Statistical Science
2001, Vol. 16, No. 2, 101–133

Interval Estimation for a Binomial Proportion

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

Abstract. We revisit the problem of interval estimation of a binomial proportion. The erratic behavior of the coverage probability of the standard Wald confidence interval has previously been remarked on in the literature (Blyth and Still, Agresti and Coull, Santner and others). We begin by showing that the chaotic coverage properties of the Wald interval are far more persistent than is appreciated. Furthermore, common textbook prescriptions regarding its safety are misleading and defective in several respects and cannot be trusted.

This leads us to consideration of alternative intervals. A number of natural alternatives are presented, each with its motivation and context. Each interval is examined for its coverage probability and its length. Based on this analysis, we recommend the Wilson interval or the equal-tailed Jeffreys prior interval for small n and the interval suggested in Agresti and Coull for larger n. We also provide an additional frequentist justification for use of the Jeffreys interval.

Key words and phrases: Bayes, binomial distribution, confidence intervals, coverage probability, Edgeworth expansion, expected length, Jeffreys prior, normal approximation, posterior.

1. INTRODUCTION

This article revisits one of the most basic and methodologically important problems in statistical practice, namely, interval estimation of the probability of success in a binomial distribution. There is a textbook confidence interval for this problem that has acquired nearly universal acceptance in practice. The interval, of course, is $\hat{p} \pm z_{\alpha/2}\ n^{-1/2}(\hat{p}(1-\hat{p}))^{1/2}$, where $\hat{p} = X/n$ is the sample proportion of successes, and $z_{\alpha/2}$ is the $100(1-\alpha/2)$th percentile of the standard normal distribution. The interval is easy to present and motivate and easy to compute. With the exceptions of the t test, linear regression, and ANOVA, its popularity in everyday practical statistics is virtually unmatched. The standard interval is known as the Wald interval as it comes from the Wald large sample test for the binomial case.

So at first glance, one may think that the problem is too simple and has a clear and present solution. In fact, the problem is a difficult one, with unanticipated complexities. It is widely recognized that the actual coverage probability of the standard interval is poor for p near 0 or 1. Even at the level of introductory statistics texts, the standard interval is often presented with the caveat that it should be used only when $n \cdot \min(p, 1-p)$ is at least 5 (or 10). Examination of the popular texts reveals that the qualifications with which the standard interval is presented are varied, but they all reflect the concern about poor coverage when p is near the boundaries.

In a series of interesting recent articles, it has also been pointed out that the coverage properties of the standard interval can be erratically poor even if p is not near the boundaries; see, for instance, Vollset (1993), Santner (1998), Agresti and Coull (1998), and Newcombe (1998). Slightly older literature includes Ghosh (1979), Cressie (1980) and Blyth and Still (1983). Agresti and Coull (1998)

Lawrence D. Brown is Professor of Statistics, The Wharton School, University of Pennsylvania, 3000 Steinberg Hall-Dietrich Hall, 3620 Locust Walk, Philadelphia, Pennsylvania 19104-6302. T. Tony Cai is Assistant Professor of Statistics, The Wharton School, University of Pennsylvania, 3000 Steinberg Hall-Dietrich Hall, 3620 Locust Walk, Philadelphia, Pennsylvania 19104-6302. Anirban DasGupta is Professor, Department of Statistics, Purdue University, 1399 Mathematical Science Bldg., West Lafayette, Indiana 47907-1399

particularly consider the nominal 95% case and show the erratic and poor behavior of the standard interval's coverage probability for small n even when p is not near the boundaries. See their Figure 4 for the cases $n = 5$ and 10.

We will show in this article that the eccentric behavior of the standard interval's coverage probability is far deeper than has been explained or is appreciated by statisticians at large. We will show that the popular prescriptions the standard interval comes with are defective in several respects and are not to be trusted. In addition, we will motivate, present and analyze several alternatives to the standard interval for a general confidence level. We will ultimately make recommendations about choosing a specific interval for practical use, separately for different intervals of values of n. It will be seen that for small n (40 or less), our recommendation differs from the recommendation Agresti and Coull (1998) made for the nominal 95% case. To facilitate greater appreciation of the seriousness of the problem, we have kept the technical content of this article at a minimal level. The companion article, Brown, Cai and DasGupta (1999), presents the associated theoretical calculations on Edgeworth expansions of the various intervals' coverage probabilities and asymptotic expansions for their expected lengths.

In Section 2, we first present a series of examples on the degree of severity of the chaotic behavior of the standard interval's coverage probability. The chaotic behavior does not go away even when n is quite large and p is not near the boundaries. For instance, when n is 100, the actual coverage probability of the nominal 95% standard interval is 0.952 if p is 0.106, but only 0.911 if p is 0.107. The behavior of the coverage probability can be even more erratic as a function of n. If the true p is 0.5, the actual coverage of the nominal 95% interval is 0.953 at the rather small sample size $n = 17$, but falls to 0.919 at the much larger sample size $n = 40$.

This eccentric behavior can get downright extreme in certain practically important problems. For instance, consider defective proportions in industrial quality control problems. There it would be quite common to have a true p that is small. If the true p is 0.005, then the coverage probability of the nominal 95% interval increases monotonically in n all the way up to $n = 591$ to the level 0.945, only to drop down to 0.792 if n is 592. This unlucky spell continues for a while, and then the coverage bounces back to 0.948 when n is 953, but dramatically falls to 0.852 when n is 954. Subsequent unlucky spells start off at $n = 1279$, 1583 and on and on. It should be widely known that the coverage of the standard interval can be significantly lower at quite large sample sizes, and this happens in an unpredictable and rather random way.

Continuing, also in Section 2 we list a set of common prescriptions that standard texts present while discussing the standard interval. We show what the deficiencies are in some of these prescriptions. Proposition 1 and the subsequent Table 3 illustrate the defects of these common prescriptions.

In Sections 3 and 4, we present our alternative intervals. For the purpose of a sharper focus we present these alternative intervals in two categories. First we present in Section 3 a selected set of three intervals that clearly stand out in our subsequent analysis; we present them as our "recommended intervals." Separately, we present several other intervals in Section 4 that arise as clear candidates for consideration as a part of a comprehensive examination, but do not stand out in the actual analysis.

The short list of recommended intervals contains the score interval, an interval recently suggested in Agresti and Coull (1998), and the equal tailed interval resulting from the natural noninformative Jeffreys prior for a binomial proportion. The score interval for the binomial case seems to have been introduced in Wilson (1927); so we call it the Wilson interval. Agresti and Coull (1998) suggested, for the special nominal 95% case, the interval $\tilde{p} \pm z_{0.025}\tilde{n}^{-1/2}(\tilde{p}(1-\tilde{p}))^{1/2}$, where $\tilde{n} = n+4$ and $\tilde{p} = (X+2)/(n+4)$; this is an adjusted Wald interval that formally adds two successes and two failures to the observed counts and then uses the standard method. Our second interval is the appropriate version of this interval for a general confidence level; we call it the Agresti–Coull interval. By a slight abuse of terminology, we call our third interval, namely the equal-tailed interval corresponding to the Jeffreys prior, the Jeffreys interval.

In Section 3, we also present our findings on the performances of our "recommended" intervals. As always, two key considerations are their coverage properties and parsimony as measured by expected length. Simplicity of presentation is also sometimes an issue, for example, in the context of classroom presentation at an elementary level. On consideration of these factors, we came to the conclusion that for small n (40 or less), we recommend that either the Wilson or the Jeffreys prior interval should be used. They are very similar, and either may be used depending on taste. The Wilson interval has a closed-form formula. The Jeffreys interval does not. One can expect that there would be resistance to using the Jeffreys interval solely due to this reason. We therefore provide a table simply listing the

limits of the Jeffreys interval for n up to 30 and in addition also give closed form and very accurate approximations to the limits. These approximations do not need any additional software.

For larger n ($n > 40$), the Wilson, the Jeffreys and the Agresti–Coull interval are all very similar, and so for such n, due to its simplest form, we come to the conclusion that the Agresti–Coull interval should be recommended. Even for smaller sample sizes, the Agresti–Coull interval is strongly preferable to the standard one and so might be the choice where simplicity is a paramount objective.

The additional intervals we considered are two slight modifications of the Wilson and the Jeffreys intervals, the Clopper–Pearson "exact" interval, the arcsine interval, the logit interval, the actual Jeffreys HPD interval and the likelihood ratio interval. The modified versions of the Wilson and the Jeffreys intervals correct disturbing downward spikes in the coverages of the original intervals very close to the two boundaries. The other alternative intervals have earned some prominence in the literature for one reason or another. We had to apply a certain amount of discretion in choosing these additional intervals as part of our investigation. Since we wish to direct the main part of our conversation to the three "recommended" intervals, only a brief summary of the performances of these additional intervals is presented along with the introduction of each interval. As part of these quick summaries, we indicate why we decided against including them among the recommended intervals.

We strongly recommend that introductory texts in statistics present one or more of these recommended alternative intervals, in preference to the standard one. The slight sacrifice in simplicity would be more than worthwhile. The conclusions we make are given additional theoretical support by the results in Brown, Cai and DasGupta (1999). Analogous results for other one parameter discrete families are presented in Brown, Cai and DasGupta (2000).

2. THE STANDARD INTERVAL

When constructing a confidence interval we usually wish the actual coverage probability to be close to the nominal confidence level. Because of the discrete nature of the binomial distribution we cannot always achieve the exact nominal confidence level unless a randomized procedure is used. Thus our objective is to construct nonrandomized confidence intervals for p such that the coverage probability $P_p(p \in CI) \approx 1 - \alpha$ where α is some prespecified value between 0 and 1. We will use the notation $C(p, n) = P_p(p \in CI), 0 < p < 1$, for the coverage probability.

A standard confidence interval for p based on normal approximation has gained universal recommendation in the introductory statistics textbooks and in statistical practice. The interval is known to guarantee that for any fixed $p \in (0, 1), C(p, n) \to 1 - \alpha$ as $n \to \infty$.

Let $\phi(z)$ and $\Phi(z)$ be the standard normal density and distribution functions, respectively. Throughout the paper we denote $\kappa \equiv z_{\alpha/2} = \Phi^{-1}(1 - \alpha/2)$, $\hat{p} = X/n$ and $\hat{q} = 1 - \hat{p}$. The standard normal approximation confidence interval CI_s is given by

$$CI_s = \hat{p} \pm \kappa\, n^{-1/2}(\hat{p}\hat{q})^{1/2}. \tag{1}$$

This interval is obtained by inverting the acceptance region of the well known Wald large-sample normal test for a general problem:

$$|(\hat{\theta} - \theta)/\widehat{se}(\hat{\theta})| \le \kappa, \tag{2}$$

where θ is a generic parameter, $\hat{\theta}$ is the maximum likelihood estimate of θ and $\widehat{se}(\hat{\theta})$ is the estimated standard error of $\hat{\theta}$. In the binomial case, we have $\theta = p$, $\hat{\theta} = X/n$ and $\widehat{se}(\hat{\theta}) = (\hat{p}\hat{q})^{1/2}n^{-1/2}$.

The standard interval is easy to calculate and is heuristically appealing. In introductory statistics texts and courses, the confidence interval CI_s is usually presented along with some heuristic justification based on the central limit theorem. Most students and users no doubt believe that the larger the number n, the better the normal approximation, and thus the closer the actual coverage would be to the nominal level $1 - \alpha$. Further, they would believe that the coverage probabilities of this method are close to the nominal value, except possibly when n is "small" or p is "near" 0 or 1. We will show how completely both of these beliefs are false. Let us take a close look at how the standard interval CI_s really performs.

2.1 Lucky n, Lucky p

An interesting phenomenon for the standard interval is that the actual coverage probability of the confidence interval contains nonnegligible oscillation as both p and n vary. There exist some "lucky" pairs (p, n) such that the actual coverage probability $C(p, n)$ is very close to or larger than the nominal level. On the other hand, there also exist "unlucky" pairs (p, n) such that the corresponding $C(p, n)$ is much smaller than the nominal level. The phenomenon of oscillation is both in n, for fixed p, and in p, for fixed n. Furthermore, drastic changes in coverage occur in nearby p for fixed n and in nearby n for fixed p. Let us look at five simple but instructive examples.

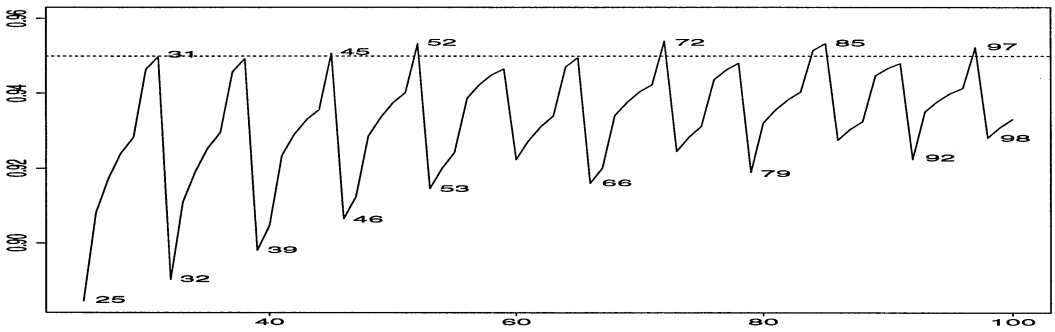

FIG. 1. *Standard interval; oscillation phenomenon for fixed* $p = 0.2$ *and variable* $n = 25$ *to* 100.

The probabilities reported in the following plots and tables, as well as those appearing later in this paper, are the result of direct probability calculations produced in S-PLUS. In all cases their numerical accuracy considerably exceeds the number of significant figures reported and/or the accuracy visually obtainable from the plots. (Plots for variable p are the probabilities for a fine grid of values of p, e.g., 2000 equally spaced values of p for the plots in Figure 5.)

EXAMPLE 1. Figure 1 plots the coverage probability of the nominal 95% standard interval for $p = 0.2$. The number of trials n varies from 25 to 100. It is clear from the plot that the oscillation is significant and the coverage probability does not steadily get closer to the nominal confidence level as n increases. For instance, $C(0.2, 30) = 0.946$ and $C(0.2, 98) = 0.928$. So, as hard as it is to believe, the coverage probability is significantly closer to 0.95 when $n = 30$ than when $n = 98$. We see that the true coverage probability behaves contrary to conventional wisdom in a very significant way.

EXAMPLE 2. Now consider the case of $p = 0.5$. Since $p = 0.5$, conventional wisdom might suggest to an unsuspecting user that all will be well if n is about 20. We evaluate the exact coverage probability of the 95% standard interval for $10 \le n \le 50$. In Table 1, we list the values of "lucky" n [defined as $C(p, n) \ge 0.95$] and the values of "unlucky" n [defined for specificity as $C(p, n) \le 0.92$]. The conclusions presented in Table 2 are surprising. We note that when $n = 17$ the coverage probability is 0.951, but the coverage probability equals 0.904 when $n = 18$. Indeed, the unlucky values of n arise suddenly. Although p is 0.5, the coverage is still only 0.919 at $n = 40$. This illustrates the inconsistency, unpredictability and poor performance of the standard interval.

EXAMPLE 3. Now let us move p really close to the boundary, say $p = 0.005$. We mention in the introduction that such p are relevant in certain practical applications. Since p is so small, now one may fully expect that the coverage probability of the standard interval is poor. Figure 2 and Table 2.2 show that there are still surprises and indeed we now begin to see a whole new kind of erratic behavior. The oscillation of the coverage probability does not show until rather large n. Indeed, the coverage probability makes a slow ascent all the way until $n = 591$, and then dramatically drops to 0.792 when $n = 592$. Figure 2 shows that thereafter the oscillation manifests in full force, in contrast to Examples 1 and 2, where the oscillation started early on. Subsequent "unlucky" values of n again arise in the same unpredictable way, as one can see from Table 2.2.

2.2 Inadequate Coverage

The results in Examples 1 to 3 already show that the standard interval can have coverage noticeably smaller than its nominal value even for values of n and of $np(1 - p)$ that are not small. This subsec-

TABLE 1
Standard interval; lucky n *and unlucky* n *for* $10 \le n \le 50$ *and* $p = 0.5$

Lucky n	17	20	25	30	35	37	42	44	49
$C(0.5, n)$	0.951	0.959	0.957	.957	0.959	0.953	0.956	0.951	0.956
Unlucky n	10	12	13	15	18	23	28	33	40
$C(0.5, n)$	0.891	0.854	0.908	0.882	0.904	0.907	0.913	0.920	0.919

TABLE 2
Standard interval; late arrival of unlucky n for small p

Unlucky n	592	954	1279	1583	1876
$C(0.005, n)$	0.792	0.852	0.875	0.889	0.898

tion contains two more examples that display further instances of the inadequacy of the standard interval.

EXAMPLE 4. Figure 3 plots the coverage probability of the nominal 95% standard interval with fixed $n = 100$ and variable p. It can be seen from Figure 3 that in spite of the "large" sample size, significant change in coverage probability occurs in nearby p. The magnitude of oscillation increases significantly as p moves toward 0 or 1. Except for values of p quite near $p = 0.5$, the general trend of this plot is noticeably below the nominal coverage value of 0.95.

EXAMPLE 5. Figure 4 shows the coverage probability of the nominal 99% standard interval with $n = 20$ and variable p from 0 to 1. Besides the oscillation phenomenon similar to Figure 3, a striking fact in this case is that the coverage never reaches the nominal level. The coverage probability is *always* smaller than 0.99, and in fact on the average the coverage is only 0.883. Our evaluations show that for all $n \leq 45$, the coverage of the 99% standard interval is strictly smaller than the nominal level for all $0 < p < 1$.

It is evident from the preceding presentation that the actual coverage probability of the standard interval can differ significantly from the nominal confidence level for moderate and even large sample sizes. We will later demonstrate that there are other confidence intervals that perform much better in this regard. See Figure 5 for such a comparison. The error in coverage comes from two sources: discreteness and skewness in the underlying binomial distribution. For a two-sided interval, the rounding error due to discreteness is dominant, and the error due to skewness is somewhat secondary, but still important for even moderately large n. (See Brown, Cai and DasGupta, 1999, for more details.) Note that the situation is different for one-sided intervals. There, the error caused by the skewness can be larger than the rounding error. See Hall (1982) for a detailed discussion on one-sided confidence intervals.

The oscillation in the coverage probability is caused by the discreteness of the binomial distribution, more precisely, the lattice structure of the binomial distribution. The noticeable oscillations are unavoidable for any nonrandomized procedure, although some of the competing procedures in Section 3 can be seen to have somewhat smaller oscillations than the standard procedure. See the text of Casella and Berger (1990) for introductory discussion of the oscillation in such a context.

The erratic and unsatisfactory coverage properties of the standard interval have often been remarked on, but curiously still do not seem to be widely appreciated among statisticians. See, for example, Ghosh (1979), Blyth and Still (1983) and Agresti and Coull (1998). Blyth and Still (1983) also show that the continuity-corrected version still has the same disadvantages.

2.3 Textbook Qualifications

The normal approximation used to justify the standard confidence interval for p can be significantly in error. The error is most evident when the true p is close to 0 or 1. See Lehmann (1999). In fact, it is easy to show that, for any fixed n, the

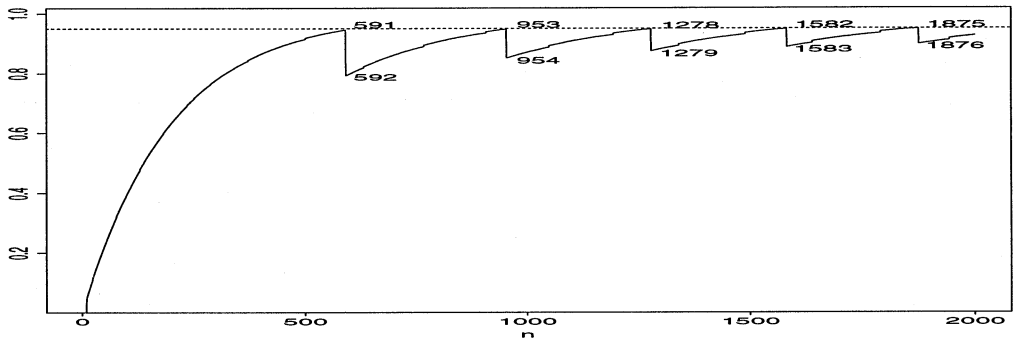

FIG. 2. *Standard interval; oscillation in coverage for small p.*

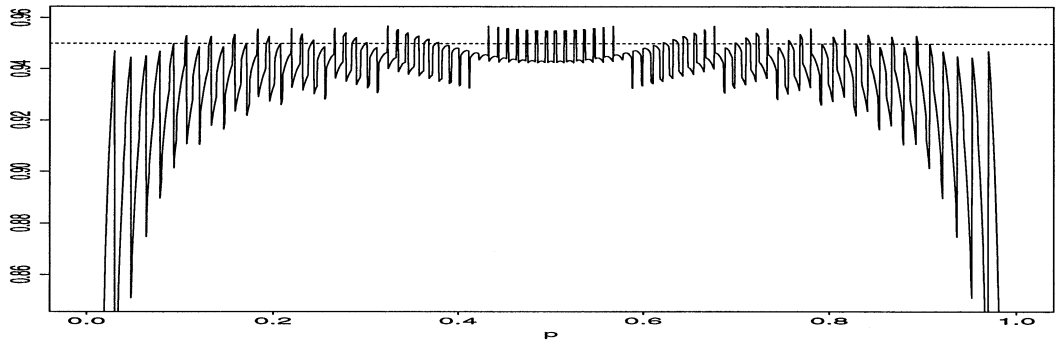

FIG. 3. *Standard interval; oscillation phenomenon for fixed* $n = 100$ *and variable* p.

confidence coefficient $C(p, n) \to 0$ as $p \to 0$ or 1. Therefore, most major problems arise as regards coverage probability when p is near the boundaries.

Poor coverage probabilities for p near 0 or 1 are widely remarked on, and generally, in the popular texts, a brief sentence is added qualifying when to use the standard confidence interval for p. It is interesting to see what these qualifications are. A sample of 11 popular texts gives the following qualifications:

The confidence interval may be used if:

1. $np, n(1-p)$ are ≥ 5 (or 10);
2. $np(1-p) \geq 5$ (or 10);
3. $n\hat{p}, n(1-\hat{p})$ are ≥ 5 (or 10);
4. $\hat{p} \pm 3\sqrt{\hat{p}(1-\hat{p})/n}$ does not contain 0 or 1;
5. n quite large;
6. $n \geq 50$ unless p is very small.

It seems clear that the authors are attempting to say that the standard interval may be used if the central limit approximation is accurate. These prescriptions are defective in several respects. In the estimation problem, (1) and (2) are not verifiable. Even when these conditions are satisfied, we see, for instance, from Table 1 in the previous section, that there is no guarantee that the true coverage probability is close to the nominal confidence level. For example, when $n = 40$ and $p = 0.5$, one has $np = n(1-p) = 20$ and $np(1-p) = 10$, so clearly either of the conditions (1) and (2) is satisfied. However, from Table 1, the true coverage probability in this case equals 0.919 which is certainly unsatisfactory for a confidence interval at nominal level 0.95.

The qualification (5) is useless and (6) is patently misleading; (3) and (4) are certainly verifiable, but they are also useless because in the context of frequentist coverage probabilities, a data-based prescription does not have a meaning. The point is that the standard interval clearly has serious problems and the influential texts caution the readers about that. However, the caution does not appear to serve its purpose, for a variety of reasons.

Here is a result that shows that sometimes the qualifications are not correct even in the limit as $n \to \infty$.

PROPOSITION 1. *Let* $\gamma > 0$. *For the standard confidence interval,*

$$\lim_{n\to\infty} \inf_{p:\, np,\, n(1-p)\geq\gamma} C(p, n) \leq P(a_\gamma < \text{Poisson}(\gamma) \leq b_\gamma), \tag{3}$$

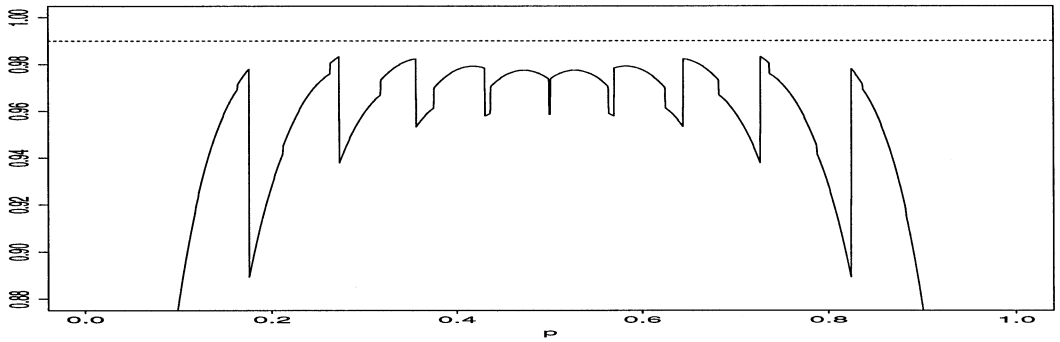

FIG. 4. *Coverage of the nominal* 99% *standard interval for fixed* $n = 20$ *and variable* p.

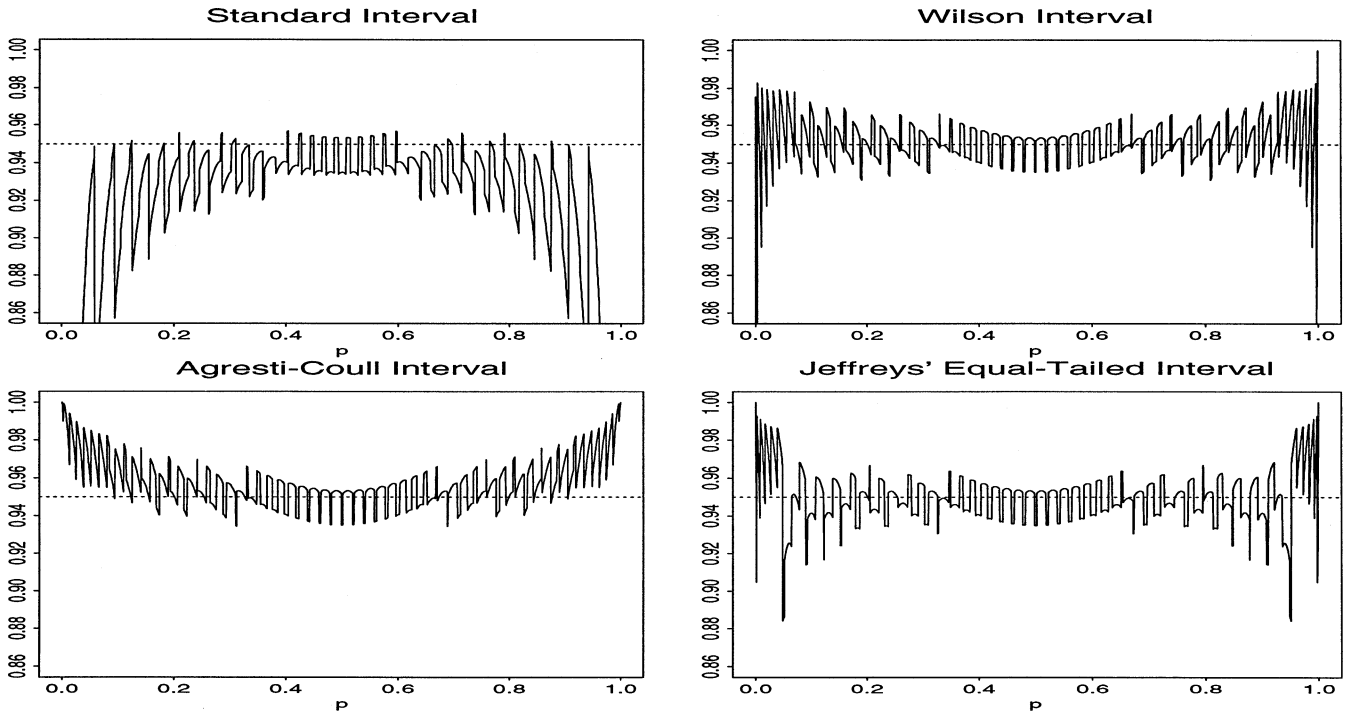

FIG. 5. *Coverage probability for* $n = 50$.

TABLE 3
Standard interval; bound (3) *on limiting minimum coverage when* $np, n(1-p) \geq \gamma$

γ	5	7	10
$\lim_{n\to\infty} \inf_{p:\, np,\, n(1-p)\geq\gamma} C(p,n)$	0.875	0.913	0.926

where a_γ *and* b_γ *are the integer parts of*

$$(\kappa^2 + 2\gamma \pm \kappa\sqrt{\kappa^2 + 4\gamma})/2,$$

where the $-$ *sign goes with* a_γ *and the* $+$ *sign with* b_γ.

The proposition follows from the fact that the sequence of $\text{Bin}(n, \gamma/n)$ distributions converges weakly to the Poisson(γ) distribution and so the limit of the infimum is at most the Poisson probability in the proposition by an easy calculation.

Let us use Proposition 1 to investigate the validity of qualifications (1) and (2) in the list above. The nominal confidence level in Table 3 below is 0.95.

TABLE 4
Values of λ_x *for the modified lower bound for the Wilson interval*

$1-\alpha$	$x=1$	$x=2$	$x=3$
0.90	0.105	0.532	1.102
0.95	0.051	0.355	0.818
0.99	0.010	0.149	0.436

It is clear that qualification (1) does not work at all and (2) is marginal. There are similar problems with qualifications (3) and (4).

3. RECOMMENDED ALTERNATIVE INTERVALS

From the evidence gathered in Section 2, it seems clear that the standard interval is just too risky. This brings us to the consideration of alternative intervals. We now analyze several such alternatives, each with its motivation. A few other intervals are also mentioned for their theoretical importance. Among these intervals we feel three stand out in their comparative performance. These are labeled separately as the "recommended intervals".

3.1 Recommended Intervals

3.1.1 *The Wilson interval.* An alternative to the standard interval is the confidence interval based on inverting the test in equation (2) that uses the null standard error $(pq)^{1/2}n^{-1/2}$ instead of the estimated standard error $(\hat{p}\hat{q})^{1/2}n^{-1/2}$. This confidence interval has the form

$$(4)\quad CI_W = \frac{X + \kappa^2/2}{n + \kappa^2} \pm \frac{\kappa n^{1/2}}{n + \kappa^2}(\hat{p}\hat{q} + \kappa^2/(4n))^{1/2}.$$

This interval was apparently introduced by Wilson (1927) and we will call this interval the Wilson interval.

The Wilson interval has theoretical appeal. The interval is the inversion of the CLT approximation

to the family of equal tail tests of H_0: $p = p_0$. Hence, one accepts H_0 based on the CLT approximation if and only if p_0 is in this interval. As Wilson showed, the argument involves the solution of a quadratic equation; or see Tamhane and Dunlop (2000, Exercise 9.39).

3.1.2 *The Agresti–Coull interval.* The standard interval CI_s is simple and easy to remember. For the purposes of classroom presentation and use in texts, it may be nice to have an alternative that has the familiar form $\hat{p} \pm z\sqrt{\hat{p}(1-\hat{p})/n}$, with a better and new choice of $\hat{p}$ rather than $\hat{p} = X/n$. This can be accomplished by using the center of the Wilson region in place of $\hat{p}$. Denote $\tilde{X} = X + \kappa^2/2$ and $\tilde{n} = n + \kappa^2$. Let $\tilde{p} = \tilde{X}/\tilde{n}$ and $\tilde{q} = 1 - \tilde{p}$. Define the confidence interval CI_{AC} for p by

$$CI_{AC} = \tilde{p} \pm \kappa(\tilde{p}\tilde{q})^{1/2}\tilde{n}^{-1/2}. \tag{5}$$

Both the Agresti–Coull and the Wilson interval are centered on the same value, $\tilde{p}$. It is easy to check that the Agresti–Coull intervals are never shorter than the Wilson intervals. For the case when $\alpha = 0.05$, if we use the value 2 instead of 1.96 for κ, this interval is the "add 2 successes and 2 failures" interval in Agresti and Coull (1998). For this reason, we call it the Agresti–Coull interval. To the best of our knowledge, Samuels and Witmer (1999) is the first introductory statistics textbook that recommends the use of this interval. See Figure 5 for the coverage of this interval. See also Figure 6 for its average coverage probability.

3.1.3 *Jeffreys interval.* Beta distributions are the standard conjugate priors for binomial distributions and it is quite common to use beta priors for inference on p (see Berger, 1985).

Suppose $X \sim \text{Bin}(n, p)$ and suppose p has a prior distribution $\text{Beta}(a_1, a_2)$; then the posterior distribution of p is $\text{Beta}(X + a_1, n - X + a_2)$. Thus a $100(1-\alpha)\%$ equal-tailed Bayesian interval is given by

$$[B(\alpha/2; X + a_1, n - X + a_2), \\ B(1-\alpha/2; X + a_1, n - X + a_2)],$$

where $B(\alpha; m_1, m_2)$ denotes the α quantile of a $\text{Beta}(m_1, m_2)$ distribution.

The well-known Jeffreys prior and the uniform prior are each a beta distribution. The noninformative Jeffreys prior is of particular interest to us. Historically, Bayes procedures under noninformative priors have a track record of good frequentist properties; see Wasserman (1991). In this problem the Jeffreys prior is $\text{Beta}(1/2, 1/2)$ which has the density function

$$f(p) = \pi^{-1}p^{-1/2}(1-p)^{-1/2}.$$

The $100(1-\alpha)\%$ equal-tailed Jeffreys prior interval is defined as

$$CI_J = [L_J(x), U_J(x)], \tag{6}$$

where $L_J(0) = 0, U_J(n) = 1$ and otherwise

$$L_J(x) = B(\alpha/2; X + 1/2, n - X + 1/2), \tag{7}$$

$$U_J(x) = B(1-\alpha/2; X + 1/2, n - X + 1/2). \tag{8}$$

The interval is formed by taking the central $1-\alpha$ posterior probability interval. This leaves $\alpha/2$ posterior probability in each omitted tail. The exception is for $x = 0(n)$ where the lower (upper) limits are modified to avoid the undesirable result that the coverage probability $C(p, n) \to 0$ as $p \to 0$ or 1.

The actual endpoints of the interval need to be numerically computed. This is very easy to do using softwares such as Minitab, S-PLUS or Mathematica. In Table 5 we have provided the limits for the case of the Jeffreys prior for $7 \le n \le 30$.

The endpoints of the Jeffreys prior interval are the $\alpha/2$ and $1-\alpha/2$ quantiles of the $\text{Beta}(x+1/2, n-x+1/2)$ distribution. The psychological resistance among some to using the interval is because of the inability to compute the endpoints at ease without software.

We provide two avenues to resolving this problem. One is Table 5 at the end of the paper. The second is a computable approximation to the limits of the Jeffreys prior interval, one that is computable with just a normal table. This approximation is obtained after some algebra from the general approximation to a Beta quantile given in page 945 in Abramowitz and Stegun (1970).

The lower limit of the $100(1-\alpha)\%$ Jeffreys prior interval is approximately

$$\frac{x + 1/2}{n + 1 + (n - x + 1/2)(e^{2\omega} - 1)}, \tag{9}$$

where

$$\omega = \frac{\kappa\sqrt{4\hat{p}\hat{q}}/n + (\kappa^2 - 3)/(6n^2)}{4\hat{p}\hat{q}} \\ + \frac{(1/2 - \hat{p})(\hat{p}\hat{q}(\kappa^2 + 2) - 1/n)}{6n(\hat{p}\hat{q})^2}.$$

The upper limit may be approximated by the same expression with κ replaced by $-\kappa$ in ω. The simple approximation given above is remarkably accurate. Berry (1996, page 222) suggests using a simpler normal approximation, but this will not be sufficiently accurate unless $n\hat{p}(1-\hat{p})$ is rather large.

TABLE 5
95% Limits of the Jeffreys prior interval

x	$n=7$		$n=8$		$n=9$		$n=10$		$n=11$		$n=12$	
0	0	0.292	0	0.262	0	0.238	0	0.217	0	0.200	0	0.185
1	0.016	0.501	0.014	0.454	0.012	0.414	0.011	0.381	0.010	0.353	0.009	0.328
2	0.065	0.648	0.056	0.592	0.049	0.544	0.044	0.503	0.040	0.467	0.036	0.436
3	0.139	0.766	0.119	0.705	0.104	0.652	0.093	0.606	0.084	0.565	0.076	0.529
4	0.234	0.861	0.199	0.801	0.173	0.746	0.153	0.696	0.137	0.652	0.124	0.612
5					0.254	0.827	0.224	0.776	0.200	0.730	0.180	0.688
6									0.270	0.800	0.243	0.757

x	$n=13$		$n=14$		$n=15$		$n=16$		$n=17$		$n=18$	
0	0	0.173	0	0.162	0	0.152	0	0.143	0	0.136	0	0.129
1	0.008	0.307	0.008	0.288	0.007	0.272	0.007	0.257	0.006	0.244	0.006	0.232
2	0.033	0.409	0.031	0.385	0.029	0.363	0.027	0.344	0.025	0.327	0.024	0.311
3	0.070	0.497	0.064	0.469	0.060	0.444	0.056	0.421	0.052	0.400	0.049	0.381
4	0.114	0.577	0.105	0.545	0.097	0.517	0.091	0.491	0.085	0.467	0.080	0.446
5	0.165	0.650	0.152	0.616	0.140	0.584	0.131	0.556	0.122	0.530	0.115	0.506
6	0.221	0.717	0.203	0.681	0.188	0.647	0.174	0.617	0.163	0.589	0.153	0.563
7	0.283	0.779	0.259	0.741	0.239	0.706	0.222	0.674	0.207	0.644	0.194	0.617
8					0.294	0.761	0.272	0.728	0.254	0.697	0.237	0.668
9									0.303	0.746	0.284	0.716

x	$n=19$		$n=20$		$n=21$		$n=22$		$n=23$		$n=24$	
0	0	0.122	0	0.117	0	0.112	0	0.107	0	0.102	0	0.098
1	0.006	0.221	0.005	0.211	0.005	0.202	0.005	0.193	0.005	0.186	0.004	0.179
2	0.022	0.297	0.021	0.284	0.020	0.272	0.019	0.261	0.018	0.251	0.018	0.241
3	0.047	0.364	0.044	0.349	0.042	0.334	0.040	0.321	0.038	0.309	0.036	0.297
4	0.076	0.426	0.072	0.408	0.068	0.392	0.065	0.376	0.062	0.362	0.059	0.349
5	0.108	0.484	0.102	0.464	0.097	0.446	0.092	0.429	0.088	0.413	0.084	0.398
6	0.144	0.539	0.136	0.517	0.129	0.497	0.123	0.478	0.117	0.461	0.112	0.444
7	0.182	0.591	0.172	0.568	0.163	0.546	0.155	0.526	0.148	0.507	0.141	0.489
8	0.223	0.641	0.211	0.616	0.199	0.593	0.189	0.571	0.180	0.551	0.172	0.532
9	0.266	0.688	0.251	0.662	0.237	0.638	0.225	0.615	0.214	0.594	0.204	0.574
10	0.312	0.734	0.293	0.707	0.277	0.681	0.263	0.657	0.250	0.635	0.238	0.614
11					0.319	0.723	0.302	0.698	0.287	0.675	0.273	0.653
12									0.325	0.713	0.310	0.690

x	$n=25$		$n=26$		$n=27$		$n=28$		$n=29$		$n=30$	
0	0	0.095	0	0.091	0	0.088	0	0.085	0	0.082	0	0.080
1	0.004	0.172	0.004	0.166	0.004	0.160	0.004	0.155	0.004	0.150	0.004	0.145
2	0.017	0.233	0.016	0.225	0.016	0.217	0.015	0.210	0.015	0.203	0.014	0.197
3	0.035	0.287	0.034	0.277	0.032	0.268	0.031	0.259	0.030	0.251	0.029	0.243
4	0.056	0.337	0.054	0.325	0.052	0.315	0.050	0.305	0.048	0.295	0.047	0.286
5	0.081	0.384	0.077	0.371	0.074	0.359	0.072	0.348	0.069	0.337	0.067	0.327
6	0.107	0.429	0.102	0.415	0.098	0.402	0.095	0.389	0.091	0.378	0.088	0.367
7	0.135	0.473	0.129	0.457	0.124	0.443	0.119	0.429	0.115	0.416	0.111	0.404
8	0.164	0.515	0.158	0.498	0.151	0.482	0.145	0.468	0.140	0.454	0.135	0.441
9	0.195	0.555	0.187	0.537	0.180	0.521	0.172	0.505	0.166	0.490	0.160	0.476
10	0.228	0.594	0.218	0.576	0.209	0.558	0.201	0.542	0.193	0.526	0.186	0.511
11	0.261	0.632	0.250	0.613	0.239	0.594	0.230	0.577	0.221	0.560	0.213	0.545
12	0.295	0.669	0.282	0.649	0.271	0.630	0.260	0.611	0.250	0.594	0.240	0.578
13	0.331	0.705	0.316	0.684	0.303	0.664	0.291	0.645	0.279	0.627	0.269	0.610
14					0.336	0.697	0.322	0.678	0.310	0.659	0.298	0.641
15									0.341	0.690	0.328	0.672

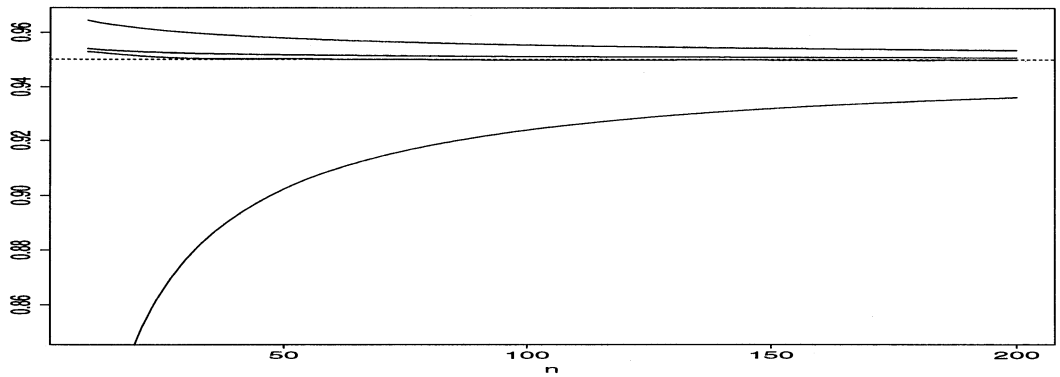

FIG. 6. *Comparison of the average coverage probabilities. From top to bottom: the Agresti–Coull interval* CI_{AC}, *the Wilson interval* CI_W, *the Jeffreys prior interval* CI_J *and the standard interval* CI_s. *The nominal confidence level is* 0.95.

In Figure 5 we plot the coverage probability of the standard interval, the Wilson interval, the Agresti–Coull interval and the Jeffreys interval for $n = 50$ and $\alpha = 0.05$.

3.2 Coverage Probability

In this and the next subsections, we compare the performance of the standard interval and the three recommended intervals in terms of their coverage probability and length.

Coverage of the Wilson interval fluctuates acceptably near $1 - \alpha$, except for p very near 0 or 1. It might be helpful to consult Figure 5 again. It can be shown that, when $1 - \alpha = 0.95$,

$$\lim_{n\to\infty} \inf_{\gamma\geq 1} C\left(\frac{\gamma}{n},\ n\right) = 0.92,$$

$$\lim_{n\to\infty} \inf_{\gamma\geq 5} C\left(\frac{\gamma}{n},\ n\right) = 0.936$$

and

$$\lim_{n\to\infty} \inf_{\gamma\geq 10} C\left(\frac{\gamma}{n},\ n\right) = 0.938$$

for the Wilson interval. In comparison, these three values for the standard interval are 0.860, 0.870, and 0.905, respectively, obviously considerably smaller.

The modification CI_{M-W} presented in Section 4.1.1 removes the first few deep downward spikes of the coverage function for CI_W. The resulting coverage function is overall somewhat conservative for p very near 0 or 1. Both CI_W and CI_{M-W} have the same coverage functions away from 0 or 1.

The Agresti–Coull interval has good minimum coverage probability. The coverage probability of the interval is quite conservative for p very close to 0 or 1. In comparison to the Wilson interval it is more conservative, especially for small n. This is not surprising because, as we have noted, CI_{AC} always contains CI_W as a proper subinterval.

The coverage of the Jeffreys interval is qualitatively similar to that of CI_W over most of the parameter space $[0, 1]$. In addition, as we will see in Section 4.3, CI_J has an appealing connection to the mid-P corrected version of the Clopper–Pearson "exact" intervals. These are very similar to CI_J, over most of the range, and have similar appealing properties. CI_J is a serious and credible candidate for practical use. The coverage has an unfortunate fairly deep spike near $p = 0$ and, symmetrically, another near $p = 1$. However, the simple modification of CI_J presented in Section 4.1.2 removes these two deep downward spikes. The modified Jeffreys interval CI_{M-J} performs well.

Let us also evaluate the intervals in terms of their average coverage probability, the average being over p. Figure 6 demonstrates the striking difference in the average coverage probability among four intervals: the Agresti–Coull interval, the Wilson interval the Jeffreys prior interval and the standard interval. The standard interval performs poorly. The interval CI_{AC} is slightly conservative in terms of average coverage probability. Both the Wilson interval and the Jeffreys prior interval have excellent performance in terms of the average coverage probability; that of the Jeffreys prior interval is, if anything, slightly superior. The average coverage of the Jeffreys interval is really very close to the nominal level even for quite small n. This is quite impressive.

Figure 7 displays the mean absolute errors, $\int_0^1 |C(p, n) - (1 - \alpha)|\, dp$, for $n = 10$ to 25, and $n = 26$ to 40. It is clear from the plots that among the four intervals, CI_W, CI_{AC} and CI_J are comparable, but the mean absolute errors of CI_s are significantly larger.

3.3 Expected Length

Besides coverage, length is also very important in evaluation of a confidence interval. We compare

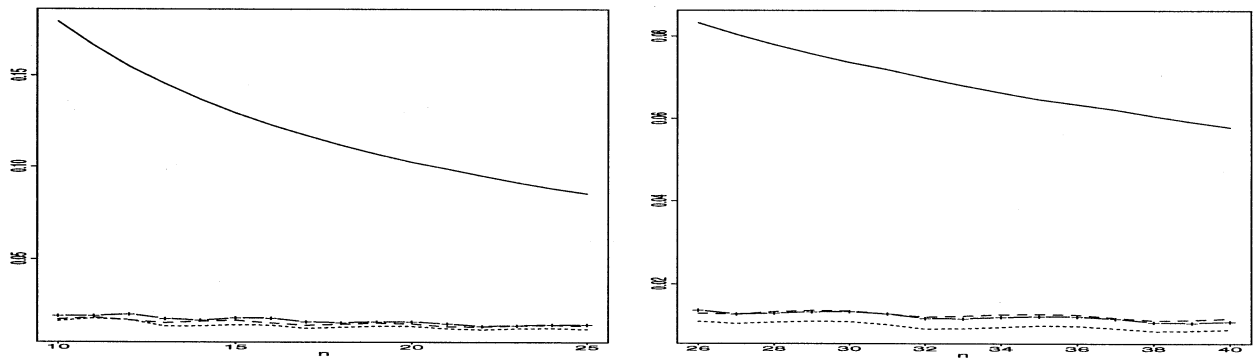

FIG. 7. *The mean absolute errors of the coverage of the standard (solid), the Agresti–Coull (dashed), the Jeffreys (+) and the Wilson (dotted) intervals for* $n = 10$ *to* 25 *and* $n = 26$ *to* 40.

both the expected length and the average expected length of the intervals. By definition,

$$\begin{aligned}&\text{Expected length}\\ &\quad = E_{n,p}(\text{length}(CI))\\ &\quad = \sum_{x=0}^{n}(U(x,n) - L(x,n))\binom{n}{x}p^x(1-p)^{n-x},\end{aligned}$$

where U and L are the upper and lower limits of the confidence interval CI, respectively. The average expected length is just the integral $\int_0^1 E_{n,p}(\text{length}(CI))\,dp$.

We plot in Figure 8 the expected lengths of the four intervals for $n = 25$ and $\alpha = 0.05$. In this case, CI_W is the shortest when $0.210 \le p \le 0.790$, CI_J is the shortest when $0.133 \le p \le 0.210$ or $0.790 \le p \le 0.867$, and CI_s is the shortest when $p \le 0.133$ or $p \ge 0.867$. It is no surprise that the standard interval is the shortest when p is near the boundaries. CI_s is not really in contention as a credible choice for such values of p because of its poor coverage properties in that region. Similar qualitative phenomena hold for other values of n.

Figure 9 shows the average expected lengths of the four intervals for $n = 10$ to 25 and $n = 26$ to 40. Interestingly, the comparison is clear and consistent as n changes. Always, the standard interval and the Wilson interval CI_W have almost identical average expected length; the Jeffreys interval CI_J is comparable to the Wilson interval, and in fact CI_J is slightly more parsimonious. But the difference is not of practical relevance. However, especially when n is small, the average expected length of CI_{AC} is noticeably larger than that of CI_J and CI_W. In fact, for n till about 20, the average expected length of CI_{AC} is larger than that of CI_J by 0.04 to 0.02, and this difference can be of definite practical relevance. The difference starts to wear off when n is larger than 30 or so.

4. OTHER ALTERNATIVE INTERVALS

Several other intervals deserve consideration, either due to their historical value or their theoretical properties. In the interest of space, we had to exercise some personal judgment in deciding which additional intervals should be presented.

4.1 Boundary modification

The coverage probabilities of the Wilson interval and the Jeffreys interval fluctuate acceptably near

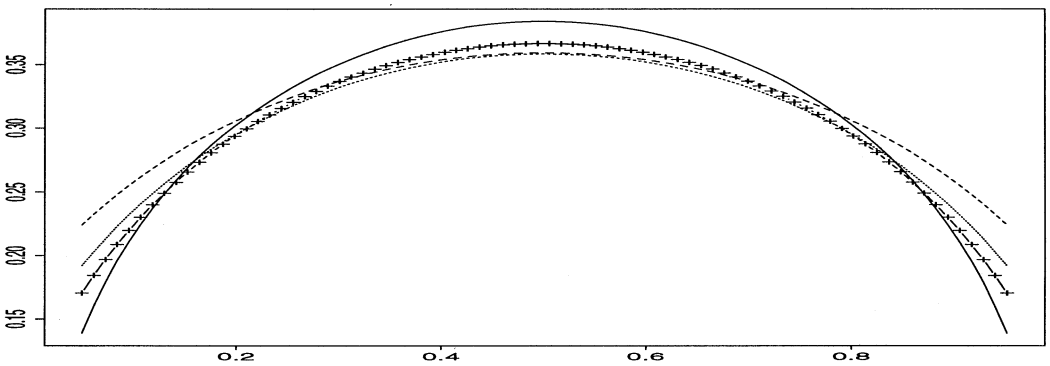

FIG. 8. *The expected lengths of the standard (solid), the Wilson (dotted), the Agresti–Coull (dashed) and the Jeffreys (+) intervals for* $n = 25$ *and* $\alpha = 0.05$.

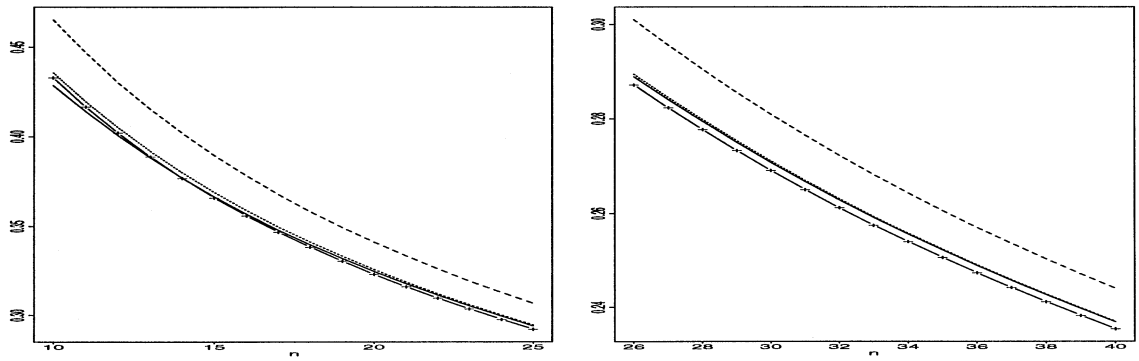

FIG. 9. *The average expected lengths of the standard* (*solid*), *the Wilson* (*dotted*), *the Agresti–Coull* (*dashed*) *and the Jeffreys* (+) *intervals for* $n = 10$ *to* 25 *and* $n = 26$ *to* 40.

$1-\alpha$ for p not very close to 0 or 1. Simple modifications can be made to remove a few deep downward spikes of their coverage near the boundaries; see Figure 5.

4.1.1 *Modified Wilson interval.* The lower bound of the Wilson interval is formed by inverting a CLT approximation. The coverage has downward spikes when p is very near 0 or 1. These spikes exist for all n and α. For example, it can be shown that, when $1-\alpha = 0.95$ and $p = 0.1765/n$,

$$\lim_{n\to\infty} P_p(p \in CI_W) = 0.838$$

and when $1-\alpha = 0.99$ and $p = 0.1174/n$, $\lim_{n\to\infty} P_p(p \in CI_W) = 0.889$. The particular numerical values (0.1174, 0.1765) are relevant only to the extent that divided by n, they approximate the location of these deep downward spikes.

The spikes can be removed by using a one-sided Poisson approximation for x close to 0 or n. Suppose we modify the lower bound for $x = 1, \ldots, x^*$. For a fixed $1 \le x \le x^*$, the lower bound of CI_W should be replaced by a lower bound of λ_x/n where λ_x solves

$$e^{-\lambda}(\lambda^0/0! + \lambda^1/1! + \cdots + \lambda^{x-1}/(x-1)!) = 1-\alpha. \tag{10}$$

A symmetric prescription needs to be followed to modify the upper bound for x very near n. The value of x^* should be small. Values which work reasonably well for $1-\alpha = 0.95$ are

$$x^* = 2 \text{ for } n < 50 \text{ and } x^* = 3 \text{ for } 51 \le n \le 100+.$$

Using the relationship between the Poisson and χ^2 distributions,

$$P(Y \le x) = P(\chi^2_{2(1+x)} \le 2\lambda)$$

where $Y \sim \text{Poisson}(\lambda)$, one can also formally express λ_x in (10) in terms of the χ^2 quantiles: $\lambda_x = (1/2)\chi^2_{2x,\alpha}$, where $\chi^2_{2x,\alpha}$ denotes the 100αth percentile of the χ^2 distribution with $2x$ degrees of freedom. Table 4 gives the values of λ_x for selected values of x and α.

For example, consider the case $1-\alpha = 0.95$ and $x = 2$. The lower bound of CI_W is $\approx 0.548/(n+4)$. The modified Wilson interval replaces this by a lower bound of λ/n where $\lambda = (1/2)\,\chi^2_{4,0.05}$. Thus,

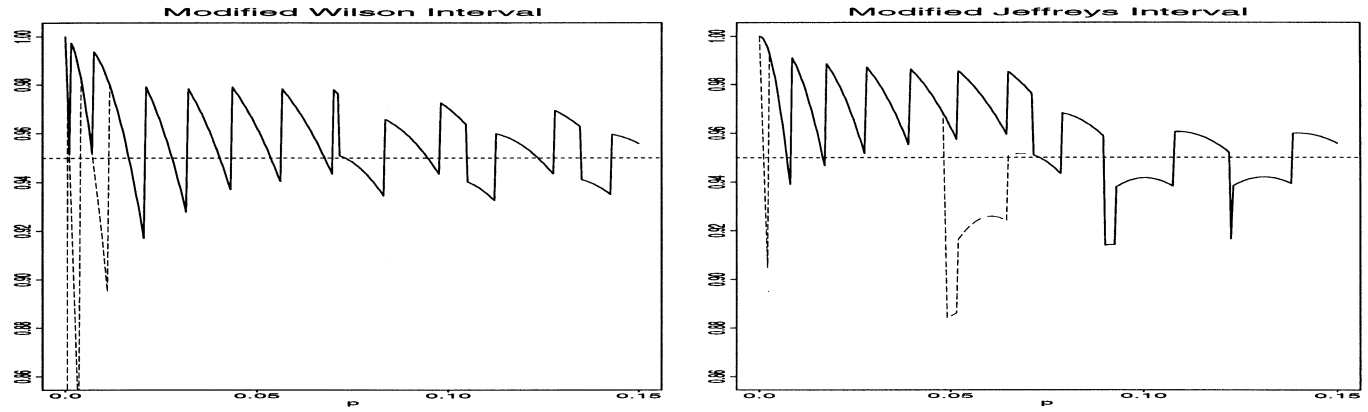

FIG. 10. *Coverage probability for* $n = 50$ *and* $p \in (0, 0.15)$. *The plots are symmetric about* $p = 0.5$ *and the coverage of the modified intervals* (*solid line*) *is the same as that of the corresponding interval without modification* (*dashed line*) *for* $p \in [0.15, 0.85]$.

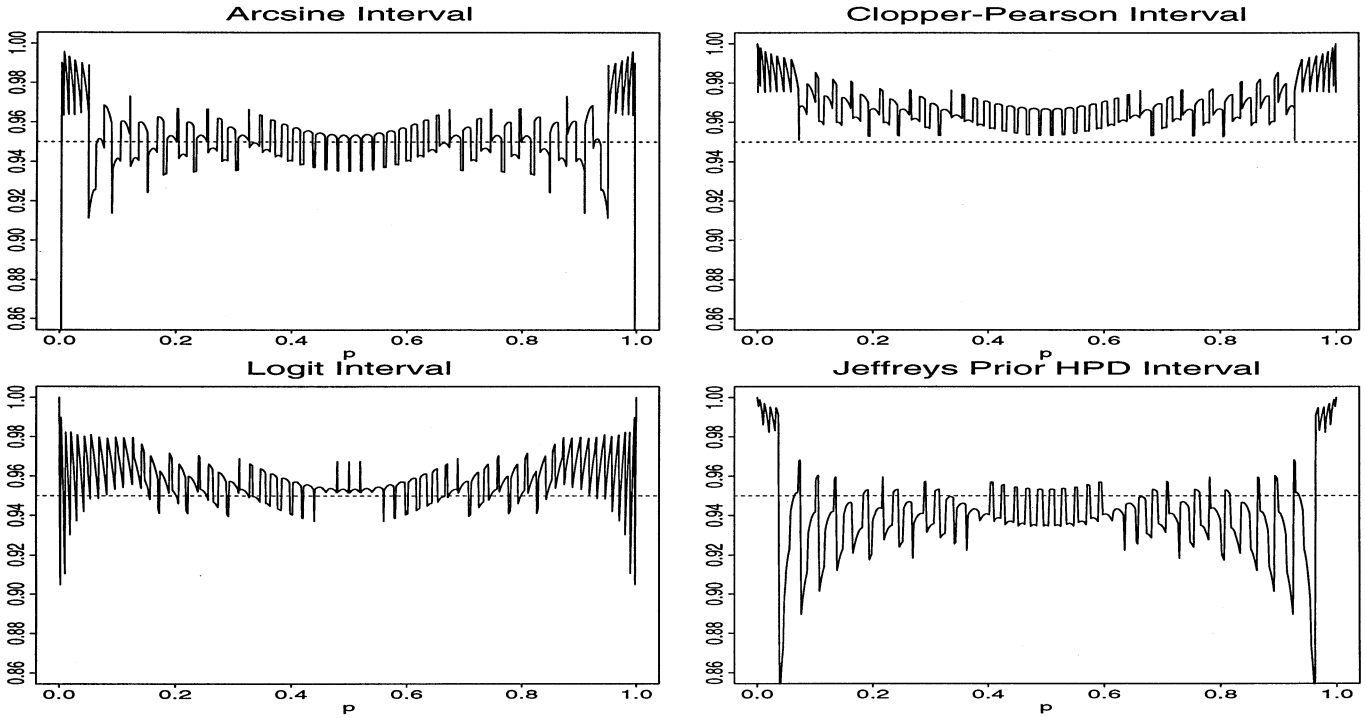

FIG. 11. *Coverage probability of other alternative intervals for* $n = 50$.

from a χ^2 table, for $x = 2$ the new lower bound is $0.355/n$.

We denote this modified Wilson interval by CI_{M-W}. See Figure 10 for its coverage.

4.1.2 *Modified Jeffreys interval.* Evidently, CI_J has an appealing Bayesian interpretation, and, its coverage properties are appealing again except for a very narrow downward coverage spike fairly near 0 and 1 (see Figure 5). The unfortunate downward spikes in the coverage function result because $U_J(0)$ is too small and symmetrically $L_J(n)$ is too large. To remedy this, one may revise these two specific limits as

$$U_{M-J}(0) = p_l \quad \text{and} \quad L_{M-J}(n) = 1 - p_l,$$

where p_l satisfies $(1 - p_l)^n = \alpha/2$ or equivalently $p_l = 1 - (\alpha/2)^{1/n}$.

We also made a slight, ad hoc alteration of $L_J(1)$ and set

$$L_{M-J}(1) = 0 \quad \text{and} \quad U_{M-J}(n-1) = 1.$$

In all other cases, $L_{M-J} = L_J$ and $U_{M-J} = U_J$. We denote the modified Jeffreys interval by CI_{M-J}. This modification removes the two steep downward spikes and the performance of the interval is improved. See Figure 10.

4.2 Other intervals

4.2.1 *The Clopper–Pearson interval.* The Clopper–Pearson interval is the inversion of the equal-tail binomial test rather than its normal approximation. Some authors refer to this as the "exact" procedure because of its derivation from the binomial distribution. If $X = x$ is observed, then the Clopper–Pearson (1934) interval is defined by $CI_{CP} = [L_{CP}(x), U_{CP}(x)]$, where $L_{CP}(x)$ and $U_{CP}(x)$ are, respectively, the solutions in p to the equations

$$P_p(X \geq x) = \alpha/2 \quad \text{and} \quad P_p(X \leq x) = \alpha/2.$$

It is easy to show that the lower endpoint is the $\alpha/2$ quantile of a beta distribution Beta$(x, n - x + 1)$, and the upper endpoint is the $1 - \alpha/2$ quantile of a beta distribution Beta$(x + 1, n - x)$. The Clopper–Pearson interval guarantees that the actual coverage probability is always equal to or above the nominal confidence level. However, for any fixed p, the actual coverage probability can be much larger than $1-\alpha$ unless n is quite large, and thus the confidence interval is rather inaccurate in this sense. See Figure 11. The Clopper–Pearson interval is wastefully conservative and is not a good choice for practical use, unless strict adherence to the prescription $C(p, n) \geq 1-\alpha$ is demanded. Even then, better exact methods are available; see, for instance, Blyth and Still (1983) and Casella (1986).

4.2.2 *The arcsine interval.* Another interval is based on a widely used variance stabilizing transformation for the binomial distribution [see, e.g., Bickel and Doksum, 1977: $T(\hat{p}) = \arcsin(\hat{p}^{1/2})$]. This variance stabilization is based on the delta method and is, of course, only an asymptotic one. Anscombe (1948) showed that replacing $\hat{p}$ by $\check{p} = (X + 3/8)/(n + 3/4)$ gives better variance stabilization; furthermore

$$2n^{1/2}[\arcsin(\check{p}^{1/2}) - \arcsin(p^{1/2})] \to N(0, 1) \quad \text{as } n \to \infty.$$

This leads to an approximate $100(1-\alpha)\%$ confidence interval for p,

$$CI_{Arc} = \left[\sin^2(\arcsin(\check{p}^{1/2}) - \tfrac{1}{2}\kappa n^{-1/2}), \sin^2(\arcsin(\check{p}^{1/2}) + \tfrac{1}{2}\kappa n^{-1/2})\right]. \tag{11}$$

See Figure 11 for the coverage probability of this interval for $n = 50$. This interval performs reasonably well for p not too close to 0 or 1. The coverage has steep downward spikes near the two edges; in fact it is easy to see that the coverage drops to zero when p is sufficiently close to the boundary (see Figure 11). The mean absolute error of the coverage of CI_{Arc} is significantly larger than those of CI_W, CI_{AC} and CI_J. We note that our evaluations show that the performance of the arcsine interval with the standard $\hat{p}$ in place of $\check{p}$ in (11) is much worse than that of CI_{Arc}.

4.2.3 *The logit interval.* The logit interval is obtained by inverting a Wald type interval for the log odds $\lambda = \log(\frac{p}{1-p})$; (see Stone, 1995). The MLE of λ (for $0 < X < n$) is

$$\hat{\lambda} = \log\left(\frac{\hat{p}}{1-\hat{p}}\right) = \log\left(\frac{X}{n-X}\right),$$

which is the so-called empirical logit transform. The variance of $\hat{\lambda}$, by an application of the delta theorem, can be estimated by

$$\widehat{V} = \frac{n}{X(n-X)}.$$

This leads to an approximate $100(1-\alpha)\%$ confidence interval for λ,

$$CI(\lambda) = [\lambda_l, \lambda_u] = [\hat{\lambda} - \kappa\widehat{V}^{1/2}, \hat{\lambda} + \kappa\widehat{V}^{1/2}]. \tag{12}$$

The logit interval for p is obtained by inverting the interval (12),

$$CI_{\text{Logit}} = \left[\frac{e^{\lambda_l}}{1+e^{\lambda_l}}, \frac{e^{\lambda_u}}{1+e^{\lambda_u}}\right]. \tag{13}$$

The interval (13) has been suggested, for example, in Stone (1995, page 667). Figure 11 plots the coverage of the logit interval for $n = 50$. This interval performs quite well in terms of coverage for p away from 0 or 1. But the interval is unnecessarily long; in fact its expected length is larger than that of the Clopper–Pearson exact interval.

REMARK. Anscombe (1956) suggested that $\hat{\lambda} = \log(\frac{X+1/2}{n-X+1/2})$ is a better estimate of λ; see also Cox and Snell (1989) and Santner and Duffy (1989). The variance of Anscombe's $\hat{\lambda}$ may be estimated by

$$\widehat{V} = \frac{(n+1)(n+2)}{n(X+1)(n-X+1)}.$$

A new logit interval can be constructed using the new estimates $\hat{\lambda}$ and $\widehat{V}$. Our evaluations show that the new logit interval is overall shorter than CI_{Logit} in (13). But the coverage of the new interval is not satisfactory.

4.2.4 *The Bayesian HPD interval.* An exact Bayesian solution would involve using the HPD intervals instead of our equal-tails proposal. However, HPD intervals are much harder to compute and do not do as well in terms of coverage probability. See Figure 11 and compare to the Jeffreys' equal-tailed interval in Figure 5.

4.2.5 *The likelihood ratio interval.* Along with the Wald and the Rao score intervals, the likelihood ratio method is one of the most used methods for construction of confidence intervals. It is constructed by inversion of the likelihood ratio test which accepts the null hypothesis H_0: $p = p_0$ if $-2\log(\Lambda_n) \le \kappa^2$, where Λ_n is the likelihood ratio

$$\Lambda_n = \frac{L(p_0)}{\sup_p L(p)} = \frac{p_0^X(1-p_0)^{n-X}}{(X/n)^X(1-X/n)^{n-X}},$$

L being the likelihood function. See Rao (1973). Brown, Cai and DasGupta (1999) show by analytical calculations that this interval has nice properties. However, it is slightly harder to compute. For the purpose of the present article which we view as primarily directed toward practice, we do not further analyze the likelihood ratio interval.

4.3 Connections between Jeffreys Intervals and Mid-*P* Intervals

The equal-tailed Jeffreys prior interval has some interesting connections to the Clopper–Pearson interval. As we mentioned earlier, the Clopper–

Pearson interval CI_{CP} can be written as

$$CI_{CP} = [B(\alpha/2; X, n - X + 1), \\ B(1 - \alpha/2; X + 1, n - X)].$$

It therefore follows immediately that CI_J is always contained in CI_{CP}. Thus CI_J corrects the conservativeness of CI_{CP}.

It turns out that the Jeffreys prior interval, although Bayesianly constructed, has a clear and convincing frequentist motivation. It is thus no surprise that it does well from a frequentist perspective. As we now explain, the Jeffreys prior interval CI_J can be regarded as a continuity corrected version of the Clopper–Pearson interval CI_{CP}.

The interval CI_{CP} inverts the inequality $P_p(X \le L(p)) \le \alpha/2$ to obtain the lower limit and similarly for the upper limit. Thus, for fixed x, the upper limit of the interval for p, $U_{CP}(x)$, satisfies

$$P_{U_{CP}(x)}(X \le x) \le \alpha/2, \tag{14}$$

and symmetrically for the lower limit.

This interval is very conservative; undesirably so for most practical purposes. A familiar proposal to eliminate this over-conservativeness is to instead invert

$$P_p(X \le L(p) - 1) + (1/2)P_p(X = L(p)) = \alpha/2, \tag{15}$$

This amounts to solving

$$(1/2)\{P_{U_{CP}(x)}(X \le x - 1) \\ + P_{U_{CP}(x)}(X \le x)\} = \alpha/2, \tag{16}$$

which is the same as

$$U_{\text{mid-}P}(X) = (1/2)B(1 - \alpha/2; x, n - x + 1) \\ + (1/2)B(1 - \alpha/2; x + 1, n - x) \tag{17}$$

and symmetrically for the lower endpoint. These are the "Mid-P Clopper-Pearson" intervals. They are known to have good coverage and length performance. $U_{\text{mid-}P}$ given in (17) is a weighted average of two incomplete Beta functions. The incomplete Beta function of interest, $B(1 - \alpha/2; x, n - x + 1)$, is continuous and monotone in x if we formally treat x as a continuous argument. Hence the average of the two functions defining $U_{\text{mid-}P}$ is approximately the same as the value at the halfway point, $x + 1/2$. Thus

$$U_{\text{mid-}P}(X) \approx B(1 - \alpha/2; x + 1/2, n - x + 1/2) = U_J(x),$$

exactly the upper limit for the equal-tailed Jeffreys interval. Similarly, the corresponding approximate lower endpoint is the Jeffreys' lower limit.

Another frequentist way to interpret the Jeffreys prior interval is to say that $U_J(x)$ is the upper limit for the Clopper–Pearson rule with $x - 1/2$ successes and $L_J(x)$ is the lower limit for the Clopper–Pearson rule with $x + 1/2$ successes. Strawderman and Wells (1998) contains a valuable discussion of mid-P intervals and suggests some variations based on asymptotic expansions.

5. CONCLUDING REMARKS

Interval estimation of a binomial proportion is a very basic problem in practical statistics. The standard Wald interval is in nearly universal use. We first show that the performance of this standard interval is persistently chaotic and unacceptably poor. Indeed its coverage properties defy conventional wisdom. The performance is so erratic and the qualifications given in the influential texts are so defective that the standard interval should not be used. We provide a fairly comprehensive evaluation of many natural alternative intervals. Based on this analysis, we recommend the Wilson or the equal-tailed Jeffreys prior interval for small n($n \le 40$). These two intervals are comparable in both absolute error and length for $n \le 40$, and we believe that either could be used, depending on taste.

For larger n, the Wilson, the Jeffreys and the Agresti–Coull intervals are all comparable, and the Agresti–Coull interval is the simplest to present. It is generally true in statistical practice that only those methods that are easy to describe, remember and compute are widely used. Keeping this in mind, we recommend the Agresti–Coull interval for practical use when $n \ge 40$. Even for small sample sizes, the easy-to-present Agresti–Coull interval is much preferable to the standard one.

We would be satisfied if this article contributes to a greater appreciation of the severe flaws of the popular standard interval and an agreement that it deserves not to be used at all. We also hope that the recommendations for alternative intervals will provide some closure as to what may be used in preference to the standard method.

Finally, we note that the specific choices of the values of n, p and α in the examples and figures are artifacts. The theoretical results in Brown, Cai and DasGupta (1999) show that qualitatively similar phenomena as regarding coverage and length hold for general n and p and common values of the coverage. (Those results there are asymptotic as $n \to \infty$, but they are also sufficiently accurate for realistically moderate n.)

APPENDIX

TABLE A.1
95% Limits of the modified Jeffreys prior interval

x	$n=7$		$n=8$		$n=9$		$n=10$		$n=11$		$n=12$	
0	0	0.410	0	0.369	0	0.336	0	0.308	0	0.285	0	0.265
1	0	0.501	0	0.454	0	0.414	0	0.381	0	0.353	0	0.328
2	0.065	0.648	0.056	0.592	0.049	0.544	0.044	0.503	0.040	0.467	0.036	0.436
3	0.139	0.766	0.119	0.705	0.104	0.652	0.093	0.606	0.084	0.565	0.076	0.529
4	0.234	0.861	0.199	0.801	0.173	0.746	0.153	0.696	0.137	0.652	0.124	0.612
5					0.254	0.827	0.224	0.776	0.200	0.730	0.180	0.688
6									0.270	0.800	0.243	0.757

x	$n=13$		$n=14$		$n=15$		$n=16$		$n=17$		$n=18$	
0	0	0.247	0	0.232	0	0.218	0	0.206	0	0.195	0	0.185
1	0	0.307	0	0.288	0	0.272	0	0.257	0	0.244	0	0.232
2	0.033	0.409	0.031	0.385	0.029	0.363	0.027	0.344	0.025	0.327	0.024	0.311
3	0.070	0.497	0.064	0.469	0.060	0.444	0.056	0.421	0.052	0.400	0.049	0.381
4	0.114	0.577	0.105	0.545	0.097	0.517	0.091	0.491	0.085	0.467	0.080	0.446
5	0.165	0.650	0.152	0.616	0.140	0.584	0.131	0.556	0.122	0.530	0.115	0.506
6	0.221	0.717	0.203	0.681	0.188	0.647	0.174	0.617	0.163	0.589	0.153	0.563
7	0.283	0.779	0.259	0.741	0.239	0.706	0.222	0.674	0.207	0.644	0.194	0.617
8					0.294	0.761	0.272	0.728	0.254	0.697	0.237	0.668
9									0.303	0.746	0.284	0.716

x	$n=19$		$n=20$		$n=21$		$n=22$		$n=23$		$n=24$	
0	0	0.176	0	0.168	0	0.161	0	0.154	0	0.148	0	0.142
1	0	0.221	0	0.211	0	0.202	0	0.193	0	0.186	0	0.179
2	0.022	0.297	0.021	0.284	0.020	0.272	0.019	0.261	0.018	0.251	0.018	0.241
3	0.047	0.364	0.044	0.349	0.042	0.334	0.040	0.321	0.038	0.309	0.036	0.297
4	0.076	0.426	0.072	0.408	0.068	0.392	0.065	0.376	0.062	0.362	0.059	0.349
5	0.108	0.484	0.102	0.464	0.097	0.446	0.092	0.429	0.088	0.413	0.084	0.398
6	0.144	0.539	0.136	0.517	0.129	0.497	0.123	0.478	0.117	0.461	0.112	0.444
7	0.182	0.591	0.172	0.568	0.163	0.546	0.155	0.526	0.148	0.507	0.141	0.489
8	0.223	0.641	0.211	0.616	0.199	0.593	0.189	0.571	0.180	0.551	0.172	0.532
9	0.266	0.688	0.251	0.662	0.237	0.638	0.225	0.615	0.214	0.594	0.204	0.574
10	0.312	0.734	0.293	0.707	0.277	0.681	0.263	0.657	0.250	0.635	0.238	0.614
11					0.319	0.723	0.302	0.698	0.287	0.675	0.273	0.653
12									0.325	0.713	0.310	0.690

x	$n=25$		$n=26$		$n=27$		$n=28$		$n=29$		$n=30$	
0	0	0.137	0	0.132	0	0.128	0	0.123	0	0.119	0	0.116
1	0	0.172	0	0.166	0	0.160	0	0.155	0	0.150	0	0.145
2	0.017	0.233	0.016	0.225	0.016	0.217	0.015	0.210	0.015	0.203	0.014	0.197
3	0.035	0.287	0.034	0.277	0.032	0.268	0.031	0.259	0.030	0.251	0.029	0.243
4	0.056	0.337	0.054	0.325	0.052	0.315	0.050	0.305	0.048	0.295	0.047	0.286
5	0.081	0.384	0.077	0.371	0.074	0.359	0.072	0.348	0.069	0.337	0.067	0.327
6	0.107	0.429	0.102	0.415	0.098	0.402	0.095	0.389	0.091	0.378	0.088	0.367
7	0.135	0.473	0.129	0.457	0.124	0.443	0.119	0.429	0.115	0.416	0.111	0.404
8	0.164	0.515	0.158	0.498	0.151	0.482	0.145	0.468	0.140	0.454	0.135	0.441
9	0.195	0.555	0.187	0.537	0.180	0.521	0.172	0.505	0.166	0.490	0.160	0.476
10	0.228	0.594	0.218	0.576	0.209	0.558	0.201	0.542	0.193	0.526	0.186	0.511
11	0.261	0.632	0.250	0.613	0.239	0.594	0.230	0.577	0.221	0.560	0.213	0.545
12	0.295	0.669	0.282	0.649	0.271	0.630	0.260	0.611	0.250	0.594	0.240	0.578
13	0.331	0.705	0.316	0.684	0.303	0.664	0.291	0.645	0.279	0.627	0.269	0.610
14					0.336	0.697	0.322	0.678	0.310	0.659	0.298	0.641
15									0.341	0.690	0.328	0.672

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*).

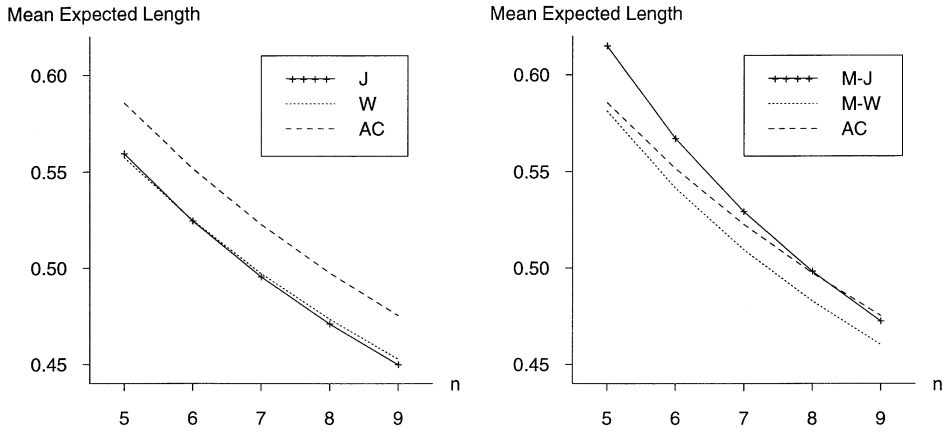

FIG. 1. *A Comparison of mean expected lengths for the nominal* 95% *Jeffreys* (*J*), *Wilson* (*W*), *Modified Jeffreys* (*M-J*), *Modified Wilson* (*M-W*), *and Agresti–Coull* (*AC*) *intervals for* $n = 5, 6, 7, 8, 9$.

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 $z_{\alpha/2}^2$ 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 $\tilde{n} = n + z_{\alpha/2}^2$ 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

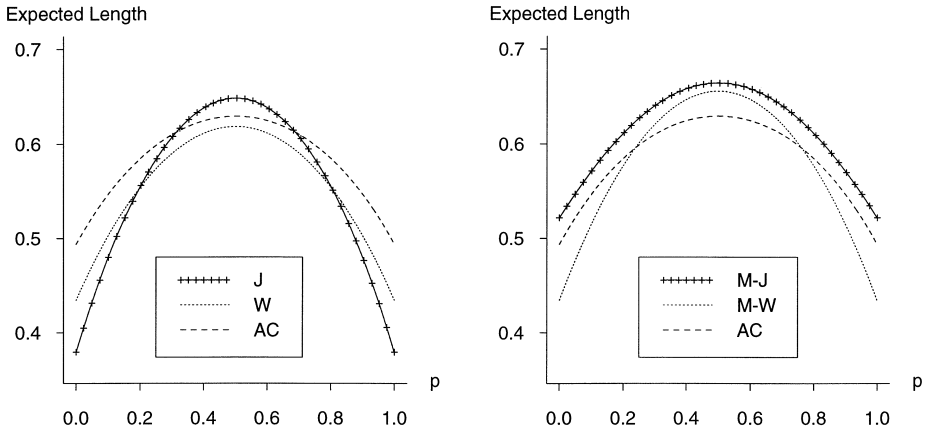

FIG. 2. *A comparison of expected lengths for the nominal* 95% *Jeffreys* (*J*), *Wilson* (*W*), *Modified Jeffreys* (*M-J*), *Modified Wilson* (*M-W*), *and Agresti–Coull* (*AC*) *intervals for* $n = 5$.

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*).

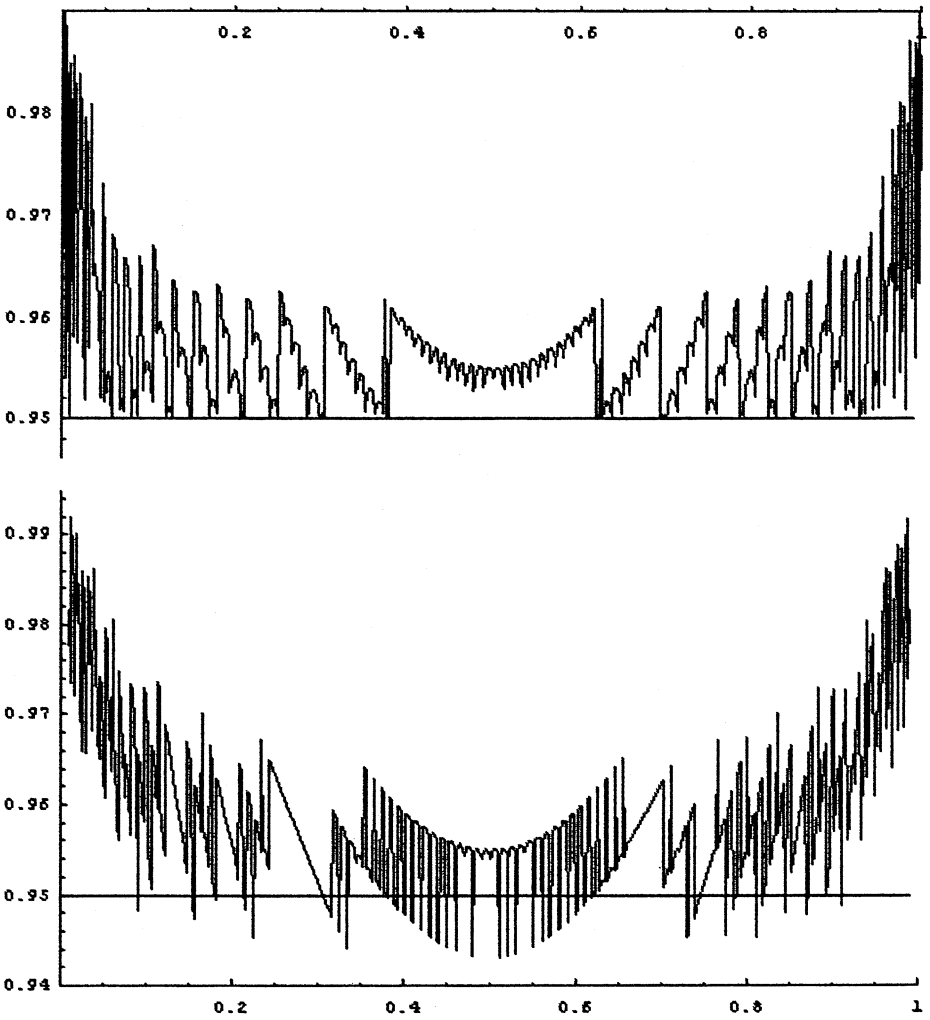

FIG. 1. *Coverage probabilities of the Blyth-Still interval* (*upper*) *and Agresti-Coull interval* (*lower*) *for* $n = 100$ *and* $1 - \alpha = 0.95$.

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.

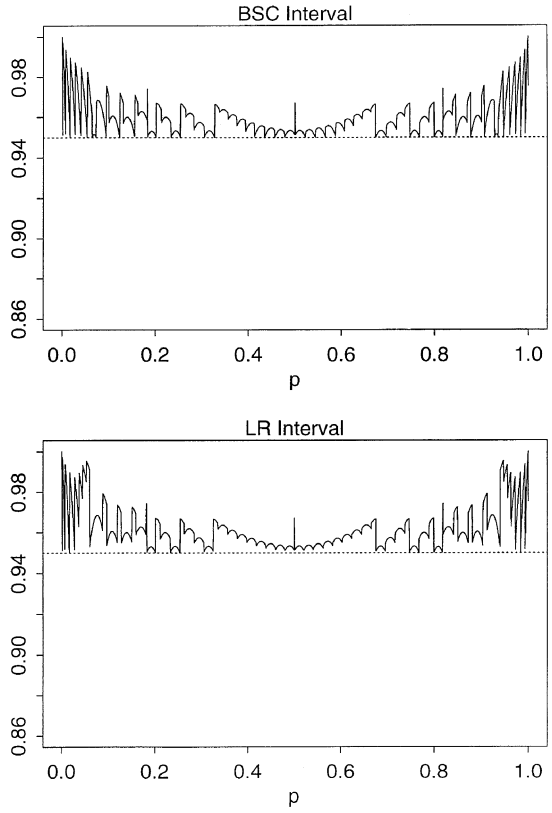

FIG. 1. *Actual coverage probabilities for BSC and LR intervals as a function of* $p(n = 50)$. *Compare to author's Figures* 5, 10 *and* 11.

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.

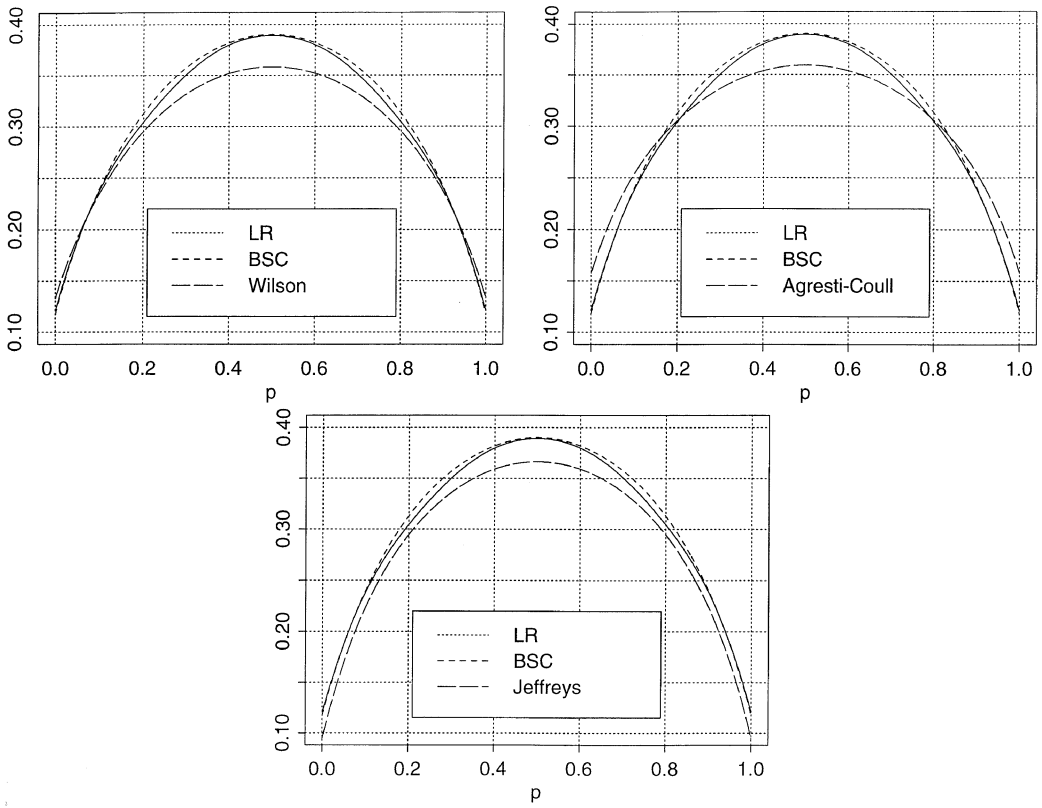

FIG. 2. *Expected lengths of BSC and LR intervals as a function of p compared, respectively, to Wilson, Agresti–Coull and Jeffreys intervals* ($n = 25$). *Compare to authors' Figure* 8.

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).

is $N(\overline{X}_n, 1/n)$. Accordingly, writing z_α for the upper $100\alpha\%$ point of the $N(0,1)$ distribution,

$$P(\theta \le \overline{X}_n + z_\alpha n^{-1/2} | \overline{X}_n) = 1 - \alpha = P(\theta \le \overline{X}_n + z_\alpha n^{-1/2} | \theta)$$

and this is an example of perfect matching. Now if $\hat{\theta}_n$ is the MLE of θ, under suitable regularity conditions, $\hat{\theta}_n|\theta$ is asymptotically (as $n \to \infty$) $N(\theta, I^{-1}(\theta))$, where $I(\theta)$ is the Fisher Information number. With the transformation $g(\theta) = \int^\theta I^{1/2}(t)$, by the delta method, $g(\hat{\theta}_n)$ is asymptotically $N(g(\theta), 1)$. Now, intuitively one expects the uniform prior $\pi(\theta) \propto c$ as the asymptotic matching prior for $g(\theta)$. Transforming back to the original parameter, Jeffreys prior is a probability matching prior for θ. Of course, this requires an invariance of probability matching priors, a fact which is rigorously established in Datta and Ghosh (1996). Thus a uniform prior for $\arcsin(\theta^{1/2})$, where θ is the binomial proportion, leads to Jeffreys Beta (1/2, 1/2) prior for θ. When θ is the Poisson parameter, the uniform prior for $\theta^{1/2}$ leads to Jeffreys' prior $\theta^{-1/2}$ for θ.

In a more formal set-up, let $X_1, \ldots, X_n$ be iid conditional on some real-valued θ. Let $\theta_\pi^{1-\alpha}(X_1, \ldots, X_n)$ denote a posterior $(1-\alpha)$th quantile for θ under the prior π. Then π is said to be a first-order probability matching prior if

$$P(\theta \le \theta_\pi^{1-\alpha}(X_1, \ldots, X_n)|\theta) = 1 - \alpha + o(n^{-1/2}). \tag{1}$$

This definition is due to Welch and Peers (1963) who showed by solving a differential equation that Jeffreys prior is the unique first-order probability matching prior in this case. Strictly speaking, Welch and Peers proved this result only for continuous distributions. Ghosh (1994) pointed out a suitable modification of criterion (1) which would lead to the same conclusion for discrete distributions. Also, for small and moderate samples, due to discreteness, one needs some modifications of Jeffreys interval as done so successfully by BCD.

This idea of probability matching can be extended even in the presence of nuisance parameters. Suppose that $\theta = (\theta_1, \ldots, \theta_p)^T$, where θ_1 is the parameter of interest, while $(\theta_2, \ldots, \theta_p)^T$ is the nuisance parameter. Writing $I(\theta) = ((I_{jk}))$ as the Fisher information matrix, if θ_1 is orthogonal to $(\theta_2, \ldots, \theta_p)^T$ in the sense of Cox and Reid (1987), that is, $I_{1k} = 0$ for all $k = 2, \ldots, p$, extending the previous intuitive argument, $\pi(\theta) \propto I_{11}^{1/2}(\theta)$ is a probability matching prior. Indeed, this prior belongs to the general class of first-order probability matching priors

$$\pi(\theta) \propto I_{11}^{1/2}(\theta) h(\theta_2, \ldots, \theta_p)$$

as derived in Tibshirani (1989). Here $h(\cdot)$ is an arbitrary function differentiable in its arguments.

In general, matching priors have a long success story in providing frequentist confidence intervals, especially in complex problems, for example, the Behrens–Fisher or the common mean estimation problems where frequentist methods run into difficulty. Though asymptotic, the matching property seems to hold for small and moderate sample sizes as well for many important statistical problems. One such example is Garvan and Ghosh (1997) where such priors were found for general dispersion models as given in Jorgensen (1997). It may be worthwhile developing these priors in the presence of nuisance parameters for other discrete cases as well, for example when the parameter of interest is the difference of two binomial proportions, or the log-odds ratio in a 2×2 contingency table.

Having argued so strongly in favor of matching priors, I wonder, though, whether there is any special need for such priors in this particular problem of binomial proportions. It appears that any Beta (a, a) prior will do well in this case. As noted in this paper, by shrinking the MLE X/n toward the prior mean 1/2, one achieves a better centering for the construction of confidence intervals. The two diametrically opposite priors Beta (2, 2) (symmetric concave with maximum at 1/2 which provides the Agresti–Coull interval) and Jeffreys prior Beta (1/2, 1/2) (symmetric convex with minimum at 1/2) seem to be equally good for recentering. Indeed, I wonder whether any Beta (α, β) prior which shrinks the MLE toward the prior mean $\alpha/(\alpha + \beta)$ becomes appropriate for recentering.

The problem of construction of confidence intervals for binomial proportions occurs in first courses in statistics as well as in day-to-day consulting. While I am strongly in favor of replacing Wald intervals by the new ones for the latter, I am not quite sure how easy it will be to motivate these new intervals for the former. The notion of shrinking can be explained adequately only to a few strong students in introductory statistics courses. One possible solution for the classroom may be to bring in the notion of continuity correction and somewhat heuristcally ask students to work with $(X + \frac{1}{2}, n - X + \frac{1}{2})$ instead of $(X, n - X)$. In this way, one centers around $(X + \frac{1}{2})/(n + 1)$ a la Jeffreys prior.

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*).

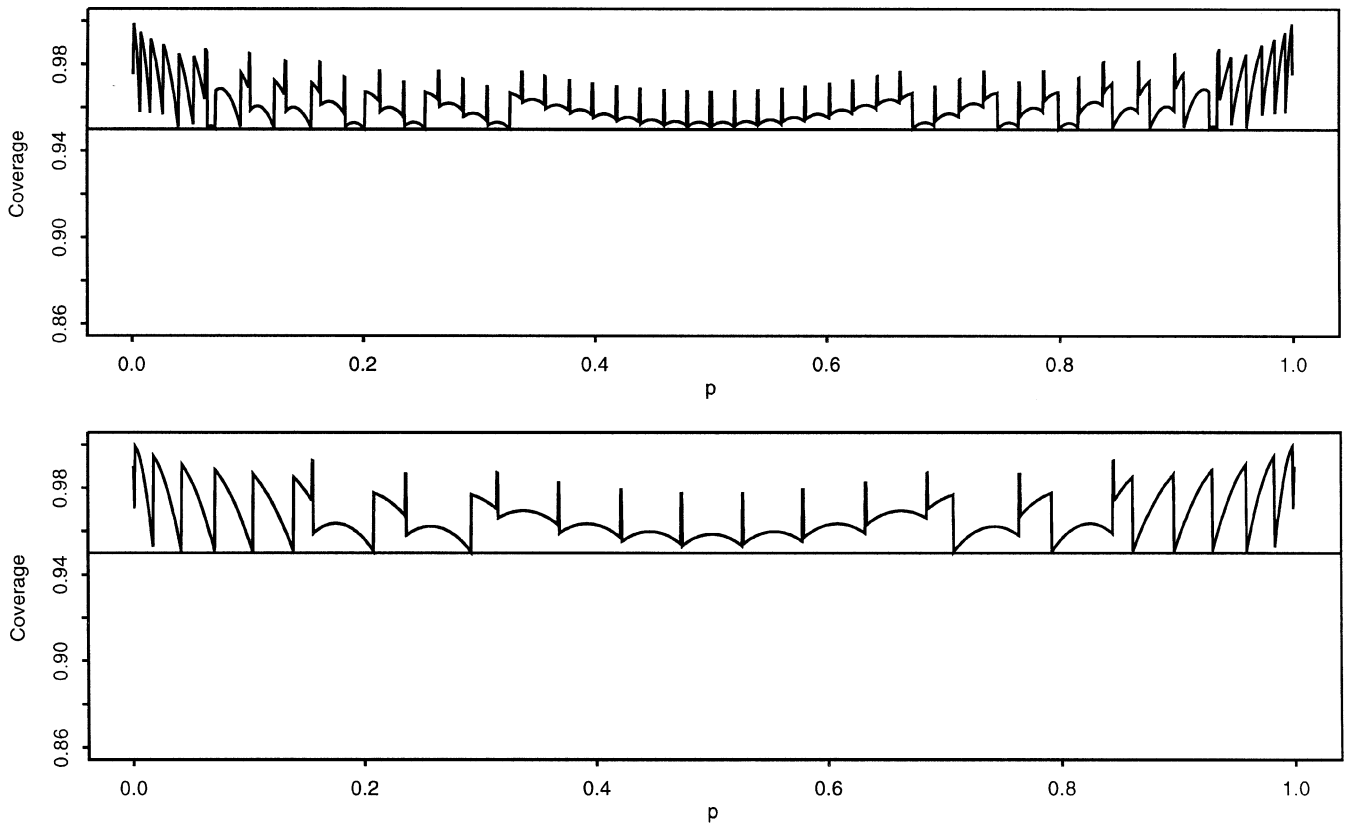

FIG. 1. *Coverage of nominal* 95% *symmetric Duffy–Santner* p *intervals for* $n = 20$ (*bottom panel*) *and* $n = 50$ (*top panel*).

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.

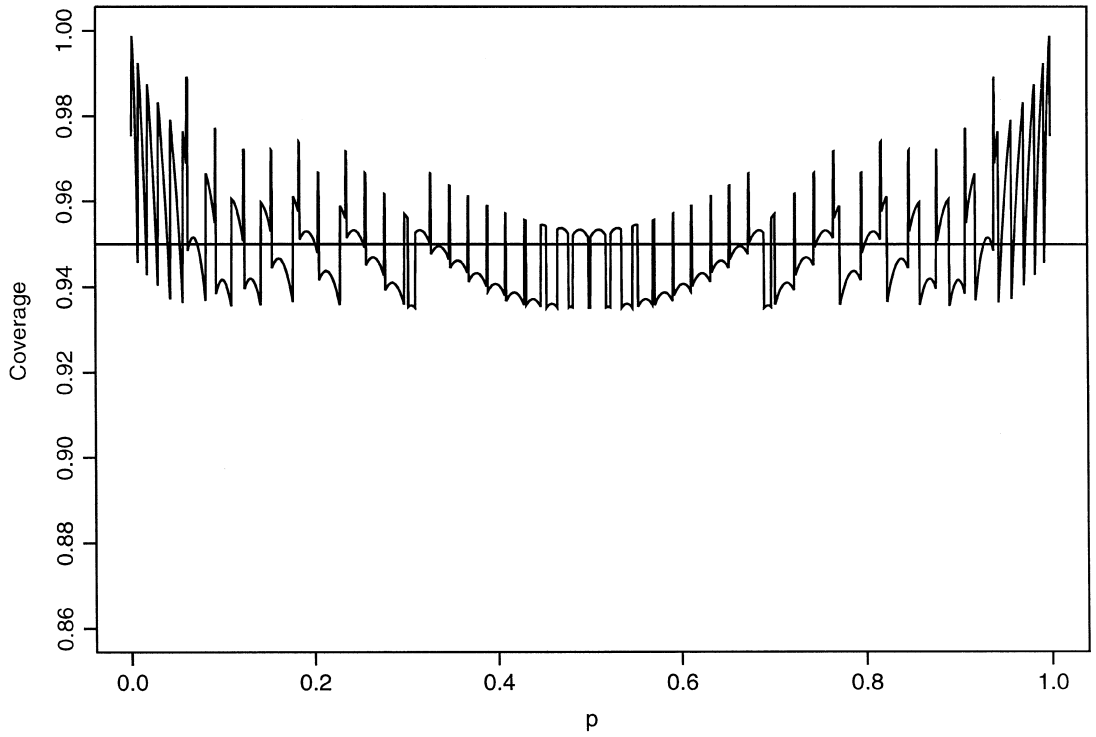

FIG. 2. *Coverage of nominal* 93.5% *symmetric Duffy–Santner* p *intervals for* $n = 50$.

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 $p_0 - p_i, 1 \le i \le T$ 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

this interval. We hope the continuing debate will add to a greater understanding of the problem, and we welcome the chance to contribute to this debate.

A. It seems that the primary source of disagreement is based on differences in interpretation of the coverage goals for confidence intervals. We will begin by presenting our point of view on this fundamental issue.
We will then turn to a number of other issues, as summarized in the following list:

B. Simplicity is important.
C. Expected length is also important.
D. Santner's proposal.
E. Should a continuity correction be used?
F. The Wald interval also performs poorly in other problems.
G. The two-sample binomial problem.
H. Probability-matching procedures.
I. Results from asymptotic theory.

A. Professors Casella, Corcoran and Mehta come out in favor of making coverage errors always fall only on the conservative side. This is a traditional point of view. However, we have taken a different perspective in our treatment. It seems more consistent with contemporary statistical practice to expect that a $\gamma\%$ confidence interval should cover the true value *approximately* $\gamma\%$ of the time. The approximation should be built on sound, relevant statistical calculations, and it should be as accurate as the situation allows.

We note in this regard that most statistical models are only felt to be approximately valid as representations of the true situation. Hence the resulting coverage properties from those models are at best only approximately accurate. Furthermore, a broad range of modern procedures is supported only by asymptotic or Monte-Carlo calculations, and so again coverage can at best only be approximately the nominal value. As statisticians we do the best within these constraints to produce procedures whose coverage comes close to the nominal value. In these contexts when we claim $\gamma\%$ coverage we clearly intend to convey that the coverage is close to $\gamma\%$, rather than to guarantee it is at least $\gamma\%$.

We grant that the binomial model has a somewhat special character relative to this general discussion. There are practical contexts where one can feel confident this model holds with very high precision. Furthermore, asymptotics are not *required* in order to construct practical procedures or evaluate their properties, although asymptotic calculations can be useful in both regards. But the discreteness of the problem introduces a related barrier to the construction of satisfactory procedures. This forces one to again decide whether $\gamma\%$ should mean "approximately $\gamma\%$," as it does in most other contemporary applications, or "at least $\gamma\%$" as can be obtained with the Blyth–Still procedure or the Cloppe–Pearson procedure. An obvious price of the latter approach is in its decreased precision, as measured by the increased expected length of the intervals.

B. All the discussants agree that elementary motivation and simplicity of computation are important attributes in the classroom context. We of course agree. If these considerations are paramount then the Agresti–Coull procedure is ideal. If the need for simplicity can be relaxed even a little, then we prefer the Wilson procedure: it is only slightly harder to compute, its coverage is clearly closer to the nominal value across a wider range of values of p, and it can be easier to motivate since its derivation is totally consistent with Neyman–Pearson theory. Other procedures such as Jeffreys or the mid-P Clopper–Pearson interval become plausible competitors whenever computer software can be substituted for the possibility of hand derivation and computation.

Corcoran and Mehta take a rather extreme position when they write, "Simplicity and ease of computation have *no role* to play in statistical practice [italics ours]." We agree that the ability to perform computations by hand should be of little, if any, relevance in practice. But conceptual simplicity, parsimony and consistency with general theory remain important secondary conditions to choose among procedures with acceptable coverage and precision.

These considerations will reappear in our discussion of Santner's Blyth–Still proposal. They also leave us feeling somewhat ambivalent about the boundary-modified procedures we have presented in our Section 4.1. Agresti and Coull correctly imply that other boundary corrections could have been tried and that our choice is thus somewhat ad hoc. (The correction to Wilson can perhaps be defended on the principle of substituting a Poisson approximation for a Gaussian one where the former is clearly more accurate; but we see no such fundamental motivation for our correction to the Jeffreys interval.)

C. Several discussants commented on the precision of various proposals in terms of expected length of the resulting intervals. We strongly concur that precision is the important balancing criterion vis-á-vis coverage. We wish only to note that there exist other measures of precision than interval expected length. In particular, one may investigate the probability of covering wrong values. In a

charming identity worth noting, Pratt (1961) shows the connection of this approach to that of expected length. Calculations on coverage of wrong values of p in the binomial case will be presented in DasGupta (2001). This article also discusses a number of additional issues and presents further analytical calculations, including a Pearson tilting similar to the chi-square tilts advised in Hall (1983).

Corcoran and Mehta's Figure 2 compares average length of three of our proposals with Blyth–Still and with their likelihood ratio procedure. We note first that their LB procedure is not the same as ours. Theirs is based on numerically computed exact percentiles of the fixed sample likelihood ratio statistic. We suspect this is roughly equivalent to adjustment of the chi-squared percentile by a Bartlett correction. Ours is based on the traditional asymptotic chi-squared formula for the distribution of the likelihood ratio statistic. Consequently, their procedure has conservative coverage, whereas ours has coverage fluctuating around the nominal value. They assert that the difference in expected length is "negligible." How much difference qualifies as negligible is an arguable, subjective evaluation. But we note that in their plot their intervals can be on average about 8% or 10% longer than Jeffreys or Wilson intervals, respectively. This seems to us a nonnegligible difference. Actually, we suspect their preference for their LR and BSC intervals rests primarily on their overriding preference for conservativity in coverage whereas, as we have discussed above, our intervals are designed to attain approximately the desired nominal value.

D. Santner proposes an interesting variant of the original Blyth–Still proposal. As we understand it, he suggests producing nominal $\gamma\%$ intervals by constructing the $\gamma^*\%$ Blyth–Still intervals, with $\gamma^*\%$ chosen so that the average coverage of the resulting intervals is approximately the nominal value, $\gamma\%$. The coverage plot for this procedure compares well with that for Wilson or Jeffreys in our Figure 5. Perhaps the expected interval length for this procedure also compares well, although Santner does not say so. However, we still do not favor his proposal. It is conceptually more complicated and requires a specially designed computer program, particularly if one wishes to compute $\gamma^*\%$ with any degree of accuracy. It thus fails with respect to the criterion of scientific parsimony in relation to other proposals that appear to have at least competitive performance characteristics.

E. Casella suggests the possibility of performing a continuity correction on the score statistic prior to constructing a confidence interval. We do not agree with this proposal from any perspective. These "continuity-corrected Wilson" intervals have extremely conservative coverage properties, though they may not in principle be guaranteed to be everywhere conservative. But even if one's goal, unlike ours, is to produce conservative intervals, these intervals will be very inefficient at their normal level relative to Blyth–Still or even Clopper–Pearson. In Figure 1 below, we plot the coverage of the Wilson interval with and without a continuity correction for $n = 25$ and $\alpha = 0.05$, and the corresponding expected lengths. It is seems clear that the loss in precision more than neutralizes the improvements in coverage and that the nominal coverage of 95% is misleading from any perspective.

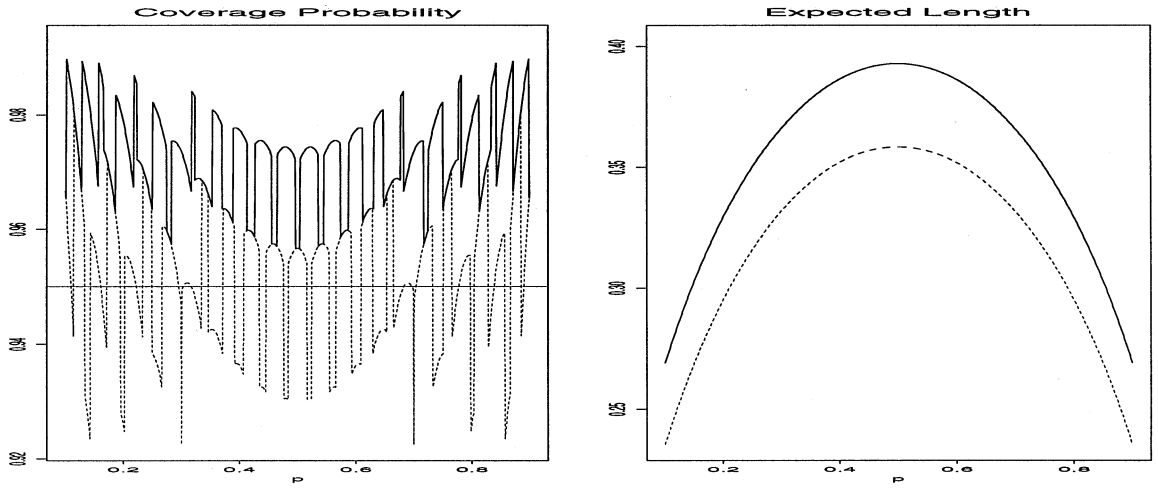

FIG. 1. *Comparison of the coverage probabilities and expected lengths of the Wilson (dotted) and continuity-corrected Wilson (solid) intervals for* $n = 25$ *and* $\alpha = 0.05$.

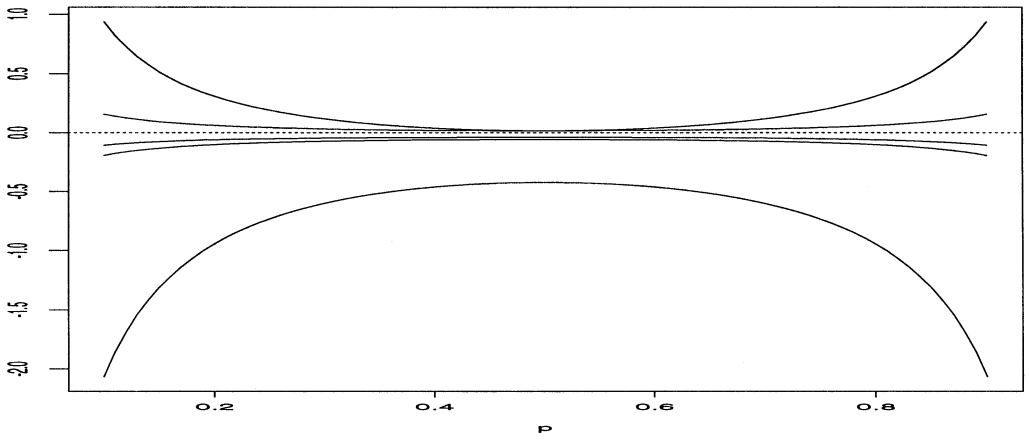

FIG. 2. *Comparison of the systematic coverage biases. The y-axis is* $nS_n(p)$. *From top to bottom*: *the systematic coverage biases of the Agresti–Coull*, *Wilson*, *Jeffreys*, *likelihood ratio and Wald intervals*, *with* $n = 50$ *and* $\alpha = 0.05$.

F. Agresti and Coull ask if the dismal performance of the Wald interval manifests itself in other problems, including nordiscrete cases. Indeed it does. In other lattice cases such as the Poisson and negative binomial, both the considerable negative coverage bias and inefficiency in length persist. These features also show up in some continuous exponential family cases. See Brown, Cai and DasGupta (2000b) for details.

In the three important discrete cases, the binomial, Poisson and negative binomial, there is in fact some conformity in regard to which methods work well in general. Both the likelihood ratio interval (using the asymptotic chi-squared limits) and the equal-tailed Jeffreys interval perform admirably in all of these problems with regard to coverage and expected length. Perhaps there is an underlying theoretical reason for the parallel behavior of these two intervals constructed from very different foundational principles, and this seems worth further study.

G. Some discussants very logically inquire about the situation in the two-sample binomial situation. Curiously, in a way, the Wald interval in the two-sample case for the difference of proportions is less problematic than in the one-sample case. It can nevertheless be somewhat improved. Agresti and Caffo (2000) present a proposal for this problem, and Brown and Li (2001) discuss some others.

H. The discussion by Ghosh raises several interesting issues. The definition of "first-order probability matching" extends in the obvious way to any set of upper confidence limits; not just those corresponding to Bayesian intervals. There is also an obvious extension to lower confidence limits. This probability matching is a one-sided criterion. Thus a family of two-sided intervals $[L_n, U_n]$ will be first-order probability matching if

$$Pr_p(p \le L_n) = \alpha/2 + o(n^{-1/2}) = Pr_p(p \ge U_n).$$

As Ghosh notes, this definition cannot usefully be literally applied to the binomial problem here, because the asymptotic expansions always have a discrete oscillation term that is $O(n^{-1/2})$. However, one can correct the definition.

One way to do so involves writing asymptotic expressions for the probabilities of interest that can be divided into a "smooth" part, S, and an "oscillating" part, Osc, that averages to $O(n^{-3/2})$ with respect to any smooth density supported within (0, 1). Readers could consult BCD (2000a) for more details about such expansions. Thus, in much generality one could write

$$\begin{aligned} &Pr_p(p \le L_n) \\ &\quad = \alpha/2 + S_{L_n}(p) + Osc_{L_n}(p) + O(n^{-1}), \end{aligned} \tag{1}$$

where $S_{L_n}(p) = O(n^{-1/2})$, and $Osc_{L_n}(p)$ has the property informally described above. We would then say that the procedure is first-order probability matching if $S_{L_n}(p) = o(n^{-1/2})$, with an analogous expression for the upper limit, U_n.

In this sense the equal-tail Jeffreys procedure is probability matching. We believe that the mid-P Clopper–Pearson intervals also have this asymptotic property. But several of the other proposals, including the Wald, the Wilson and the likelihood ratio intervals are not first-order probability matching. See Cai (2001) for exact and asymptotic calculations on one-sided confidence intervals and hypothesis testing in the discrete distributions.

The failure of this one-sided, first-order property, however, has no obvious bearing on the coverage properties of the two-sided procedures considered in the paper. That is because, for any of our procedures,

$$S_{L_n}(p) + S_{U_n}(p) = 0 + O(n^{-1}), \tag{2}$$

even when the individual terms on the left are only $O(n^{-1/2})$. All the procedures thus make compensating one-sided errors, to $O(n^{-1})$, even when they are not accurate to this degree as one-sided procedures.

This situation raises the question as to whether it is desirable to add as a secondary criterion for two-sided procedures that they also provide accurate one-sided statements, at least to the probability matching $O(n^{-1/2})$. While Ghosh argues strongly for the probability matching property, his argument does not seem to take into account the cancellation inherent in (2). We have heard some others argue in favor of such a requirement and some argue against it. We do not wish to take a strong position on this issue now. Perhaps it depends somewhat on the practical context—if in that context the confidence bounds may be interpreted and used in a one-sided fashion as well as the two-sided one, then perhaps probability matching is called for.

I. Ghosh's comments are a reminder that asymptotic theory is useful for this problem, even though exact calculations here are entirely feasible and convenient. But, as Ghosh notes, asymptotic expressions can be startingly accurate for moderate sample sizes. Asymptotics can thus provide valid insights that are not easily drawn from a series of exact calculations. For example, the two-sided intervals also obey an expression analogous to (1),

$$Pr_p(L_n \le p \le U_n) = 1 - \alpha + S_n(p) + Osc_n(p) + O(n^{-3/2}). \tag{3}$$

The term $S_n(p)$ is $O(n^{-1})$ and provides a useful expression for the smooth center of the oscillatory coverage plot. (See Theorem 6 of BCD (2000a) for a precise justification.) The following plot for $n = 50$ compares $S_n(p)$ for five confidence procedures. It shows how the Wilson, Jeffreys and chi-squared likelihood ratio procedures all have coverage that well approximates the nominal value, with Wilson being slightly more conservative than the other two.

As we see it our article articulated three primary goals: to demonstrate unambiguously that the Wald interval performs extremely poorly; to point out that none of the common prescriptions on when the interval is satisfactory are correct and to put forward some recommendations on what is to be used in its place. On the basis of the discussion we feel gratified that we have satisfactorily met the first two of these goals. As Professor Casella notes, the debate about alternatives in this timeless problem will linger on, as it should. We thank the discussants again for a lucid and engaging discussion of a number of relevant issues. We are grateful for the opportunity to have learned so much from these distinguished colleagues.

ADDITIONAL REFERENCES

AGRESTI, A. and CAFFO, B. (2000). Simple and effective confidence intervals for proportions and differences of proportions result from adding two successes and two failures. *Amer. Statist.* **54**. To appear.

AITKIN, M., ANDERSON, D., FRANCIS, B. and HINDE, J. (1989). *Statistical Modelling in GLIM*. Oxford Univ. Press.

BOOS, D. D. and HUGHES-OLIVER, J. M. (2000). How large does n have to be for Z and t intervals? *Amer. Statist.* **54** 121–128.

BROWN, L. D., CAI, T. and DASGUPTA, A. (2000a). Confidence intervals for a binomial proportion and asymptotic expansions. *Ann. Statist.* To appear.

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

BROWN, L. D. and LI, X. (2001). Confidence intervals for the difference of two binomial proportions. Unpublished manuscript.

CAI, T. (2001). One-sided confidence intervals and hypothesis testing in discrete distributions. Preprint.

COE, P. R. and TAMHANE, A. C. (1993). Exact repeated confidence intervals for Bernoulli parameters in a group sequential clinical trial. *Controlled Clinical Trials* **14** 19–29.

COX, D. R. and REID, N. (1987). Orthogonal parameters and approximate conditional inference (with discussion). *J. Roy. Statist. Soc. Ser. B* **49** 113–147.

DASGUPTA, A. (2001). Some further results in the binomial interval estimation problem. Preprint.

DATTA, G. S. and GHOSH, M. (1996). On the invariance of noninformative priors. *Ann. Statist.* **24** 141–159.

DUFFY, D. and SANTNER, T. J. (1987). Confidence intervals for a binomial parameter based on multistage tests. *Biometrics* **43** 81–94.

FISHER, R. A. (1956). *Statistical Methods for Scientific Inference*. Oliver and Boyd, Edinburgh.

GART, J. J. (1966). Alternative analyses of contingency tables. *J. Roy. Statist. Soc. Ser. B* **28** 164–179.

GARVAN, C. W. and GHOSH, M. (1997). Noninformative priors for dispersion models. *Biometrika* **84** 976–982.

GHOSH, J. K. (1994). *Higher Order Asymptotics*. IMS, Hayward, CA.

HALL, P. (1983). Chi-squared approximations to the distribution of a sum of independent random variables. *Ann. Statist.* **11** 1028–1036.

JENNISON, C. and TURNBULL, B. W. (1983). Confidence intervals for a binomial parameter following a multistage test with application to MIL-STD 105D and medical trials. *Technometrics*, **25** 49–58.

JORGENSEN, B. (1997). *The Theory of Dispersion Models*. CRC Chapman and Hall, London.

LAPLACE, P. S. (1812). *Théorie Analytique des Probabilités*. Courcier, Paris.

LARSON, H. J. (1974). *Introduction to Probability Theory and Statistical Inference*, 2nd ed. Wiley, New York.

PRATT, J. W. (1961). Length of confidence intervals. *J. Amer. Statist. Assoc.* **56** 549–567.

RINDSKOPF, D. (2000). Letter to the editor. *Amer. Statist.* **54** 88.

RUBIN, D. B. and SCHENKER, N. (1987). Logit-based interval estimation for binomial data using the Jeffreys prior. *Sociological Methodology* **17** 131–144.

STERNE, T. E. (1954). Some remarks on confidence or fiducial limits. *Biometrika* **41** 275–278.

TIBSHIRANI, R. (1989). Noninformative priors for one parameter of many. *Biometrika* **76** 604–608.

WELCH, B. L. and PEERS, H. W. (1963). On formula for confidence points based on intergrals of weighted likelihoods. *J. Roy. Statist. Ser. B* **25** 318–329.

YAMAGAMI, S. and SANTNER, T. J. (1993). Invariant small sample confidence intervals for the difference of two success probabilities. *Comm. Statist. Simul. Comput.* **22** 33–59.