ACKNOWLEDGMENTS

We thank Xuefeng Li for performing some helpful computations and Jim Berger, David Moore, Steve Samuels, Bill Studden and Ron Thisted for useful conversations. We also thank the Editors and two anonymous referees for their thorough and constructive comments. Supported by grants from the National Science Foundation and the National Security Agency.

REFERENCES

ABRAMOWITZ, M. and STEGUN, I. A. (1970). *Handbook of Mathematical Functions.* Dover, New York.

AGRESTI, A. and COULL, B. A. (1998). Approximate is better than "exact" for interval estimation of binomial proportions. *Amer. Statist.* **52** 119–126.

ANSCOMBE, F. J. (1948). The transformation of Poisson, binomial and negative binomial data. *Biometrika* **35** 246–254.

ANSCOMBE, F. J. (1956). On estimating binomial response relations. *Biometrika* **43** 461–464.

BERGER, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*, 2nd ed. Springer, New York.

BERRY, D. A. (1996). *Statistics: A Bayesian Perspective.* Wadsworth, Belmont, CA.

BICKEL, P. and DOKSUM, K. (1977). *Mathematical Statistics.* Prentice-Hall, Englewood Cliffs, NJ.

BLYTH, C. R. and STILL, H. A. (1983). Binomial confidence intervals. *J. Amer. Statist. Assoc.* **78** 108–116.

BROWN, L. D., CAI, T. and DASGUPTA, A. (1999). Confidence intervals for a binomial proportion and asymptotic expansions. *Ann. Statist* to appear.

BROWN, L. D., CAI, T. and DASGUPTA, A. (2000). Interval estimation in discrete exponential family. Technical report, Dept. Statistics. Univ. Pennsylvania.

CASELLA, G. (1986). Refining binomial confidence intervals *Canad. J. Statist.* **14** 113–129.

CASELLA, G. and BERGER, R. L. (1990). *Statistical Inference.* Wadsworth & Brooks/Cole, Belmont, CA.

CLOPPER, C. J. and PEARSON, E. S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika* **26** 404–413.

COX, D. R. and SNELL, E. J. (1989). *Analysis of Binary Data*, 2nd ed. Chapman and Hall, London.

CRESSIE, N. (1980). A finely tuned continuity correction. *Ann. Inst. Statist. Math.* **30** 435–442.

GHOSH, B. K. (1979). A comparison of some approximate confidence intervals for the binomial parameter *J. Amer. Statist. Assoc.* **74** 894–900.

HALL, P. (1982). Improving the normal approximation when constructing one-sided confidence intervals for binomial or Poisson parameters. *Biometrika* **69** 647–652.

LEHMANN, E. L. (1999). *Elements of Large-Sample Theory.* Springer, New York.

NEWCOMBE, R. G. (1998). Two-sided confidence intervals for the single proportion; comparison of several methods. *Statistics in Medicine* **17** 857–872.

RAO, C. R. (1973). *Linear Statistical Inference and Its Applications.* Wiley, New York.

SAMUELS, M. L. and WITMER, J. W. (1999). *Statistics for the Life Sciences*, 2nd ed. Prentice Hall, Englewood Cliffs, NJ.

SANTNER, T. J. (1998). A note on teaching binomial confidence intervals. *Teaching Statistics* **20** 20–23.

SANTNER, T. J. and DUFFY, D. E. (1989). *The Statistical Analysis of Discrete Data.* Springer, Berlin.

STONE, C. J. (1995). *A Course in Probability and Statistics.* Duxbury, Belmont, CA.

STRAWDERMAN, R. L. and WELLS, M. T. (1998). Approximately exact inference for the common odds ratio in several 2×2 tables (with discussion). *J. Amer. Statist. Assoc.* **93** 1294–1320.

TAMHANE, A. C. and DUNLOP, D. D. (2000). *Statistics and Data Analysis from Elementary to Intermediate.* Prentice Hall, Englewood Cliffs, NJ.

VOLLSET, S. E. (1993). Confidence intervals for a binomial proportion. *Statistics in Medicine* **12** 809–824.

WASSERMAN, L. (1991). An inferential interpretation of default priors. Technical report, Carnegie-Mellon Univ.

WILSON, E. B. (1927). Probable inference, the law of succession, and statistical inference. *J. Amer. Statist. Assoc.* **22** 209–212.

Comment

Alan Agresti and Brent A. Coull

In this very interesting article, Professors Brown, Cai and DasGupta (BCD) have shown that discreteness can cause havoc for much larger sample sizes that one would expect. The popular (Wald) confidence interval for a binomial parameter p has been known for some time to behave poorly, but readers will surely be surprised that this can happen for such large n values.

Interval estimation of a binomial parameter is deceptively simple, as there are not even any nuisance parameters. The gold standard would seem to be a method such as the Clopper–Pearson, based on inverting an "exact" test using the binomial dis-

Alan Agresti is Distinguished Professor, Department of Statistics, University of Florida, Gainesville, Florida 32611-8545 (*e-mail: aa@stat.ufl.edu*). *Brent A. Coull is Assistant Professor, Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts* 02115 (*e-mail: bcoull@hsph.harvard.edu*).

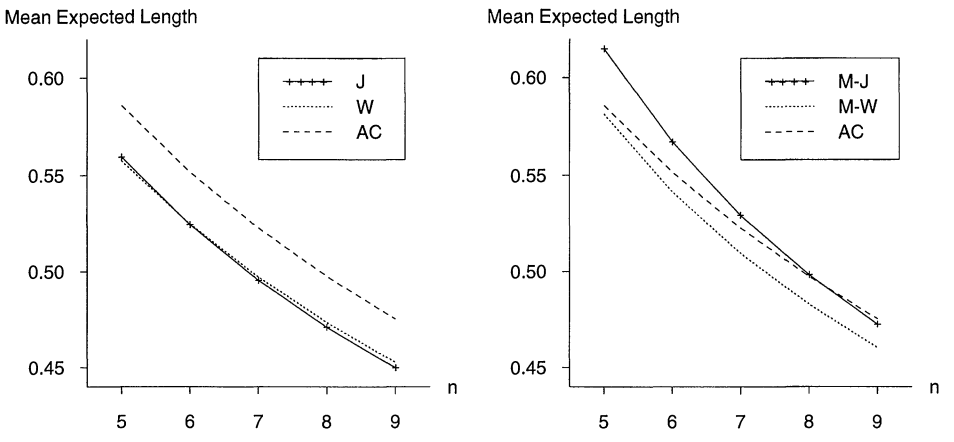

FIG. 1. *A Comparison of mean expected lengths for the nominal 95% Jeffreys (J), Wilson (W), Modified Jeffreys (M-J), Modified Wilson (M-W), and Agresti–Coull (AC) intervals for* $n = 5, 6, 7, 8, 9$.

tribution rather than an approximate test using the normal. Because of discreteness, however, this method is too conservative. A more practical, nearly gold standard for this and other discrete problems seems to be based on inverting a two-sided test using the exact distribution but with the *mid-P value*. Similarly, with large-sample methods it is better *not* to use a continuity correction, as otherwise it approximates exact inference based on an ordinary P-value, resulting in conservative behavior. Interestingly, BCD note that the Jeffreys interval (CI_J) approximates the mid-P value correction of the Clopper–Pearson interval. See Gart (1966) for related remarks about the use of $\frac{1}{2}$ additions to numbers of successes and failures before using frequentist methods.

1. METHODS FOR ELEMENTARY STATISTICS COURSES

It's unfortunate that the Wald interval for p is so seriously deficient, because in addition to being the simplest interval it is the obvious one to teach in elementary statistics courses. By contrast, the Wilson interval (CI_W) performs surprisingly well even for small n. Since it is too complex for many such courses, however, our motivation for the "Agresti–Coull interval" (CI_{AC}) was to provide a simple approximation for CI_W. Formula (4) in BCD shows that the midpoint $\tilde{p}$ for CI_W is a weighted average of $\hat{p}$ and 1/2 that equals the sample proportion after adding $z^2_{\alpha/2}$ pseudo observations, half of each type; the square of the coefficient of $z_{\alpha/2}$ is the same weighted average of the variance of a sample proportion when $p = \hat{p}$ and when $p = 1/2$, using $\tilde{n} = n + z^2_{\alpha/2}$ in place of n. The CI_{AC} uses the CI_W midpoint, but its squared coefficient of $z_{\alpha/2}$ is the variance $\tilde{p}\tilde{q}/\tilde{n}$ at the weighted average $\tilde{p}$ rather than the weighted average of the variances. The resulting interval $\tilde{p} \pm z_{\alpha/2}(\tilde{p}\tilde{q}/\tilde{n})^{1/2}$ is wider than CI_W (by Jensen's inequality), in particular being conservative for p near 0 and 1 where CI_W can suffer poor coverage probabilities.

Regarding textbook qualifications on sample size for using the Wald interval, skewness considerations and the Edgeworth expansion suggest that guidelines for n should depend on p through $(1 - 2p)^2/[p(1-p)]$. See, for instance, Boos and Hughes-Oliver (2000). But this does not account for the effects of discreteness, and as BCD point out, guidelines in terms of p are not verifiable. For elementary course teaching there is no obvious alternative (such as t methods) for smaller n, so we think it is sensible to teach a single method that behaves reasonably well for all n, as do the Wilson, Jeffreys and Agresti–Coull intervals.

2. IMPROVED PERFORMANCE WITH BOUNDARY MODIFICATIONS

BCD showed that one can improve the behavior of the Wilson and Jeffreys intervals for p near 0 and 1 by modifying the endpoints for CI_W when $x = 1, 2, n-2, n-1$ (and $x = 3$ and $n-3$ for $n > 50$) and for CI_J when $x = 0, 1, n-1, n$. Once one permits the modification of methods near the sample space boundary, other methods may perform decently besides the three recommended in this article.

For instance, Newcombe (1998) showed that when $0 < x < n$ the Wilson interval CI_W and the Wald logit interval have the same midpoint on the logit scale. In fact, Newcombe has shown (personal communication, 1999) that the logit interval necessarily

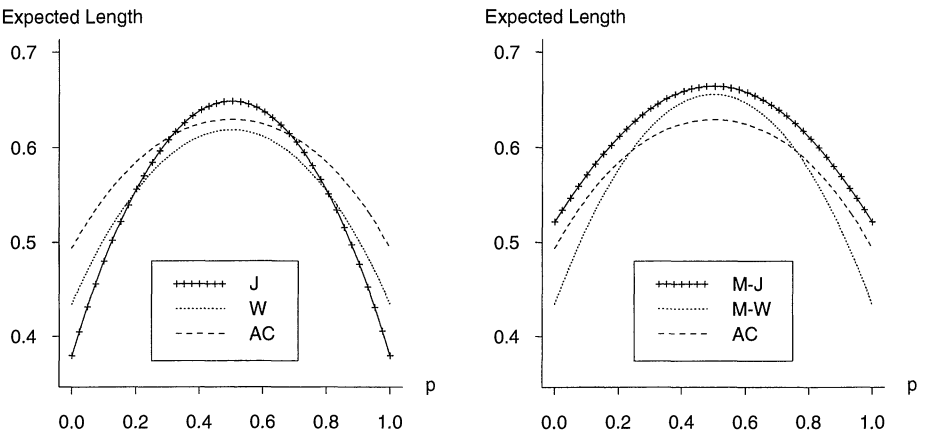

FIG. 2. *A comparison of expected lengths for the nominal 95% Jeffreys (J), Wilson (W), Modified Jeffreys (M-J), Modified Wilson (M-W), and Agresti–Coull (AC) intervals for* $n = 5$.

contains CI_W. The logit interval is the uninformative one $[0, 1]$ when $x = 0$ or $x = n$, but substituting the Clopper–Pearson limits in those cases yields coverage probability functions that resemble those for CI_W and CI_{AC}, although considerably more conservative for small n. Rubin and Schenker (1987) recommended the logit interval after $\frac{1}{2}$ additions to numbers of successes and failures, motivating it as a normal approximation to the posterior distribution of the logit parameter after using the Jeffreys prior. However, this modification has coverage probabilities that are unacceptably small for p near 0 and 1 (See Vollset, 1993). Presumably some other boundary modification will result in a happy medium. In a letter to the editor about Agresti and Coull (1998), Rindskopf (2000) argued in favor of the logit interval partly because of its connection with logit modeling. We have not used this method for teaching in elementary courses, since logit intervals do not extend to intervals for the difference of proportions and (like CI_W and CI_J) they are rather complex for that level.

For practical use and for teaching in more advanced courses, some statisticians may prefer the likelihood ratio interval, since conceptually it is simple and the method also applies in a general model-building framework. An advantage compared to the Wald approach is its invariance to the choice of scale, resulting, for instance, both from the original scale and the logit. BCD do not say much about this interval, since it is harder to compute. However, it is easy to obtain with standard statistical software (e.g., in SAS, using the LRCI option in PROC GENMOD for a model containing only an intercept term and assuming a binomial response with logit or identity link function). Graphs in Vollset (1993) suggest that the boundary-modified likelihood ratio interval also behaves reasonably well, although conservative for p near 0 and 1.

For elementary course teaching, a disadvantage of all such intervals using boundary modifications is that making exceptions from a general, simple recipe distracts students from the simple concept of taking the estimate plus and minus a normal score multiple of a standard error. (Of course, this concept is not sufficient for serious statistical work, but some over simplification and compromise is necessary at that level.) Even with CI_{AC}, instructors may find it preferable to give a recipe with the same number of added pseudo observations for all α, instead of $z^2_{\alpha/2}$. Reasonably good performance seems to result, especially for small α, from the value $4 \approx z^2_{0.025}$ used in the 95% CI_{AC} interval (i.e., the "add two successes and two failures" interval). Agresti and Caffo (2000) discussed this and showed that adding four pseudo observations also dramatically improves the Wald two-sample interval for comparing proportions, although again at the cost of rather severe conservativeness when both parameters are near 0 or near 1.

3. ALTERNATIVE WIDTH COMPARISON

In comparing the expected lengths of the three recommended intervals, BCD note that the comparison is clear and consistent as n changes, with the average expected length being noticeably larger for CI_{AC} than CI_J and CI_W. Thus, in their concluding remarks, they recommend CI_J and CI_W for small n. However, since BCD recommend modifying CI_J and CI_W to eliminate severe downward spikes of coverage probabilities, we believe that a

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).