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).