lower bound is 0 and for $X = 1$ the lower bound is $1 - (1 - \alpha)^{1/n}$ [the solution to $P(X = 0) = 1 - \alpha$].

3. USE YOUR TOOLS

The essential message that I take away from the work of BCD is that an approximate/formula-based approach to constructing a binomial confidence interval is bound to have essential flaws. However, this is a situation where brute force computing will do the trick. The construction of a $1 - \alpha$ binomial confidence interval is a discrete optimization problem that is easily programmed. So why not use the tools that we have available? If the problem will yield to brute force computation, then we should use that solution.

Blyth and Still (1983) showed how to compute exact intervals through numerical inversion of tests, and Casella (1986) showed how to compute exact intervals by refining conservative intervals. So for any value of n and α, we can compute an exact, shortest $1 - \alpha$ confidence interval that will not display any of the pathological behavior illustrated by BCD. As an example, Figure 1 shows the Agresti–Coull interval along with the Blyth–Still interval for $n = 100$ and $1 - \alpha = 0.95$. While the Agresti–Coull interval fails to maintain 0.95 coverage in the middle p region, the Blyth–Still interval always maintains 0.95 coverage. What is more surprising, however, is that the Blyth–Still interval displays much less variation in its coverage probability, especially near the endpoints. Thus, the simplistic numerical algorithm produces an excellent interval, one that both maintains its guaranteed coverage and reduces oscillation in the coverage probabilities.

ACKNOWLEDGMENT

Supported by NSF Grant DMS-99-71586.

Comment

Chris Corcoran and Cyrus Mehta

We thank the authors for a very accessible and thorough discussion of this practical problem. With the availability of modern computational tools, we have an unprecedented opportunity to carefully evaluate standard statistical procedures in this manner. The results of such work are invaluable to teachers and practitioners of statistics everywhere. We particularly appreciate the attention paid by the authors to the generally oversimplified and inadequate recommendations made by statistical texts regarding when to use normal approximations in analyzing binary data. As their work has plainly shown, even in the simple case of a single binomial proportion, the discreteness of the data makes the use of some asymptotic procedures tenuous, even when the underlying probability lies away from the boundary or when the sample size is relatively large.

The authors have evaluated various confidence intervals with respect to their coverage properties and average lengths. Implicit in their evaluation is the premise that overcoverage is just as bad as undercoverage. We disagree with the authors on this fundamental issue. If, because of the discreteness of the test statistic, the desired confidence level cannot be attained, one would ordinarily prefer overcoverage to undercoverage. Wouldn't you prefer to hire a fortune teller whose track record exceeds expectations to one whose track record is unable to live up to its claim of accuracy? With the exception of the Clopper–Pearson interval, none of the intervals discussed by the authors lives up to its claim of 95% accuracy throughout the range of p. Yet the authors dismiss this interval on the grounds that it is "wastefully conservative." Perhaps so, but they do not address the issue of how the wastefulness is manifested.

What penalty do we incur for furnishing confidence intervals that are more truthful than was required of them? Presumably we pay for the conservatism by an increase in the length of the confidence interval. We thought it would be a useful exercise

Chris Corcoran is Assistant Professor, Department of Mathematics and Statistics, Utah State University, 3900 old Main Hill, Logon, Utah, 84322-3900 (e-mail: corcoran@math.usu.edu). Cyrus Mehta is Professor, Department of Biostatistics, Harvard School of Public Health, 655 Huntington Avenue Boston, Massachusetts 02115 and is with Cytel Software Corporation, 675 Massachusetts Avenue, Cambridge, Massachusetts 02319.

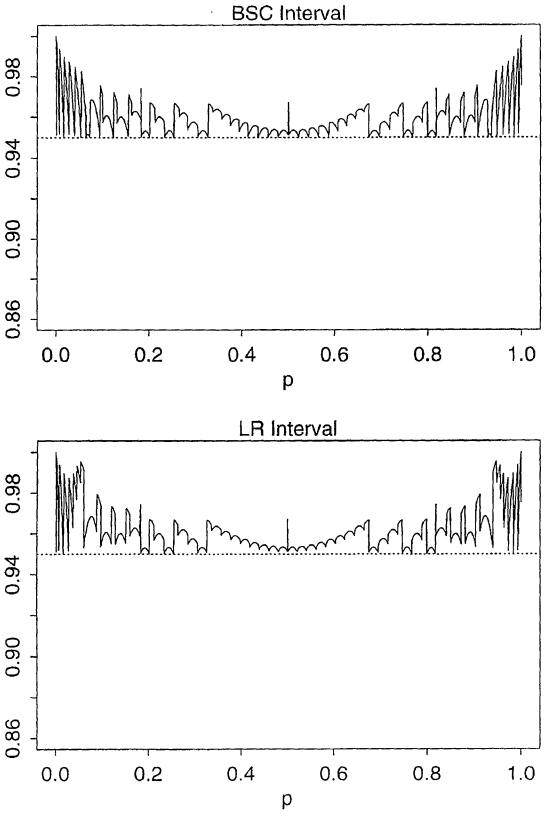

FIG. 1. *Actual coverage probabilities for BSC and LR intervals as a function of* $p(n = 50)$. *Compare to author's Figures* 5, 10 *and* 11.

to actually investigate the magnitude of this penalty for two confidence interval procedures that are guaranteed to provide the desired coverage but are not as conservative as Clopper–Pearson. Figure 1 displays the true coverage probabilities for the nominal 95% Blyth–Still–Casella (see Blyth and Still, 1983; Casella, 1984) confidence interval (BSC interval) and the 95% confidence interval obtained by inverting the exact likelihood ratio test (LR interval; the inversion follows that shown by Aitken, Anderson, Francis and Hinde, 1989, pages 112–118).

There is no value of p for which the coverage of the BSC and LR intervals falls below 95%. Their coverage probabilities are, however, much closer to 95% than would be obtained by the Clopper–Pearson procedure, as is evident from the authors' Figure 11. Thus one could say that these two intervals are uniformly better than the Clopper–Pearson interval.

We next investigate the penalty to be paid for the guaranteed coverage in terms of increased length of the BSC and LR intervals relative to the Wilson, Agresti–Coull, or Jeffreys intervals recommended by the authors. This is shown by Figure 2.

In fact the BSC and LR intervals are actually shorter than Agresti–Coull for $p < 0.2$ or $p > 0.8$, and shorter than the Wilson interval for $p < 0.1$ and $p > 0.9$. The only interval that is uniformly shorter than BSC and LR is the Jeffreys interval. Most of the time the difference in lengths is negligible, and in the worst case (at $p = 0.5$) the Jeffreys interval is only shorter by 0.025 units. Of the three asymptotic methods recommended by the authors, the Jeffreys interval yields the lowest average probability of coverage, with significantly greater potential relative undercoverage in the $(0.05, 0.20)$ and $(0.80, 0.95)$ regions of the parameter space. Considering this, one must question the rationale for preferring Jeffreys to either BSC or LR.

The authors argue for simplicity and ease of computation. This argument is valid for the teaching of statistics, where the instructor must balance simplicity with accuracy. As the authors point out, it is customary to teach the standard interval in introductory courses because the formula is straightforward and the central limit theorem provides a good heuristic for motivating the normal approximation. However, the evidence shows that the standard method is woefully inadequate. Teaching statistical novices about a Clopper–Pearson type interval is conceptually difficult, particularly because exact intervals are impossible to compute by hand. As the Agresti–Coull interval preserves the confidence level most successfully among the three recommended alternative intervals, we believe that this feature when coupled with its straightforward computation (particularly when $\alpha = 0.05$) makes this approach ideal for the classroom.

Simplicity and ease of computation have no role to play in statistical practice. With the advent of powerful microcomputers, researchers no longer resort to hand calculations when analyzing data. While the need for simplicity applies to the classroom, in applications we primarily desire reliable, accurate solutions, as there is no significant difference in the computational overhead required by the authors' recommended intervals when compared to the BSC and LR methods. From this perspective, the BSC and LR intervals have a substantial advantage relative to the various asymptotic intervals presented by the authors. They guarantee coverage at a relatively low cost in increased length. In fact, the BSC interval is already implemented in StatXact (1998) and is therefore readily accessible to practitioners.

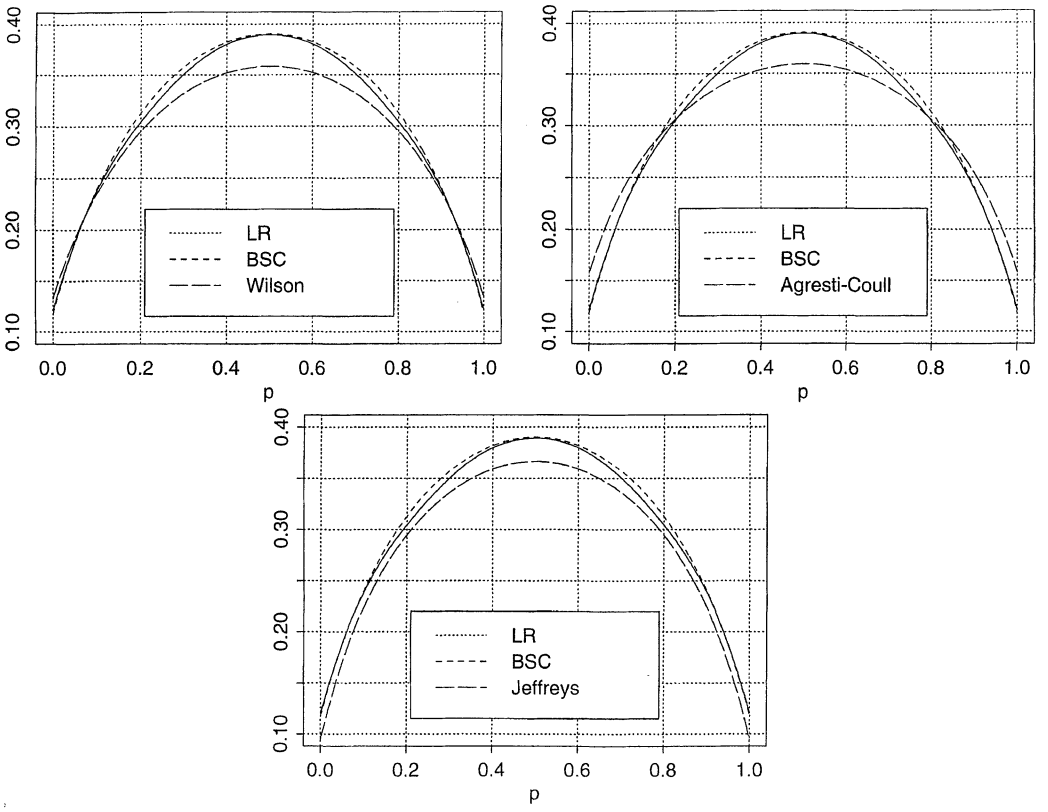

FIG. 2. *Expected lengths of BSC and LR intervals as a function of p compared, respectively, to Wilson, Agresti–Coull and Jeffreys intervals* ($n = 25$). *Compare to authors' Figure* 8.

Comment

Malay Ghosh

This is indeed a very valuable article which brings out very clearly some of the inherent difficulties associated with confidence intervals for parameters of interest in discrete distributions. Professors Brown, Cai and Dasgupta (henceforth BCD) are to be complimented for their comprehensive and thought-provoking discussion about the "chaotic" behavior of the Wald interval for the binomial proportion and an appraisal of some of the alternatives that have been proposed.

My remarks will primarily be confined to the discussion of Bayesian methods introduced in this paper. BCD have demonstrated very clearly that the modified Jeffreys equal-tailed interval works well in this problem and recommend it as a possible contender to the Wilson interval for $n \leq 40$.

There is a deep-rooted optimality associated with Jeffreys prior as the unique *first-order probability matching prior* for a real-valued parameter of interest with no nuisance parameter. Roughly speaking, a probability matching prior for a real-valued parameter is one for which the coverage probability of a one-sided Baysian credible interval is asymptotically equal to its frequentist counterpart. Before giving a formal definition of such priors, we provide an intuitive explanation of why Jeffreys prior is a matching prior. To this end, we begin with the fact that if $X_1, \ldots, X_n$ are iid $N(\theta, 1)$, then $\overline{X}_n = \sum_{i=1}^n X_i/n$ is the MLE of θ. With the uniform prior $\pi(\theta) \propto c$ (a constant), the posterior of θ

Malay Ghosh is Distinguished Professor, Department of Statistics, University of Florida, Gainesville, Florida 32611-8545 (e-mail: ghoshm@stat.ufl.edu).