Comment

Roger L. Berger

"We believe that the LR criterion remains a generally reasonable first option for non-Bayesian parametric hypothesis-testing problems."

"[The LR criterion is] a very general method, one that is almost always applicable, and is also optimal in some cases."

The first quote is from the preceding paper by Perlman and Wu which I will refer to as PW. The second quote is from Casella and Berger (1990, page 346). There is a good deal of agreement here about the usefulness of likelihood ratio tests (LRTs). So what has prompted PW to feel the need to defend the LR criterion so vigorously?

The question is whether, in some problems, another test might be preferable to the LRT. Despite calling the LRT a "generally" reasonable first option, PW really seem to argue that the LRT is the primary option, to be abandoned only in very extraordinary circumstances. On the other hand, near the end of their Section 10 they do say, "It would be of interest to characterize those problems where the LRT is or is not successful," and they say, "The LR criterion is not infallible." This indicates to me that PW would be willing to use some other test besides the LRT in some circumstances. I will assume this to be true in the remainder of my comments.

1. WHAT CRITERION TO JUDGE TESTS?

1.1 α-Admissibility

I think PW agree with me that, after a LRT is derived in some problem, it needs to be scrutinized to determine if the LR criterion did produce a good test in this particular problem. Then the question is, "What criteria should be used to judge the LRT?" In the articles by other authors and me to which PW refer, the criterion is clear, α-admissibility. More precisely, if two tests are both level-α, and the power of the first test is greater than the power of the second test everywhere on the alternative, then the first test is preferred. It was never my intent to assert that α-admissibility is the only reasonable criterion. I do not believe this is true. But α-admissibility is a well-understood criterion that has been considered by statisticians for over sixty years. I find that it is easily understood and reasonable also to my colleagues who are scientists in other areas. I think it is a reasonable way to compare error probabilities of tests. In my papers, I have simply pointed out that, if one uses this classical criterion, tests that are superior to the LRT can be found in some problems. If the reader rejects this method of comparing tests, then he or she will have little interest in my results.

In any case, the criterion for comparing tests must be stated clearly, first, then applied to the problem at hand. The Emperor should not kill the messenger because he does not like the message. But this is exactly what PW propose. They say in Section 2 that if the α-admissibility criterion delivers the wrong message, that the LRT is inferior, then it is the criterion that should be abandoned. Kill the messenger for delivering the wrong message.

So clearly, PW do not want to use α-admissibility to determine if a LRT is reasonable in a particular problem. What criterion will they use? Unfortunately, the answer is unclear. In this article they use numerous criteria for different problems. It is not explained why one criterion is used in one problem and another criterion is used in another. It seemed that, for each problem, the criterion was used that would put the LRT in the best light for that problem. This was very unconvincing to me. I hope that in their rejoinder to these comments, PW will clearly state what criterion they use, after deriving a LRT, to determine if it is a reasonable test for the problem at hand. I will now comment on some of the various criteria that PW use to compare tests.

1.2 Decision Theoretic Admissibility

Frequently, PW use decision theoretic admissibility (d-admissibility) to defend LRTs. For several examples they point out that the LRT is d-admissible and that α-inadmissibility does not imply d-inadmissibility. Through the first nine sections, I thought that PW's criterion was this: derive the LRT; if it is d-admissible, it is a good test. But then in the first example in Section 10, they describe a LRT that is inadmissible. Does this mean the LRT

Roger L. Berger is Professor, Department of Statistics, North Carolina State University, Raleigh, North Carolina 27695-8203 (e-mail: berger@stat.ncsu.edu).

is not good? No, they engage in "mathematical acrobatics" to decide the LRT is appropriate despite its inadmissibility. Once again, when the criterion (d-admissibility) delivers the wrong message, that the LRT is inferior, PW quickly abandon the criterion rather than the LRT.

I think that d-admissibility is a reasonable criterion. Certainly, if one test dominates another in the d-admissible sense, the first test is preferred. But, in these testing problems as in most problems, the difficulty is that d-admissibility is not a very discerning criterion. Many tests are d-admissible. As PW note, in many problems both the New Tests and the LRT are d-admissible. So this criterion does not choose which is better. Here α-admissibility gives a way of comparing two tests that might both be d-admissible.

The first example in Section 10 might lead one to conclude that some criterion based on the ancillary principle is appropriate for comparing tests. But, arguments involving the ancillary principle are usually unconvincing to me. Difficulties in implementing the ancillary principle, such as those discussed by Cox and Hinkley (1974, pages 33 and following) in discussing Basu (1964), lead me to conclude that it is not a generally applicable principle for the comparison of statistical procedures.

1.3 Bayesian Criterion

Although PW often say they are recommending the LRT as a non-Bayesian test, they use a Bayesian criterion to defend the LRT several times. In Sections 4 through 6 they discuss priors against which the LRT or the New Tests are Bayes or approximately Bayes. Then they defend the LRT by asserting that its prior is more reasonable than the prior for the New Test. To use this criterion to decide if a test is reasonable, the user must decide what prior is reasonable, even to choose a non-Bayesian test. If the user is going to go through this effort of determining a reasonable prior, I think he should bite the Bayesian bullet, choose his prior and use a Bayesian test for this prior. This Bayesian test will not be the LRT for any easily expressed prior in the examples in Sections 4 through 6.

Actually, PW pose a harder problem than a standard Bayesian analysis. It is more difficult to start with a test and then determine for what prior it is Bayes than to find a Bayes test for a given prior. PW did not usually state specific priors, only general features of priors. I sometimes had difficulty following their reasoning as to why a test was approximately Bayes versus a prior with some features. Because they are proposing that one should determine what prior a test is Bayes against to determine if the test is good, it would have been instructive if they had worked out the details for at least some examples. For example, for testing (18) they say the LRT is approximately Bayes for a prior that puts equal weight on L_1 and L_{23} while test (21) is approximately Bayes for a prior that puts unequal weight on these two sets. Surely, not all priors with these equal and unequal weights will work. I would be interested in the details of how they determined these features for the priors.

1.4 Significantly Better Fit

At the end of Section 2, PW state

> ... the question to be addressed by a statistical hypothesis test ($\equiv$ significance test) is the following: based on the observed data, does the family of observed distributions represented by the alternative hypothesis H_1 fit (or support, or explain) the observed data *significantly* better than the family represented by the null hypothesis H_0? Only if the fit is *significantly* better should we reject H_0 in favor of H_1. The tests discussed in this paper are evaluated on the basis of this (we believe) generally accepted criterion.

PW say this is *the criterion* they will use to evaluate tests. But I do not see a criterion for comparing tests here! The quote expresses the question to be addressed by a hypothesis test, but the quote gives no indication of how one judges if one test answers this question better than another. It does not give a definition of what is meant by "the fit is significantly better." How can one judge if one test statistic measures this fit better than another? PW hint at an answer in Section 3. After defining a LRT as a test that rejects H_0 if $\Lambda(X) \geq c$, they say that because they require significantly better support, c must be greater than 1. This suggests that $\Lambda(X)$ measures the support in H_1 versus the support in H_0. Are they saying $\Lambda(X)$ is the only test statistic that properly measures this support? What about χ^2 statistics, score statistics, Wald statistics and so on? I hope that in their rejoinder to these comments, PW will define how the above quote defines a criterion for comparing tests.

1.5 Closeness to H_0

In Sections 4–6, PW discuss in great detail whether certain sample points should be in the rejection region or the acceptance region. In these sections the criterion seems to be based on the Euclidean distance from the sample point ($\equiv$ the MLE) to H_0. It is not clear to me if PW consider this Euclidean distance criterion to be a generally

applicable criterion or whether they are using it in these normal mean problems because Euclidean distance gives the same ordering to the sample points as the LRT statistic. If it is the former, I do not know an argument for why Euclidean distance is always appropriate and how this criterion is implemented in the presence of nuisance parameters. If Euclidean distance is used as a perfect surrogate for the LRT statistic, then there is not much to be discussed. Having defined the LRT statistic as the appropriate measure, there is no way PW can be convinced that the LRT is inappropriate!

1.6 Appropriate Power Functions

The criterion used in Endnote 6 is perhaps the most indecipherable to me. PW speak of undefined concepts like "distributions most difficult to distinguish from H_1." Then they declare it "entirely appropriate" that the power of the LRT be less than α at $\mu = (0, 0)$. Is this a defense of the LRT? As PW themselves recall, every level-α test (except the trivial test), including the New Tests that have been proposed, have power less than α at $\mu = (0, 0)$. Are their power functions "entirely appropriate," also? Or are PW asserting that the LRT's power of α^p at $(0, 0)$ is the one, true "entirely appropriate" power and other powers, such as $\alpha^{p-1}/2$, although less than α, are inappropriate?

2. COMMON SENSE, INTUITION, SCIENTIFIC APPROPRIATENESS

In the previous section, I have tried to list some of the many criteria PW used to defend LRTs. I have expressed my hope that PW will choose and defend the one criterion they recommend for comparison of tests. However, I do not think this is really PW's goal. I think the criteria they really wish to espouse are the vague, subjective and undefined concepts of common sense, intuition and scientific appropriateness. Such subjective concepts seem to me to be of little use in an objective comparison of tests, but, in the spirit of PW, let me tell my own fantasy.

Two graduate students and their dissertation advisor come to PW for help in analyzing their data. These three will each give a talk about their scientific work at a professional symposium. The students will each talk about their work, and then the advisor will give an overview of the work in her lab. The following conversations ensue.

On Day 1, the first student (S1) comes for a consulting appointment. After a discussion of the experiment PW and S1 agree that the student's data can be modeled as $X_1 \sim N_1(\mu_1, 1)$, and the scientific question to be addressed can be answered by testing H_0: $\mu_1 = 0$ versus H_1: $\mu_1 \neq 0$.

S1: The data I obtained in my experiment were $X_1 = 2.00$.

PW: Because $X_1^2 = 4.00 > 3.89 = \chi^2_{1,0.05}$, you have found significant evidence that $\mu_1 \neq 0$.

On Day 2, the second student (S2) comes for a consulting appointment. After a discussion of the experiment, PW and S2 agree that the student's data can be modeled as $(X_2, X_3) \sim N_2((\mu_2, \mu_3), I_2)$, and the scientific question to be addressed can be answered by testing H_0: $(\mu_2, \mu_3) = (0, 0)$ versus H_1: $(\mu_2, \mu_3) \neq (0, 0)$.

S2: The data I obtained in my experiment were $(X_2, X_3) = (\sqrt{6.00}, 0)$.

PW: Because $X_2^2 + X_3^2 = 6.00 > 5.99 = \chi^2_{2,0.05}$, you have found significant evidence that $(\mu_2, \mu_3) \neq (0, 0)$.

On Day 3, the advisor (AD) comes for a consulting appointment.

AD: My students have reported to me on their meetings with you. I just wanted to verify that, based on the data $(X_1, X_2, X_3) = (2.00, \sqrt{6.00}, 0)$ collected in my lab, I can conclude that $\mu_1 \neq 0$ and $(\mu_2, \mu_3) \neq (0, 0)$.

PW: Oh, no! You must use common sense and intuition to reach the scientifically acceptable conclusion. Because $X_1^2 = 4.00 < 5.99 = \chi^2_{2,0.05}$, you cannot conclude $\mu_1 \neq 0$.

AD: That stupid S1! He said that because $X_1^2 = 4.00 > 3.89 = \chi^2_{1,0.05}$, we had found significant evidence that $\mu_1 \neq 0$.

PW: No, S1 can compare X_1^2 to 3.89. But, because you will also discuss (X_2, X_3), you must compare X_1^2 to 5.99.

AD: I get it. This is one of those multiple testing problems I learned about in my stats class in graduate school. Because I am doing multiple tests, I have to use a bigger critical value.

PW: No, you can test H_0: $\mu_1 = 0$ or $(\mu_2, \mu_3) = (0, 0)$ and reject this H_0 because $X_1^2 = 4.00 > 3.89$ and $X_2^2 + X_3^2 = 6.00 > 5.99$. This test has a type I error probability of $\alpha = 0.05$, but it would not be the common sensical, intuitive and scientifically acceptable conclusion.

I leave it to the reader to decide. Will AD leave PW's office appreciating the simplicity and clarity of statistical analysis? Or will she think it is ridiculous that at the symposium S1 and S2 will speak on

the significance of their results, then she will rise and speak on the nonsignificance of the same data?

Other authors have criticized some New Tests because their rejection regions are nonmonotone or nonconnected or because they contain sample points for which the MLE is in H_0. However, PW are the first likelihoodlums that we have encountered who are so die-hard that they prefer the LRT (20) to the simple, size α, uniformly more powerful, intersection–union combination of the individual LRTs in (21). Some readers may be interested in the more complete discussion of combining individual LRTs that can be found in Berger (1997). Saikali and Berger (1998) discuss a problem related to (20) and (21), that of testing two normal means using samples of unequal sizes.

3. CONCLUSION

PW argue that intuition and common sense have a role in determining good statistical procedures. Intuition can be very useful, even essential, in guiding one toward the correct solution to a problem, but intuition should *never* be the criterion used to justify a solution. Scientific and statistical conclusions must be based on objective, well defined, verifiable and consistently applied criteria. They must never be based on personal, subjective criteria. If we are indeed teaching our students that intuition is the primary criterion in scientific inquiry, then a fundamental reassessment of the mission of mathematical statistics is urgently needed.

Comment

D. R. Cox

Michael Perlman and Lang Wu have produced some striking examples where application of the formal theory of optimal tests leads to procedures which they regard as unacceptable. I entirely agree with that interpretation.

In Section 9 they produce a list of quotations essentially to the effect that the notions of Neyman–Pearson theory are inapplicable, at least in many situations of the interpretation of data, and that what the authors broadly call Fisher's position is to be strongly supported. While again I broadly agree, I do not think this makes the Neyman–Pearson approach unimportant for inferential problems, only that some care in interpretation and application is needed. I tried to argue this in the later parts of Cox (1958) and in a different form in Cox (1977) and the following brief notes are in amplification and extension of those remarks. I suggest that many of the differences between different viewpoints get much less if we take the position that:

1. The notions of error rates, acceptance and rejection of hypotheses and so on give certain quantities hypothetical physical intepretations and are not instructions on how to apply the methods. For example, we do not have to choose an α and firmly reject or not according to the specified rule. Such hypothetical procedures, however, clarify the meaning of p-values as measuring instruments in a way entirely in line with the common device of defining quantities operationally via measuring procedures which are in fact not used directly. Think, for example, of the definition of the acceleration due to gravity measured at sea level under the peak of Mt. Everest.
2. Some very plausible operational requirements, for example similarity of tests, are appealing but if insisted on exactly may have unexpected and unpleasant consequences. Other requirements, like the unbiased property of tests, are more clearly somewhat arbitrary and provisional. That such requirements lead to trouble is not in itself fatal to the whole approach of assessing statistical procedures via hypothetical operational properties. They reflect both on the care needed in formulating and interpreting optimality requirements, in fact in generality, and in this particular context probably also on the need to impose conditions of a quasi-logical character before looking for sensitivity.
3. To be relevant to the problems under analysis and to satisfy Fisher's definition of probability

D. R. Cox is an Honorary Fellow, Nuffield College, Oxford OX1 1NF, United Kingdom (*e-mail: david.cox@nuffield.oxford.ac.uk*).

(Fisher, 1956, page 33), calculations must be exactly or approximately conditional, this being a more fundamental concept than sensitivity.

4. The object of significance tests is to calculate approximately p-values.
5. We may wish to assess procedures that are not in a technical sense optimal either because none such exist or because of considerations such as transparency or robustness. Neyman–Pearson arguments are clearly very fruitful for this.

In the particular context of significance tests it may be helpful to draw some further distinctions. First, the following formulation assumes null hypotheses specifying point values for parameters, in general with nuisance parameters. It does not regard, for example, $\theta < \theta_0$ as an appropriate null hypothesis. This seems a rather clear distinction between Fisherian and some Neyman–Pearson formulations. There are then a number of possibilities including the following:

- Only two possibilities are contemplated, two simple hypotheses, $\theta = \theta_0$ or $\theta = \theta_1$ regarded as essentially on an equal footing, one called the null hypothesis and the other the alternative. There are quite strong arguments for using the observed likelihood ratio as it stands. This formulation does not seem to cover well the typical situation in which significance tests are applied. It is, however, close to simple forms of discriminant analysis.
- Only a null hypothesis is firmly established. We think about the kinds of alternative against which we want sensitivity and formulate these provisionally via alternative models, for example via a parametric family defined by $\theta > \theta_0$. These are not necessarily to be taken as a base for interpretation. A realistic example is testing linearity of regression by including a quadratic term. If we find nonlinearity, we may well not use the quadratic model but transform either or both the response or explanatory variables or fit a nonlinear model in line with subject matter considerations. The Neyman–Pearson theory is ideal for discussions of this general, somewhat exploratory, situation. It seems preferable to Fisher's notion of a pure significance test, although in the last analysis it is often essentially equivalent to it, the former requiring choice of alternatives, the latter the choice of a test statistic.
- We have, say, a full parametric family, all of them serious candidates as a base for interpretation, and a special value θ_0, say. Often but not always this situation is best addressed as a problem of interval estimation, especially if the null hypothesis is what is sometimes called a dividing hypothesis.

A final general point is that to be useful in realistic problems, a general theory has to deal effectively with situations in which in some sense only approximate answers are achievable.

Of course the relation of all this with Bayesian theory is another story. Note, however, that the last of the above three formulations is reasonably directly formulated in Bayesian terms.

Comment

Michael P. McDermott and Yining Wang

We congratulate Professors Perlman and Wu on a very interesting article that brings together many examples of multiparameter hypothesis testing problems in the literature for which the likelihood ratio test (LRT) has been long thought to be "deficient." The article raises many philosophical issues regarding how to approach hypothesis testing problems and, as the authors point out, echoes the famous Fisher–Neyman debate. In general, we agree with much of what Perlman and Wu (and, yes, Fisher) have to say, particularly with regard to the notion of using "statistical common sense" in choosing a test. In addition, the authors' description of the nature of the prior distributions required to render the New Tests Bayes (or approximately so) is very helpful in conveying the authors' point about the practical utility (or lack thereof) of some of these tests. On the other hand, while Perlman and Wu appear to equate "appropriate inferences" with those provided by the LRT, we are not convinced

Michael P. McDermott is Associate Professor, Department of Biostatistics, University of Rochester, 601 Elmwood Avenue, Box 630, Rochester, New York 14642 (e-mail: mikem@metro.bst.rochester.edu). Yining Wang is at Schering-Plough Research Institute, 2015 Galloping Hill Road, K-15-2, 2315, Kenilworth, New Jersey 07033-0539 (e-mail: wayne.wang@spcorp.com).

that it is easy to characterize what is meant by an "appropriate inference" or, for that matter, "statistical common sense." Critical regions that clearly violate "statistical common sense" are often easy to spot; however, it is not so easy to draw the line between tests that are and are not violators.

Our interest in the problem discussed in Section 4, namely that of testing hypotheses concerning linear inequalities, was stimulated by its mathematical, rather than practical, aspects (McDermott and Wang, 1999). We agree with the general point that adding regions to the critical region of the LRT that are not connected or that are in close proximity to the origin (or that are even "inside the null hypothesis") yields tests that are objectionable from the standpoint of a practitioner. For this reason, we were careful not to label the LRT for this problem as "deficient."

On the other hand, adding a region that *is* connected to the critical region of the LRT and *is not* in close proximity to the origin will yield a more powerful test and, to some, may not be practically objectionable. For example, the critical region of the test first proposed by Liu and Berger (1995) for the problem discussed in Section 4 is shown in Figure 1 as the union of the shaded and unshaded regions. Note that the scale in this figure (transformed using $\Phi(\ )$) differs from that in Perlman and Wu's Figure 1. One could take the critical region to be only the shaded region $A_0 \cup A_1$ to avoid including points in close proximity to the origin (on the X_1–X_2 scale). Two natural questions arise: (1) How does one determine the lower boundaries of this added region? (2) Isn't this New Test still objectionable on the grounds that the rejection region is nonmonotone? We don't have an objective basis for determining an answer to the first question; it would seem to depend on the researcher's notion of "common sense." Perlman and Wu would argue that the critical region of the LRT provides an objective and satisfactory boundary. Their criticisms of a more powerful New Test having a slightly expanded critical region (as above) would include the lack of monotonicity and the need for "mathematical acrobatics."

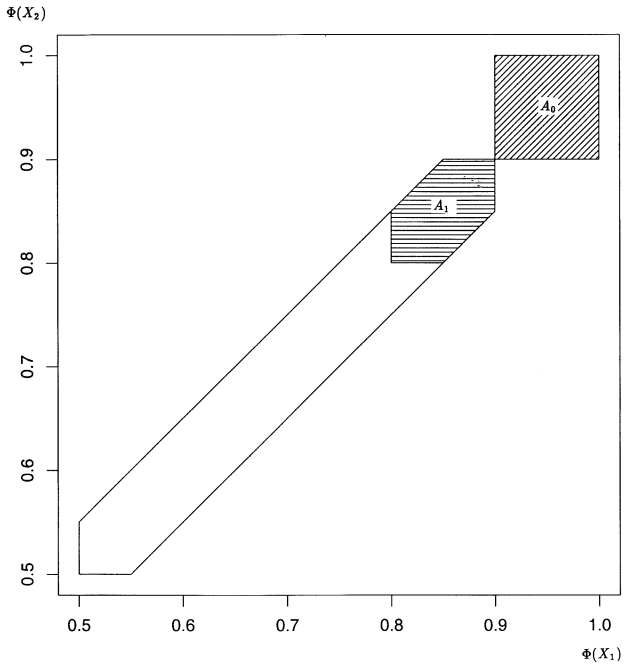

FIG. 1. *Truncated critical region of Liu and Berger's* (1995) *test. The union of the shaded and unshaded regions is the critical region proposed by Liu and Berger* (1995), *and* A_0 *is the critical region of the LRT. The shaded region* $A_0 \cup A_1$ *is the truncated critical region.*

With regard to the issue of monotonicity, Laska and Meisner (1989) make the case that

> ...restriction to monotone functions seems natural. An experimenter who obtains values of each component test statistic at least as large as those observed by his colleague would not find reasonable a test procedure that results in his colleague rejecting H_0 but denies such rejection to him. Since regulatory agencies are subject to judicial review, allowing nonmonotone procedures cannot help but lead to accusations of unfairness. Imagine how the courts would view the fairness of the FDA if the outcome (a, a) rejected H_0 but the outcome (∞, a) did not.

While this argument is convincing, it can also be argued that if the outcome (a, a) does not violate the experimenter's "common sense" (or that of the regulatory agency) in terms of providing evidence against H_0, then one should not deny "the colleague" the opportunity to reject H_0 as long as he or she (and the regulatory agency) is willing to risk a small increase in the probability of falsely doing so. Of course, the LRT would not reject H_0 in either case [(a, a) or (∞, a)], but "the colleague" may view that as an "inappropriate inference" that violates his or her notion of "common sense."

Other weaknesses of nonmonotone tests support the argument against their use in practice. One cannot sensibly define a p-value associated with such a test. Also, as demonstrated in Figure 2 below for Liu and Berger's (1995) test, there may be sample points for which one rejects H_0 at the 5% significance level but fails to reject H_0 at the 10% level [interestingly, Berger's (1989) Test I does not have this property]. The use of a test exhibiting such behavior in practice is difficult to justify. To the extent that a researcher (or agency) is uncomfortable with drawing conclusions from a nonmono-

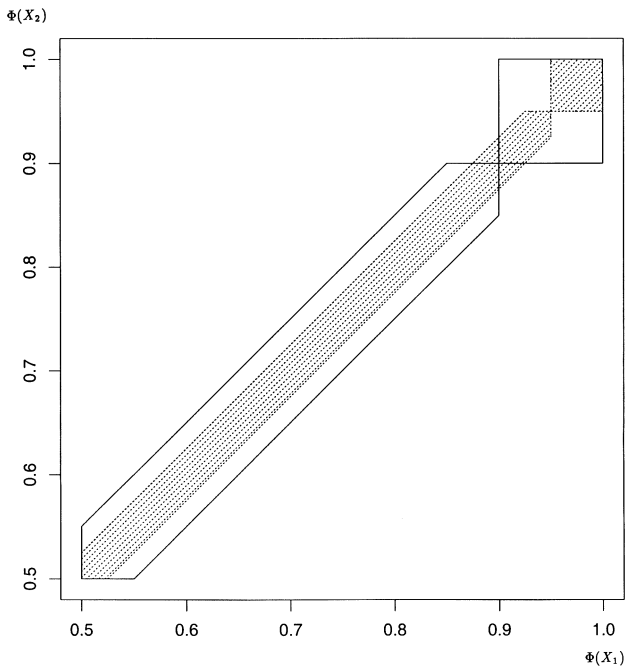

FIG. 2. *Critical regions of Liu and Berger's* (1995) *test for* $\alpha = 0.10$ (*solid lines*) *and for* $\alpha = 0.05$ (*dashed lines, shaded*). *Note that portions of the latter region are not contained in the former region, so that for some sample points it is possible to reject* H_0 *at the % significance level yet fail to reject* H_0 *at the* 10% *level.*

tone test, he or she will side with Perlman and Wu on this issue.

The objection of Perlman and Wu to the need for "mathematical acrobatics," which was raised in the context of the bioequivalence problem in Section 7, is less convincing. Acknowledging some well-founded criticisms of their unbiased test for the bioequivalence problem, Brown, Hwang and Munk (1997) proposed a modification of their test (C_T) that truncates the rejection region. This modification is similar in spirit to the truncation proposed in Figure 1 above with the important exception that C_T retains a desirable monotonicity property. Perlman and Wu state that "The need for such mathematical acrobatics reinforces our contention that the quest for (nearly) unbiased tests more powerful than the LRT is misguided." We don't see the basis for this statement; the authors have not offered a substantive criticism of this test. They argue that the LRT "already makes perfectly appropriate inferences" and that "if this is deemed unsatisfactory, then the solution is very simple: more observations are needed, not Better New Tests." This, however, is not a strong argument *against* the C_T procedure of Brown, Hwang and Munk (1997). We believe that this example is evidence against the authors' opinion that "the goal of constructing tests that are less biased and everywhere more powerful than the LRT is without intrinsic merit." The C_T test of Brown, Hwang and Munk (1997) does not appear to provide "unwarranted and inappropriate inferences," nor does it appear to be "practically unacceptable."

In general, the New Tests are associated with an increased probability of falsely rejecting H_0. However, this probability is still bounded below a fixed significance level, and the New Tests also have increased probability of correctly rejecting H_0. The relative merits of this trade-off are certainly debatable when there is not universal agreement regarding what constitutes a rejection region that represents "statistical common sense." Defining this region as the rejection region of the LRT, as Perlman and Wu would advocate, would add an element of circular reasoning to the authors' argument.

The "conditionally anticonservative" property of our conditional LRT (Wang and McDermott, 1998) revealed by Perlman and Wu in Section 8 raises other interesting issues. If p is large and the observed V assumes a particular form, the conditional LRT will nearly always reject H_0 for sample points consistent with the alternative hypothesis but close (in terms of Mahalanobis distance) to the origin. The use of the term "anticonservative" here implies that this is an undesirable property, but this is not necessarily the case. In fact, it is often not difficult to identify a statistic upon which to condition in order to render a test "conditionally anti-conservative," even severely so. This is illustrated in the following example.

EXAMPLE. Consider the problem of testing H_0: $\mu = 0$ versus H_1: $\mu > 0$ based on a random sample of n observations from a $N(\mu, 1)$ distribution. The uniformly most powerful test rejects H_0 when $\sqrt{n}\,\overline{X} > 1.645$, assuming $\alpha = 0.05$. If one conditions on the sum of the first $n - 1$ observations,

$$P_{H_0}\left(\sqrt{n}\,\overline{X} > 1.645 \,\middle|\, \sum_{i=1}^{n-1} X_i = k\right) = 1 - \Phi(1.645\sqrt{n} - k),$$

which tends to 1 as $k \to \infty$.

Obviously the "conditionally anticonservative" property is not necessarily undesirable. A key aspect of this example is the fact that large values of the statistic being used for conditioning are themselves evidence against H_0. The same is true of the statistic K [the number of positive components of $\pi_S(X; \mathscr{O})$], as discussed by Bartholomew (1961) and Perlman and Wu (1999). When p is large, the fact that *all* of the components of the sample mean vector are positive is itself evidence against H_0. It is not so clear that rejecting H_0 for such sample

points is an inappropriate decision, even if the sample mean vector is close to the origin. Indeed, the LRT for the same problem, but with Σ completely known (Kudô, 1963), exhibits similar "conditionally anticonservative" behavior (the conditional LRT and the LRT for known Σ are asymptotically equivalent). For example, when $\alpha = 0.05$, $\Sigma = I_p$ and $p = 100$, the null probability that Kudô's LRT rejects H_0 given $K = p$ is equal to 0.991. We conjecture that for certain other values of Σ, a smaller value of p will suffice to render this conditional probability to be close to 1. We believe that this behavior of the LRT is reasonable and that, for the problem involving completely unknown Σ, the conditional LRT is the preferred test. All of this reinforces our belief that the search for tests that are less biased and (generally) more powerful than the LRT is *not* misguided.

Finally, it is worth noting that there has been a movement away from hypothesis testing and toward parameter estimation (point and interval/region) in applied research (Gardner and Altman, 1986; The Standards of Reporting Trials Group, 1994). Relatively little research has been devoted to this aspect of inference under order restrictions, perhaps because it is a difficult area. It would be interesting and informative to see, for example, what the shapes of the confidence regions would be as obtained from inverting the LRT and the New Tests for the problems considered in this paper.

ACKNOWLEDGMENT

The authors thank colleagues at the University of Rochester, Department of Biostatistics, for helpful discussions.

Rejoinder

Michael D. Perlman and Lang Wu

We are grateful to the Editor and the discussants for their interest in, and contributions to, these important issues. If our paper has achieved nothing other than to serve as a vehicle for bringing Professor Cox's eminently sensible views on hypothesis testing to the pages of *Statistical Science*, then we deem our efforts to be highly successful. In particular, his item 2 expresses our views perhaps even more aptly than we have done ourselves. His reminders of the need for care "in formulating and interpreting optimality requirements" and for imposing "conditions of a quasi-logical character before looking for sensitivity" epitomize what we have called "statistical common sense."

Berger and McDermott and Wang, themselves being Creators of New Tests, do not entirely agree with our position, but we have found many of their comments to be constructive and have modified some of our assertions accordingly, as will be indicated below. Nonetheless, our message remains clear.

Science, above all, must be consistent with itself and with the real world. Science progresses when inconsistencies are discovered in an existing theory. Newtonian mechanics served adequately for two hundred years, but its inconsistencies on the subatomic and cosmological scales led to the development of quantum mechanics and general relativity. Analogously, while Neyman–Pearson testing theory has been one of the most widely applied and successful tools for statisticians in the past six decades, it has been almost as widely recognized that it can lead to troubling inconsistencies, even from a non-Bayesian perspective. As indicated in Section 9 we are far from the first to note this, but we feel that the examples that we have collected here, especially those in Sections 4, 5 and 7, are more dramatic and convincing than those known previously.

We reiterate our case briefly here. Even in relatively uncomplicated testing problems involving only normal (Gaussian) distributions, application of the classical NP criterion of a more (or most) powerful size α test and the related criterion of α-admissibility (as opposed to d-admissibility) may lead to New Tests that are transparently untenable, as in the examples of Sections 4, 5 and 7 (more on this below). Even the Creators admit their misgivings, for example, Berger's retreat from his Test II to the less powerful Test I (Section 4), Brown, Hwang and Munk's retreat from their unbiased test to the less powerful truncated version (Section 7). Their doubts (note: based on statistical common sense!) show that they agree, at least implicitly, with our and our many predecessors' contention that *the NP criterion is not absolute* and that its applicability in a

testing problem must be governed by the reasonableness of the resulting test.

We fully agree with L. D. Brown, who wrote to us as follows:

> I don't believe the Emperor needs new clothes —I agree the LRT is a valid default procedure. Any proposed alternatives should be inspected according to standards of Common Sense. (On occasion, the LRT also deserves to be criticized by those standards, as well.)

We freely admit that neither we nor any other non-Bayesians have formulated a precise definition of statistical "reasonableness" or "common sense," but like the Emperor, most statisticians know them when they see them. Berger himself has very good statistical intuition, as evidenced by his example in Section 2 of his discussion. While aimed (effectively, we concede below) at the example in our Section 6, when applied to the testing problem (6) in our Section 4, his intuition provides the final nail in the coffin for his New Tests.

For, suppose that Berger's students S1 and S2 observe independent normal variables $X_1 \sim N_1(\mu_1, 1)$ and $X_2 \sim N_1(\mu_2, 1)$, respectively. S1 wishes to test $\mu_1 \le 0$ versus $\mu_1 > 0$, while S2 wishes to test $\mu_2 \le 0$ versus $\mu_2 > 0$. On Day 1, S1 comes to Berger and says, "I have observed $X_1 = -1$," to which Berger replies, "You have not found significant evidence that $\mu_1 > 0$." On Day 2, S2 comes to Berger and says, "I have observed $X_2 = -1$," to which Berger replies, "You have not found significant evidence that $\mu_2 > 0$." On Day 3, AD comes to Berger and says, "Based on the data $(X_1, X_2) = (-1, -1)$ collected in my lab, I cannot conclude that $\mu_1 > 0$ or $\mu_2 > 0$. This is disappointing, since I was hoping to find at least one significant effect." But Berger immediately replies, "Cheer up! By applying my New Test II, which is everywhere more powerful than the LRT while maintaining the same size, based on these data we can conclude that *both* $\mu_1 > 0$ *and* $\mu_2 > 0$!" The AD leaves Berger's office with a renewed respect for modern statistical theory.

Clearly, such a New Test is the statistical equivalent of cold fusion; no statistician would recommend its use. We suggest that the reader, especially any young researcher tempted to create her/his own New Test, keep this example in mind as he/she weighs Berger's statement that the NP criterion of α-admissibility "is a well-understood criterion that has been considered by statisticians for over sixty years . . . . If the reader rejects this method of comparing tests, then he or she will have little interest in my results." To quote another Berger (Jim): "Statistics looks very bad when it recommends a conclusion that clearly contradicts common sense." (Berger and Wolpert, 1988, page 141).

Those who know us know our healthy respect for mathematical theory. But we must question the value of statistical research "stimulated by its mathematical, rather than practical, aspects" [MW] when such work produces impractical procedures that are then promoted (fortunately unsuccessfully) to the applied community. Perhaps the strongest criticism of the advocacy of statistical procedures that may be mathematically optimal but that would be unacceptable to scientific investigators was stated by Fraser and Reid (cf. Brown, 1990, page 507): "[The] statistician who advertises the [scientifically unacceptable procedure] is guilty of professional misconduct."

Perhaps anticlimatically, but as requested by Berger (Roger), let us examine the statistical basis for our "common sense" criticisms of the New Tests more closely. For the Gaussian examples of Sections 4, 5 and 7, where the sample and parameter spaces can be identified, the *rejection* regions of the New Tests include sample points either (a) in the alternative hypothesis but arbitrarily close[25] to the null hypothesis (Sections 4 and 7), (b) closer to the null hypothesis than to the alternative (Section 5) or (c) actually inside the null hypothesis (Sections 4 and 7).

As Cox writes, " . . . the object of significance tests is to calculate approximately p-values." Because the p-value associated with a testing procedure is a device for assessing the fit of the data to the null hypothesis relative to the alternative hypothesis, tests should reflect the quality of this fit. Because, for normal distributions, sample points satisfying (a), (b) or (c) do not fit the alternative better than the null hypothesis, no one[26] would reasonably conclude that such sample points support the alternative significantly better than the null hypothesis, but this is what the New Tests would have us do.

The NP criterion and α-admissibility, which here favor these New Tests, simply fail in these examples. This is more than mere "common sense" or "intuition," it is statistical common sense based on experience with statistical theory and practice and is shared by many eminent statisticians, such as those cited in Section 9.

It should be noted that we are using "fit" and "distance" only in a qualitative, ordinal sense. That is, we take it as self-evident that (at least for normal models) if the data fit the null hypothesis as well as, or better than, they fit the alternative hypothesis, it cannot be claimed that they provide statistical evidence favoring the alternative. We are *not* asserting that distance itself always should be used directly as the test statistic, for as Berger

correctly notes, this would be equivalent to saying that the LRT *always* should be used. In fact, as noted below, we now accept Berger's argument that his IUT is preferable to the LRT for problem (18) in Section 6.

Berger says that it was never his intent "to assert that α-admissibility is the only reasonable criterion." But the hard reality is that α-admissibility is often unreasonable, not, as Berger would have us say, because it may contradict the LRT, but as we actually said and again demonstrate in the AD's example seven paragraphs above, because it may contradict the evidence provided by the data.

We believe, like Fisher and many others, that the purpose of statistical testing is to assess this evidence, not to maximize a mathematical criterion. (Recall Cox's statement four paragraphs above.) If the AD's data are insufficient to detect significant effects, like any good scientist her recourse should be to obtain more data and/or redesign her experiments, not to resort to New Tests. To paraphrase Cox's item 1, *power*, *size*, *and bias are only incidental tools for*, *not the ultimate goals of*, *statistical testing in a scientific context*. This is the crux of our difference with Berger, McDermott and Wang and their colleagues.

Classical NP testing theory works well in its original domain, for example, simple point hypotheses (perhaps with nuisance parameters) or interval hypotheses for one-parameter monotone likelihood ratio families. But because the examples in Sections 4, 5 and 7 demonstrate that the NP criterion is unreliable for testing problems outside this domain, the New Tests in Sections 6 and 8 cannot be deemed "superior" merely because they prevail under the NP criterion and α-admissibility.

Having established this, we now concede that for problem (18) of our Section 6, Berger's defense of his test (21) is convincing (see Section 2 of his discussion). Our criticism was too hasty, for (21) does not exhibit unreasonable behavior of type (a), (b) or (c) described seven paragraphs above.[27] As Berger's discussion shows, (21) is the intersection–union test (IUT) for (18) (cf. Berger and Sinclair, 1982; Casella and Berger, 1990; Berger and Hsu, 1996), another generally reasonable approach to hypothesis testing.

The failure of the LRT for (18) is even more dramatic if the dimension is large. Suppose that $X \equiv (X_1, X_2, \ldots, X_p)$ and $\mu \equiv (\mu_1, \mu_2, \ldots, \mu_p)$ in (17) are now p-dimensional and that the subspaces L_1, L_{23} in (18) and (19) are replaced by

$$L_1 \equiv \{\mu \mid \mu_2 = \cdots = \mu_p = 0\},$$
$$L_{2\cdots p} \equiv \{\mu \mid \mu_1 = 0\},$$

respectively, orthogonal linear subspaces of dimensions 1 and $p-1$. The size α LRT and IUT are now given by (20) and (21) with $\chi^2_{2,\alpha}$ replaced by $\chi^2_{p-1,\alpha}$. Now consider a sample point $(X_1, X_2, \ldots, X_p)$ such that $X_2^2 + \cdots + X_p^2 \gg p$, so that we may safely conclude that $\mu \notin L_1$. Then the original testing problem (18) reduces to that of testing $\mu_1 = 0$ versus $\mu_1 \neq 0$. Because X_1 is independent of $(X_2, \ldots, X_p)$, we may base our inference on X_1 alone and reject $\mu_1 = 0$ (i.e., reject H_0) if $X_1^2 > \chi^2_{1,\alpha}$. This agrees with the IUT, whereas the LRT would have us reject $\mu_1 = 0$ if $X_1^2 > \chi^2_{p-1,\alpha}$, which is much too conservative. Thus we see that the *acceptance* region of the LRT includes sample points located (d) inside the alternative while very far from the null hypothesis. We deem this just as unreasonable as the properties (a), (b), (c) of the rejection regions of the New Tests that we have criticized so vigorously.[28]

We reemphasize that Berger's derivation and defense of his IUT (21) [and our criticism of the LRT (20)] rest upon the reasonableness of the IUT approach, not on its superiority under the NP criterion. In fact, we no longer refer to (21) as a New Test, for certainly Berger could construct a Newer Test that is everywhere more powerful than (21) by enlarging its rejection region to include points arbitrarily close to the origin, along the lines of his New Test I for (6) and his New Test for (10), both in our Section 4.

Furthermore, if Berger indeed believes in the IUT approach, then why did he not also adopt that approach for problem (6) in Section 4, where *the IUT coincides with the LRT*, *not with his New Test I or II*?

In fact, Berger consistently ignores his own call for "consistently applied criteria" rather than "personal, subjective criteria." Suppose for a moment that, even in the aftermath of our AD example, Berger maintains his faith in the NP criterion and α-admissibility. Now suppose that he attempts to reconcile his advocacy of the New Tests with his advocacy of union–intersection tests (UIT) in Casella and Berger (1990, page 379):

> Since the LRT is uniformly more powerful than the UIT in Theorem 8.3.4, we might ask why we should use the UIT. One reason is that the UIT has a smaller Type I Error probability [everywhere on the null hypothesis].

If "New Test" is substituted for "LRT" and "LRT" for "UIT," then this statement remains entirely valid, hence denies the very basis of the New Tests. Thus Berger directly contradicts himself.

What do we "likelihoodlums" actually believe about the LRT? In our paper we clearly state our

conviction that the LR criterion is "*a* generally reasonable first option" for hypothesis testing problems but "is not infallible." At the beginning of his discussion, however, Berger says that we "really argue that the LRT is *the* primary option" (emphases added). Taking issue with this statement that we did not make, he goes on to ask why we ignore other generally reasonable first options such as the chi-square, Wald and Rao score test statistics. (We would add the IUT/UIT methods to this list of generally reasonable approaches.) We emphasize that nowhere did we exclude any generally reasonable approach to testing, nor did we intend to do so.

In fact, for any testing problems involving normal distributions with known covariance matrices, such as those in our Sections 4, 5 and 6, the Wald and Rao tests are always identical to the LRT, while for such problems with unknown variance or covariance matrix, such as those in Sections 7 and 8, the three tests closely resemble each other and are dissimilar to the New Tests. Furthermore, when they are applicable, IUT and UIT tests often coincide with (Sections 4 and 7) or closely resemble (Section 8), the LRT rather than the New Tests. By asserting that his New Tests render the LRT inferior, Berger is in fact asserting that they render the Wald, Rao and IUT/UIT tests inferior as well. We doubt that he intended to wield such a blunt instrument.

If we have managed to convince McDermott and Wang that the NP criterion does not represent the primary goal of statistical testing, then their "conditional LRT" for problem (25) in Section 8 has no prima facie advantage over the LRT and the two tests must be compared according to some notion of reasonableness. Focusing on the case where $C = \mathscr{O}$, the nonnegative orthant, MW say in their discussion above:

> When [the dimension] p is large, the fact that *all* of the components of the sample mean vector are positive [i.e., when $K = p$] is itself evidence against H_0 [$\mu = 0$]. It is not so clear that rejecting H_0 for such sample points is an inappropriate decision, even if the sample mean vector is close to the origin.

This seems transparently clear to us, since a sample mean close to the origin simply cannot provide compelling evidence for the alternative hypothesis, but let us examine this issue more closely.

We agree with MW that, in general, conditional anticonservative behavior of a test is not undesirable per se. A notable exception occurs if the conditioning statistic T is ancillary or approximately ancillary (cf. Cox, 1958). In MW's Example, their conditioning statistic $T \equiv \sum_{i=1}^{n-1} X_i$ is highly nonancillary so this exception does not apply and as they properly note, the conditional anticonservative behavior is fully appropriate: large values of T already provide strong evidence for H_1 and the conditional probability of rejecting H_0 is appropriately large in this case.

The situation is similar for the LRT for problem (25) in our Section 8 when Σ is completely known, say $\Sigma = I_p$, and $T = K$, the number of positive components of the projection of X onto the nonnegative orthant $\mathscr{O}$ in $\mathbf{R}^p$. MW perceptively note that when p is large the LRT is also conditionally anticonservative in this case. Again this behavior is not undesirable, because K is highly nonancillary: as MW would assert, the fact that *all* of the components of the sample mean vector are positive (i.e., $K = p$) is itself evidence against H_0, because

$$P_{0, I_p}[K = p] = 1/2^p$$

is very small for large p.

However, when Σ is completely unknown, the case actually treated in Section 8, it is important to note that while K is still not ancillary, it is now much less informative. In this case, MW's assertion that "the fact that *all* of the components of the sample mean vector are positive (i.e., $K = p$) is itself evidence against H_0" is incorrect, because now

$$\sup_{\Sigma} P_{0, \Sigma}[K = p] = 1/2$$

no matter how large p might be. (Apply Lemma 2.1 of Perlman and Wu, 2000a). Thus the conditional anticonservative behavior of MW's test is less defensible in this case.

As with our other examples, however, our main purpose merely was to demonstrate that the rejection region of the MW test contains sample points that, while they do not satisfy (a), (b) or (c) above, are (e) not significantly farther from the origin than would be expected under H_0. In fact, (53) and (54) show that if $n - p$ is large, then the critical value $c_\alpha(v)$ of MW's test can be as small as

$$(n - p + 1)^{-1} \chi^2_{1, 2\alpha} \approx (n - p + 1)^{-1} O(1),$$

whereas it follows from (37) that the critical value c_α of the LRT is no smaller than

$$(n - p + 1)^{-1} \chi^2_{p-1, \alpha} \approx (n - p + 1)^{-1} \left[p + O\left(\sqrt{p}\right) \right]$$

if p is also large. But when $K = p$, the MW and LRT statistics both reduce to the sample Mahalanobis distance $XS^{-1}X'$, whose null distribution

when $n - p$ is large is

$$XS^{-1}X' \sim (n - p + 1)^{-1}\chi_p^2 \approx (n - p + 1)^{-1}\left[p + O\left(\sqrt{p}\right)\right]$$

both unconditionally and conditionally given $K = p$. Thus, for large p and $n - p$, MW's test tells us to reject H_0 for certain sample points that actually fit H_0 extremely well.

By contrast, for the case $\Sigma = I_p$ (completely known) the null distribution of the LRT statistic is a mixture of the chi-square distributions $\chi_0^2 \equiv 0$, $\chi_1^2, \ldots, \chi_p^2$ with weights proportional to the binomial coefficients $\binom{p}{0}, \binom{p}{1}, \cdots, \binom{p}{p}$ hence symmetric about $p/2$, so the critical value is $p/2 + O(\sqrt{p})$. Therefore, unlike the critical value of MW's test as discussed above, this at least increases with p at the proper rate, since now under the null hypothesis,

$$XX' \sim \chi_p^2 \approx p + O\left(\sqrt{p}\right)$$

for large p, both unconditionally and conditionally given $K = p$. Reduction of the critical value from $p + O(\sqrt{p})$ to $p/2 + O(\sqrt{p})$ may be appropriate because, as noted above, occurrence of the event $K = p$ gives more evidence against the null hypothesis in this case.

Returning now to the testing problem (6) in our Section 4, McDermott and Wang, citing Laska and Meisner (1989), join our criticism of the New Tests, based on the nonmonotonicity of the rejection regions (RR). Although we entirely agree with MW, it is of some interest to examine the criterion of monotonicity in a decision-theoretic framework.

When the null and/or alternative hypotheses are not points or linear subspaces, tests that are proper Bayes and therefore d-admissible may not possess monotone acceptance or rejection regions. For problem (6), consider a three-point prior distribution for (μ_1, μ_2) that assigns mass $(\pi, \pi, 1 - 2\pi)$ to $(1, 0) \in H_0$, $(0, 1) \in H_0$, and $(\nu, \nu) \in H_1$, respectively, where $\nu > 0$. Then for $\nu \geq 1$, the RR of the Bayes test is monotonically increasing: if (x_1, x_2) belongs to the RR and $x_1' \geq x_1$, $x_2' \geq x_2$, then (x_1', x_2') also belongs to the RR. For $\nu = 1/2$, however, neither the RR nor the AR is monotone; the AR is a symmetric strip centered about the line $x_1 = x_2$ and the RR is its complement. Furthermore, as $\nu \downarrow 0$ it is the AR, not the RR, that approaches a monotonically increasing limit. Thus a monotonically increasing RR is desirable only under the *additional* assumption (probably justified) that (μ_1, μ_2) tends to assume larger values under H_1 than under H_0, but this is not necessarily specified by (6) alone.

In conclusion, we have no Grand Unified Testing Theory to present, no universal mathematical criterion for judging all tests in all testing problems as Berger asks of us, and think it unlikely that such a criterion exists. Depending on one's point of view, this nonexistence may be seen as either a strength or a weakness of statistics, but in any event it is better to have no universal criterion than cling to an inappropriate one. We hope that we have alerted statisticians to the dangers inherent in uncritical application of the NP criterion, and, more generally, convinced them to join Fisher, Cox and many others in carefully weighing the scientific relevance and logical consistency of any mathematical criterion proposed for statistical theory.

ADDITIONAL REFERENCES

BARTHOLOMEW, D. J. (1961). A test of homogeneity of means under restricted alternatives (with discussion). *J. Roy. Statist. Soc. Ser. B* **23** 239–281.

BASU, D. (1964). Recovery of ancillary information. *Sankhyā* **26** 3–16.

BERGER, R. L. (1997). Likelihood ratio tests and intersection-union tests. In *Advances in Statistical Decision Theory and Applications* (S. Panchapakesan and N. Balakrishnan, eds.) 225–237. Birkhäuser, Boston.

CASELLA, G. and BERGER, R. L. (1990). *Statistical Inference*. Duxbury, Belmont, CA.

COX, D. R. (1958). Some problems connected with statistical inference. *Ann. Math. Statist.* **29** 357–372.

COX, D. R. (1977). The role of significance tests (with discussion). *Scand. J. Statist.* **4** 49–70.

COX, D. R. and HINKLEY, D. V. (1974). *Theoretical Statistics*. Chapman and Hall, London.

FISHER, R. A. (1956). *Statistical Methods and Scientific Inference*. Oliver and Boyd, Edinburgh.

GARDNER, M. J. and ALTMAN, D. G. (1986). Confidence intervals rather than p-values: estimation rather than hypothesis testing. *British Med. J.* **292** 746–750.

KUDÔ, A. (1963). A multivariate analogue of the one-sided test. *Biometrika* **50** 403–418.

SAIKALI, K. G. and BERGER, R. L. (1998). More powerful tests for the sign testing problem. Technical Report 2511, Inst. Statist., North Carolina State Univ.

STANDARDS OF REPORTING TRIALS GROUP. (1994). A proposal for structured reporting of randomized controlled trials. *J. Amer. Med. Assoc.* **272** 1926–1931.