A Coverage Probability Approach to Finding an Optimal Binomial Confidence Procedure

Mark F. SCHILLING and Jimmy A. DOI

The problem of finding confidence intervals for the success parameter of a binomial experiment has a long history, and a myriad of procedures have been developed. Most exploit the duality between hypothesis testing and confidence regions and are typically based on large sample approximations. We instead employ a direct approach that attempts to determine the optimal coverage probability function a binomial confidence procedure can have from the exact underlying binomial distributions, which in turn *defines* the associated procedure. We show that a graphical perspective provides much insight into the problem. Both procedures whose coverage never falls below the declared confidence level and those that achieve that level only approximately are analyzed. We introduce the Length/Coverage Optimal method, a variant of Sterne's procedure that minimizes average length while maximizing coverage among all length minimizing procedures, and show that it is superior in important ways to existing procedures.

KEY WORDS: Binomial confidence intervals; Exact confidence interval; Length minimizing.

1. INTRODUCTION

Obtaining a confidence interval for a binomial success parameter is one of the most common and basic of statistical problems. A great number of solutions have been proposed in the 80-plus years since the original development of confidence intervals, continuing even into the 21st century. Yet after all this time and study, still no consensus has emerged.

One reason is that two distinct standards have been used—one which requires the confidence procedure to adhere to the "strict" classical requirement for coverage: $\inf_p P(p \in \mathrm{CI}) \geq 1–\alpha$ where p is the binomial parameter, the other which allows this requirement to be satisfied only approximately. Here, we present a new method that is optimal with respect to length and coverage for the strict case, and provide an approximate version that outperforms existing procedures for the approximate coverage criterion. We adopt Casella's (1986) terminology in defining a *confidence procedure* for a given sample size n as a collection of $n + 1$ intervals, one for each possible value of the binomial random variable X. Thus,we do not consider randomized confidence intervals here.

The prevailing confidence procedure presented in elementary textbooks, known as the *Wald method*, is not a strict method. The Wald interval, $\hat{p} \pm z_{\alpha/2}(\hat{p}(1-\hat{p})/n)^{1/2}$, is one of a large collection that are based on a normal approximation, with various conditions on minimum sample size. In addition to leaving unaddressed the instances where those conditions are not met, the coverage probability of the Wald interval is often far below the claimed level—even for cases that are well within the guidelines. See Brown, Cai, and DasGupta (2001) for a thorough analysis.

Many alternatives to the Wald method have been developed, most based on concepts central to statistical theory such as the normal approximation to the binomial, likelihood, the score function, Bayesian methods, and so forth. Ultimately, all that matters to those who use confidence intervals in the real world is performance. From this standpoint, in our view, there are two prime criteria: Are adequate coverage probabilities achieved throughout the entire range of parameter values? That is, does the confidence procedure deliver what it promises? And does the procedure produce confidence intervals that are as narrow as possible? Certainly there are other points of view regarding the primacy of these two criteria. See, for example, Vos and Hudson (2005, 2008) for an alternate perspective.

In the evaluation of confidence procedures for practical use, we regard the two overriding considerations to be length minimization and maximal coverage, in that order. What is desired is an interval estimate of the parameter value that is as precise as possible and can be regarded as very likely to contain it. We believe that practitioners often tend to think loosely of a confidence interval as a virtual de facto guarantee that the parameter lies within the interval. Thus, maximizing coverage without an increase in average length is beneficial, since overshooting the

Mark F. Schilling, Department of Mathematics, California State University Northridge, Northridge, CA 91330 (E-mail: *mark.schilling@csun.edu*). Jimmy A. Doi, Department of Statistics, California Polytechnic State University, San Luis Obispo, CA 93407 (E-mail: *jdoi@calpoly.edu*). The authors thank Bret Holladay for his invaluable assistance with the programming and figures needed for this article, as well as the associate editor and referees for their insightful comments which helped improve this article.

DOI: 10.1080/00031305.2014.899274

stated confidence level puts practice more in line with this common interpretation.

Historically, generating confidence intervals by means of a formula was also a valuable goal. That is a highly anachronistic view today, given that *exact* confidence intervals (those based directly on the binomial distribution) are easily generated and available via statistical software and the Internet. Several other properties *are* desirable; since in our opinion these are secondary in importance, we will defer addressing them until after considering coverage and length.

2. BINOMIAL COVERAGE PROBABILITY FUNCTIONS

For given p, the coverage probability CP(p) of a binomial confidence procedure is the chance of observing a number of successes X for which the associated confidence interval contains p. The set of all CP(p), $0 \leq p \leq 1$, is called the *coverage probability function* (*cpf*) of the confidence procedure. For any subset Ω of $\{0, 1, 2, \ldots, n\}$, we define the *acceptance curve* $\mathrm{AC}_{\Omega}(p)$ to be the Ω-likelihood function; that is, $\mathrm{AC}_{\Omega}(p) = P(X \in \Omega) = \sum_{x\in\Omega}\binom{n}{x}p^{x}(1-p)^{n-x}$ considered as a function of p. The coverage probability function is necessarily composed of some collection of portions of acceptance curves, and *stipulating the cpf is equivalent to specifying the confidence procedure* (up to a set of measure zero). Thus, the binomial confidence interval problem can be approached directly from consideration of how to choose acceptance curves for each p.

For a sensible confidence procedure, a value of x that is reasonably likely to occur for a given p should contain that p in the confidence interval for x, while an unlikely value of x should not. Since any binomial probability distribution is unimodal, the values of x that should exclude p will be found in one or both tails of the distribution. We can therefore restrict consideration to acceptance curves associated with the sets $\Omega_{lu} = \{l, l+1, l+2, \ldots, u-1, u\}$, which are called *acceptance intervals* (Blyth and Still 1983).

In what follows, we write AC(l–u) for $\mathrm{AC}_{\Omega}(p)$ when $\Omega = \Omega_{lu}$. We call the difference $u-l$ the *span* of the acceptance curve. Figure 1 shows labeled portions of all such curves for $n = 8$ that are above or not much below 90%. For a strict 90% confidence procedure, the cpf must be built from segments that lie entirely above the line at 0.90, while a reasonable approximate 90% confidence procedure could include portions of those curves shown in Figure 1 that fall somewhat below 0.90.

2.1 Type O and Type I Acceptance Curves

Note that $\mathrm{AC}_{\Omega_{0u}}(p) \geq P(X = 0) = (1 - p)^{n}$ and $\mathrm{AC}_{\Omega_{ln}}(p) \geq P(X = n) = p^{n}$; all such acceptance curves are continuous and have a maximum of 1, therefore each exceeds $1-\alpha$ for p sufficiently near 0 and 1, respectively. We call curves of the form $\mathrm{AC}_{\Omega_{0u}}(p)$ and $\mathrm{AC}_{\Omega_{ln}}(p)$ *Type O* acceptance curves. If we wish to build a cpf for a strict 90% confidence procedure for $n = 8$, we can see from Figure 1 that for small p only the Type O acceptance curves AC(0–0), AC(0–1), . . ., AC(0–8) are in play. As p increases, the AC(0–0)

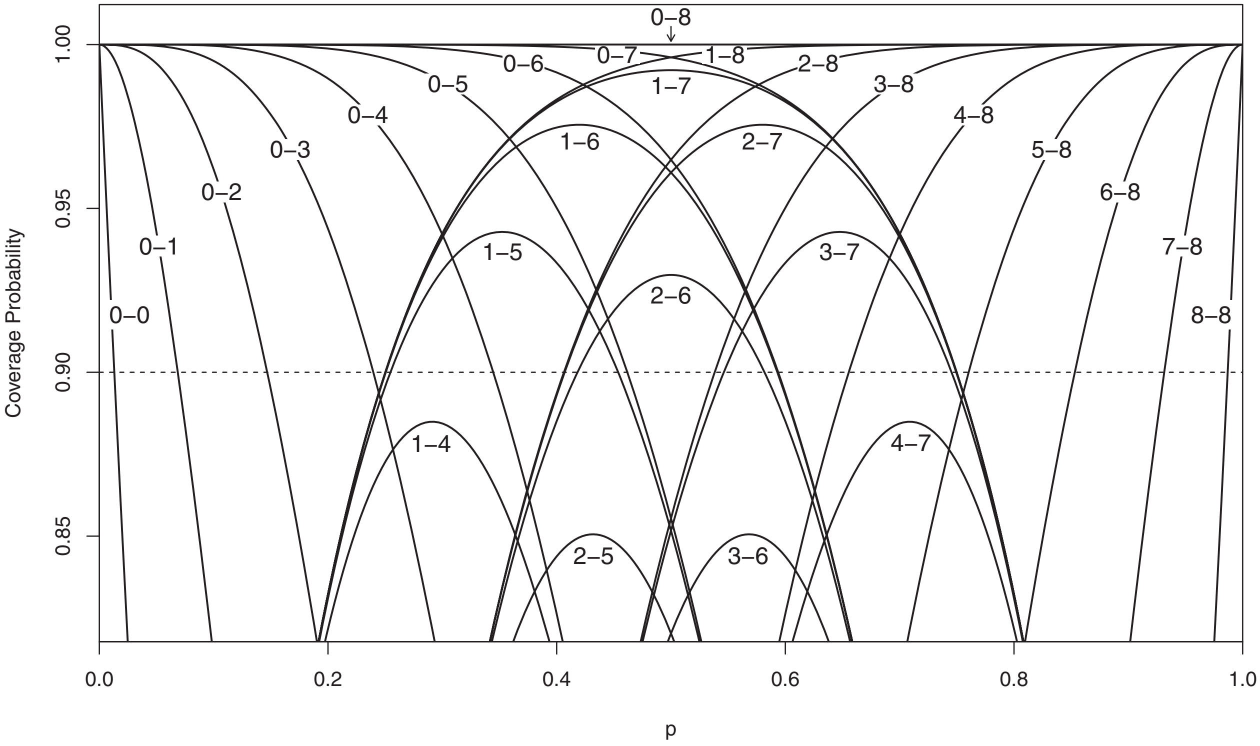

Figure 1. Portions of acceptance curves eligible for strict and reasonable approximate 90% confidence procedures for $n = 8$. Each curve AC(l–u) is labeled with its l and u values.

curve drops out, then AC(0–1) does, and so on. After $p \approx 0.26$, acceptance curves for which $l > 0$ are eligible for selection. Then after $p \approx 0.74$, only the Type O curves AC(0–8), AC(1–8), . . ., AC(8–8) can be used.

We define *Type I* acceptance curves to be those associated with acceptance intervals for which both $l > 0$ and $u < n$. It is easy to show that such curves are continuous, unimodal, and satisfy $\lim_{p\to 0^+} \mathrm{AC}(p) = \lim_{p\to 1^-} \mathrm{AC}(p) = 0$. Thus for p near 0 and 1, a binomial coverage probability function must be comprised entirely of Type O acceptance curves; elsewhere, a cpf will nearly always consist entirely of Type I acceptance curves. It is easy to distinguish the Type O and Type I curves in Figure 1 as the monotone and nonmonotone curves, respectively.

To see the role acceptance curves play in an established confidence procedure, consider the "adjusted Wald" method proposed by Agresti and Coull (1998) for the case $n = 8$, $1-\alpha = 90\%$. This procedure simply adds a set number of successes and failures (possibly noninteger) to the data, much in the manner of a Bayesian prior, before computing the Wald interval. Adjusted Wald is not a strict confidence procedure, and therefore includes portions of acceptance curves that fall below 90%. However, the adjusted Wald's cpf aligns much more closely with the nominal confidence level than that of the ordinary Wald procedure. Figure 2(a) displays the cpf for the 90% adjusted Wald procedure for $n = 8$, overlaid on the acceptance curves this procedure uses. (*Technical note*. In constructing cpf graphs for this article, we treat confidence intervals as half open, of the form $[l_x, u_x)$, except for $x = n$, where we take the interval to be closed. This avoids spikes in the graphs at values of p which represent the upper limit of one acceptance interval and the lower limit of another. We sometimes include vertical line segments between consecutive acceptance curves for better visualization of the cpfs.)

Figure 2(a) shows an inner section composed of seven Type I pieces, surrounded by Type O regions on each side. This structure occurs for any reasonable confidence procedure unless n is very small or the confidence level is very high, in which case there is no Type I region. Figure 2(b) shows the interplay between the coverage probability function, the specific acceptance curves involved, and the resulting confidence intervals.

2.2 Constraints on Acceptance Curve Choices

The adjusted Wald cpf shown in Figures 2(a) and (b) is symmetrical around $p = 0.5$, and as a result the associated confidence intervals satisfy the *equivariance* property (Blyth and Still 1983):

$$\text{If } x \text{ generates the confidence interval } [l_x, u_x]\,, \text{ then}$$
$$n - x \text{ yields the confidence interval } [1 - u_x, 1 - l_x]\,.$$

Equivariance is essential for a binomial confidence procedure, since switching "success" and "failure" should yield an equivalent outcome. A confidence procedure is equivariant if and only if it has a symmetrical cpf, thus an acceptable procedure must satisfy CP(p) = CP(1–p).

Next, observe that as p increases from 0 to 1 in Figure 2(b), the values of l and u for the acceptance curves used are nondecreasing. Looking back at Figure 1, we can see that in principle a confidence procedure could be constructed that did not have this property. For example, the cpf could jump from AC(1–6) to AC(1–5) for a time before having to return to a curve for which $u \geq 6$. But then the confidence set for $x = 6$ would not be an interval, as it would have a gap where p corresponds to AC(1–5). To avoid gaps, the sequences of l and u values must each be nondecreasing in x.

2.3 Some Typical Coverage Probability Functions

Figures 3(a)–(c) show the cpfs of the Wald, adjusted Wald, and Clopper–Pearson procedures, respectively, for the case $n = 20$, $1 - \alpha = 95\%$. A Clopper–Pearson confidence interval (Clopper and Pearson 1934) is obtained by solving each of the equations $P(X \geq x) = \alpha/2$ and $P(X \leq x) = \alpha/2$ for p, the solution to the first equation being l_x and the solution to the second

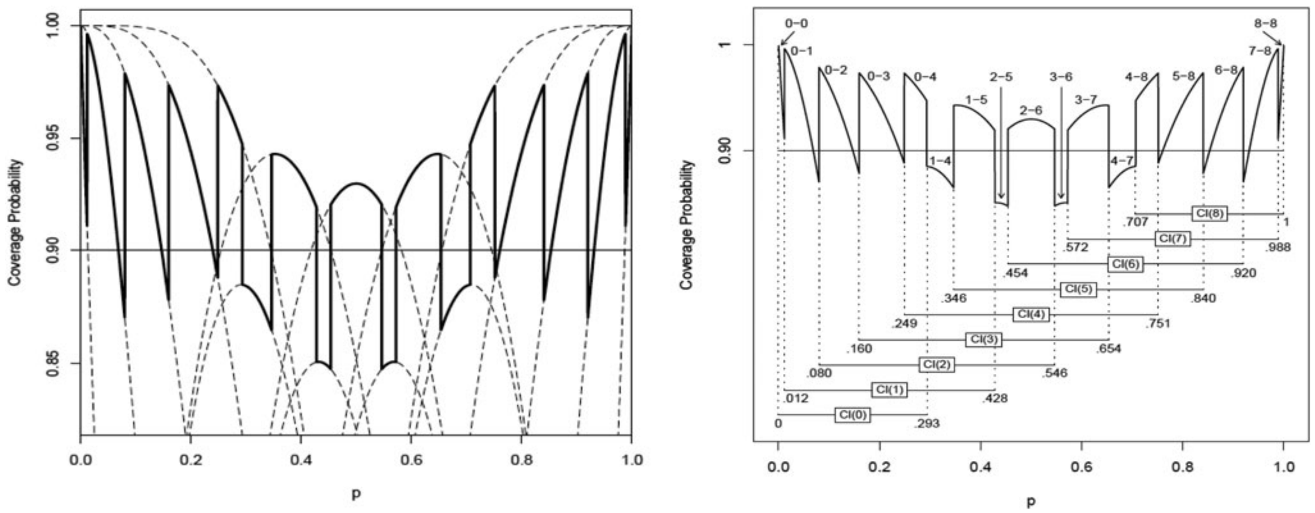

Figure 2. (a) Adjusted Wald coverage probability function and associated acceptance curves ($n = 8$, nominal confidence level = 90%). (b) Interplay between the cpf, the specific acceptance curves involved, and the resulting confidence intervals for the adjusted Wald 90% procedure.

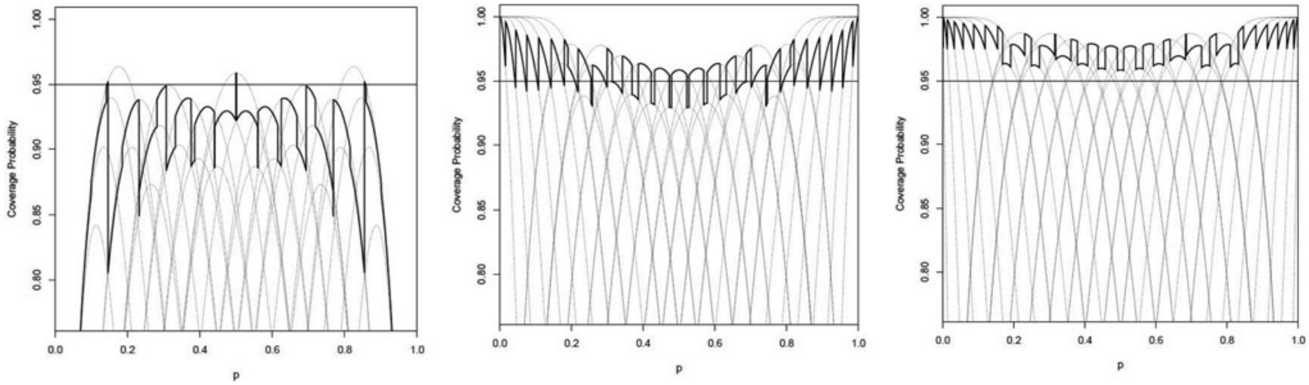

Figure 3. (a)–(c). Coverage probability plots for the (a) Wald, (b) adjusted Wald, and (c) Clopper–Pearson procedures for $n = 20$, $1-\alpha = 95\%$.

being u_x. (For $x = 0$ and $x = n$ there are no solutions to the first and second equations, respectively; the Clopper–Pearson method adopts $l_0 = 0$ and $u_n = 1$ for these two cases.) Note first that, as mentioned earlier, the Wald cpf falls far below the confidence level in several places; in fact at $p = 0$ and 1, the Wald cpf drops to 0. Furthermore, nearly the entire cpf lies below the confidence level. The adjusted Wald fares much better, but still falls below the confidence level for many values of p. The Clopper–Pearson procedure, in contrast, produces a cpf that is excessively high, often considerably above the confidence level. Of itself this is not a drawback, but it comes at a significant cost: the corresponding confidence intervals are substantially larger than necessary.

All three of these popular procedures share the feature that in the large central region of the plot, the cpf shifts very frequently between different Type I acceptance curves. Although such behavior invariably manifests itself for formula based binomial confidence procedures, such large and frequent fluctuations in the cpf are a reflection of an inefficient method.

3. LENGTH MINIMIZING STRICT CONFIDENCE PROCEDURES

The primary measure of the quality of a confidence procedure that achieves satisfactory coverage, whether in the strict sense or some appropriately defined approximate sense, is in our view the shortness of its confidence intervals. There are two criteria that are commonly used—average length and expected length. Average length is simply the arithmetic mean of the lengths of the $n + 1$ confidence intervals that a binomial confidence procedure can produce. Expected length, in contrast, is a function of p, for each p weighting each of the $n + 1$ possible confidence intervals by the probabilities of the associated values of x. If expected length is integrated with respect to p over the entire parameter space [0,1], the result is average length. That is, average length is a functional of expected length that provides an overall assessment of expected length. Consequently, we focus only on average length and often refer to it simply as *length*.

Figure 1 illustrates that there are many choices that can be made regarding the sequence of acceptance curve segments that can be used to construct the cpf of a legitimate binomial confidence procedure. It follows from Crow (1956) and Casella (1986), however, that selecting at each p an acceptance curve with minimal span among those curves with coverage above $1-\alpha$ is necessary and sufficient for a strict procedure to have minimum average length.

Figure 4 shows all portions of acceptance curves that have minimal span for $n = 8$, $1-\alpha = 90\%$. Note that in the Type O regions the acceptance curves with minimal span are unique. The locations where the left-side Type O curves intersect the confidence level therefore establish the lower limits of the confidence intervals for the first several values of x at the highest possible values of p. Each such confidence limit can easily be shown to be the α-quantile of a beta distribution with parameters x and $n - x + 1$. Corresponding statements apply to the upper limits for the highest values of x, by equivariance.

In the region of the graph where portions of Type I curves are eligible for selection (being ≥ 1–α), there are many intervals in which two or more acceptance curves with the same minimal span exist. These intervals play a key role in the quest for an optimal procedure.

3.1 The Search for an Optimal Binomial Confidence Procedure

Our goal is to find a strict binomial confidence procedure that maximizes coverage among all length minimizing procedures. Initially the recipe appears simple: select at each p the highest

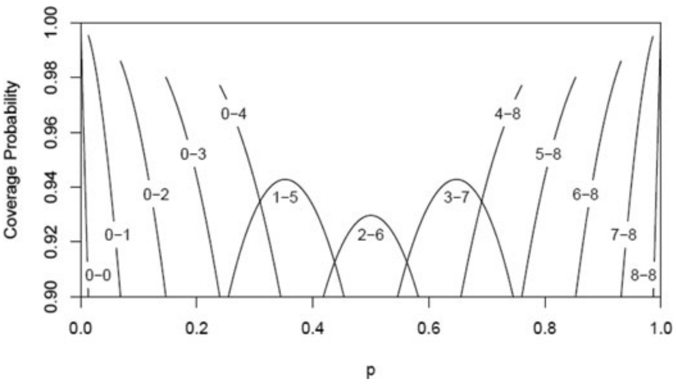

Figure 4. All portions of acceptance curves available for a strict length minimizing procedure for $n = 8$, $1-\alpha = 90\%$.

acceptance curve of minimal span for that p. (In the case when two acceptance curves of minimal span each have maximal height, we assign to p the one with the higher l and u values to maintain half-open intervals of the form [,).) This turns out to be equivalent to a method proposed by Sterne (1954) under an entirely different rationale. In the language of acceptance curves, Sterne's approach was to construct acceptance intervals for each p by entering the values of x in decreasing order of their chance of being observed until $\mathrm{AC}_{\Omega_{l\,u}}(p) \geq 1-\alpha$. It follows immediately that no other set of x values with as few or fewer elements can have higher total probability; furthermore the unimodality of binomial distributions implies that for any p Sterne's set of values must be consecutive. This establishes the equivalence with the recipe given above.

For any R in $\{0, 1, 2, \ldots, n\}$, form the upper envelope function $N_R = N_R(p)$ of all acceptance curves for which $u - l = R$. We call N_R the *necklace of span R.* The cpf of Sterne's method lies entirely on necklaces. Specifically, for each p, the lowest lying necklace that is above $1-\alpha$ is selected. Figure 5 shows the case $n = 20$, $1-\alpha = 95\%$. Sterne's cpf is comprised of portions of the necklaces with spans ≤ 8. An immediate consequence of the above strategy is that each point where $\mathrm{CP}(p) = 1-\alpha$ establishes a single confidence limit, and each necklace cusp point represents both a lower confidence limit for one x and an upper confidence limit for another x.

Unfortunately, Crow (1956) pointed out that this approach sometimes produces confidence sets for some values of x that are not intervals—that is, they contain gaps, as the necessary monotonicity in l and u described in Section 2.2 fails to hold. We now show precisely when and why this happens by identifying an essential relationship between Type I curves. We then investigate remedies.

One would expect that the peaks of the Type I acceptance curves must move to the right as l and/or u increase, as held in the case shown in Figure 1. We confirm in Proposition 1 that this is true in general; the proof is given in the Appendix. Let $p_M(l, u; n)$ be the unique value of p that maximizes AC(l–u) for given sample size n:

Proposition 1. For fixed n and $l > 0$, $p_M(l, u; n)$ is an increasing function of u for $l \leq u \leq n-1$, and for fixed n and $u < n$, $p_M(l, u; n)$ is an increasing function of l for $1 \leq l \leq u$.

Recall that to obtain confidence *intervals* (having no gaps), the values of l and u for the sequence of acceptance curves used in a confidence procedure must be nondecreasing. As has already been shown, the Type O curves *AC*(0–u) and *AC*(l–n) each have a maximum of 1, occurring at $p = 0$ and $p = 1$ respectively. From this fact together with Proposition 1, it follows that whenever the cpf transitions from one acceptance curve to another as p increases, the maximum of the new acceptance curve cannot occur to the left of the previous one's or a gap will result.

Figure 6 illustrates how this requirement may be violated for Sterne's method. Moving from left to right, the cpf travels along the first four Type O curves before moving to necklace N_4 and using first AC(0–4), then briefly AC(1–5). At approximately p = 0.141, however, AC(1–5) drops below the confidence level and the necklace with minimum span now becomes N_5. Since AC(1–5) drops out *before* the first cusp of N_5, the Sterne cpf next uses AC(0–5), a curve whose maximum is to the left of AC(1–5)'s, and the value of l decreases in this transition. Consequently, the 90% confidence set for $x = 0$ has a gap from approximately 0.127 to 0.141, consisting of the p interval in which the cpf lies on AC(1–5). We see that if a portion of the cpf lies on an acceptance curve AC(l–u) that rises just slightly above the confidence level but then drops below the confidence level and out of contention before (for $p < 0.5$) the location of the cusp of the necklace just above (joining AC((l−1)–u) to AC(l–(u+1))), then a gap results, as Sterne's cpf transfers to an acceptance curve with a lower value of l. Figure 6 in fact shows two such gaps, the second arising from the fact that AC(2–7) drops

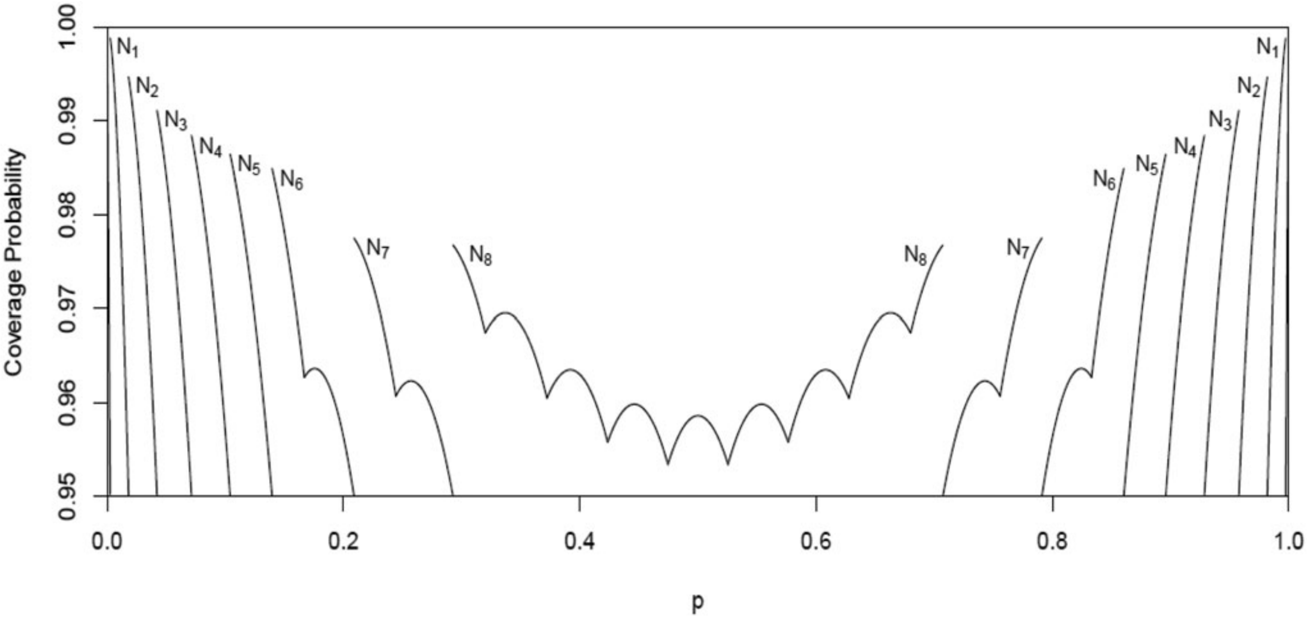

Figure 5. Cpf of Sterne's confidence procedure for $n = 20$, $1-\alpha = 95\%$. The barely visible portions near $p = 0$ and $p = 1$ belong to the necklace N_0.

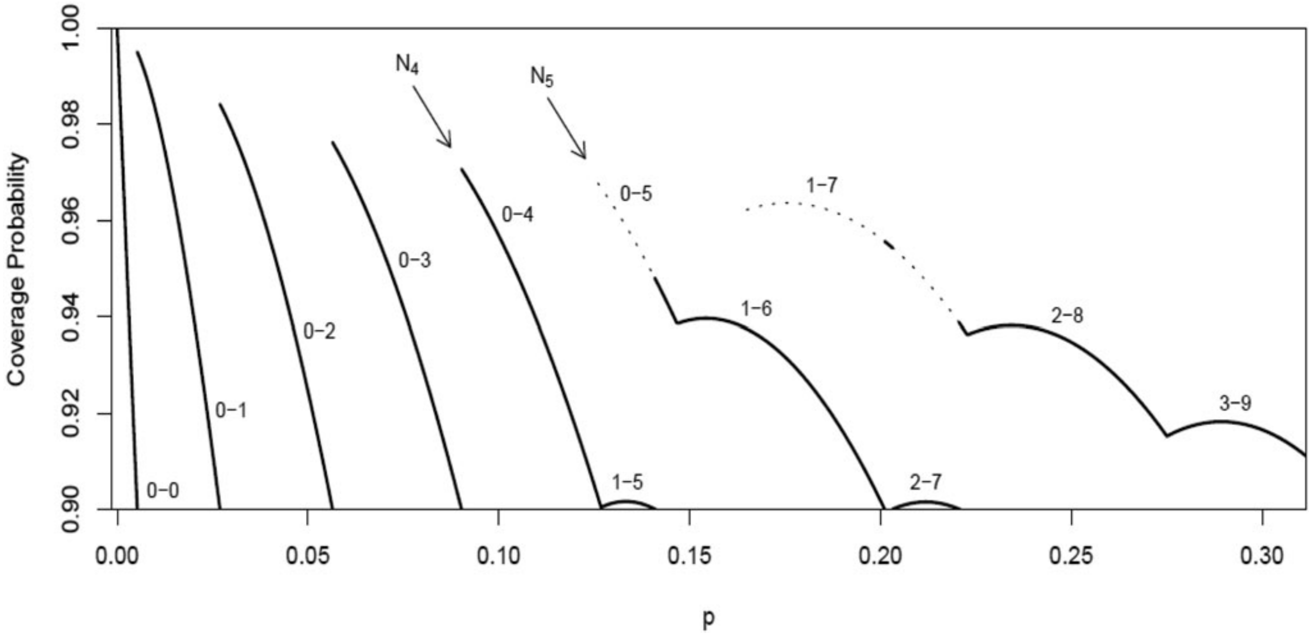

Figure 6. Portion of Sterne's cpf for $n = 20$, $1-\alpha = 90\%$ showing how gaps can occur.

below the confidence level before the cusp joining AC(1–7) to AC(2–8).

The collection of all length minimizing strict binomial confidence procedures—that is, those that do not produce gaps—was identified by Casella (1986). These procedures can now be characterized as those deriving from every possible cpf that is composed entirely of pieces of the acceptance curves of minimal span but never moves from one curve to another whose peak is to the left of its own.

The fact that Sterne's approach often fails to produce confidence *intervals* (and perhaps because in the 1950s there was resistance to nonformula based methods) may explain why his method was not widely adopted. Crow (1956) and Blyth and Still (1983) proposed remedies for the gap problem, each based on altering the transition rule whenever there is a choice between acceptance curves of minimum span to ensure that each of the sequences of l and u values are monotone increasing. However, Crow's method has a critical drawback in that its interval endpoints in (0,1) are often the same for different x. For example, for $n = 20$, $1-\alpha = 95\%$ the confidence intervals for $x = 9$, 10, and 11 are respectively (0.222, 0.707), (0.293, 0.707), and (0.293, 0.778), surely an unappealing result. Crow's procedure also has significantly lower coverage than Sterne's method. Blyth and Still's method does not have the shortcoming of shared lower or upper endpoints and performs quite well overall, but its coverage is lower than for Sterne's method for many p as its cpf does not lie entirely on the necklaces.

3.2 The LCO Procedure

Both Crow's and Blyth and Still's methods eliminate the gaps that Sterne's procedure sometimes generates, but sacrifice coverage to do so. There are other ways to handle the gap problem, however. One is to simply "fill the gaps" (Reiczigel 2003, Hirji 2006). For example, in the first case shown in Figure 6, the Sterne confidence set for $x = 0$, $[0, 0.127) \cup [0.141, 0.147)$, yields the "gap filled" confidence interval [0, 0.147). Obviously this remedy fails to achieve length minimization, however.

We now present an alternative strategy that avoids gaps and still produces a strict length minimizing procedure, having the additional property of maximizing coverage among all such procedures. We call the resulting procedure and intervals *Length/Coverage Optimal*, or LCO for short. The confidence intervals produced by the LCO method are identical to those of Sterne's procedure when the latter does not produce gaps; however, we investigated all 300 cases with $n \leq 100$ and $1-\alpha = 90\%$, 95%, or 99% and found gaps for some x in approximately 40% of them.

When the highest acceptance curve of minimal span, say AC($(l-1)$–u), produces a gap for a particular x, the LCO solution for $p < 0.5$ is to substitute for p in that gap the next alternative acceptance curve of minimal span available, AC(l–(u+1)). See Figure 7; for $n = 20$, $1-\alpha = 90\%$ using AC(1–6) in place of AC(0–5) removes the second subinterval $0.141 \leq p < 0.147$ from the confidence set for $x = 0$ and transfers those p to the confidence interval for $x = 6$. This eliminates the gap for $x = 0$. The gap that occurs for $x = 1$ is similarly resolved by moving directly from AC(2–7) to AC(2–8) rather than first to AC(1–7).

To summarize, the cpf of the LCO method lies on the highest acceptance curve of minimal span except in the relatively rare cases where it is necessary to resolve a gap, where it substitutes the next highest available acceptance curve of the same span. Thus, LCO maximizes coverage among all strict, length minimizing procedures.

Here is a formal description of our algorithm for determining LCO confidence intervals:

Step 1. For each p from 0 to 0.5 incremented in steps of size Δp, let $AC_p(l-u)$ denote the acceptance curve achieving the maximum value (greater than or equal to $1-\alpha$) among

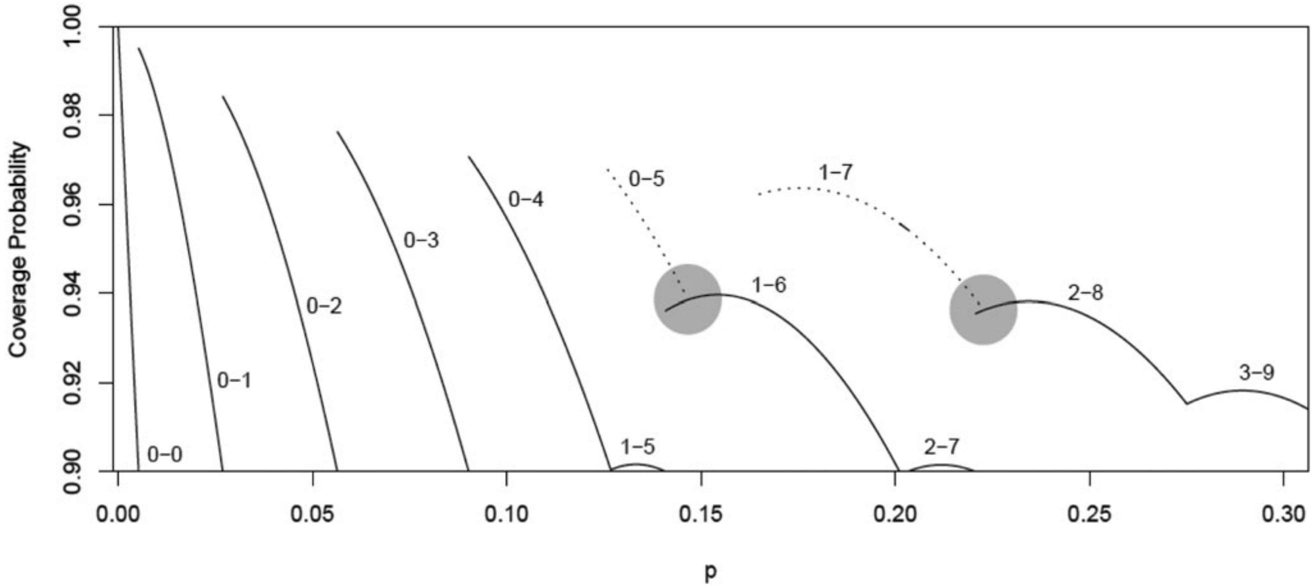

Figure 7. LCO cpf for $n = 20$, $1-\alpha = 90\%$. The gray circles show where the LCO cpf eliminates the gaps caused by using the highest available acceptance curves of minimal span.

all curves of minimal span at p. If multiple minimal span curves assume the maximum, choose $\mathrm{AC}_p(l\text{–}u)$ to be the curve with the largest value of l. Assign p to the confidence intervals for each integer x in $[l, u]$, except as provided in Step 2.

Step 2. Whenever $\mathrm{AC}_p(l\text{–}u')$ and $\mathrm{AC}_{p+\Delta p}(l'\text{–}u')$ in Step 1 are such that $l' < l$, let k be the largest integer such that Step 1 yields $\mathrm{AC}_{p+k\Delta p}(l'\text{–}u')$. Reassign $p, p+\Delta p, \ldots, p+k\Delta p$ to the confidence intervals for each integer x in $[l'+1, u'+1]$.

Step 3. The assignment of an acceptance curve for each $p > 0.5$ (and therefore the completion of all confidence intervals) follows from the symmetry requirement $\mathrm{CP}(p) = \mathrm{CP}(1\text{–}p)$, which is needed for equivariance.

The complete set of LCO confidence intervals for $1 \le n \le 100$ at confidence levels 90%, 95%, and 99% (based on a grid size of $\Delta p = 10^{-6}$ and rounded to five decimal places), along with R code for the LCO algorithm, is posted at *http://www.calpoly.edu/~jdoi/LCO*.

4. OTHER DESIRABLE PROPERTIES

Aside from minimal length and maximal coverage, there are several other properties that one would want a binomial confidence procedure to have, although we regard these as much less important than length and coverage because they do not pertain solely to the specific binomial experiment that a practitioner is analyzing. We explore several of these below.

4.1 Monotonicity in *x* and *n*

Blyth and Still (1983) listed two properties that reflect how we would expect binomial confidence intervals to behave. First, if x is increased by one (for fixed n) we would expect that both limits of the confidence interval would also increase. Second, given x and n, if an additional trial resulted in success (failure), both limits of the confidence interval should increase (decrease).

The first of these monotonicity properties fails for Crow's procedure, as indicated earlier. The second property does not hold generally for any reasonable confidence procedure since for all n the lower limit for $x = 0$ is 0 and the upper limit for $x = n$ is 1. Furthermore in practice binomial confidence intervals are rounded, normally to either two or three decimal places, and at such levels of rounding, different sample sizes each with the same number of successes x often yield the same endpoint. Thus, it seems reasonable to relax the second monotonicity property to say that when an additional trial produces a success (failure), neither confidence limit should decrease (increase).

4.2 Nesting

Another desirable property is *nesting*: if two confidence intervals having different levels are computed from the same data, the confidence interval with the higher confidence level should contain the one with the lower confidence level. For nesting to occur at all confidence levels, as the level increases the lower limits for each x must be nonincreasing and the upper limits must be nondecreasing. Now consider how the nesting issue manifests itself in the plot of the cpf. We use two specific cases to illustrate, but the argument is essentially the same for any length minimizing procedure.

Look first at Figure 5 and imagine steadily raising the bottom boundary of Figure 5—that is, the confidence level—and consider how the confidence intervals change as the level increases. The intersections of the Type O curves with the confidence level will move outward, which results in nesting for endpoints determined in those regions. In the Type I region, none of the interval endpoints determined by the 12 cusps change until the confidence level starts to pass above them. For example, the second cusp from the left (along N_7) marks the transition from AC(1–8) to AC(2–9), thereby determining u_1 and l_9 for confidence levels at and below this cusp. When the confidence level rises just above the cusp, the procedure must replace the portion of the cpf that is now below $1-\alpha$ with a segment from the higher necklace N_8. The only available segment that maintains monotonicity in l and u is AC(1–9). As the confidence level rises, the interval on which AC(1–9) is used progressively widens. This still preserves nesting, as u_1 moves to the right and l_9 moves to the left as the confidence level increases.

The above arguments show that nesting normally occurs as the confidence level rises. Unfortunately, a problem arises for confidence levels that are close to the peak height of an acceptance curve. An example is the case shown in Figure 7, where the 90% confidence level is slightly lower than the peak of AC(1–5). Note that the cpf of a length minimizing procedure must use the entire portion of AC(1–5) shown, as no other acceptance curve of minimal span is available. To avoid a gap in the confidence interval for $x = 0$, after using this segment the cpf must avoid a transition from AC(1–5) to AC(0–5) and instead must move to AC(1–6), the only alternative acceptance curve of minimal span. Thus for any length minimizing procedure at 90% confidence, the right end of the visible portion of AC(1–5) represents l_6. Now consider the higher confidence level 93.85% that intersects the cusp where AC(0–5) meets AC(1–6). See Figure 8. The cpf of any length minimizing procedure for this level would necessarily follow the necklace N_5 near this cusp, thereby determining the lower confidence limit for $x = 6$ as the point where AC(0–5) transitions to AC(1–6), shown in Figure 8 as l_6^*. However, l_6^* is *larger* than the value of l_6 previously determined for 90% confidence, hence nesting is violated.

Blaker (2000) proved that no length minimizing confidence procedure obeys the nesting property for all x. Figure 8 reveals that the reason for this unfortunate result is that the peaks of any necklace are offset horizontally from the cusps of the necklace lying just above it (except for n odd at $p = 0.5$). We verify this analytically in the Appendix. Figure 8 shows that the peak of AC(1–5) is to the left of the cusp between AC(0–5) and AC(1–6). Although length-minimizing procedures do not generally have to transition at the cusps, when the confidence level is equal to the height of a cusp they *must* do so, leading to a nesting violation.

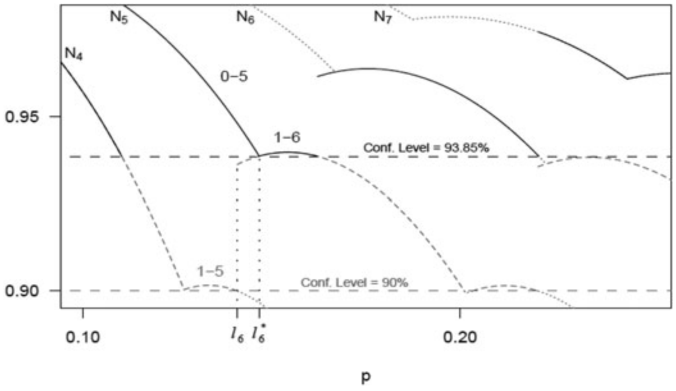

Figure 8. Sections of LCO cpfs for $n = 20$. Comparing the cpf for two different confidence levels shows how nesting can be violated. The solid curves show the cpf for $1-\alpha = 93.85\%$; the dashed curves show portions of the cpf for $1-\alpha = 90\%$. Nesting is violated since the lower confidence limit for $x = 6$ at level 93.85%, l_6^*, lies to the right of the corresponding limit for level 90%, l_6.

Surprisingly, therefore, both the existence of confidence sets with gaps and of nonnested length minimizing confidence procedures arise from the *same* phenomenon. Both pathologies involve confidence levels that pass just below the peak of an acceptance curve. The values of p that cause a nesting violation in the case illustrated above, those between $l_6 = 0.141$ and $l_6^* = 0.147$ in Figure 8, represent precisely the second interval in Sterne's confidence set for $x = 0$.

5. COMPARISON OF LCO AND TWO POPULAR EXACT METHODS

In this section, we briefly introduce two popular exact methods and compare their performance with the LCO procedure:

5.1 Blyth–Still–Casella Method

As indicated earlier, Casella (1986) determined the complete collection of length minimizing binomial procedures. The *Blyth–Still–Casella* procedure (BSC), available in the StatXact$^{\text{TM}}$ software, is the one member of this set that has been widely used, particularly by biomedical researchers.

5.2 Blaker's Method

For each p, determine the tail probability $T_p(x) = \min(P(X \leq x), P(X \geq x))$. Then, p is included in the confidence interval for x if and only if $P(X\colon T_p(X) \leq T_p(x)) > \alpha$. The primary advantage of Blaker's method, and its raison d'être, is that two confidence intervals based on the same data but at different confidence levels are guaranteed to be nested. Consequently, Blaker's procedure is not length minimizing.

5.3 Comparison Results

Both LCO and BSC are length minimizing; Blaker's method is not. The excess average length of Blaker's method is fairly small, never being more than 0.62% for $1 \leq n \leq 100$ and confidence levels 90%, 95%, and 99%.

While Blaker's method produces nested intervals in every case, for BSC and LCO, comparing all 90% and 95% intervals and all 95% and 99% intervals reveals that there are *no* cases where BSC intervals are not nested and only two out of all 10,300 comparisons where nesting is violated for LCO—and then only slightly ($n = 21$; $l_6(0.90) = 0.132 > l_6(0.95) = 0.130$ and its equivariant counterpart for $x = 15$). Thus, nesting violations are a virtual nonissue for all three methods, at least at the most commonly used confidence levels.

All three methods are monotonic in x: for fixed n, both confidence limits always increase in the case of an additional success. As for monotonicity in n, BSC, Blaker's method and LCO each have a very small number of exceptions. For LCO, the proportion of comparisons in which an additional trial resulting in a failure gives a higher rather than a lower confidence limit is 0.05% for lower confidence limits and 0.89% for upper confidence limits. BSC and Blaker's method have even fewer violations. These

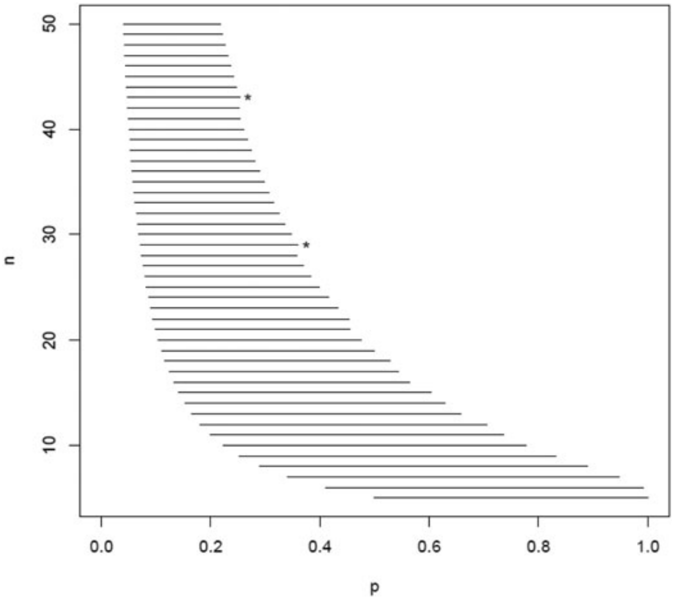

Figure 9. Monotonicity behavior of LCO 95% confidence intervals for $x = 5$, $n \leq 50$. Cases where monotonicity is violated are indicated by asterisks. There are no violations for $43 < n \leq 100$; only 30% of all x values have any violations of monotonicity for $n \leq 100$.

rates are based on confidence limits at three decimal places. The sizes of the violations for all three methods are generally quite small; the largest order reversal of endpoints for LCO is only 0.013. Among all LCO confidence intervals at 90%, 95%, and 99%, more than 70% of the x values between 0 and 100 have no violations of monotonicity in n for $n \leq 100$. Figure 9 shows how close to monotonic in n the LCO intervals are for a typical case among the 30% where violations *do* occur.

As strict confidence procedures, Blaker, BSC, and LCO each meet or exceed $1-\alpha$ for all p. Coverage can differ significantly, however. Figure 10 gives a comparison of LCO and BSC for the case $n = 12$, $1-\alpha = 90\%$. LCO's coverage is considerably higher in several regions of the parameter space. For $p \approx 0.184$ and 0.816, for example, BSC will fail to contain the true parameter nearly twice as often as LCO.

The overall coverage performance of a specific binomial confidence procedure can be assessed by its mean coverage $M = \int_0^1 \mathrm{CP}(p)dp$. Mean coverage can be thought of as the posterior mean coverage with respect to a uniform prior on p. Figure 11 reflects the consistent edge that LCO has in mean coverage versus BSC. LCO and Blaker's method have nearly equal mean coverage for all three confidence levels in practically every case. The reason that Blaker is competitive with LCO in coverage, however, is that Blaker's method is not length minimizing.

6. APPROXIMATE BINOMIAL CONFIDENCE PROCEDURES

We call a confidence procedure an *approximate* procedure when the coverage of the procedure falls below the stated confidence level, but not by a large amount (at least for most p). Approximate binomial confidence procedures are widely used, although it is not clear that users are always aware that they violate the strict definition of a confidence procedure. Allowing the cpf to fall below the nominal confidence level typically results in smaller confidence intervals than strict procedures produce.

6.1 Assessing the Performance of an Approximate Binomial Confidence Procedure

The judgment of what might be considered a "best" approximate confidence procedure is more subjective than for strict procedures. Again though, short intervals and high coverage, in that order, should be primary goals. The two coverage criteria

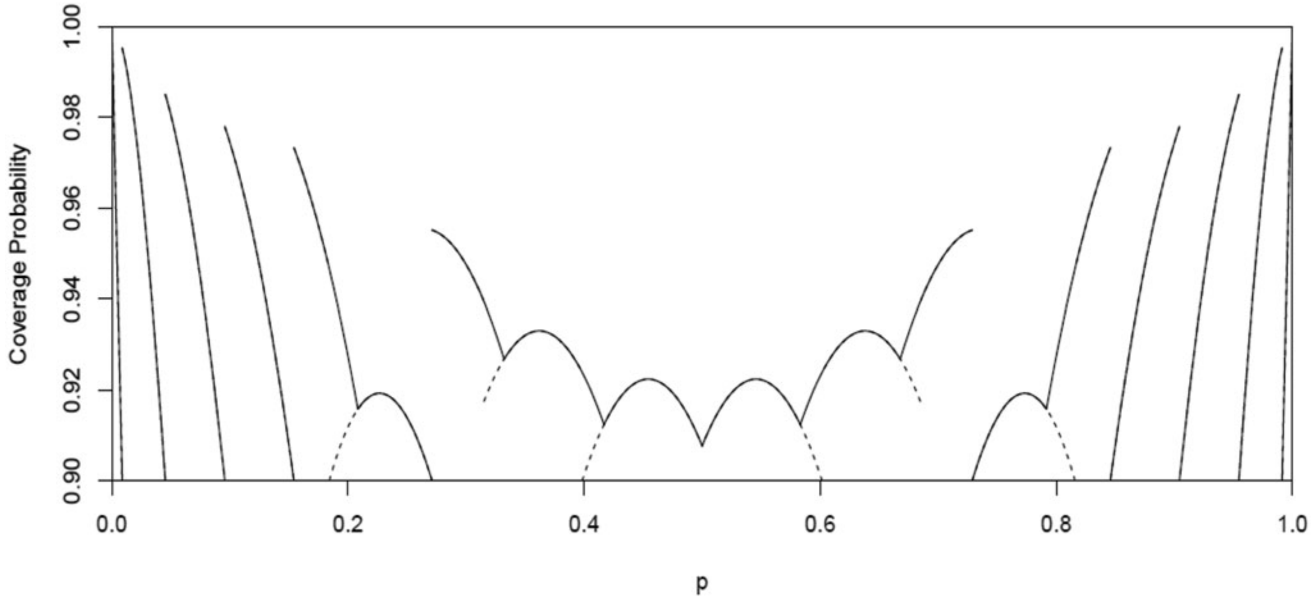

Figure 10. Coverage probability functions for the LCO and Blyth–Still–Casella Confidence Procedures, $n = 12$, $1-\alpha = 90\%$. Solid lines represent the cpf for LCO; dashed line segments show the cpf for BSC where it differs from LCO.

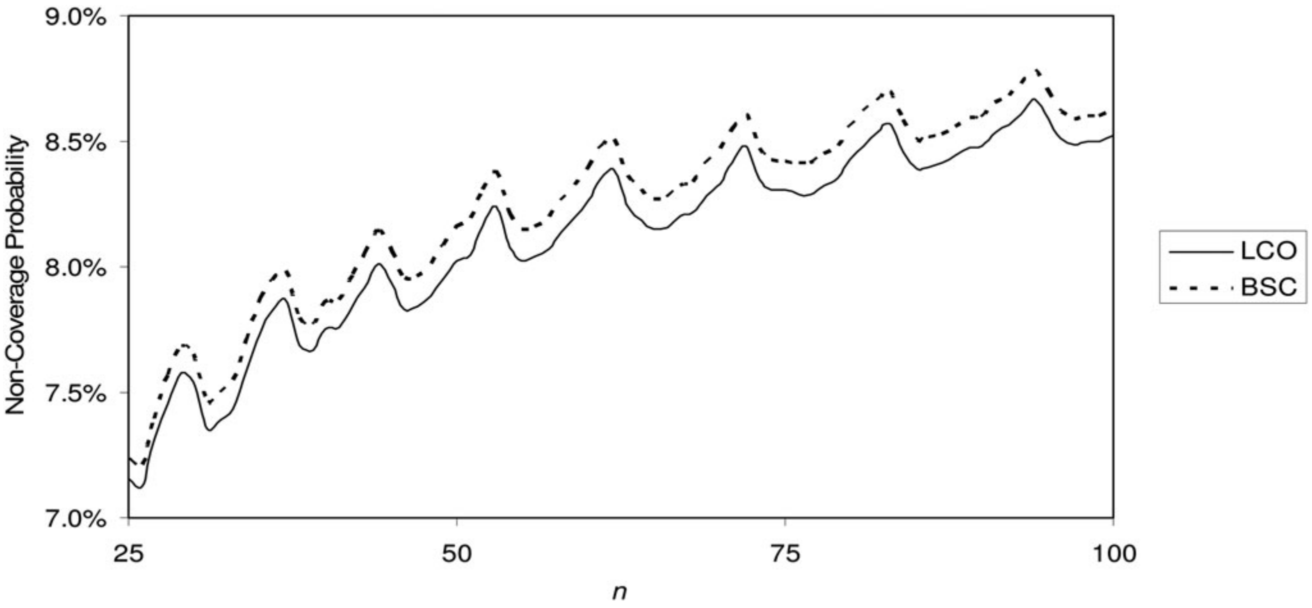

Figure 11. Mean noncoverage for LCO and BSC 90% confidence procedures.

most approximate procedures have been judged by are minimum and mean coverage.

Considering first minimum coverage, suppose some value $1-\alpha' < 1-\alpha$ is regarded as an acceptable minimum. If subject to this standard the goal is to minimize average confidence interval length, then one should choose a length-minimizing procedure at level $1-\alpha'$. Achieving high coverage wherever it does not incur a cost in length is even more desirable for approximate confidence procedures than for strict ones, since this will minimize the amount of the parameter space in which the cpf falls below $1-\alpha$ and maximize overall mean coverage. Consequently, the LCO procedure (now at level $1-\alpha'$) is again an ideal choice.

Another possible criterion is integrated absolute error, $A = \int_0^1 |\mathrm{CP}(p) - (1-\alpha)|dp$ (see, e.g., Brown, Cai and DasGupta 2001); this measure assesses how close the cpf stays to the nominal confidence level. Integrated squared error may also be used. These criteria value equally a situation when the cpf is below $1-\alpha$ and an instance when it is above, yet those two cases have entirely different ramifications, as the latter represents inadequate coverage, while the former merely represents a "bonus"—extra coverage not promised by the stated confidence level. A procedure which achieves coverage well above the confidence level for a large portion of the parameter space should not be penalized.

We propose instead to measure only the overall extent to which a procedure falls below the stated level with $D = \int_0^1 [1-\alpha-\mathrm{CP}(p)]I[\mathrm{CP}(p) < 1-\alpha]dp$, which we call the *deficit* of the procedure. D computes the total area captured below the confidence level and above the cpf (see Figure 12).

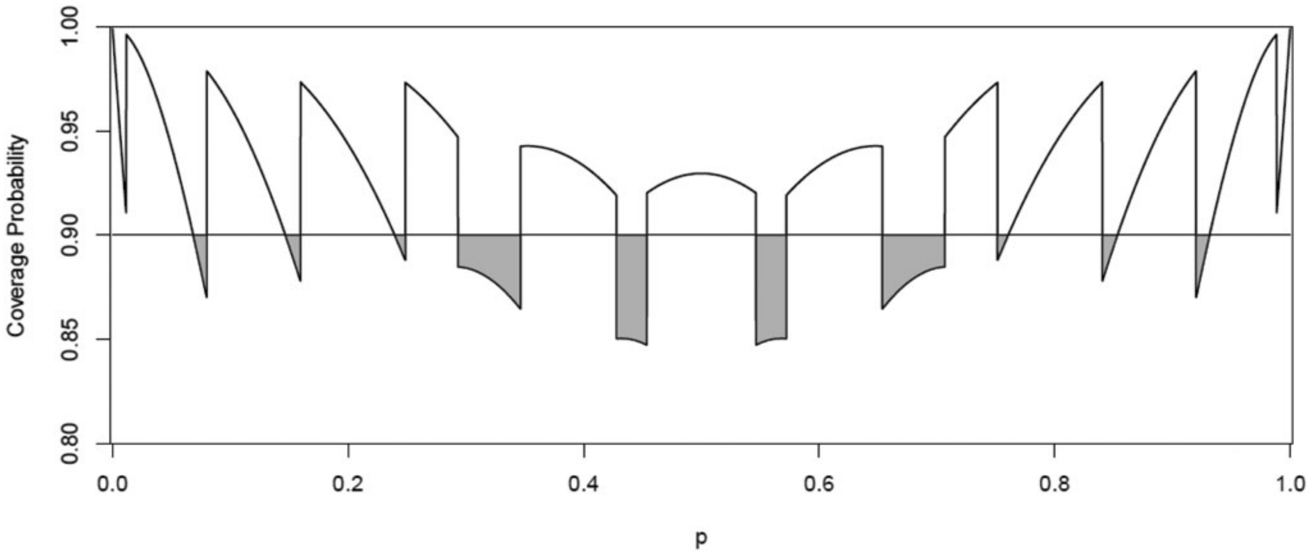

Figure 12. The deficit of the adjusted Wald procedure for $n = 8$, $1-\alpha = 90\%$.

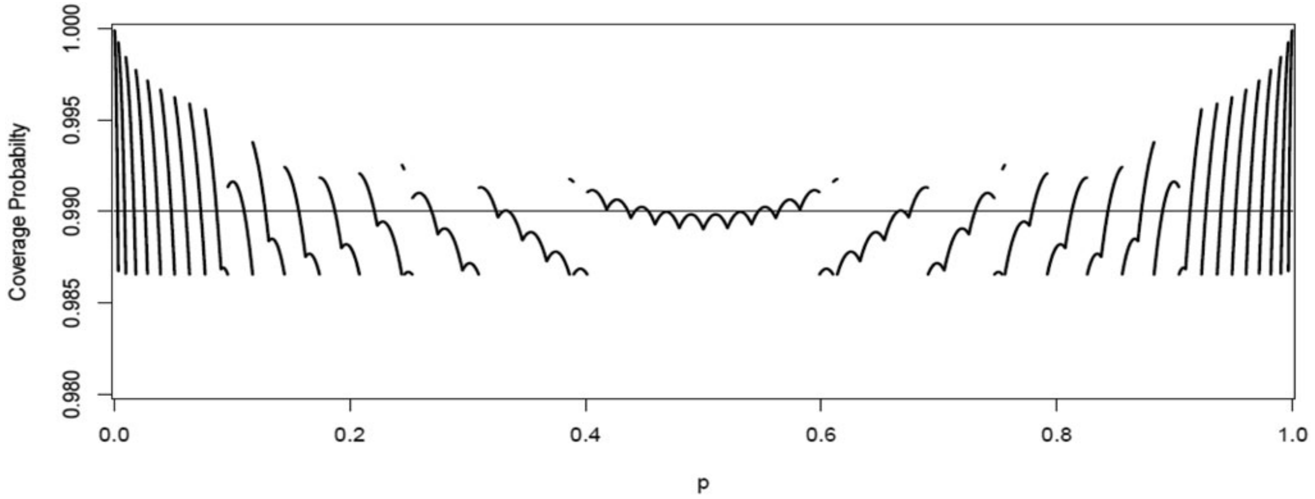

Figure 13. Cpf for the LCO mean coverage adjusted procedure, $n = 50$, $1-\alpha = 99\%$.

If a uniform prior is placed on p, then D represents the chance that the confidence interval will fail to cover the true value of p *because* of the fact that the procedure is approximate rather than strict. D together with minimum coverage provides a good summary of the degree to which an approximate procedure falls short of being a strict one.

One way to create an approximate confidence procedure is to recalibrate a strict procedure to a lower actual level so that its mean coverage equals $1-\alpha$ (see Reiczigel 2003). Figure 13 shows the cpf of an LCO procedure adjusted to have mean coverage 99%. Note the flat "bottom"—that is, minimum coverage is achieved at a very large number of points, which produces small intervals.

6.2 A Comparison of Several Popular Approximate Confidence Procedures

In this section, we compare the performance of the LCO procedure adjusted to have mean coverage $1-\alpha$ with that of several well known approximate procedures: Wald, Adjusted Wald, Jeffreys and Mid-P (see Brown, Cai and DasGupta 2001), and Wilson's score (Wilson (1927)). Table 1 gives results for average length, minimum coverage, mean coverage, and deficit for $n = 20$; results were similar for other n. We show Wald in parentheses because its coverage properties are so poor that it does not possess the qualities we would want for a satisfactory approximate confidence procedure.

At each confidence level, the adjusted LCO procedure achieves the shortest intervals of any method aside from Wald, and is even competitive with Wald at both the 90% and 95% levels despite Wald's low coverage. Since coverage and length are closely related, procedures that give shorter intervals should yield lower coverage. Table 1 shows this is generally the case; for example, adjusted Wald scores high on all measures of coverage but produces relatively long intervals, whereas Jeffreys produces fairly short intervals but has relatively low minimum and mean coverage and a high deficit. Adjusted LCO however, despite producing intervals nearly as short as Wald's, performs quite well on all three measures. In fact, adjusted LCO generally outperforms all of the other procedures in coverage relative to length. The most pronounced advantage adjusted LCO has, however, is for minimum coverage. Only adjusted Wald roughly matches adjusted LCO on this measure; however adjusted Wald produces the longest intervals, as noted by Brown, Cai and DasGupta (2001). The superior performance of the adjusted LCO procedure on minimum coverage is a direct result of its cpf's flat "bottom," a property that commonly used approximate procedures do not possess.

7. STRICT OR APPROXIMATE?

In applications, little attention has been paid to whether a binomial confidence procedure is strict or approximate. This is somewhat understandable given that for many other statistical

Table 1. Minimum and mean coverage and deficit of six approximate binomial confidence procedures for $n = 20$ (values shown are in percent)

	90%				95%				99%			
Conf Level:	Avg L	Min Cov	Mean Cov	Deficit	Avg L.	Min Cov	Mean Cov	Deficit	Avg L.	Min Cov	Mean Cov	Deficit
Adj. LCO	0.269	85.90	90.00	0.93	0.319	92.91	95.00	0.64	0.412	98.40	99.00	0.14
Jeffreys	0.273	82.04	90.17	1.19	0.323	89.34	95.11	0.75	0.417	96.59	99.04	0.17
Wilson	0.275	79.77	90.70	0.78	0.325	83.66	95.30	0.53	0.417	88.84	98.84	0.30
Mid-P	0.283	84.11	91.74	0.46	0.335	92.93	96.11	0.26	0.431	98.68	99.32	0.04
Adj. Wald	0.284	86.67	91.95	0.36	0.337	92.92	96.18	0.16	0.435	98.08	99.22	0.07
(Wald)	0.268	0.00	80.54	9.51	0.316	0.00	84.58	10.42	0.403	0.00	88.28	10.72

Table 2. Reduction in average confidence interval lengths using adjusted LCO procedure rather than strict LCO procedure

Confidence level	$n = 5$	10	20	50	100
90%	14.2%	13.5%	9.1%	6.0%	4.5%
95%	13.7%	10.8%	7.6%	4.7%	3.6%
99%	9.6%	6.7%	6.0%	3.8%	2.8%

situations a procedure having a constant or nearly constant cpf of level $1 - \alpha$ can be easily achieved. For the one sample binomial problem, however, the differences between strict and approximate are significant enough to warrant attention.

Approximate procedures violate the formal definition of a strict confidence procedure, with the reward being intervals of reduced length. How much is gained by bending the rules? Table 2 gives for various n and $1-\alpha$ the reduction in average confidence interval lengths for strict and approximate (mean coverage $= 1-\alpha$) LCO confidence procedures. The reduction in mean length from using approximate mean coverage adjusted intervals is greatest for small n and lower confidence levels. Table 2 shows that length savings are substantial in many cases. Even for $n = 50$ and $1-\alpha = 95\%$, where the 4.7% reduction in mean length is relatively small, an approximate procedure achieves intervals as short on average as those of a strict procedure that uses 10% more data.

Table 2 lends support to Agresti and Coull's (1998) contention that "approximate is better than exact" when it comes to one sample binomial confidence procedures. While an individual in a "single use" situation might prefer the stringent coverage guarantee of a strict procedure, a person or agency that computes many confidence intervals can feel reasonably confident that using mean coverage adjusted approximate procedures will result in the overall proportion of those intervals that cover the parameter matching well to the nominal level.

Given the substantial discrepancies between the actual cpf and the stated confidence level for the one sample binomial problem and other situations involving data with substantial discreteness, we propose a "truth in advertising" qualifier when reporting confidence intervals for such situations. If an approximate procedure is used, full disclosure would mean reporting that fact along with the minimum coverage level. If a strict confidence procedure is used, both the nominal and average level of coverage could be reported. For example, for one sample binomial data with $n = 30$ one could report that the LCO procedure was employed to obtain a 90% confidence interval (mean coverage: 92.5%) for the success parameter. As mean coverage is typically much higher than the confidence level, a better sense of the degree of confidence would then be conveyed than by just providing the nominal level.

8. SUMMARY

Since the LCO procedure is superior to other length-minimizing procedures in combining length minimization with maximal coverage, we recommend its use. When used as an approximate confidence procedure LCO has an additional advantage in that it maintains especially high minimum coverage. As for the question of strict versus approximate procedures, Brown, Cai and DasGupta (2001), concurring with Agresti and Coull (1998), argued that in modern statistical practice, approximate procedures should be preferred. In any case, we believe that at least for the one sample binomial problem the user should make clear which type is being used, given their substantial difference in performance (coverage and length).

APPENDIX

Proof of Proposition 1

Differentiating $\mathrm{AC}_{\Omega_{lu}}(p)$ for Type I curves provides the value at which $\mathrm{AC}_{\Omega_{lu}}(p)$ achieves its maximum:

$$p_M(l, u; n) = \left(1 + \left[\binom{n-1}{u} \Big/ \binom{n-1}{l-1}\right]^{1/(u-l+1)}\right)^{-1}. \quad \text{(A.1)}$$

To establish the first part of the proposition, for given $u < n-1$ write the expression inside the brackets in (A.1) as $Q_{n;l,u} = \binom{n-1}{u}/\binom{n-1}{l-1} = \prod_{i=0}^{u-l} \frac{n-l-i}{u-i}$. Since each term in the repeated product is greater than $\frac{n-u-1}{u+1}$, we have $Q_{n;l,u} > (\frac{n-u-1}{u+1})^{u-l+1}$; thus

$$(Q_{n;l,u})^{u-l+2} > (Q_{n;l,u})^{u-l+1}\left(\frac{n-u-1}{u+1}\right)^{u-l+1} = \left[Q_{n;l,u}\left(\frac{n-u-1}{u+1}\right)\right]^{u-l+1} = (Q_{n;l,u+1})^{u-l+1}.$$

Taking the $(u–l+1)(u–l+2)$th root of both sides of the inequality then yields

$$(Q_{n;l,u})^{1/(u-l+1)} > (Q_{n;l,u+1})^{1/((u+1)-l+1)}.$$

It follows immediately that $p_M(l, u; n) < p_M(l, u+1; n)$, proving the first part of the lemma. The second part follows similarly. □

Calculation of the Location of Necklace Cusps

Let $p_C(l, u; n)$ denote the cusp at the intersection of AC$((l–1)–(u–1))$ and AC$(l–u)$. Setting $P(l–1 \le X \le u–1) = P(l \le X \le u)$ yields immediately

$$P(X = l-1) = P(X = u), \text{ or } \binom{n}{l-1} p^{l-1}(1-p)^{n-l+1} = \binom{n}{u} P^u (1-p)^{n-u},$$

which gives $(\frac{p}{1-p})^{u-l+1} = \binom{n}{l-1}/\binom{n}{u}$. Solving for p yields the location of the cusp:

$$P_C(1, u; n) = \left(1 + \left[\binom{n}{u} \Big/ \binom{n}{l-1}\right]^{1/(u-l+1)}\right)^{-1}. \quad \text{(A.2)}$$

Interestingly, the form of (A.2) is just the same as (A.1), and in fact $p_C(l, u; n) = p_M(l, u; n+1)$.

Now let us compare $p_C(l, u; n)$ to $p_M(l, u-1; n)$, the location of the peak of AC$(l–(u–1))$, which is the peak most nearly in line with the cusp. We can do this by studying the ratio of their kernels

$$\frac{Q_{n+1;l,u}}{Q_{n;l,u-1}} = \frac{n+1-l}{u}.$$

This ratio is not equal to 1 unless $l + u = n + 1$, in which case$p_M(l, u-1; n) = 0.5$, thus in all other cases the peak of AC(l–(u−1)) is not aligned with the cusp joining AC((l−1)–(u−1)) and AC(l–u). It is not difficult to show that whenever $l + u < n + 1$, we have $p_M(l, u-1; n) < p_C(l, u; n) < 0.5$, and when $l + u > n + 1$, $p_M(l, u-1; n) > p_C(l, u; n) > 0.5$.

[Received June 2013. Revised February 2014.]

REFERENCES

Agresti, A., and Coull, B. A. (1998), "Approximate Is Better Than "Exact" for Interval Estimation of Binomial Proportions," *The American Statistician*, 52, 119–126. [135,144]

Blaker, H. (2000), "Confidence Curves and Improved Exact Confidence Intervals for Discrete Distributions," *Canadian Journal of Statistics*, 28, 783–798. [140]

Blyth, C. R., and Still, H. A. (1983), "Binomial Confidence Intervals," *Journal of the American Statistical Association*, 78, 108–116. [134,135,138,139]

Brown, L. T., Cai, T. T., and DasGupta, A. (2001), "Interval Estimation for a Binomial Proportion," *Statistical Science*, 16, 101–133. [144]

Brown, L. T., Cai, T. T., and DasGupta, A. (2002), "Confidence Intervals for a Binomial Proportion and Asymptotic Expansions," *The Annals of Statistics*, 30, 160–201. [133,142,143]

Casella, G. (1986), "Refining Binomial Confidence Intervals," *Canadian Journal of Statistics*, 14, 113–129. [133,136,138,140]

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. [135]

Crow, E. L. (1956), "Confidence Intervals for a Proportion," *Biometrika*, 43, 423–435. [136,137,138]

Hirji, K. F. (2006), *Exact Analysis of Discrete Data*, New York: Chapman and Hall. [138]

Reiczigel, J. (2003), "Confidence Intervals for the Binomial Parameter: Some New Considerations," *Statistics in Medicine*, 22, 611–621. [138,143]

Sterne, T. E. (1954), "Some Remarks on Confidence or Fiducial Limits," *Biometrika*, 41, 275–278. [137]

Vos, P. W., and Hudson, S. (2005), "Evaluation Criteria for Discrete Confidence Intervals: Beyond Coverage and Length," *The American Statistician*, 59, 137–142. [133]

Vos, P. W., and Hudson, S. (2008), "Problems With Binomial Two-Sided Tests and the Associated Confidence Intervals," *Australian & New Zealand Journal of Statistics*, 50, 81–89. [133]

Wilson, E. B. (1927), "Probable Inference, the Law of Succession, and Statistical Inference," *Journal of the American Statistical Association*, 22, 209–212. [143]