Bulletin of the Psychonomic Society
1984, 22 (5), 463-466

Antinonrobustness: A case study in the sociology of science

JAMES V. BRADLEY
New Mexico State University, Las Cruces, New Mexico

A quarter-century ago, during a period when belief in the robustness of classical tests on means was practically a professional shibboleth, a series of large, carefully controlled, and well-validated experiments and sampling studies (supplemented and supported by extensive mathematical derivations) dramatically showed that highly publicized claims of robustness were insufficiently qualified and that extreme nonrobustness could occur under perfectly reasonable experimental and testing conditions. When these findings were published in technical reports, they tended either to be ignored or to be so misrepresented and distorted by those who cited them as to make them appear to support, rather than question, the very claims of robustness they tended to refute. Attempts to publish these iconoclastic results in many of the most renowned professional journals were met with rejection based upon reviewer comments so illogical or fatuous as to be interpreted only as an indication of either contrived obstructionism or pathetic professional incompetence. Eventual acceptance by a few refereed journals could, in every case, be interpreted as a political-type fluke.

Bad news is seldom graciously received. In ancient times, its messenger risked his life by announcing it. Today we are more civilized and attempt to stifle the announcement rather than the announcer. This article concerns obstructionism encountered in attempts to publicize factual data casting doubts upon two widely held beliefs—that extremely nonnormal distributions are rare or nonexistent and that the classical statistical tests on means (Z, t, and F) are extremely insensitive to violations of their assumptions of normal distributions with equal variances, that is, are "robust" against violations of those assumptions.

NONNORMALITY

Extreme Nonnormality Simultaneously and Independently Encountered in Adjacent, Routine Experiments by Myself and a Colleague

A quarter-century ago, while performing a series of experiments in engineering psychology, I became increasingly nervous about the validity of the normality assumption that I was required to make in analyzing the data. To resolve my anxieties, I selected a single experimental condition from my most recent experiment, administered 2,520 trials to a single subject under this single condition, and plotted the frequency distribution of his scores. To my great surprise, the resulting distribution was far more skewed than any I had ever seen depicted in statistical textbooks. Using a different experimental condition from the same experiment, I immediately obtained another distribution of 2,520 scores. This time the distribution was even more skewed and was decidedly L-shaped. A little reflection (as well as more formal analytic considerations) revealed that the obtained distribution shapes were a logical consequence of the subject's task and, indeed, of circumstances characterizing a broad class of experiments (see Bradley 1975, 1977, 1982b, 1983). Furthermore, as soon as I announced my findings to my laboratory colleagues, one of them, Harry Jerison, showed me even more L-shaped distributions (Figure D in Bradley, 1977) that he had simultaneously been obtaining in an entirely different area of research while running subjects in an adjacent laboratory cubicle—a most unlikely coincidence if such events are highly improbable. Clearly, therefore, these L-shaped distributions could hardly be dismissed as rare oddities. (Subsequently, a number of colleagues have reported similar distributions, and I have encountered others in the literature.)

The author's mailing address is: Department of Psychology, New Mexico State University, Box 3452, Las Cruces, NM 88003.

NONROBUSTNESS

Consequent Nonrobustness

Since the normality assumption had obviously been badly violated, the question arose as to how badly that violation affected the validity of normality-assuming statistical tests. To answer this question, I performed a series of empirical sampling studies to determine the robustness of classical normality-assuming tests when sampling from populations having roughly the same L-shape as the second distribution I had obtained. These studies showed the classical normality-assuming Z, t, and F tests on means to be far less robust than I had been led to believe, and under a variety of perfectly reasonable sampling and testing conditions they seemed grossly nonrobust by any sensible criterion of robustness (Bradley 1968, 1980a, 1980b, 1980c, 1984).

Validity of Results

Since my results were decidedly iconoclastic, a question arises as to their accuracy. My empirical sampling studies were based upon sampling distributions of 10,000

or more (sometimes as many as 150,000) values of the test statistic, and the samples upon which the test statistic was based could be any of a variety of sizes, ranging from 2 to 4,096 observations. Thus, the studies were far "larger" than was customary at the time, and, furthermore, they were extremely "well-controlled," that is, reliable, accurate, carefully checked, and verified; and several of them have been replicated, always successfully, by myself (Bradley, 1968) or others (Wike & Church, 1982).

Transformability of Data to Normality

Another question raised by the studies is whether or not the L-shaped distribution can be converted to quasi-normality by applying the proper transformation of scores. I attempted this without success just after the earliest studies were done, using all of the common transformations, and more recently, Wike and Church (1982) and I (Bradley, 1982b), independently and more formally, have investigated this possibility, again without success.

ANTINONROBUSTNESS

Initial Publication

At first I published my studies in government technical reports [now readily available only from University Microfilms, (Bradley, 1968)]. However, since my studies seemed irreproachable in technique and yet suggested conclusions about robustness quite different from the prevailing beliefs (based upon far smaller and less well-controlled studies), I felt certain that I had the makings of an article on robustness that would surely be accepted by any of the best statistical or methodological journals. Accordingly, I began a 15-year career of attempting to publish my results in refereed journals and experienced a simultaneous education in the sociology of science. Up to this time, I had never failed to have an article accepted by the journal to which it was originally submitted. In my naivete, I was quite sure that my scientific colleagues would want to know the truth and to see it disseminated throughout the scientific world. It had not occurred to me that experts on normality-assuming statistics would furiously defend the source of their livelihood and the basis for their prestige, even at the expense of distorting the facts and misrepresenting my position. Although unaware of it at the time, I was on a collision course with antinonrobustness, not only on the part of journal editors and referees but also from readers of my government reports.

Some who read my government reports adopted the strategy of ignoring my findings. Eminent authors of some classical statistical textbooks were intimately aware of my results and had acknowledged their validity. Yet, in their book's discussion of robustness, they cited only those studies favorable to the claim of robustness, omitting any reference to my studies, and this situation has prevailed in a succession of new editions and revisions.

Others completely misrepresented my findings, reporting them as though they were evidence in favor of robustness. In a statement remarkable for its unnatural and deceptive wording, Donaldson (1966) alleged that "Bradley shows that F is ultimately robust for either heterogeneous within cell variances or unequal sample size, but not for both" (p. 29). More straightforwardly expressed, he was (incorrectly) saying that I had shown that, even when all sample sizes are infinite, the F test is not perfectly robust against heterogeneity of variance except in the special case in which sample sizes are equal. However, it was for the two-independent-sample t test that I had shown this, and (based upon extensive mathematical proofs that I had presented in the cited report) I had expressly denied the generalizability of this finding (perfect robustness at equal and infinite sample sizes) to the F test based upon more than two samples. Govindarajulu and Leslie (1972) stated that my sampling distributions for the one-sample t statistic and for the standardized sample mean "compared well with the theoretical distributions under normality assumption" (p. 9), even though at the testing tails some of the discrepancies were large and obvious even at $N=1,024$. Although, in the abstract of the report they referenced, I had called my results iconoclastic, Bevan, Denton, and Myers (1974) alleged that "The findings of Norton (1952), Bradley (1964), Donaldson (1968) and others . . . suggest that the F test is exceedingly robust with respect to Type I and II errors for violations of normality and homogeneity of variance of treatment population" (p. 199). In a 52-page article replete with many tables and graphs presenting numerical robustness data from studies favoring robustness, Glass, Peckham, and Sanders (1972) failed to reproduce any of the dramatic nonrobustness data from my studies, which contained a wealth of tables and graphs they could have used. Instead, they contented themselves with a merely verbal evaluation that incorrectly made it appear that I had not found much, if any, nonrobustness at the .05 or .01 levels but only "beyond .01": "Bradley (1963, 1966) created some doubt about the robustness of the t- and F-tests to violation of the normality assumption, especially at increasingly remote tail regions (that is, beyond .01) . . . . However, we are unsympathetic to dramatizations of the lack of robustness of the ANOVA by appeal to small α's" (p. 254). Later (pp. 281-282), in a masterpiece of understatement, these same authors pointed out that many of the studies they had just reported were methodologically flawed, after which they offered suggestions for the improvement of such studies. These suggestions came very close to being a summary of the methodological characteristics of my own study, but, of course, the reader was not informed of this.

Subsequent Unsuccessful Efforts To Publish In Refereed Journals

Distortion had been the fate of my government reports. Suppression greeted my attempts to publish in journals. Although I submitted articles to many of the most prestigious statistical or methodological journals in the world, their referees and editors either were pathetically ignorant on the subject of robustness or were simply determined to find any reason, however nonsensical, for rejecting an article that took issue with an established (and professionally convenient) belief. Although these same journals had published numerous methodologically inferior articles that encouraged a belief in robustness, my

large and carefully controlled study was repeatedly rejected for reasons that were specious, uninformed, frequently illogical, and often ludicrous or silly. Protests to the editor were generally to no avail; the editor's responses usually displayed as little cerebration as the comments of the referees.

A pose of cosmic wisdom was affected by reviewers who, ignoring (or ignorant of) facts to the contrary, attempted to establish truth by fiat. Although my colleague and I had both obtained L-shaped distributions in actual routine experiments, and had done so simultaneously and independently, I was repeatedly assured by referees and editors that my empirical L-shaped distribution either was spurious or occurred so rarely that it should cause no more concern than the prospect of "being struck on the head by meteors." Over and over again, I was fallaciously informed that the observations constituting the long positive tail of my L-shaped distribution were spurious "outliers" that really didn't belong to the true distribution and should therefore be discarded—leaving a "true" quasi-normal distribution. Time after time, I received the pontifical assurance that my distribution was "very artificial," "unusual," "very extreme," and "rarely encountered in practice." (It is interesting in this regard to note that the overwhelming preponderance of distributions sampled in published studies favorable to robustness were not obtained empirically in an actual experiment, as mine was, but were simply "artificial" mathematical density functions.) Although I had already tried virtually all of the common transformations to normality, without success (including those mentioned in the quotation to follow), it was repeatedly stated or implied by disdainfully overconfident referees, who obviously had not tried them, that they would solve the entire problem: "Bradley's example of reaction times [they were *not* reaction times] is usually dealt with using a log transformation or the arcsin of the square root. Both work well, and are well known with [sic] psychology."

Although experimenters are frequently, if not usually, ignorant of the shape of the sampled population, I often received, as justification for rejecting my manuscript, comments such as "these days no knowledgeable statistician should use the normal theory tests when there is evidence of a long tail" (yet, actually, the exaggerated claims of robustness that I was trying to counteract encourage the experimenter to do precisely that).

Although it takes only a single black swan to refute the generalization that all swans are white, and although I was primarily disputing exaggerated and insufficiently qualified claims implying virtually universal robustness (i.e., a universal generalization), I was criticized repeatedly for sampling from only a single population shape: "A Monte Carlo study of one population is hard to justify." (Actually, for the one-sample Z test I have robustness data for 24 different sampled populations (Bradley, 1973), and there is a distressing and impressive degree of nonrobustness in at least half of the cases.) Ironically, although I was merely disputing reckless generalizations about robustness, I was sometimes falsely accused of making sweeping generalizations about the prevalence of nonrobustness. Finally, overgeneralization about robustness was shrugged off as a problem unworthy of journal space by a reviewer for a well-known and prestigious journal, who wrote "oversimplification is a 'problem' in most textbooks in all research areas. It is not the purpose of [this journal] to publish small [sic] counter-examples to textbook dogma." (Unfortunately, the textbook dogma was also oversimplified journal dogma.)

No one would claim that one should discover a method of abolishing radioactivity before pointing out the dangers of radioactive waste. Yet I was often criticized for not solving the problem I was pointing out. Actually, the neat solution is to use nonparametric statistics, but my critics wanted me to produce a "parametric" solution. In effect, they wanted me to make parametric statistics invulnerable to my criticisms before being permitted to express them—at which time, of course, they would be irrelevant.

All of the foregoing is inane enough. However, some reviewers' comments were so illogical or professionally unenlightened as to boggle the mind. Despite the fact that the degree of heterogeneity of variance that I had investigated often produced or contributed to drastic nonrobustness, a reviewer for one of the most prestigious statistical journals in the world objected that "A ratio of 2 to 1 of Variances is hardly a serious violation of hetrogeneity [sic]" (it was actually a ratio of 4 to 1 and a violation of homogeneity), which, if true, means that I had found extreme nonrobustness even for trivial violations of assumptions and makes my point a fortiori. A referee and the editor of probably the most prestigious psychological journal concerned with methodology, apparently unaware of the well-known fact that, as sample sizes increase, Z, t, and F necessarily tend to become increasingly robust against nonnormality (and become perfectly robust when sample sizes become infinite), used the reflection of this trend in my data as a reason for rejecting my article. The editor wrote "First, the consultant points out, your main conclusion [it wasn't a conclusion of any kind] is that as sample sizes increase to 50 or 100 even when drawing samples from a non-normal distribution, the probability of rejection of H as [sic] now approaches alpha." Thus, perfection at infinity, and an approach to it as n increases, is made to justify the toleration and complacent acceptance of extreme imperfection at actual experimental sample sizes.

Many of the comments I received were conjectural, trivial, or just plain nasty. Although the number of articles on robustness is quite large (and extensive bibliographies have been compiled for them, which I have often referenced for the sake of brevity), I was falsely accused over and over again of being unaware of the existence of references that I had not explicitly included in my article (and the accusations were sometimes repeated after my denials—some referees apparently could not believe that anyone aware of the classical articles "establishing" robustness could be so rash as to present contrary evidence.) Although the Greek and Roman alphabets afford so few symbols that each one must necessarily serve many different mathematical purposes, I was trivially belabored for my choice of symbol by a referee for one of America's most illustrious statistical

journals: "I object to the use of RHO for anything by [sic] correlation. Statistics has enough trouble trying to standardize its notation without some psychologist coming in and taking one of the few agreed upon things and misusing it." Some comments were even more gratuitously nasty: "Bradley's writing is schizophrenic."

The examples given above are only a very small sample of the obstructionistic, professionally incompetent, incorrect, illogical, nonsensical, obtuse, sloppy, trivial, silly, and nasty responses that I received from editors and referees of the most illustrious publications concerned with the subject of my articles. Indeed, there are few reputable journals, appropriate for submission, that have not rejected my articles for reasons that appeared either (1) to be specious and clumsily contrived to perpetuate the prevailing trust in robustness, or (2) to be horrendously uninformed and lacking in the most elementary professional competence, or both. Surely the first type of reason is a manifestation of antinonrobustness, but perhaps some of the obtuseness and sloppiness in the second category merely reflects an unwillingness to undergo the intellectual labor of providing ostensibly sensible reasons for a course of action already decided upon on the basis of prejudice. Thus, antinonrobustness may account for much of both categories, although it is certain that not all of the nonsense created by editors and referees can be attributed to this cause. The broader subject of irresponsible publication practices has been treated elsewhere by myself and others (Bradley, 1981, 1982a; Mahoney, 1977; Peters & Ceci, 1982).

Eventual Publication by Fluke

After years of perseverance, I did eventually manage to publish some articles relevant to the subject in refereed journals. However, in every case the acceptance of the article appeared to be due to some sort of fluke, essentially political in nature, rather than to the intrinsic scientific merits of the article. These publications occurred: (1) in a journal whose founder I had favorably cited in one of my books and who had subsequently invited me to write articles for his journal (however, after a change of editors, my articles were rejected); (2) in a nonstatistical, nonmethodological journal in which I had previously published many articles on a totally different subject—for one of which I had been presented with an award by the society sponsoring the journal (however, many objections were raised against it by referees, and subsequent to publication the article was attacked on bizarre grounds in the same journal); (3) in a statistical journal that had first rejected the article on the grounds that it was entirely theoretical, rashly stating that empirical evidence might have been publishable; to what I believe was their acute vexation, I surprised them by producing the empirical evidence they had said might justify acceptance (I then received salvo after salvo of pedantic, trivial, and nonsensical objections to successive revisions of the manuscript—which terminated only after I pointed out to the editor that "the referees are now inconsistent with their own previous positions and with each other"); and (4) in a journal that had recently published an article favorable to robustness in which my work (or at least my position) had been totally misrepresented and to the editor of which I had complained that I deserved the opportunity to set the record straight. Thus, it would be a mistake to conclude that eventual acceptance of some of my manuscripts necessarily demonstrates the absence or attenuation of antinonrobustness.

REFERENCES

BEVAN, M. F., DENTON, J. Q., & MYERS, J. L. (1974). The robustness of the F test to violations of continuity and form of treatment population. *British Journal of Mathematical and Statistical Psychology*, **27**, 199-204.

BRADLEY, J. V. (1968). Studies in research methodology. *Dissertation Abstracts* (Monograph Section), **28**, 4815B-4816B. (University Microfilms Nos. 68-7445, 68-7446, and 68-7447)

BRADLEY, J. V. (1973). The central limit effect for a variety of populations and the influence of population moments. *Journal of Quality Technology*, **5**, 171-177.

BRADLEY, J. V. (1975). The optimal-pessimal paradox. *Human Factors*, **17**, 321-327.

BRADLEY, J. V. (1977). A common situation conducive to bizarre distribution shapes. *American Statistician*, **31**, 147-150.

BRADLEY, J. V. (1980a). Nonrobustness in classical tests on means and variances: A large-scale sampling study. *Bulletin of the Psychonomic Society*, **15**, 275-278.

BRADLEY, J. V. (1980b). Nonrobustness in one-sample Z and t tests: A large-scale sampling study. *Bulletin of the Psychonomic Society*, **15**, 29-32.

BRADLEY, J. V. (1980c). Nonrobustness in Z, t, and F tests at large sample sizes. *Bulletin of the Psychonomic Society*, **16**, 333-336.

BRADLEY, J. V. (1981). Pernicious publication practices. *Bulletin of the Psychonomic Society*, **18**, 31-34.

BRADLEY, J. V. (1982a). Editorial overkill. *Bulletin of the Psychonomic Society*, **19**, 271-274.

BRADLEY, J. V. (1982b). The insidious L-shaped distribution. *Bulletin of the Psychonomic Society*, **20**, 85-88.

BRADLEY, J. V. (1983). Paradox lost, paradox regained: Reply from a flagellated straw man. *Bulletion of the Psychonomic Society*, **21**, 69-72.

BRADLEY, J. V. (1984). The complexity of nonrobustness effects. *Bulletin of the Psychonomic Society*, **22**, 250-253.

DONALDSON, T. S. (1966). *Power of the F-test for nonnormal distributions and unequal error variances* (Memorandum No. RM-5072-PR). Santa Monica, CA: The Rand Corporation.

GLASS, G. V., PECKHAM, P. D., & SANDERS, J. R. (1972). Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. *Review of Educational Research*, **42**, 237-288.

GOVINDARAJULU, Z, & LESLIE, R. T. (1972). *Annotated bibliography on robustness studies of statistical procedures* (U.S. Department of Health, Education, and Welfare Publication No. (HSM) 72-1051). Rockville, MD: National Center for Health Statistics.

MAHONEY, M. J. (1977). Publication prejudices: An experimental study of confirmatory bias in the peer review system. *Cognitive Therapy and Research*, **1**, 161-175.

PETERS, D. P., & CECI, S. J. (1982). Peer-review practices of psychological journals: The fate of published articles, submitted again. *Behavioral and Brain Sciences*, **5**, 187-255.

WIKE, E. L., & CHURCH, J. D. (1982). Nonrobustness in F tests: 1. A replication and extension of Bradley's study. *Bulletin of the Psychonomic Society*, **20**, 165-167.

(Manuscript received for publication June 18, 1984.)