Comment

Thomas J. Santner

I thank the authors for their detailed look at a well-studied problem. For the Wald binomial p interval, there has not been an appreciation of the long persistence (in n) of p locations having substantially deficient achieved coverage compared with the nominal coverage. Figure 1 is indeed a picture that says a thousand words. Similarly, the asymptotic lower limit in Theorem 1 for the minimum coverage of the Wald interval is an extremely useful analytic tool to explain this phenomenon, although other authors have given fixed p approximations of the coverage probability of the Wald interval (e.g., Theorem 1 of Ghosh, 1979).

My first set of comments concern the specific binomial problem that the authors address and then the implications of their work for other important discrete data confidence interval problems.

The results in Ghosh (1979) complement the calculations of Brown, Cai and DasGupta (BCD) by pointing out that the Wald interval is "too long" in addition to being centered at the "wrong" value (the MLE as opposed to a Bayesian point estimate such is used by the Agresti–Coull interval). His Table 3 lists the probability that the Wald interval is longer than the Wilson interval for a central set of p values (from 0.20 to 0.80) and a range of sample sizes n from 20 to 200. Perhaps surprisingly, in view of its inferior coverage characteristics, the Wald interval tends to be *longer* than the Wilson interval with very high probability. Hence the Wald interval is both too long and centered at the wrong place. This is a dramatic effect of the skewness that BCD mention.

When discussing any system of intervals, one is concerned with the consistency of the answers given by the interval across multiple uses by a single researcher or by groups of users. Formally, this is the reason why various symmetry properties are required of confidence intervals. For example, in the present case, requiring that the p interval $(L(X), U(X))$ satisfy the symmetry property

$$(1)\quad (L(x), U(x)) = (1 - L(n - x), 1 - U(n - x))$$

for $x \in \{0, \ldots, n\}$ shows that investigators who reverse their definitions of success and failure will be consistent in their assessment of the likely values for p. Symmetry (1) is the minimal requirement of a binomial confidence interval. The Wilson and equal-tailed Jeffrey intervals advocated by BCD satisfy the symmetry property (1) and have coverage that is centered (when coverage is plotted versus true p) about the nominal value. They are also straightforward to motivate, even for elementary students, and simple to compute for the outcome of interest.

However, regarding p confidence intervals as the inversion of a family of acceptance regions corresponding to size α tests of H_0: $p = p_0$ versus H_A: $p \neq p_0$ for $0 < p_0 < 1$ has some substantial advantages. Indeed, Brown et al. mention this inversion technique when they remark on the desirable properties of intervals formed by inverting likelihood ratio test acceptance regions of H_0 versus H_A. In the binomial case, the acceptance region of any reasonable test of H_0: $p = p_0$ is of the form $\{L_{p_0}, \ldots, U_{p_0}\}$. These acceptance regions invert to intervals if and only if L_{p_0} and U_{p_0} are nondecreasing in p_0 (otherwise the inverted p confidence set can be a union of intervals). Of course, there are many families of size α tests that meet this nondecreasing criterion for inversion, including the very conservative test used by Clopper and Pearson (1934). For the binomial problem, Blyth and Still (1983) constructed a set of confidence intervals by selecting among size α acceptance regions those that possessed additional symmetry properties and were "small" (leading to short confidence intervals). For example, they desired that the interval should "move to the right" as x increases when n is fixed and should "move the left" as n increases when x is fixed. They also asked that their system of intervals increase monotonically in the coverage probability for fixed x and n in the sense that the higher nominal coverage interval *contain* the lower nominal coverage interval.

In addition to being less intuitive to unsophisticated statistical consumers, systems of confidence intervals formed by inversion of acceptance regions also have two other handicaps that have hindered their rise in popularity. First, they typically require that the confidence interval (essentially) be constructed for *all* possible outcomes, rather than merely the response of interest. Second, their rather brute force character means that a specialized computer program must be written to produce the acceptance sets and their inversion (the intervals).

Thomas J. Santner is Profesor, Ohio State University, 404 Cockins Hall, 1958 Neil Avenue, Columbus, Ohio 43210 (e-mail: tjs@stat.ohio-state.edu).

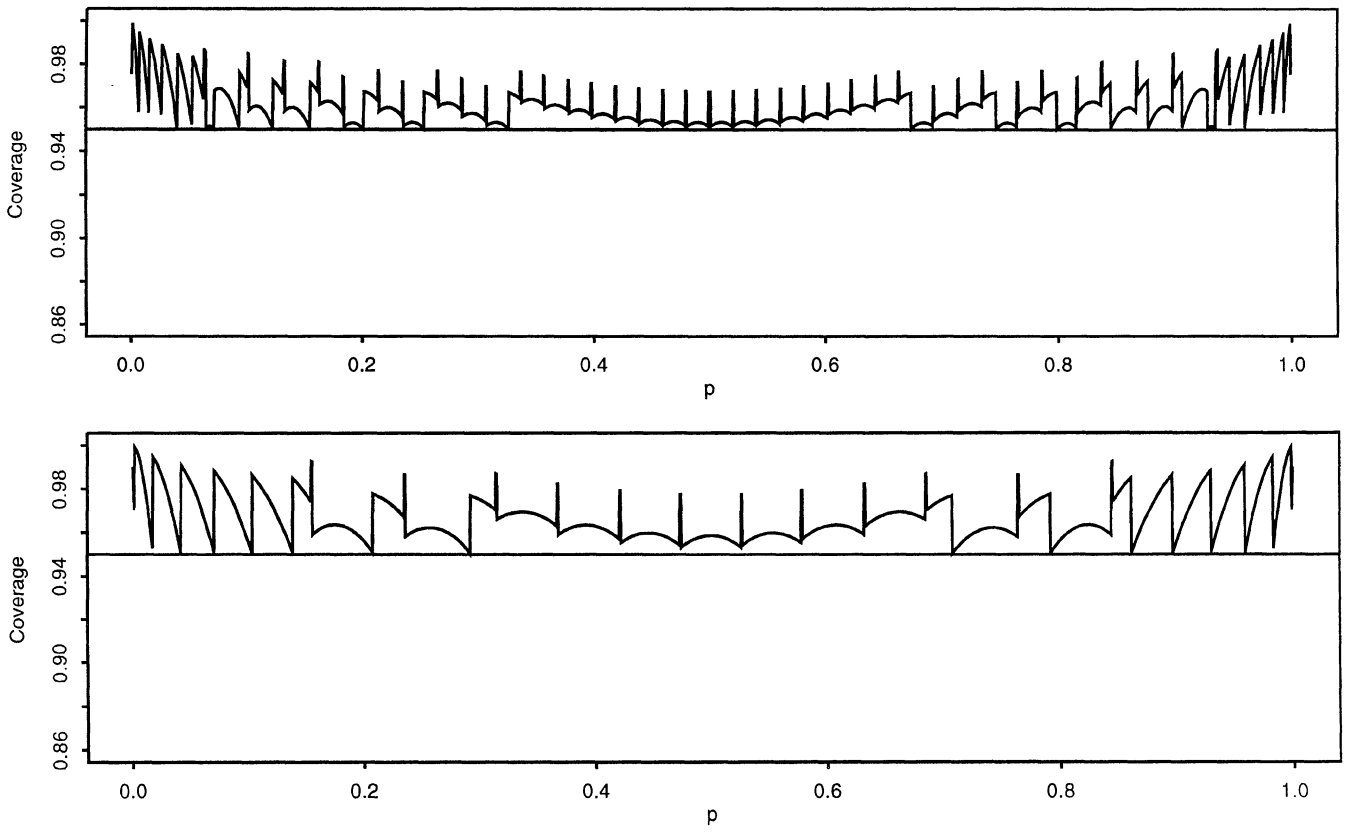

FIG. 1. *Coverage of nominal 95% symmetric Duffy–Santner p intervals for $n = 20$ (bottom panel) and $n = 50$ (top panel).*

However, the benefits of having reasonably short and suitably symmetric confidence intervals are sufficient that such intervals have been constructed for several frequently occurring problems of biostatistics. For example, Jennison and Turnbull (1983) and Duffy and Santner (1987) present acceptance set–inversion confidence intervals (both with available FORTRAN programs to implement their methods) for a binomial p based on data from a multistage clinical trial; Coe and Tamhane (1989) describe a more sophisticated set of repeated confidence intervals for $p_1 - p_2$ also based on multistage clinical trial data (and give a SAS macro to produce the intervals). Yamagami and Santner (1990) present an acceptance set–inversion confidence interval and FORTRAN program for $p_1 - p_2$ in the two-sample binomial problem. There are other examples.

To contrast with the intervals whose coverages are displayed in BCD's Figure 5 for $n = 20$ and $n = 50$, I formed the multistage intervals of Duffy and Santner that strictly attain the nominal confidence level for all p. The computation was done naively in the sense that the multistage FORTRAN program by Duffy that implements this method was applied using one stage with stopping boundaries arbitrarily set at $(a, b) = (0, 1)$ in the notation of Duffy and Santner, and a small adjustment was made to insure symmetry property (1). (The nonsymmetrical multiple stage stopping boundaries that produce the data considered in Duffy and Santner do not impose symmetry.) The coverages of these systems are shown in Figure 1. To give an idea of computing time, the $n = 50$ intervals required less than two seconds to compute on my 400 Mhz PC. To further facilitate comparison with the intervals whose coverage is displayed in Figure 5 of BCD, I computed the Duffy and Santner intervals for a slightly lower level of coverage, 93.5%, so that the *average* coverage was about the desired 95% nominal level; the coverage of this system is displayed in Figure 2 on the same vertical scale and compares favorably. It is possible to call the FORTRAN program that makes these intervals within SPLUS which makes for convenient data analysis.

I wish to mention that are a number of other small sample interval estimation problems of continuing interest to biostatisticians that may well have very reasonable small sample solutions based on analogs of the methods that BCD recommend.

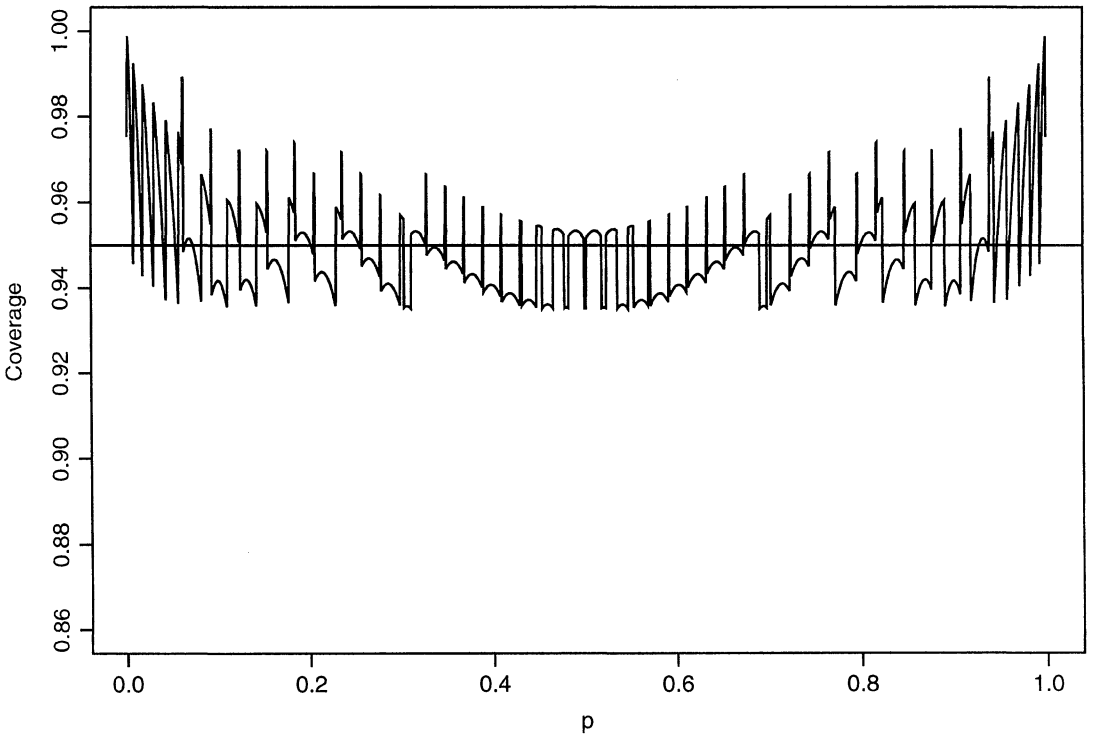

FIG. 2. *Coverage of nominal* 93.5% *symmetric Duffy–Santner* p *intervals for* $n = 50$.

Most of these would be extremely difficult to handle by the more brute force method of inverting acceptance sets. The first of these is the problem of computing simultaneous confidence intervals for $p_0 - p_i, 1 \leq i \leq T$ that arises in comparing a control binomial distribution with T treatment ones. The second concerns forming simultaneous confidence intervals for $p_i - p_j$, the cell probabilities of a multinomial distribution. In particular, the equal-tailed Jeffrey prior approach recommended by the author has strong appeal for both of these problems.

Finally, I note that the Wilson intervals seem to have received some recommendation as the method of choice in other elementary texts. In his introductory texts, Larson (1974) introduces the Wilson interval as the method of choice although he makes the vague, and indeed false, statement, as BCD show, that the user can use the Wald interval if "n is large enough." One reviewer of Santner (1998), an article that showed the coverage virtues of the Wilson interval compared with Wald-like intervals advocated by another author in the magazine *Teaching Statistics* (written for high school teachers) commented that the Wilson method was the "standard" method taught in the U.K.

Rejoinder

Lawrence D. Brown, T. Tony Cai and Anirban DasGupta

We deeply appreciate the many thoughtful and constructive remarks and suggestions made by the discussants of this paper. The discussion suggests that we were able to make a convincing case that the often-used Wald interval is far more problematic than previously believed. We are happy to see a consensus that the Wald interval deserves to be discarded, as we have recommended. It is not surprising to us to see disagreement over the specific alternative(s) to be recommended in place of