurimetrics Journal*, 33, 19-33.

Gastwirth, J.L. (2013). Should law and public policy adopt practical causality' as the appropriate criteria for deciding product liability cases and public policy? *Law, Probability and Risk*, 12, 169-18.

Gastwirth, J.L., Miao, W. and Pan, Q. (2017). Statistical issues in Kerner v. Denver: a class action disparate impact case. *Law, Probability and Risk*, 16, 35-54.

Graubard, B.I (2009) Comment on "Using the Peters-Belson method in equal employment opportunity personnel evaluations" by Sinclair and Pan. *Law Probability and Risk*, 8, 119-122.

Gray, M. (1993). Can statistics tell us what we do not want to hear? The case of complex salary structures. *Statistical Science*, 8, 144-179,

Greenhouse, J.B. (2009). Commentary: Cornfield, epidemiology and causality. *International Journal of Epidemiology*, 38, 1199-1201.

Greenhouse, S.W. (1982). Jerome Cornfield's contributions to epidemiology. *Biometrics*, 38, Supplement, 33-46.

Guo Z, Cheng J, Lorch S. A. and Small D. S. (2014). Using an instrumental variable to test for unmeasured confounding. *Statistics in Medicine*, 33, 35283546.

Halpin, T. J., Holtzhauer, F.J., Campbell, R.J. et al. (1982). Reye's syndrome and medication use. *Journal of the American Medical Association*, 248, 687-691.

Hardie, R.M., Newton, .H, Bruce, J.C., Glasgow, J.F.T., Mowat, A.P., Stephenson, J.B.P and Hall, S.M. (1996). The changing clinical pattern of Reye's syndrome 1982-1990, *Archives of Disease in Childhood*, 74, 400-405.

Hersch, J. and Bullock, B.D. 2014). The Use and Misuse of Econometric Evidence in Employment Discrimination Cases. *Washington and Lee Law Review*, 71, 2365-2429.

Mantel, N. (1963). Chi-square tests with one degree of freedom: Extensions of the Mantel-Haenszel test. *Journal of the American Statistical Association*, 68, 690-700.

Porter, J.D.H., Robinson, P.H., Glasgow, J.F.T., Banks, J.H. and Hall, S.M. (1990). Trends in the incidence of Reye's syndrome and the use of aspirin. *Archives of Disease in Childhood*, 65, 826-829.

Public Health Service (1985). Public Health service study on Reye's syndrome and medications. *New England Journal of Medicine*, 313, 849-857.

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

Rosenbaum, P.R. and Krieger, A.M. (1990). Sensitivity analysis for two-sample permutation tests in observational studies. *Journal of the American Statistical Association*, 85, 493-498.

Rosenblum, M. (2015). Strategic evidence issues in equal employment litigation. *Touro Law Review*, 16, 1299-1317.

Sinclair, M.D. and Pan, Q. (2009). Using the Peters-Belson method in equal employment opportunity personnel evaluations. *Law, Probability and Risk*, 8, 95-117.

Trauner, D.A. (1984). Reye's syndrome, *Western Journal of Medicine*, 141, 206-209.

Yu, B. (2009). Modelling an omitted factor in employment discrimination cases. *Law, Probability and Risk*, 8, 153-158.

Zeisel, H. and Kaye, D. H. (1997). *Prove It with Figures: Empirical Methods in Litigation*, Springer-Verlag, New York, USA.

Observational Studies 4 (2018) 32-33 Submitted 3/17; Published 1/18

Learning from and Responding to Statistical Criticism

Andrew Gelman **gelman@stat.columbia.edu**
Department of Statistics
Columbia University
New York, NY 10027, USA

Irwin Bross' article, "Statistical Criticism," gives advice that is surprisingly current, given that it appeared in the journal *Cancer* nearly sixty years ago. Indeed, the only obviously dated aspects of this paper are the use of the generic male pronoun and the sense that it was still an open question whether cigarette smoking caused lung cancer.

In his article, Bross acts a *critic of criticism*, expressing support for the general form but recommending that critics go beyond hit-and-run, dogmatism, speculation, and tunnel vision. This all seems reasonable to me, but I think criticisms can also be taken at face value. If I publish a paper and someone replies with a flawed criticism, I still should be able to respond to its specifics. Indeed, there have been times when my own work has been much improved by criticism that was itself blinkered but which still revealed important and fixable flaws in my published work.

I would go further and argue that nearly all criticism has value. Again, I'll place myself in the position of the researcher whose work is being slammed. Consider the following sorts of statistical criticism, aligned in roughly decreasing order of quality:

- A thorough, comprehensive reassessment. Of course this is valuable: if an expert goes to the trouble of (a) finding problems in my work, (b) demonstrating that my errors were consequential, and (c) providing an alternative, then this is a clear step forward.

- A narrow but precise correction. If I have made a mistake in data processing or analysis, or if I have missed some alternative explanation for my findings, I would like to know. Even if it turns out that my error did not affect my main conclusions, it will be helpful to myself and to future researchers to fix the immediate problem

- Identification of a potential problem. What if someone criticizes one of my published papers by suggesting a problem without demonstrating its relevance? This can be annoying, but I don't see the problem with the publication of such a criticism: Readers should be made aware of this potential problem, and future researchers can explore it.

- Confusion. All too often, criticism reveals a misunderstanding on the part of the critic. But this can have value too, in revealing that I have failed to communicate some point in my original article. We can't hope to anticipate all possible misreadings of our work, but it is good to take advantage of opportunities to clarify.

- Hack jobs. Bross was writing about cancer studies, an area where cigarette companies paid big money for several decades to highly-credentialed M.D.'s and Ph.D.'s to crit-

©2018 Andrew Gelman.

> icize epidemiological research using any and all arguments at hand. It can be hard to deal with criticism that is motivated by a desire to muddy the waters rather than to get at the truth. Hack criticism can indeed have negative value, and the problem here is not so much in the criticism itself – after all, even a hack can make a good point, and hacks will use legitimate arguments where available. Rather, the hack problem comes in the critical process: a critic who aims at truth should welcome a strong response, while a hack will be motivated to avoid any productive resolution.

Another way to see the value of post-publication criticism, even when it is imperfect, is to consider the role of *pre*-publication review. It is perfectly acceptable for a peer reviewer to raise a narrow point, to speculate, or to point out a potential data flaw without demonstrating that the problem in question is consequential. Referees are encouraged to point out potential concerns, and it is the duty of the author of the paper to either correct the problems or to demonstrate their unimportance. Somehow, though, the burden of proof shifts from the author (in the pre-publication stage) to the critic (after the paper has been published). It is not clear to me that either of these burdens is appropriate. I would prefer a smoother integration of scientific review at all stages, with pre-publication reports made public and post-publication reports being appended to published articles.

Overall, I am inclined to paraphrase Al Smith and reply to Bross that the ills of criticism can be cured by more criticism. That said, I recognize that any system based on open exchange can be hijacked by hacks, trolls, and other insincere actors. The key issues in dealing with such people are economic and political, not statistical, but we still need to be able to learn from and respond to statistical criticisms, whatever their source.

Acknowledgments

We thank Dylan Small for inviting this paper.

Observational Studies 4 (2018) 34-41 Submitted 1/18; Published 1/18

The Tenability of Counterhypotheses: A comment on Bross' discussion of statistical criticism

Jennifer Hill **jennifer.hill@nyu.edu**
Department of Applied Statistics, Social Science, and the Humanities
New York University
New York, New York, 10003

Katherine J. Hoggatt **katherine.hoggatt@va.gov**
VA Health Services Research and Development (HSR&D)
Center for the Study of Healthcare Innovation, Implementation & Policy
VA Greater Los Angeles Healthcare System
Los Angeles, CA, 90073

1. Introduction

We enjoyed reading the commentary by Bross about the appropriate role for a statistician as critic (Bross, 1960). It was an important discussion to initiate at the time the article was written, particularly in light of the highly contentious and scientifically critical debate about the link between smoking and cancer, which involved some of the leading statistical minds of the era (Cornfield, 1954; Cornfield et al., 1959). We believe discussions about the role of a statistical critic are equally if not more relevant today given that our ability to casually critique the work of research "proponents," to use Bross' term, and to disseminate such comments broadly in unrefereed venues has increased exponentially since the time that Bross was writing. Given the complexity and breadth of the issues involved, we focus our discussion on Brosss contention that a critic should have a tenable counter-hypothesis. We further position our comments within the context of causal inference where some additional subtleties arise with regard to satisfying this requirement.

2. Tenable counter-hypotheses

For a critic's counter-hypothesis to be tenable, Bross maintains that "a minimal requirement would be that the effects predicted from the critic's hypothesis should be in line with the actual data, at least in direction and order of magnitude." This seems reasonable in the abstract, but there is not a well-defined criterion for meeting this requirement. For example, how should we achieve this goal if we wish to estimate a causal effect from observational data when even reliably estimating the direction of effects requires making strong, often untestable, assumptions? To illustrate these issues, we consider the common scenario where a proponent is investigating a hypothesis about a causal effect of an exposure on an outcome using observational data. We describe how "dogmatic" statistical criticism (e.g., that you cannot infer causation from observational data) can lead to further methodological errors

©2018 Jennifer Hill and Katherine J. Hoggatt.

and fail to shed light on whether a proponent's hypothesis or a critic's counter-hypothesis should be considered more credible. Finally, we discuss how sensitivity analysis may provide a path forward.

2.1 Observational Data

Let us consider a specific situation in which the proponent's initial hypothesis is about a causal effect. For example, suppose that a proponent claims that exposure to the measles, mumps, and rubella (MMR) vaccine increases the occurrence of autism. Let us also assume for the moment that no randomized or natural experiment is available. According to Bross' criterion, a critic who disagreed with the proponent's claim would need to posit a counter-hypothesis, such as that the MMR vaccine has no effect on autism incidence. Further, according to Bross, this counter-hypothesis should be supported by, or at least not inconsistent with, existing observational data. Let us assume that the critic has access to a reasonably-sized, child-level observational dataset with accurate measurements of what vaccines the child received (and when), subsequent developmental assessment resulting in an autism diagnosis, and pre-treatment measurements of potential confounders. What kinds of estimates from this dataset might we be willing to accept as supportive of a counter-hypothesis?

Even ignoring the (not insubstantial) issues around statistical and practical significance (Berger and Selke, 1987; Gelman and Loken, 2013; Wasserstein and Lazar, 2016), major issues loom. We know that any given estimate of $\mathrm{E}[Y \mid Z, X]$ (where Y is the outcome, Z is the treatment, and X is a vector of potential confounding covariates unaffected by the treatment) is unlikely to be an unbiased estimate of the true estimand, e.g. $\mathrm{E}[Y(1) - Y(0)]$ (where $Y(0)$ and $Y(1)$ are potential outcomes with the typical definitions, as in Rubin, 1978). There are several reasons for this, but first and foremost it is unlikely that we have satisfied the so-called ignorability assumption, $Y(1), Y(0) \perp Z|X$. Colloquially speaking (and ignoring some technical subtleties, see, for instance Greenland et al., 1999), this means it is unlikely in most observational studies that we have measured all confounding covariates.

In the absence of a design that creates this independence structure, we are left to consider how estimates of the causal estimand behave when we only control for subsets of X that are insufficient to guarantee ignorability. For example, suppose the truth is that vaccines decrease the incidence of autism (even though the marginal association is positive). This could lead to a situation where analyses that include a proper subset of the sufficient set confounders yield estimates that are not only biased but of the wrong sign. Such a situation (which arguably is not terribly rare) would make it all too easy to use the data as "evidence" that supports a variety of different counter-hypotheses; that is, it would be easy to show that the estimands corresponding to a counter-hypothesis of a positive effect are "in line with the data."

A more confusing situation arises when we posit a point hypothesis (as opposed to the directional hypothesis above), such as that the true effect of MMR vaccine on autism is 0 (for a discussion of the evidence that supports this claim see Plotkin et al., 2009). Supporting such a counter-hypothesis would be complicated, not only because estimates from analyses that do not satisfy ignorability might have different signs and magnitudes, but also because it is unclear what it means to support a point null hypothesis.

2.2 Randomization to the rescue?

How can we proceed if it is uncertain or unlikely that ignorability is satisfied? An overly simplistic solution to the problem might be to require that only evidence from randomized experiments be accepted to support a counter-hypothesis. After all, in its pure form the randomized experiment justifies the assumption of (strong) ignorability. Bross (1960) and other contemporaries (including, notably, (Cornfield, 1954)) expressed frustration however that statisticians were using the "gold standard" of the randomized experiment as a cudgel to beat down all attempts to make a causal claim using observational data. In fact Bross (1960) highlights this practice in the article in the section on "dogmatic criticism".

We are concerned by reflexive dogmatic criticism as well. One problem with the emphasis on a controlled or natural experiment is that we may ignore evidence about causal effects if that evidence is derived from non-randomized experiments. This tunnel vision can be particularly problematic when investigating research questions that do not lend themselves to randomized experiments for ethical or logistical reasons. An additional problem is that we may end up overstating the infallibility of randomized (controlled or naturally occuring) experiments that occur in practice, no matter their vulnerabilities. Many complications can and often do arise that would preclude a researcher from making a causal claim, even in the context of a randomized experiment, without making additional assumptions. These complications include but are not limited to missing data, noncompliance, measurement error, and grouped data structures. Combinations of these issues are common and are even more difficult to handle (for example, see Barnard et al., 2003; Reardon and Raudenbush, 2013). In situations where the randomized experiment is free from such complications or when additional required assumptions seem plausible (e.g. the exclusion restriction in a randomized experiment with noncompliance or a missing at random assumption to recover missing data), randomized experiments are nonetheless almost always limited in their generalizability (Stuart et al., 2015).

Even given these well-known limitations, there persists a belief that a study with randomization is necessarily a more rigorous approach to a causal inquiry than a study without this feature (Imai et al., 2008). This confidence tends to extend to so-called natural experiments as well (see, for example, Duncan et al., 2004), including methods such as instrumental variables, regression discontinuity, and even, oddly, fixed effects models for identifying causal effects. Yet we know that when the assumptions of these methods fail to hold, things can go badly quickly (Angrist et al., 1996; Reardon and Raudenbush, 2013; Martens et al., 2006; Middleton et al., 2016). Moreover, even when one of these methodologies works well, it will, like randomized experiments, tend to yield estimates that apply most directly to narrow slices of the observation sample, and additional assumptions are necessary to generalize these local average treatment effect (LATE) estimates to a broader population (Hoggatt and Greenland, 2014).

The upshot is that when making causal inferences, most analyses, whether they use data from observational studies or randomized experiments, will rely on some sort of untestable assumptions. If we are requiring that a counter-hypothesis be tenable, it seems the criteria should include a reasonable assessment of the plausibility of such assumptions. However, if we are comparing competing sets of untestable assumptions (corresponding to the proponent's original analysis and the critic's analysis in support of a counter-hypothesis) how should we assess which of the sets of assumptions are most plausible? Would it be better, for instance, to use an instrumental variables approach where the instrument is weak and the exclusion restriction is questionable or to use an observational study where we are uncertain that we have measured all confounders?

2.3 Sensitivity Analysis: A way forward?

One way to tackle this problem is to promote increased use of sensitivity analyses, by which we mean any of a variety of approaches that explore the sensitivity of our estimates to violations of key assumptions of our analysis. Rather than making a binary decision about which counter-hypotheses are tenable, the goal would be for critics (and perhaps proponents if they are acting as their own critics) to provide a range of estimates that are derived from different sets of assumptions supporting the proponent's analyses. Bross was one of the first scholars to propose this strategy in the context of health research (Bross, 1966, 1967), and much of the early sensitivity analysis literature focused on methods to address possible departures from ignorability assumption (for example Cornfield et al., 1959).

Yet today, more than 50 years after Bross wrote his commentary, sensitivity analysis is seldom used in applied empirical research. This is true despite that the fact that there has been increased focus in the methodological literature in recent years on approaches to assess sensitivity to departures from the ignorability assumption in simple observational studies (see, for example, Rosenbaum and Rubin, 983a; Rosenbaum, 1987; Greenland, 1996; Gastwirth et al., 1998; Rosenbaum, 2002; Imbens, 2003; McCandless et al., 2007; Rosenbaum, 2010; Harada, 2013; Carnegie et al., 2016; Dorie et al., 2016). Moreover some simple sensitivity analyses can be done with a basic spreadsheet program, and software packages are available for more complex applications (for example, Gangl, 2004; Keele, 2010; Carnegie et al., 2015).

It is true, however, that there has been less of a focus on developing methods and software to explore the sensitivity to assumptions required for other types of causal analyses including instrumental variables, mediation, fixed effects, and regression discontinuity (exceptions include Imbens and Rubin, 1997; Small, 2007; Imai et al., 2010; Middleton et al., 2016; McCandless and Somers, 2017). Even more rare are publications that compare two competing identification strategies (for an interesting example of this see DiPrete and Gangl, 2004) or that simultaneously address two different types of assumptions in one analysis (for example Dorie et al., 2016). Certainly more work is needed to create user-friendly, interpretable approaches that can be applied in a variety of circumstances.

Furthermore, the results from a sensitivity analysis will be more useful when that analysis incorporates both statistical expertise and a subject matter expert's prior knowledge A truly interdisciplinary approach to sensitivity analysis could, for example, incorporate subject-matter-specific models for the data generating process, information about the types of likely unobserved confounders, and the most plausible direction and magnitude of (conditional) associations between the unobserved confounders and the treatment and outcome. It could also address Bross' call for the statistical critic to supply more specific and tenable counter-hypotheses. Unfortunately, requiring collaboration between investigators, who are subject matter experts, and statisticians, who can translate this knowledge into parameters for a formal, quantitative sensitivity analysis, may create an additional hurdle to broader use.

We argue that the practical barriers to adoption of sensitivity analysis are not merely technical, however. Equally lacking are clear incentives to make sensitivity analysis a routine part of empirical research. It is understandable that research proponents would be reluctant to incorporate additional analyses that may make their findings less credible. For example, "failure testing" to assess quantitatively whether a violation of ignorability could "explain away" an observed association will often show that the existence of such a confounder is possible, if not plausible. More sophisticated applications of sensitivity analysis may be better suited to the objectives of research proponents when these methods quantify how causal effect estimates change depending on a wide range of specific assumptions (for example see Carnegie et al., 2016), thus formalizing the process of assessing which counter-hypotheses are most tenable.

External incentives to promote sensitivity analysis may also be needed, and research gatekeepers (such as editors and reviewers) have an important role to play. For example, editorial standards could promote the use and transparent reporting of results from sensitivity analyses in the peer-reviewed literature. Referees of papers could request statistical reviews and encourage or even require that research proponents include sensitivity analysis for key assumptions. Journal editors could also require that statistical criticism be published in discussion articles with clearly stated counter-hypotheses and sensitivity analysis as appropriate. Efforts to encourage data sharing can also promote more rigorous evaluation of counter-hypotheses using a research proponent's own data.

3. Conclusion

Today, as when Bross wrote his commentary, it can seem as if the only job of the statistical critic is to point out problems that could occur, regardless of plausibility or likely impact. A downside of this kind of hit and run criticism, where the mere observation of a flaw in a study's methodology is enough to discount the study's findings, is that it can foster a double standard whereby a research proponent must rule out every conceivable alternative hypothesis to justify a study's findings but a critic need only suggest a counter-hypothesis to undermine them. This double standard may lead to knee-jerk dismissal of findings based on observational data and overconfidence in randomization as a design feature. We agree with Bross that progress may require the statistical critic to stand on more equal footing with the research proponent. Incentivizing or requiring this has ramifications for issues as broad as the standards for peer review in journal publication, data sharing policies, and establishment of criteria for evaluating the empirical support for scientific hypotheses.

Acknowledgments

We would like to acknowledge support for this project from VA HSR&D (IIR 15-436) as well as by Institute of Education Sciences grant R305D110037. The views expressed within are solely those of the authors, and do not necessarily represent the views of the Department of Veterans Affairs or of the United States government.

References

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91:444–472.

Barnard, J., Frangakis, C., Hill, J. L., and Rubin, D. B. (2003). A principal stratification approach to broken randomized experiments: A case study of vouchers in new york city. *Journal of the American Statistical Association*, 98:299–323.

Berger, J. and Selke, T. (1987). Testing a point null hypothesis: The irreconcilability of p values and evidence. *Journal of the American Statistical Association*, 82:112–122.

Bross, I. D. (1960). Statistical criticism. *Cancer*, 13:394400.

Bross, I. D. (1966). Spurious effects from an extraneous variable. *Journal of Chronic Diseases*, 19(6):637–647.

Bross, I. D. (1967). Pertinency of an extraneous variable. *Journal of Chronic Diseases*, 20(7):487–495.

Carnegie, N. B., Harada, M., Dorie, V., and Hill, J. (2015). *treatSens: Sensitivity Analysis for Causal Inference.* R package version 2.0, accessed 07/13/2015. Available from: http://CRAN.R-project.org/package=treatSens.

Carnegie, N. B., Harada, M., and Hill, J. (2016). Assessing sensitivity to unmeasured confounding using a simulated potential confounder. *Journal of Research on Educational Effectiveness*, 9:395–420.

Cornfield, J. (1954). Questions and answers: Statistical relationships and proof in medicine. *The American Statistician*, 22:173–203.

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.

DiPrete, T. and Gangl, M. (2004). Assessing bias in the estimation of causal effects: Rosenbaum bounds on matching estimators and instrumental variables estimation with imperfect instruments. *Sociological Methodology*, 34:271–310.

Dorie, V., Carnegie, N. B., Harada, M., and Hill, J. (2016). A flexible, interpretable framework for assessing sensitivity to unmeasured confounding. *Statistics and Medicine*, 35:3453–3470.

Duncan, G. J., Magnuson, K. A., and Ludwig, J. (2004). The endogeneity problem in developmental studies. *Research in Human Development*, 1(1-2):59–80. Available from: https://doi.org/10.1080/15427609.2004.9683330.

Gangl, M. (2004). Rbounds: Stata module to perform rosenbaum sensitivity analysis for average treatment effects on the treated. Available from: https://EconPapers.repec.org/RePEc:boc:bocode:s438301.

Gastwirth, J. L., Krieger, A. M., and Rosenbaum, P. R. (1998). Dual and simultaneous sensitivity analysis for matched pairs. *Biometrika*, 85(4):907–920.

Gelman, A. and Loken, E. (2013). The garden of forking paths: Why multiple comparisons can be a problem, even when there is no fishing expedition or p-hacking and the research hypothesis was posited ahead of time. Technical report, Columbia University.

Greenland, S. (1996). Basic methods for sensitivity analysis of biases. *International Journal of Epidemiology*, 25(6):1107–1116.

Greenland, S., Robins, J. M., and Pearl, J. (1999). Confounding and collapsibility in causal inference. *Statistical Science*, 14(1):29–46.

Harada, M. (2013). Generalized sensitivity analysis. Technical report, New York University, New York, NY.

Hoggatt, K. J. and Greenland, S. (2014). Extending organizational schema for causal effects (commentary to accompany gatto, campbell, and schwartz). *Epidemiology*, 25(1):98–102.

Imai, K., Keele, L., and Tingley, D. (2010). A general approach to causal mediation analysis. *Psychological Methods*, 15(4):309–334.

Imai, K., King, G., and Stuart, E. (2008). Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society, Series A*, 171(2):481–502.

Imbens, G. (2003). Sensitivity to exogeneity assumptions in program evaluation. In *The American Economic Review: Papers and Proceedings of the One Hundred Fifteenth Annual Meeting of the American Economic Association*, volume 93, pages 126–132, New York, NY. American Economic Association.

Imbens, G. W. and Rubin, D. B. (1997). Bayesian inference for causal effects in randomized experiments with noncompliance. *The Annals of Statistics*, 25:305–327.

Keele, L. (2010). An overview of rbounds: An r package for rosenbaum bounds sensitivity analysis with matched data. Technical report, Columbus, OH.

Martens, E., Pestman, W., de Boer, A., Belitser, S., and Klungel, O. (2006). Instrumental variables: application and limitations. *Epidemiology*, 17(3):260–267.

McCandless, L. C., Gustafson, P., and Levy, A. (2007). Bayesian sensitivity analysis for unmeasured confounding in observational studies. *Statistics in Medicine*, 26(11):2331–2347.

McCandless, L. C. and Somers, J. (2017). Bayesian sensitivity analysis for unmeasured confounding in causal mediation analysis. *Statistical Methods in Medical Research*, in press.

Middleton, J., Scott, M., Diakow, R., and Hill, J. (2016). Bias amplification and bias unmasking. *Political Analysis*, 24:307–323.

Plotkin, S., Gerber, J., and Offit, P. (2009). Vaccines and autism: A tale of shifting hypotheses. *Clinical Infectious Diseases*, 48(4):456–461.

Reardon, S. and Raudenbush, S. (2013). Under what assumptions do site-by-treatment instruments identify average causal effects? *Sociological Methods and Research*, 42(2):143–163.

Rosenbaum, P. R. (1987). Sensitivity analysis for certain permutation inferences in matched observational studies. *Biometrika*, 74(1):13–26.

Rosenbaum, P. R. (2002). *Observational Studies*. Springer, New York.

Rosenbaum, P. R. (2010). Design sensitivity and efficiency in observational studies. *Journal of the American Statistical Association*, 105:692–702.

Rosenbaum, P. R. and Rubin, D. B. (1983a). Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society. Series B (Methodological)*, 45(2):212–218.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics*, 6:34–58.

Small, D. (2007). Sensitivity analysis for instrumental variables regression with overidentifying restrictions. *Journal of the American Statistical Association*, 102:1049–1058.

Stuart, E., Bradshaw, C. P., and Leaf, P. (2015). Assessing the generalizability of randomized trial results to target populations. *Prevention Science*, 16:475–485.

Wasserstein, R. L. and Lazar, N. A. (2016). The ASA's statement on p-values: Context, process, and purpose. *American Statistician*, 70:129–133.

Observational Studies 1 (2018) 42-56 Submitted 7/17; Published 1/18

Judging Statistical Criticism

Daniel E. Ho **dho@law.stanford.edu**
Stanford Law School
Stanford, CA 94305, U.S.A.

Abstract

Bross (1960) proposes rules for statistical criticism, chiefly that critics bear the responsibility of proving the tenability of a counterhypothesis. This Comment makes three points. First, the higher the tenability standard, the more statisticians will be drawn into the local ground rules of a substantive field. Bross feared this prospect, yet his work exemplifies it. Second, more content needs to be given to the tenability standard across domains. Proving tenability may be untenable, for instance, when data is unavailable. Third, Bross's proposal ultimately led him to espouse a quasi-judicial "adversary science" proceeding to resolve controversial issues of public policy (Bross, 1980). But Bross's own involvement in a pilot at the Nuclear Regulatory Commission illustrates the difficulties with a "science court" model, with adversarialism potentially exacerbating rather than muting political conflict. I illustrate these points with the common setting of statistical evidence in an antidiscrimination suit, using data from the University of Texas at Austin School of Law. Ultimately, Bross's work raises profound questions about the institutions for judging statistical criticism.

1. Introduction

The hydrogen bomb. Nuclear power. Genetically modified organisms. In the 1970s, Washington DC was abuzz with a newfangled way to resolve the most contentious public policy issues of the day: "Science Court" (Kantrowitz, 1967; Mazur, 1973). The White House organized a Task Force. The Department of Commerce, the National Science Foundation, and the American Association for the Advancement of Science co-sponsored a conference, convening scientists and policymakers to design the court — neutral scientific judges, scientific advocates presenting their case with opportunity for cross-examination, and a verdict on scientific facts — giving the green light for an experiment. Allan Mazur, one of the two principal proponents of science court, reminisced, "lunching at a small table with a Nobel laureate, the President's science advisor, and Margaret Mead was testosterone inducing" (Mazur, 1993, p. 165). One journal dubbed it "the ultimate in peer review" (Seagrave, 1976) and others referred to the idea as a "supreme court of science" (Mazur, 1993, p. 164). When asked for suitable topics, EPA Administrator Russell Train suggested global warming (Leeper, 1976, pp. 718-19).

As quickly as the model gained traction, it engendered sharp criticism. Detractors charged that there was a mismatch between adversarialism and the scientific process. Commingling science and advocacy would encourage abuse of science (Matheny and Williams, 1981). Others argued that the idea of a court judgment clashed with science as a communal enterprise. Ecologist Barry Commoner, for instance, decried the "attempt to reintroduce

©2018 Daniel E. Ho.

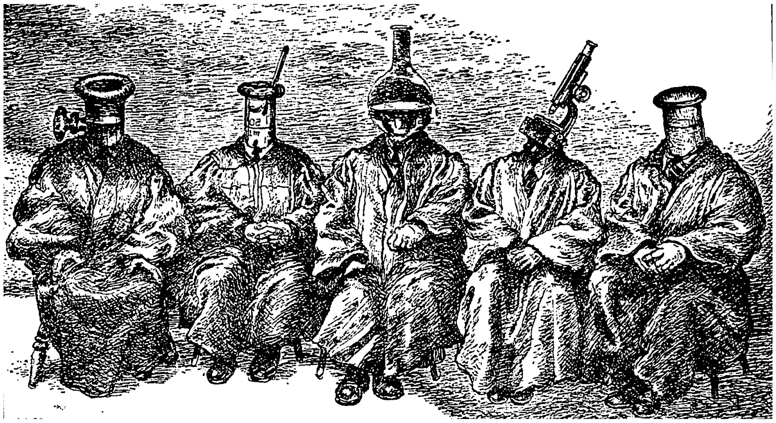

Figure 1: Illustration of the "Science Court" appearing in the *New York Times.* Thanks to Carlos Llerena Aguirre for permission to excerpt.

authoritarianism in science" (Seagrave, 1976, p. 378). The *New York Times* ran a cartoon of lab instruments donning judicial robes (Figure 1).[1] Philip Abelson, editor of *Science*, questioned its likely efficacy: "You could put a bunch of scientists in white robes and they could . . . make a solemn judgment of truth. And a lot of people would still think the devil is lurking out there in the Bermuda Triangle" (Seagrave, 1976, p. 379).

Irwin Bross himself was a fan. Indeed, Bross participated in one of the few pilots sponsored by the Nuclear Regulatory Commission (NRC), arguing that the scientific evidence proved negative health effects of low-level radiation exposure. Bross's advocacy for the science court, at least in his conception, was a natural extension of the ideas in "Statistical Criticism."[2] While Bross's advice was phrased in terms of professional ethics and norms in 1960, it led him to advocate for institutions to judge statistical and scientific criticism. His involvement also illustrates how adversarial science and science court ultimately fizzled.

2. The Virtues of Statistical Criticism

I am grateful for the invitation to comment on the re-publication of Bross's "Statistical Criticism" in *Observational Studies.* The article raises profound issues of scientific truth-seeking and how we should judge statistical criticism. Bross lamented the "superficial and sophomoric" statistical criticism levied against tobacco-cancer studies. He provided

[1] John Noble Wilford, Science Considers Its Own 'Court,' N.Y. Times, Feb. 29, 1976, at 140.

[2] Citing to Bross (1960), Bross (1980) makes the case as follows:

> A large number of standard put-downs are in wide use in academia. These put-downs are criticisms by innuendo rather than statements that show specific errors in the data, methods, or findings. What can be done to raise the depressingly low standards of statistical criticism and scientific controversy that now exist? What can be done to stop this kind of nonsense from delaying or blocking essential public health action? Is there any way to resolve these statistical issues more quickly and scientifically? One device that has promise is called adversary science. This device adapts to scientific questions the adversary procedure used in courtroom trials.

.

a typology of statistical criticism: (a) the hit-and-run, which points to a flaw without developing a counterhypothesis; (b) the dogmatic, which appeals to statistical theory to categorically dismiss bodies of work; (c) the speculative, which proposes a counterhypothesis but makes no attempt to reconcile it with extant evidence; and (d) the tubular, which fails to see evidence contrary to the favored hypothesis (tubular / tunnel vision). Bross argued persuasively that the responsible critic should bear the responsibility for proving the tenability of a counterhypothesis.

To illustrate this practice, Bross revisited smoking and disease data analyzed by Berkson (1958).[3] Berkson was critical of the link between smoking and cancer. He observed that because smoking appeared to be associated with a wide range of diseases, selection bias may confound smoking studies. Wrote Berkson: "I find it quite incredible that smoking should cause all these diseases" (p. 32).

Bross believed Berkson to have fallen prey to tunnel vision. To illustrate how one might prove the tenability of selection bias, Bross formalized a kind of placebo test by distinguishing "specific diseases," where etiological evidence supports a link to chemical components of tobacco, and "nonspecific diseases," where no etiological link is evident. Conducting separate tests for each category, Bross showed how Berkson's observation appeared incorrect. While specific diseases were statistically significantly correlated with smoking, nonspecific diseases were not. On its own terms, Berkson's criticism fell short of proving tenability.

Bross's contribution stands up well. Substitute any contentious policy issue (taxes and economic growth, guns and crime, universal health insurance and cost) with "smoking and lung cancer", and many of the same observations hold. The quality of statistical criticism can remain poor, "obscur[ing] a scientific discussion rather than clarify[ing] it" (Bross, 1960, p. 394). Perhaps because they are easier than full reanalyses, hit-and-run, dogmatic, speculative, and tunnel vision critiques persist.[4] Tunnel vision and motivated reasoning continue to lead to divergent conclusions on factual inferences (see, e.g., Kahan et al., 2012). Just as Ronald Fisher discounted epidemiological data on smoking (Bross, 1960, p. 396), some present day observers "raise[] randomization to the level of dogma" (id.), without being willing to contemplate any observational, descriptive, or qualitative evidence (Cook, 2015). While randomization is understandably the gold standard for causal inference, such dogma is unhelpful in areas where randomization is infeasible or unethical and where natural experiments are sparse.

3. Unresolved Difficulties

I offer three comments on Bross's contribution. The first is about the tension internal to Bross's "Statistical Criticism," when the tenability standard increasingly requires the statistician to engage with local ground rules. The second is about the meaning of tenability across contexts (e.g., when data is unavailable or when it is uncertain how to weight observational and experimental evidence). The third is about the path of "adversary science" that "Statistical Criticism" paved for Bross.

[3] The data originally come from Doll and Hill (1956).

[4] Writing in the 1990s, Mazur (1993, p. 169) opined that "the perception of undisciplined, raucous and chaotic technical controversy has dissipated" since the 1970s.

3.1 The Gravity of Local Ground Rules

Bross advocated that critics go beyond simply raising objections. Instead, critics should develop a counterhypothesis and prove its tenability with existing data and knowledge. While proving the tenability of a counterhypothesis is indeed worthwhile, this responsibility also potentially conflicts with Bross's admonition for the statistician to stay close to her domain. "A statistician should be especially careful . . . in the domain of the subject matter field — he is functioning as an epidemiologist or sociologist or psychiatrist . . . rather than as a statistician." In Bross's view, this was acutely the case for substantive rather than methodological counterhypotheses, "since 'local' ground rules . . . come into play." Yet the tension pervades the large swath of efforts to prove the tenability of a counterhypothesis.

Consider the Berkson example. Applying a permutation test to specific and nonspecific diseases may *seem* methodological. At the very least Bross didn't seem to classify Berkson's as a "substantive hypothesis" raising concern over the limits of the statistician's domain, suggesting that "analytical tools" can guard against Berkson's tunnel vision. But the only way for Bross to distinguish disease types (e.g., coronary thrombosis, cardiovascular disease, other respiratory disease) was to resort to local ground rules. The statistician necessarily must engage with the etiological medical evidence to understand the plausibility of the mechanism. Bross classified 15 disease categories into (a) specific, (b) questionable, and (c) nonspecific diseases. Yet how is the statistician supposed to reach such a decision without entering the substantive domain? For instance, how are we to know that "other" cardiovascular disease should be classified as nonspecific, as Bross classifies, when smoking has since been shown to affect cardiovascular disease on a range of measures (e.g., Critchley and Capewell, 2003)? What coronary events and diagnoses were included by Berkson in this category and how do we know they are not plausibly related to smoking? Bross's classification is not necessarily wrong, but his examination of Berkson's counterhypothesis illustrates that proof of tenability requires engagement with local ground rules. And the higher the standard for tenability, the greater this tension.

In later writings, as he was drawn further into the substance of the radiation debate, Bross fleshed out these concerns, posing provocatively whether statisticians should serve as scientists or be relegated to "shoe clerks." Bross worried that the path of least resistance would be to serve as a shoe clerk, by which he meant simply pleasing the customer to earn a commission (e.g., running the power calculation, fitting the model). But as scientists, applied statisticians must engage with substantive problems and criticisms (Ho and Rubin, 2011 ("To ground the assumptions [of causal inference], substantive knowledge and research are required."); Rubin, 2008 ("[N]o amount of fancy analysis can salvage an inadequate data base unless there is substantial scientific knowledge to support heroic assumptions.")). And Bross worried that doing so would raise the potential for conflicts in collaborative research settings. Administrators, for instance, may desire particular outcomes: "telling the truth can be very hazardous when it contradicts an administrator's view of things" (Bross, 1974, p. 127).

Put differently, tenability pushes one away from serving as a shoe clerk.

3.2 The Tenability of Tenability

It is unclear how tenability would operate across different contexts. What is the criterion by which the critic (e.g., Fisher) *should* weight observational data? In some areas, the observational data may be so limited that Fisher's dismissal of a body of evidence may in fact be warranted. The literature on the causal effect of legalized capital punishment on crime, for instance, is fraught with so many methodological challenges (e.g., highly nonrandom adoption of capital punishment and capital prosecution) that it is not obvious whether *any* observational design can ever replicate the hypothetical experiment (see Donohue and Wolfers, 2005). Can we give more content to tenability by formally incorporating prior knowledge about counterhypotheses? Bross implicitly did so when finding that the genetic hypothesis was untenable because of the rise of the male death rate.

And what does tenability require of a critic when the data may be unavailable to conduct alternative tests? Berkson published the relevant data on half a page of the *Journal of the American Statistical Association*, conveniently available for Bross's reanalysis (see Berkson, 1958, p. 34). When the underlying data is more complex and not publicly available, such reanalysis may not be as feasible. Given the rise of proprietary datasets in the age of "big data," critics may be less able to engage in the reasoned reanalysis and criticism that Bross espouses of such data.

Just as Ronald Fisher should not have dismissed the body of epidemiological data for want of randomization, the body of criticism should not be dismissed for want of reanalysis.

3.3 The Limits of Adversary Science

My third comment pertains to Bross's appeal to law when seeking a set of evidentiary rules. "Statistical Criticism" appeals to law in calling for evidentiary rules. The proponent has the "burden of proof" for a hypothesis. The critic has the burden to prove that a counterhypothesis is tenable. When a tenable counterhypothesis is shown, the proponent must show it to be wrong. Indeed, this process closely mirrors the legal framework in employment discrimination (disparate treatment) cases: the plaintiff must establish a prima facia claim of discrimination by preponderance of the evidence; the defendant then has the burden of rebutting the prima facia case; and the plaintiff prevails by showing that this rebuttal is wrong.[5]

In later work, Bross went further and argued that the solution for "rais[ing] the depressingly low standards of statistical criticism" lies in "adversary science" borrowed directly from the courtroom (aka "science court") (Bross, 1980, p. 37). Kantrowitz (1967) first proposed a science court, and credited Bross as providing one of several motivating proposals in the Interim Report of the Task Force for the Science Court experiment (Kantrowitz, 1977, pp. 332, 340). Kantrowitz articulated three guiding principles. First, there should be sharp separation of value judgments from judgments of scientific fact. Second, neutral, independent, scientific judges (with no prior work on the issue) would preside, with advocates presenting evidence on either side, with opportunity for cross-examination. Third, the court should issue a published decision on the state of scientific fact.

[5] McDonnell Douglas v. Green, 411 U.S. 792 (1973). The legal standard is more precise by specifying the standard of proof and by distinguishing burdens of proof and production.

While much has been written about the conceptual merits and challenges of science court (e.g., Aakhus, 1999; Bazelon, 1976; Burk, 1993; Martin, 1977; Matheny and Williams, 1981; Mazur, 1993), Bross's experience offers us a concrete sense of how well the institution might foster statistical dialogue. In 1978, Bross participated in an NRC public hearing that aimed to pilot the "science court" with the subject of health effects of low-level radiation. Bross tasked himself with "establish[ing] a . . . prima facie case that there are serious human health hazards from dosages of ionizing radiation in the range between 100 millirads and 10 rads" (American Chemical Society, 1978). Bross also articulated the task for Harvard epidemiologist Kenneth Rothman: "My opponent must take the position contrary to mine that there is no hazard" (id.). Echoing "Statistical Criticism," Bross stated, "It is not enough for him to argue that there might be questions or doubts . . . or that there alternative interpretations" (id.). In his view, the hearing — with a neutral chair, formal presentations, and a form of cross-examination — was a success, permitting a "clear public answer" to emerge (Bross, 1980, p. 37).

Yet while Bross wrote positively of his experience, contemporaneous reports suggest that the NRC hearing was a poor exemplar of science court. No NRC judges were actually present. (Two weeks before the hearing, the Supreme Court had struck down a lower court ruling urging the NRC to create more genuine dialogue on nuclear safety,[6] possibly explaining the lack of interest in this dialogue.) Rothman, whom the NRC had engaged to assess Bross's evidence, refused to engage in adversarialism. Perhaps he objected to Bross's charge to prove a negative (that "there is no hazard"). On his account, Rothman's role was to provide an independent, unbiased review. The lack of agreement on whether the hearing was adversarial epitomizes the normative clash of policy advocacy and scientific inquiry. To make matters worse, the NRC had actually engaged Rothman to specifically evaluate Bross's report, seeing Rothman's lack of expertise in radiation as a virtue. In that sense, Rothman, lacking subject matter expertise, was less peer reviewer / adversary than science court judge. While he agreed to some extent with Bross on potential dangers of radiation, he claimed that Bross had used the data twice: both to develop and test hypotheses. Rather than a proceeding about the broad evidence base about radiation risks, the hearing focused on a specific assessment of Bross's study and reanalysis. This was far from impersonal science. Concluded Rothman: "I cannot agree that his findings warrant any revision in our thinking about the health consequences of radiation exposure."

This NRC experience is consistent with how science courts, thrust into highly politicized issues, foundered. Matheny and Williams (1981) studied the proposal for a science court to resolve a power line dispute in Minnesota. Rather than mitigating conflict, the science court turned it into a "political hot potato" (p. 355). As many had feared, separating value judgments from scientific judgments proved challenging. Even when separated, Judge Bazelon feared that the science court would obscure the ultimate importance of value judgments (Bazelon, 1976). When used for delay, the proceeding itself became political. Asked about the science court proposal for licensing two nuclear power plants, a Con Edison representative complained about yet another barrier in the regulatory process. "We had five years of hearings . . . [It's] a PR kind of thing, and maybe a science court would help" but "it just adds another layer to what we have to deal with already" (Seagrave, 1976, p. 379-80).

[6] Vermont Yankee v. Natural Resources Defense Council, 435 U.S. 519 (1978).

Concern about unwieldy procedural requirements animated the Supreme Court to reverse the lower court's remand for more process in nuclear licensing. And even the lower court that mandated more process expressed deep trepidation about the suitability for quasi-judicial adversarialism at the agency level. "Factual issues in hybrid proceedings tend to be complex scientific or technical ones involving mathematical or experimental data peculiarly inappropriate for trial-type procedures."[7]

These factors ultimately contributed to science court's demise. "Like a sky rocket, [the science court] got a lot of attention as it ascended but just as quickly fell downward to crash and burn" (Mazur, 1993, p. 161).

4. Empirical Illustration: Adversary Science in Employment Discrimination

To further illustrate these points, we consider a common setting of statistical evidence offered by opposing experts in an employment discrimination suit. This setting has formal rules for the admissibility of evidence[8] and, as mentioned above, places burdens of proof that largely mirror Bross's proposed process. Statistical evidence often plays a large role in such cases and analyses and datasets are required to be submitted to opposing parties, with opportunity for deposition, testimony, and cross-examination.

The specific example comes from the University of Texas at Austin School of Law. In December 2011, Dean Lawrence Sager resigned, amidst allegations of improper use of a foundation to compensate faculty members, including claims of gender discrimination. While there was no formal litigation surrounding these claims, the allegations were covered widely in the media and the data are representative of the kind of case that could end up in trial.

Table 1 provides descriptive statistics for the dataset on 63 faculty members, compiled from public records. For simplicity of exposition and because such techniques rarely enter the courtroom, we do not consider more advanced, and arguably appropriate, methods for causal inference here (e.g., matching methods, panel techniques, causal intermediation). The conventional posture is that each side's expert offers statistical evidence, most commonly linear regression models. Typical debates are about the sample definition, measurement, and the proper specification, each potentially suggesting or contradicting discrimination.

The first row of Table 1 shows a statistically significant salary difference between male and female faculty members. Male faculty members earn, on average, \$35k more than female faculty members (p-value = 0.02). Potential covariates are listed below the outcomes in Table 1, suggesting considerable differences along gender lines. (As we will see below, whether they are true covariates (i.e., unaffected by the treatment) depends on the substantive theory of the case.) Men have been teaching on average eight years longer, reflecting the diversification of the legal profession over the past few decades (Chused, 1988). Women, on the other hand, are more likely to have held positions as clerks to federal judges. Because a common concern in estimating gender discrimination is about "productivity," we aug-

[7] Natural Resources Defense Council v. Nuclear Regulatory Commission, 547 F.2d 633, 656 (D.C. Cir. 1976).

[8] See Daubert v. Merrell Dow Pharmaceuticals, 509 U.S. 579 (1993).

Outcomes	Male mean	Female mean	Difference	p-value
Salary ($1000s)	266.52	231.19	35.32	0.02
Forgivable loan ($1000s)	71.74	8.82	62.92	0.00
Covariates				
Articles	29.17	21.06	8.12	0.15
Years teaching	26.02	17.76	8.26	0.03
Endowed chair	0.61	0.35	0.26	0.08
Federal clerkship	0.35	0.71	-0.36	0.01
Doctoral degree	0.13	0.18	-0.05	0.67
n	46	17		

Table 1: Descriptive statistics for University of Texas at Austin School of Law faculty dataset. The first column represents the mean for male faculty members; the second column represents the mean for female faculty members; the third column represents the gender difference; and the fourth column presents the p-value from a t-test. Salary and forgivable loans are in $1000s, and salary is the twelve-month salary, excluding forgivable loans. n represents the number of observations.

mented this dataset with counts of the number of articles published by the faculty member from publicly available CVs. Male faculty members have written eight more articles, on average, but the difference is not statistically significant.

Imagine that the plaintiff's expert introduces a regression of salary against gender and articles. Figure 2 plots the data on articles published (logged to adjust for skewness) on the x-axis and salary on the y-axis. The lines present simple fits from the regression model, with gender (coded as 1 if male and 0 if female) and articles (logged) as predictors, with 95% confidence intervals. Red (blue) colors correspond to female (male) observations. The gender coefficient remains statistically significant, and the expert may conclude that even controlling for productivity, women are underpaid, corroborating an inference of discrimination.

Of course, many specifications, even with such a sparse covariate set, are possible. The defendant's expert may counter that this regression fails to adjust for other confounding factors (e.g., years in teaching), offering the second regression in Table 2. The gender difference is no longer statistically significant. In a narrow sense, the defendant's expert has carried out the responsibility of proving the tenability of her counterhypothesis: the gender difference in Model (1) may simply be an artifact of academic rank.

Yet in a deeper sense, the statistician is necessarily drawn into the local ground rules. Just as Bross made coding decisions of diseases based on substantive grounds, the statistician must make substantive decisions in specifying the model. Perhaps the very mechanism by which the University of Texas is discriminating against women is by awarding endowed chairs only to men, so that controlling for such covariates introduces post-treatment bias (Rosenbaum, 1984). The only way to explore this hypothesis further – particularly if the question is about gender discrimination by the dean – is to understand the substantive mechanism. Does the dean in fact exercise discretion in endowment decisions or is a committee

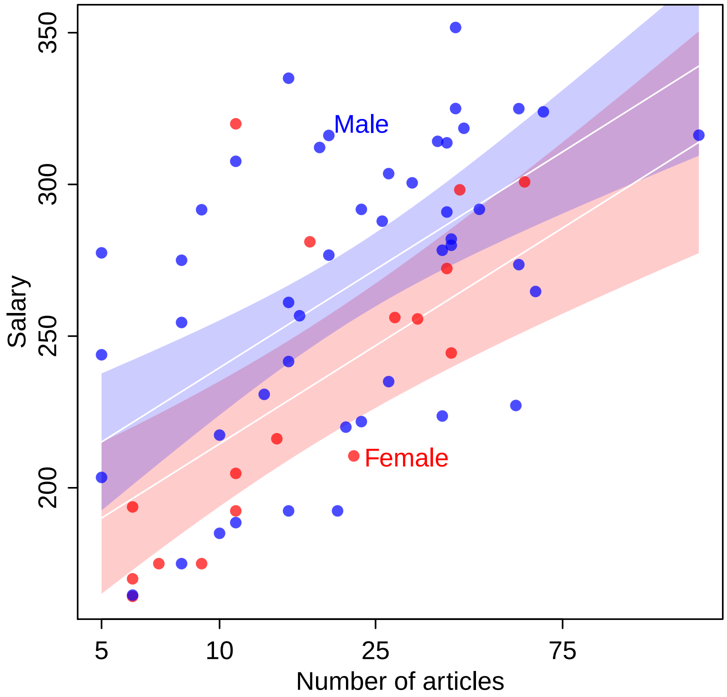

Figure 2: Correlation between number of articles (on a log scale due to skewness) and salary (in $1,000s). Red colors correspond to female faculty members, and blue colors correspond to male faculty members. 95% confidence bands from regression model are overlaid.

responsible, which might seek external letters for the promotion decision? The answer is informed at core by an understanding of the statistics of causal inference, but lies, in a sense, beyond statistics alone (Ho and Kramer, 2013; Ho and Rubin, 2011).

Consider now Models (3) and (4) of Table 2. Recall that one of the concerns that emerged was about Sager's use of a foundation, separate from the University of Texas system and not reported to the Regents, to recruit, retain, and compensate faculty members. The principal vehicle for faculty compensation constituted loans to be forgiven over a number of years. The models present the same specifications for the outcome of forgivable loan amount. Here we see that the difference increases from $64k to $91k when covariates are added and remains statistically significant in both models, suggesting that the plaintiff may have a stronger case on this outcome dimension.

To understand whether this evidence is consistent with discrimination again requires substantive knowledge. One theory of the case is that the dean used the foundation to respond specifically to outside offers to faculty (i.e., retention). On the idea that responding to such outside offers is not indicative of bias, Bross might advocate conducting separate analyses for cases of retention and recruitment. If forgivable loans are indeed merely responses to outside offers, the evidence in Models (3) and (4) may have nothing to do with

	Salary		**Forgivable loan**	
	(1)	(2)	(3)	(4)
Male	25.09**	12.70	64.07**	91.07***
	(11.67)	(9.63)	(30.91)	(27.11)
log(Articles)	35.31***	16.95***	−3.97	5.10
	(6.58)	(6.21)	(17.43)	(17.49)
Years teaching		−0.74*		−5.75***
		(0.42)		(1.17)
Endowed chair		101.57***		117.20**
		(17.11)		(48.16)
Full Professor		52.59***		132.98***
		(15.09)		(42.49)
Doctoral degree		−7.28		26.63
		(11.93)		(33.59)
Federal clerkship		−8.73		−9.52
		(9.56)		(26.90)
Constant	133.19***	147.17***	19.85	2.79
	(20.74)	(19.06)	(54.94)	(53.65)
Observations	63	63	63	63
R^2	0.39	0.67	0.07	0.43

Table 2: Linear least squares regression estimates of gender discrimination. Models (1) and (2) are for the outcome of 12-month salary in \$1000s. Models (3) and (4) or for the outcome of forgivable loans in \$1000s. Coefficient estimates are presented with standard errors in parentheses. The salary models show sensitivity to the covariate set, with a statistically significant gender difference in Model (1) becoming statistically insignificant in Model (2). In the forgivable loan models, the gender difference magnifies as more covariates are added. Whether the predictors are true "covariates" in the sense of being unaffected by the treatment of gender depends on a substantive understanding of the theory of discrimination. $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

discrimination by the dean per se. Instead, they may indicate either (a) that male faculty are more likely to seek opportunities for lateral movement, or (b) that the real parties engaging in discrimination are *other* schools by systematically making more lateral offers to male faculty.

The application also illustrates the difficult position generalist judges and juries are placed in by dueling expert reports. Precisely because experts are paid by adversarial parties, the testimony will be skewed to that side and experts may have an incentive to obscure rather than elucidate substantively important assumptions (Greiner and Rubin, 2011). Judge Richard Posner, a leading proponent of law and economics who has published econometric work, concluded, “Econometrics is such a difficult subject that it is unrealistic to expect the average judge or juror to be able to understand all the criticisms” (Posner, 1999). Yet how is a generalist judge to decide technical claims (e.g., about model fit, clustering of standard errors, principal stratification)? Even given the formal burden of proof, is it clear when the critics case has become “tenable” when the statistical evidence requires local ground knowledge? Under what conditions does such adversarial science outperform conventional statistical inquiry? Both the legal system and statistical science struggle with how to judge such statistical criticism, and the employment discrimination setting does not provide strong support for exporting adversarial science.

5. Conclusion

It’s been a pleasure to have the opportunity to reflect on Bross’s “Statistical Criticism.” The article remains as relevant now as it was back then, raising profound questions about who judges whether rules of statistical criticism were adhered to.

While the science court per se may be flawed, contrasting the judicial and peer scientific models can be valuable to developing other possibilities for institutional reform. In the court room, importing scientific neutrality by greater use of court-appointed experts may make it easier for judges and juries to incorporate complex statistical evidence (Cecil and Willging, 1994).[9] In administrative agencies, importing peer review practices may improve the reliability and scientific judgment of regulatory enforcement (Ho, 2017).

On the flipside, in some circumstances, exporting elements of adversarialism might benefit statistical learning. Professional journals may incentivize higher quality science by (a) providing reviewers with datasets and replication code (just as courts mandate of experts) and (b) potential publication of discussion from reviewers to air out differences in analysis and interpretation (just as courts require opposing expert reports to be disclosed). To the extent that science court proponents were worried about representing the full range of opinions, scientific advisory committees might encourage members to write separately where their judgment of the tenability of a counterhypothesis diverges from the committee report. Scientific consensus building may benefit from forms of “adversarial collaboration,” espoused by Daniel Kahneman: joint research by parties to resolve a debate, potentially with a neutral arbiter as part of the research team (Kahneman, 2003, pp. 729-30).

If properly designed and evaluated, these reforms would fall short of a full-blown science court, but could improve institutions for judging statistical criticism.

[9] To be sure, there are criticisms of such techniques, as wresting power from litigants and importing a foreign inquisitorial technique (see Deason, 1998). Yet the alternative may be the mess Bross bemoaned.

Late in his career, Bross struggled with that broader question. No doubt, he was feeling embattled by the criticism of his claims for low-level radiation risk. In the *American Journal of Public Health* (Bross, 1979), he decried:

> Are the conflicts-of-interest hypothetical or real? The extraordinary editorial handling of our paper provides factual evidence on this point. The hostility of the editor to our findings is evidenced by the gratuitous comment in his introductory note: 'Dr. Bross stands virtually alone in his defense of his data.'

As critiques of his conclusions mounted (see, e.g., Boice and Land, 1979; Oppenheim, 1977; Rao, 1978), drawing Bross further into the substantive territory he had earlier warned of, he grew increasingly skeptical of the peer review system as enforcing the rules of criticism. Peers can of course have dramatically divergent standards of tenability (Simon et al., 1981), but Bross went further and charged that agencies manipulated peer review in low-level radiation "to suppress, vilify, or cut off the funding of the little scientists" (Walker, 2000, p. 95). Responding to the inability of peer review to avert the Summerlin scandal — where William Summerlin had faked tissue-culture skin transplants at the Sloan-Kettering Institute — Bross famously wrote in the *New York Review of Books*: "Big science is bad science."[10] Finding himself in the minority view on radiation risks may also have bolstered Bross's enthusiasm for adversarial science. But the judicial system itself, from which he had so heavily borrowed in proposing adversarial science, rebuffed him too. The Second Circuit affirmed a dismissal of Bross's suit against the Veterans Administration pertaining to his radiation research: "the interest of a scientist like Dr. Bross in seeking professional and governmental recognition of his views, although unquestionably genuine, [does not] fall within the [law's] zone of interest."[11]

Acknowledgments

Thanks to Dylan Small for inviting this commentary, to Zoe Ashood, Sandy Nader, Sam Sherman, and Dylan Small for comments, and to Carlos Llerena Aguirre for permission to reprint the Science Court image.

[10] Irwin D.J. Bross, A Better Mouse Trap, N.Y. Rev. Books, June 10, 1976.

[11] Bross v. Turnage, 889 F.2d 1256, 1257 (2d Cir. 1989) (the specific law was the Veterans' Dioxin and Radiation Exposure Compensation Standards Act).

References

Aakhus, M. (1999). Science court: A case study in designing discourse to manage policy controversy. *Knowledge, Technology & Policy*, 12(2):20–37.

American Chemical Society (1978). Nrc sponsors low-level radiation hazard debate. *Chemical & Engineering News*, 56(16):14. Available from: http://dx.doi.org/10.1021/cen-v056n016.p014.

Bazelon, D. L. (1976). Coping with technology through the legal process. *Cornell Law Review*, 62:817–832.

Berkson, J. (1958). Smoking and lung cancer: Some observations on two recent reports. *Journal of the American Statistical Association*, 53(281):28–38.

Boice, J. D. and Land, C. E. (1979). Adult leukemia following diagnostic x-rays? (review of report by bross, ball, and falen on a tri-state leukemia survey). *American Journal of Public Health*, 69(2):137–145.

Bross, I. D. (1960). Statistical criticism. *Cancer*, 13(2):394–400.

Bross, I. D. (1974). The role of the statistician: Scientist or shoe clerk. *The American Statistician*, 28(4):126–127.

Bross, I. D. (1979). Protection of the public health against radiation hazards. *American Journal of Public Health*, 69(6):609–610.

Bross, I. D. (1980). When speaking to washington, tell the truth, the whole truth, and nothing but the truth, and do so intelligibly. *The American Statistician*, 34(1):34–38.

Burk, D. L. (1993). When scientists act like lawyers: The problem of adversary science. *Jurimetrics*, 33(3):363–376.

Cecil, J. S. and Willging, T. E. (1994). Court-appointed experts. *Reference Manual on Scientific Evidence*, 527–573.

Chused, R. H. (1988). The hiring and retention of minorities and women on american law school faculties. *University of Pennsylvania Law Review*, 137(2):537–569. Available from: http://www.jstor.org/stable/3312253.

Cook, T. D. (2015). The inheritance bequeathed to william g. cochran that he willed forward and left for others to will forward again: The limits of observational studies that seek to mimic randomized experiments. *Observational Studies*, 1:141–164.

Critchley, J. A. and Capewell, S. (2003). Mortality risk reduction associated with smoking cessation in patients with coronary heart disease: A systematic review. *JAMA*, 290(1):86–97.

Deason, E. E. (1998). Court-appointed expert witnesses: Scientific positivism meets bias and deference. *Orregon Law Review*, 77:59–156.

Doll, R. and Hill, A. B. (1956). Lung cancer and other causes of death in relation to smoking. *British Medical Journal*, 2(5001):1071.

Donohue, J. J. and Wolfers, J. (2005). Uses and abuses of empirical evidence in the death penalty debate. *Stanford Law Review*, 58(3):791–846.

Greiner, D. J. and Rubin, D. B. (2011). Causal effects of perceived immutable characteristics. *Review of Economics and Statistics*, 93(3):775–785.

Ho, D. E. (2017). Does peer review work? an experiment of experimentalism. *Stanford Law Review*, 69:1–119.

Ho, D. E. and Kramer, L. (2013). The empirical revolution in law. *Stanford Law Review*, 65:1195–1202.

Ho, D. E. and Rubin, D. B. (2011). Credible causal inference for empirical legal studies. *Annual Review of Law and Social Science*, 7:17–40.

Kahan, D. M., Hoffman, D. A., Braman, D., and Evans, D. (2012). They saw a protest: Cognitive illiberalism and the speech-conduct distinction. *Stanford Law Review*, 64:851–906.

Kahneman, D. (2003). Experiences of collaborative research. *American Psychologist*, 58(9):723–730.

Kantrowitz, A. (1967). Proposal for an institution for scientific judgment. *Science*, 156(3776):763–764.

Kantrowitz, A. (1977). The science court experiment. *Jurimetrics Journal*, 17(4):332–341.

Leeper, E. (1976). Science court 'tried,' cleared for test case. *BioScience*, pages 717–719.

Martin, J. A. (1977). The proposed 'science court'. *Michigan Law Review*, 75:1058–1091.

Matheny, A. R. and Williams, B. A. (1981). Scientific disputes and adversary procedures in policy-making: An evaluation of the science court. *Law & Policy*, 3(3):341–364.

Mazur, A. (1973). Disputes between experts. *Minerva*, 11(2):243–262.

Mazur, A. (1993). The science court: Reminiscence and retrospective. *Risk*, 4:161.

Oppenheim, B. E. (1977). Genetic damage from diagnostic radiation. *JAMA*, 238(10):1024–1025.

Posner, R. A. (1999). The law and economics of the economic expert witness. *The Journal of Economic Perspectives*, 13(2):91–99.

Rao, P. (1978). Genetic damage from diagnostic radiation. *Investigative Radiology*, 13(1):100–101.

Rosenbaum, P. R. (1984). The consequences of adjustment for a concomitant variable that has been affected by the treatment. *Journal of the Royal Statistical Society. Series A (General)*, 147(5):656–666.

Rubin, D. B. (2008). For objective causal inference, design trumps analysis. *The Annals of Applied Statistics*, 2(3):808–840.

Seagrave, S. (1976). The ultimate in peer review: Science court: Test case this year? *BioScience*, pages 377–380.

Simon, G. A., Cole, J., and Cole, S. (1981). Chance and consensus in peer review. *Science*, 214(4523):881–886.

Walker, J. S. (2000). *Permissible Dose: A History of Radiation Protection in the Twentieth Century*. University of California Press.

Observational Studies 4 (2018) 57-60 Submitted 5/17; Published 1/18

Another Ground Rule

Charles S. Reichardt **Chip.Reichardt@du.edu**
Department of Psychology
University of Denver
Denver, CO 80208 USA

Bross (1960, p. 394) wrote, "If both proponents and critics have to watch their P's and Q's, we might hope that it would be easier to achieve broad agreement on scientific issues." Bross then went on to offer a ground rule (i.e., one of the P's and Q's) explicitly for critics of research hypotheses, though he emphasized (pp. 399-400) the "same ground rules should apply to both" proponents and critics. I have two purposes. First, to quibble with Bross about his ground rule. And second, to propose another ground rule.

Bross argues that alternative explanations should be judged tenable before they are allowed to see the light of day. And for an alternative explanation to be judged tenable, it must agree with available data. Certainly this is correct. Except that agreement with available data is not an infallible indicator of the adequacy of an explanation.

Consider Darwinian evolution. Significant data argued against Darwins theory at the time it was proposed (Bryson, 2003). For example, the best available evidence was that the earth was far too young, even according to Darwin's own account, for species to have evolved per natural selection. And the fossil record was too sparse providing too little evidence of the intermediate life forms that Darwin required. Plus Darwin's theory was at odds with well accepted contemporary thinking. Even T. X. Huxley, who was one of Darwin's staunchest supporters, believed Darwin was wrong about the rapidity with which evolution took place. Huxley, a saltationist, believed evolution happened suddenly rather than gradually. So was Darwin's alternative explanation for evolution in sufficient agreement with available data, and therefore sufficiently tenable, to permit publication according to Bross?

Or consider the work of Ingaz Semmelweis. When Ingaz Semmelweis was hired as a physician at the Vienna General Hospital in 1846, as many as twenty percent of the women giving childbirth in the hospital's First Obstetrical Clinic died from puerperal fever, also known as childbed fever. Semmelweis was determined to discover the cause. He uncovered a telling clue when a colleague cut his finger while performing an autopsy and died from symptoms similar to puerperal fever. Based on that evidence, Semmelweis hypothesized that the disease was caused by contamination from "cadaverous material" and found his hypothesis could explain another mystery. The First Obstetric Clinic at Vienna General had a much higher death rate from puerperal fever than the Second Clinic. The difference? The First Clinic was attended by physicians who often performed autopsies before serving on the obstetrics ward; the Second Clinic was attended by mid-wives, who did not perform autopsies. Acting on his hypothesis, Semmelweis instituted a policy that physicians wash their hands in a solution of chlorinated lime before examining patients. Occurrences of

©2018 Charles Reichardt.

puerperal fever declined dramatically. In April 1847, before the policy of hand washing was instituted, eighteen percent of the patients in the First Clinic died from puerperal fever. A month later, after hand washing was implemented, the mortality rate dropped to two percent. The same outcomes were obtained whenever hand washing was implemented by either Semmelweis or his students.

The results of Semmelweis' hand-washing experiments became widely known. But the findings ran counter to the conventional medical beliefs current at the time. Puerperal fever was thought to be due to multiple causes including effluviums that were thought to be spread through the air. In addition, the causes of illnesses were thought to be as unique as individuals themselves and determinable only on a case by case basis. As a result, Semmelweis' unconventional proposal that a simple cause such as a lack of cleanliness could be responsible for puerperal fever was rejected out of hand by the medical profession. Indeed, physicians believed their social status precluded them from having unclean hands. In spite of Semmelweis' demonstrations of the effectiveness of hand washing in reducing puerperal fever, most doctors did not adopt the practice until the germ theory of disease was advanced by Pasteur and Lister decades later.

So Semmelweis' alternative explanation for the cause of puerperal fever was not tenable in the eyes of the proponents of the current theories. Of course, Bross would have likely responded that Semmelweis' alternative explanation was highly tenable, given Semmelweis' data, and was rejected only because of what he calls tubular vision. Nonetheless, the acceptance of Semmelweis' theory was delayed because, for whatever reasons, it was judged unacceptable at the time and thousands subsequently died unnecessarily. Would Bross' ground rule have aided and abetted the rejection of Semmelweis' conclusions?

Bross also notes that he allows speculation in a research report, just not as part of a conclusion section. Because Einstein's general theory of relativity had little data in its support at the time it was published, it was a highly speculative conclusion. Indeed, evidence that light bends in the presence of gravity, as predicted by the theory, was not provided for four years after Einstein's theory was published. And one of the theory's main tenets, the existence of gravity waves, was not confirmed until 100 years later. So would Bross have stood in the way of publishing that largely unsupported speculation?

The point is the following. Bross is correct that those who propose alternative explanations, whether proponent or critic, should be required to check their explanations against available data. But data is always incomplete and can be all too easily misinterpreted. So judging what is and is not in agreement with data can be tricky. Assessing tenability is an inexact science. As a result, researchers and editors must be circumspect in applying Bross' rule.

I might note one other caution with regard to Bross' rule. To be tenable, any one alternative explanation need not, by itself, account for all of the observed results. I've seen instances where an estimate of a treatment effect, for example, is said to be immune to a rival hypothesis because the alternative explanation was insufficient to account for the entirety of the estimate. But more than one bias can be present. And perhaps together they could account for the whole treatment effect estimate. So the tenability of alternative explanations needs to be considered en masse rather than one at a time.

Now let me consider my second purpose which is to provide another ground rule – especially for proponents, rather than critics, of explanations. Like Bross (1960), Doll and

Hill (1952) were concerned with the theory that smoking causes lung cancer. And Doll and Hill presented data that well supported that theory. The critical point, for my presentation, is that Doll and Hill included a section of their article entitled "Validity of the Results." Cochran (1965, pp. 252-253) explains the importance of such a section:

> When summarizing the results of a study that shows an association consistent with the causal hypothesis, the investigator should always list and discuss all alternative explanations of his [sic] results (including different hypotheses and biases in the results) that occur to him. This advice may sound trite, but in practice is often neglected. A model is the section "Validity of the results" by Doll and Hill (1952), in which they present and discuss six alternative explanations of their results in a study.

Certainly it is well accepted by now that proponents of scientific hypotheses should openly report weaknesses of their research and alternative explanations for their findings. For example, the American Psychological Association's (2010, p. 35) widely used Publication Manual states:

> Your interpretation of the results should take into account (a) sources of potential bias and other threats to internal validity, (b) the imprecision of measures, (c) the overall number of tests or overlap among tests, (d) the effect sizes observed, and (e) other limitations or weaknesses of the study.

But according to the APA Manual, such an accounting of limitations and weaknesses is to be placed in the discussion section of an article. In contrast, I believe, as Cochran recommends, that such critical reflections on a research study deserve their own dedicated section, as in Doll and Hill.

And not only that, reviewers and editors should insist that such a section goes to the heart of limitations and weaknesses in a research study rather than just skimming the surface. I've read reports where the limitations of a study include such obvious reflections as that the results should not be generalized beyond the population of participants and the outcome measures that were used. But the same reports ignore warnings of much more insidious concerns such as omitted variables and hidden biases. Perhaps the authors are unaware of such difficulties. If so, that is all the more reason for reviewers and editors to insist they be acknowledged forthrightly. Or perhaps the authors are aware but afraid that acknowledging such weaknesses would disqualify their research from publication. If so, reviewers and editors must explicitly disavow such disqualification when the research is otherwise of high quality – because such weaknesses are simply an inherent feature in some realms of research, such as observational studies. Even sensitivity analyses are not guaranteed to bracket the true sizes of treatment effects.

Are we doing what we should to encourage substantive researchers to abide by the practice endorsed by Cochran – one of the founders of statistical methods in observational studies? Honesty and integrity in science require that researchers be their own worst critics. We should insist that a discussion of limitations and weaknesses be the whole truth and that it appears in its own section and not as a brief mention of superficial weaknesses hidden amid substantive summarizing in a discussion section.

Acknowledgments

Thanks to Kris Preacher, Bill Revelle, and Anne Strobridge for their helpful comments on the manuscript.

References

American Psychological Association (2010). *Publication Manual* (6th edition), Washington, DC: American Psychological Association.

Bross, I. D. J. (1960). Statistical criticism. *Cancer*, 13, 394-400.

Bryson, B. (2003). *A Short History of Nearly Everything.* NY: Broadway Books.

Cochran W. G. (1965). The planning of observational studies of human populations (with discussion). *Journal of the Royal Statistical Society, Series A*, 128, 234-266.

Observational Studies 4 (2018) 61-64 Submitted 12/17; Published 1/18

Statistical Criticism, Self-Criticism and the Scientific Method

David Rindskopf **drindskopf@gc.cuny.edu**
CUNY Graduate Center
New York, NY 10016, U.S.A.

I have long admired Bross's article on statistical criticism, and in my mind it has much broader implications than those Bross chose to present. In fact, others have discussed much the same points in the context of all of scientific methodology.

I also hope that Bross would say that his rules should be applied to self-criticism as well as criticism by others. There would be less need for criticism by others if there were better self-criticism in the first place. Cochran (1965) is cited by Rosenbaum:

> When summarizing the results of a study that shows an association consistent with the causal hypothesis, the investigator should always list and discuss all alternative explanations of his results (including different hypotheses and biases in the results) that occur to him.

Such advice is valuable, but alas mere mortals (including me) are usually deficient in self-criticism. What sometimes helps is to put a piece away for a while after writing it, and then going back specifically to criticize it before putting it on display for others to criticize.

In spite of the difficulty of self-criticism, the author is best placed to criticize, having full access to all the data of the study. As data sets more often become publicly available, this advantage will diminish. Even so, many data sets have restricted access and in those cases it will remain difficult for critics to test alternative theories.

Chamberlain (1890, reprinted 1964) and Platt (1965) have similar views as Bross, but applied more broadly to scientific methods as a whole.

Chamberlain discussed the affection a scientist feels for his or her ideas, and how unconscious bias ("tubular vision" in Bross's terms) takes over:

> As soon as this parental affection takes possession of the mind, there is a rapid passage to the adoption of the theory. There is an unconscious selection and magnifying of the phenomena that fall into harmony with the theory and support it, and an unconscious neglect of those that fail of coincidence. The mind lingers with pleasure upon the facts that fall happily into the embrace of the theory, and feels a natural coldness toward those that seem refractory. Instinctively there is a special searching- out of phenomena that support it, for the mind is led by its desires. There springs up, also, an unconscious pressing of the theory to make it fit the facts, and a pressing of the facts to make them fit the theory. When these biasing tendencies set in, the mind rapidly degenerates into the partiality of paternalism. The search for facts, the observation of phenomena and their interpretation, are all dominated by affection for the

©2018 David Rindskopf.

> favored theory until it appears to its author or its advocate to have been overwhelmingly established. The theory then rapidly rises to the ruling position, and investigation, observation, and interpretation are controlled and directed by it. From an unduly favored child, it readily becomes master, and leads its author whithersoever it will. The subsequent history of that mind in respect to that theme is but the progressive dominance of a ruling idea.

Chamberlain suggests that instead researchers should develop alternative theories so that they do not maintain an allegiance to any one of them.

Platt (1964) thought that the best way to make progress in science was to eliminate alternative theories with each experiment. In that context, he said:

> In its separate elements, strong inference is just the simple and old-fashioned method of inductive inference that goes back to Francis Bacon. The steps are familiar to every college student and are practiced, off and on, by every scientist. The difference comes in their systematic application. Strong inference consists of applying the following steps to every problem in science, formally and explicitly and regularly:
>
> 1) Devising alternative hypotheses;
>
> 2) Devising a crucial experiment (or several of them), with alternative possible outcomes, each of which will, as nearly as possible, exclude one or more of the hypotheses;
>
> 3) Carrying out the experiment so as to get a clean result;
>
> 1') Recycling the procedure, making subhypotheses or sequential hypotheses to refine the possibilities that remain; and so on.

In the case of the statistician, the alternative hypotheses are alternative artifacts that might plausibly account for the findings. In parallel with Platt's reasoning for experiments, the statistician should control the most plausible alternatives first.

Platt cites one paper that contains the following line of reasoning:

> Our conclusions...might be invalid if...(i)...(ii)...or (iii)...We shall describe experiments which eliminate these alternatives.

Statisticians can use the same method, substituting "analyses" for "experiments".

Note that Platt and Chamberlain's views contradict the way science is commonly taught. Students are often taught to develop a theory or hypothesis, and then test it. They are not taught to think of alternative hypotheses until it is too late (when their experiment has failed); or in case it is not contradicted, they believe their hypothesis was confirmed, when, in fact, plausible rival hypotheses may have made the same prediction.

Donald T. Campbell often discussed plausible rival hypotheses, even devising checklists of general categories of these that one should consider in evaluating causal conclusions from a study. In the first such instance I can find, Campbell and Stanley, 1963, discussing pre-test post-test design said:

> Between O_1 [observation at time 1] and O_2 [observation at time 2] many other change-producing events may have occurred in addition to the experimenter's X [treatment]. If the pretest (O_1) and the posttest (O_2) are made on different days, then the events between may have caused the difference. To become a ***plausible*** **rival hypothesis** [italics in original], such an event should have occurred to most of the students in the group under study, say in some other class period or via a widely disseminated news story.

Note that Campbell and Stanley did not say one should defend against all rival hypotheses. They, like Bross, saw that as too broad; only plausible rival hypotheses need be checked. This is more reasonable than requiring all (an infinite number, and therefore rather difficult to execute) to be checked, but does leave some wiggle room: What is plausible?

Campbell and Stanley (1963) also stated:

> In 1923, W. A. McCall published a book entitled *How to Experiment in Education.* The present chapter aspires to achieve an up-to-date representation of the interests and considerations of that book, and for this reason will begin with an appreciation of it. In his preface McCall said: "There are excellent books and courses of instruction dealing with the statistical manipulation of experimental data, but there is little help to be found on the methods of securing adequate and proper data to which to apply statistical procedure."

Campbell and Stanley emphasized the phrase "securing adequate and proper data" in discussing this quote. This suggests a possible (sometimes probable) reason why critics could have problems making a case against a proponent: They don't have the data that would be adequate or proper to back their claims. The proponent has much more control over the data; the critic always has less control, and sometimes has little or none.

This suggests that Bross's requirement for critics may sometimes be too strict. What data are available for the critic to investigate the plausible rival hypotheses? Are they sufficient to address the necessary issues? If so, were they used correctly by the critic? If not, did the critic spell out what data would be necessary to confirm the counterhypothesis?

Rosenbaum (2002) presented an interesting quote from Fisher, who stated that criticism should not be accepted merely because it is an authority making it; an example he gives is "His controls are totally inadequate", without any elaboration of what adequate controls might be. This is interesting because Bross cites Fisher for making a similar statement in the arguments about smoking and cancer.

Bross provided a valuable service to the field by suggesting that there should be standards for criticism. We might argue about exactly how those standards should be applied, and whether they can be as strictly adhered to as he would like, but requiring critics to do more than sling random criticisms without some backing for their statements was, and still is, a reasonable standard to meet.

References

Bross, I. D. J. (1960). Statistical criticism. *Cancer*, 13, 394-400.

Chamberlin, T. C. (1890). The method of multiple working hypotheses. *Science* 15:9296. (reprinted in *Science* 148: 754759 [1965]).

Cochran, W. G. (1965) The planning of observational studies of human populations (with discussion). *Journal of the Royal Statistical Society, Series A*, 128, 134-155.

Platt, J. R. 1964. Strong inference. *Science* 146:347353.

Rosenbaum, P.R. (2002). *Observational Studies*, 2nd Edition, Springer-Verlag, New York.

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

©2018 Paul Rosenbaum and Dylan Small.

> 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 logically 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.