more fair comparison of expected lengths uses the modified versions CI_{M-J} and CI_{M-W}. We checked this but must admit that figures analogous to the BCD Figures 8 and 9 show that CI_{M-J} and CI_{M-W} maintain their expected length advantage over CI_{AC}, although it is reduced somewhat.

However, when n decreases below 10, the results change, with CI_{M-J} having greater expected width than CI_{AC} and CI_{M-W}. Our Figure 1 extends the BCD Figure 9 to values of $n < 10$, showing how the comparison differs between the ordinary intervals and the modified ones. Our Figure 2 has the format of the BCD Figure 8, but for $n = 5$ instead of 25. Admittedly, $n = 5$ is a rather extreme case, one for which the Jeffreys interval is modified unless $x = 2$ or 3 and the Wilson interval is modified unless $x = 0$ or 5, and for it CI_{AC} has coverage probabilities that can dip below 0.90. Thus, overall, the BCD recommendations about choice of method seem reasonable to us. Our own preference is to use the Wilson interval for statistical practice and CI_{AC} for teaching in elementary statistics courses.

4. EXTENSIONS

Other than near-boundary modifications, another type of fine-tuning that may help is to invert a test permitting unequal tail probabilities. This occurs naturally in exact inference that inverts a single two-tailed test, which can perform better than inverting two separate one-tailed tests (e.g., Sterne, 1954; Blyth and Still, 1983).

Finally, we are curious about the implications of the BCD results in a more general setting. How much does their message about the effects of discreteness and basing interval estimation on the Jeffreys prior or the score test rather than the Wald test extend to parameters in other discrete distributions and to two-sample comparisons? We have seen that interval estimation of the Poisson parameter benefits from inverting the score test rather than the Wald test on the count scale (Agresti and Coull, 1998).

One would not think there could be anything new to say about the Wald confidence interval for a proportion, an inferential method that must be one of the most frequently used since Laplace (1812, page 283). Likewise, the confidence interval for a proportion based on the Jeffreys prior has received attention in various forms for some time. For instance, R. A. Fisher (1956, pages 63–70) showed the similarity of a Bayesian analysis with Jeffreys prior to his fiducial approach, in a discussion that was generally critical of the confidence interval method but grudgingly admitted of limits obtained by a test inversion such as the Clopper–Pearson method, "though they fall short in logical content of the limits found by the fiducial argument, and with which they have often been confused, they do fulfil some of the desiderata of statistical inferences." Congratulations to the authors for brilliantly casting new light on the performance of these old and established methods.

Comment

George Casella

1. INTRODUCTION

Professors Brown, Cai and DasGupta (BCD) are to be congratulated for their clear and imaginative look at a seemingly timeless problem. The chaotic behavior of coverage probabilities of discrete confidence sets has always been an annoyance, resulting in intervals whose coverage probability can be vastly different from their nominal confidence level. What we now see is that for the Wald interval, an approximate interval, the chaotic behavior is relentless, as this interval will not maintain $1 - \alpha$ coverage for any value of n. Although fixes relying on ad hoc rules abound, they do not solve this fundamental defect of the Wald interval and, surprisingly, the usual safety net of asymptotics is also shown not to exist. So, as the song goes, "Bye-bye, so long, farewell" to the Wald interval.

Now that the Wald interval is out, what is in? There are probably two answers here, depending on whether one is in the classroom or the consulting room.

George Casella is Arun Varma Commemorative Term Professor and Chair, Department of Statistics, University of Florida, Gainesville, Florida 32611-8545 (e-mail: casella@stat.ufl.edu).

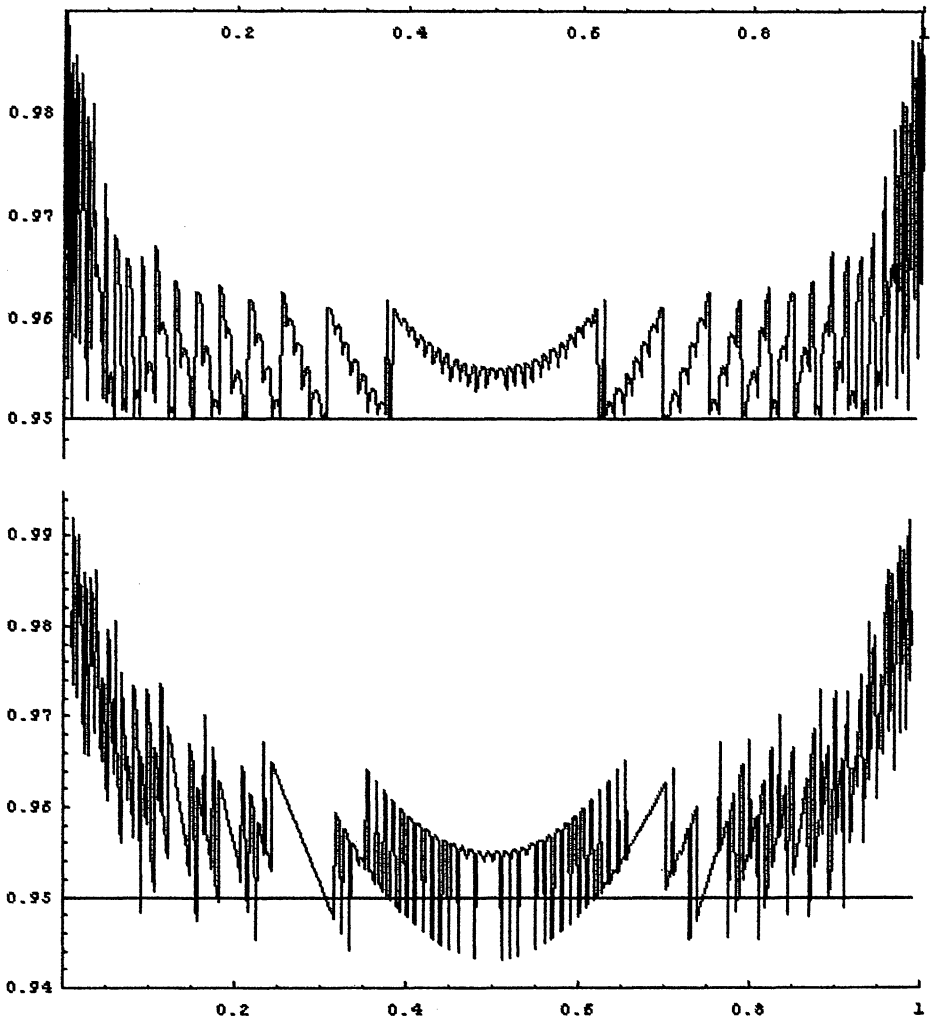

FIG. 1. *Coverage probabilities of the Blyth-Still interval (upper) and Agresti-Coull interval (lower) for* $n = 100$ *and* $1 - \alpha = 0.95$.

2. WHEN YOU SAY 95%...

In the classroom it is (still) valuable to have a formula for a confidence intervals, and I typically present the Wilson/score interval, starting from the test statistic formulation. Although this doesn't have the pleasing $\hat{p} \pm$ *something*, most students can understand the logic of test inversion. Moreover, the fact that the interval does not have a symmetric form is a valuable lesson in itself; the statistical world is not always symmetric.

However, one thing still bothers me about this interval. It is clearly not a $1 - \alpha$ interval; that is, it does not maintain its nominal coverage probability. This is a defect, and one that should not be compromised. I am uncomfortable in presenting a confidence interval that does not maintain its stated confidence; when you say 95% you should mean 95%!

But the fix here is rather simple: apply the "continuity correction" to the score interval (a technique that seems to be out of favor for reasons I do not understand). The continuity correction is easy to justify in the classroom using pictures of the normal density overlaid on the binomial mass function, and the resulting interval will now maintain its nominal level. (This last statement is not based on analytic proof, but on numerical studies.) Anyone reading Blyth (1986) cannot help being convinced that this is an excellent approximation, coming at only a slightly increased effort.

One other point that Blyth makes, which BCD do not mention, is that it is easy to get exact confidence limits at the endpoints. That is, for $X = 0$ the

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.