Received: 28 January 2022 | Revised: 3 May 2023 | Accepted: 30 June 2023

DOI: 10.1111/sjos.12672

ORIGINAL ARTICLE

Scandinavian Journal of Statistics

Locally correct confidence intervals for a binomial proportion: A new criteria for an interval estimator

Paul H. Garthwaite[1] | **Maha W. Moustafa**[2] | **Fadlalla G. Elfadaly**[1]

[1]School of Mathematics and Statistics, The Open University, Milton Keynes, UK

[2]School of Computing, Mathematics and Data Sciences, Coventry University, Coventry, UK

Correspondence
Fadlalla G. Elfadaly, School of Mathematics and Statistics, The Open University, Milton Keynes MK7 6AA, UK.
Email: fadlalla.elfadaly@open.ac.uk

Abstract

Well-recommended methods of forming "confidence intervals" for a binomial proportion give interval estimates that do not actually meet the definition of a confidence interval, in that their coverages are sometimes lower than the nominal confidence level. The methods are favoured because their intervals have a shorter average length than the Clopper–Pearson (gold-standard) method, whose intervals really are confidence intervals. As the definition of a confidence interval is not being adhered to, another criterion for forming interval estimates for a binomial proportion is needed. In this paper, we suggest a new criterion for forming one-sided intervals and equal-tail two-sided intervals. Methods which meet the criterion are said to yield *locally correct confidence intervals*. We propose a method that yields such intervals and prove that its intervals have a shorter average length than those of any other method that meets the criterion. Compared with the Clopper–Pearson method, the proposed method gives intervals with an appreciably smaller average length. For confidence levels of practical interest, the mid-p method also satisfies the new criterion and has its own optimality property. It gives locally correct confidence intervals that are only slightly wider than those of the new method.

KEYWORDS

Clopper–Pearson, coverage, discrete distribution, mid-*p*, shortest interval

1 INTRODUCTION

A good number of methods of forming a confidence interval for a binomial success parameter (p) have been studied. For example, Vollset (1993) compares thirteen methods, Newcombe (1998) compares seven methods and Brown et al. (2001) examine 11. These studies only consider methods that aim to form equal-tail confidence intervals, which can be constructed from one-sided intervals. That is, if $(l,\ u)$ is a $(1-2\alpha)$ confidence interval for p, then $(l,\ 1)$ and $(0,\ u)$ are one-sided $(1-\alpha)$ confidence intervals for p. The 'gold-standard' method for forming equal-tail confidence intervals is the Clopper–Pearson method (Jovanovic & Levy, 1997; Leemis & Trivedi, 1996). This interval estimator meets the definition of a method for forming equal-tail confidence intervals: the coverage of its one-sided $(1-\alpha)$ confidence intervals is *guaranteed* to be at least $(1-\alpha)$ for any value of p. Moreover, among those methods that meet the definition, its intervals have the shortest average length.

The drawback of the Clopper–Pearson method is that its coverage is above the nominal confidence level for almost all values of p, and for modest sample sizes this conservatism is noticeable. This is illustrated in the left-hand diagram in Figure 1, which plots the coverage of its 97.5% upper one-sided intervals against p when samples are drawn from a bin$(20, p)$ distribution. The saw-tooth pattern in the plot results from the discreteness of the sample space and arises with any method of forming confidence intervals for a binomial proportion. It can be seen that coverage is commonly above 98.5% and sometimes exceeds 99.5%.

One approach to reducing conservatism is to construct confidence intervals that are not constrained to have equal tails. Crow (1956), Blyth and Still (1983) and Casella (1986) adopt this approach and produce effective methods that reduce conservatism while maintaining the nominal confidence level. In practice, however, equal-tail intervals seem to be much preferred to two-sided intervals with unequal tails. Here we only consider one-sided intervals, and two-sided intervals with equal tails.

Agresti and Coull (1998, p. 119) argue that the conservatism of the Clopper–Pearson method makes it "…inappropriate to treat this approach as optimal for statistical practice." Instead, they advocate interval estimators that give shorter intervals for which the coverage probability is usually quite close to the nominal level, though coverage is less than that level for some values of p. An interval estimator that is often recommended under these criteria is the Wilson method (Wilson, 1927), whose coverage is plotted in the right-hand diagram in Figure 1 (again for 97.5% upper one-sided intervals for sampling from a bin$(20, p)$ distribution). For most values of p its coverage is quite close to the nominal level and, because it is sometimes liberal and generally less conservative than the Clopper–Pearson method, on average its intervals are shorter.

This pragmatic attitude underlies almost all work on methods of forming equal-tail confidence intervals for binomial proportions—interval estimators are examined that do not strictly give confidence intervals (their coverage is sometimes too small) but their coverages are reasonably close

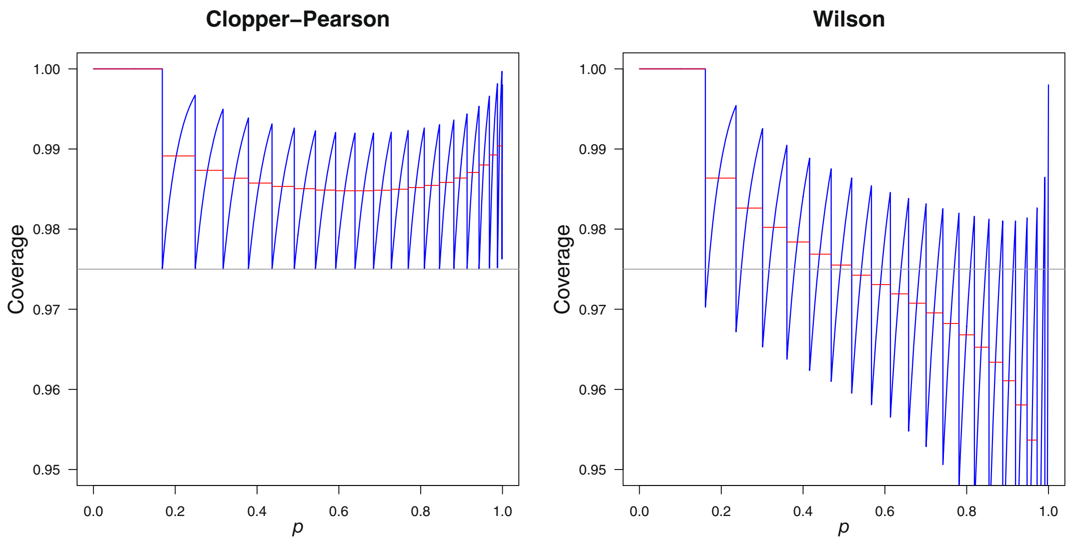

FIGURE 1 Coverage of upper one-sided 97.5% confidence intervals for the Clopper–Pearson and Wilson methods for a fixed sample size of $n = 20$ and variable p. Short horizontal (red) lines show the average coverage between consecutive spikes.

to the nominal level. Recommendations as to which interval estimator is "best" are made on the basis of coverage and the average length of their intervals. Simplicity of the methods is sometimes considered as well. However, methods giving shorter intervals tend to have lower coverage and an appropriate trade-off between accuracy of coverage and average interval length is partly a matter of opinion and recommendations as to which method is best cannot be clear-cut.

Assuming equal-tail intervals are required, the Clopper–Pearson method should be the method of choice because, under mild regularity conditions, it gives equal-tail intervals with a shorter average length than any other method that genuinely yields equal-tail confidence intervals (Buehler, 1957). The number of alternative methods that have been recommended, often by good statisticians, suggests that the definition of a confidence interval does not yield intervals that are completely satisfactory to everyone. We believe that there are benefits in relaxing the notion of a confidence interval and that a new criterion would be useful—simply requiring coverage to be "close" to the nominal level creates ambiguity as to whether one interval estimator is better than another. With an appropriate criterion, there should be some interval estimators that

(a) satisfy the new criterion,
(b) give intuitively sensible intervals,
(c) give intervals whose average length is acceptably short.

Under such a criterion, it would be reasonable to restrict attention to interval estimators with the above properties and the best estimator would be the one with the shortest average interval length. The challenge is to find an appropriate new criterion.

In Section 2, we define a criterion that meets requirements (a)–(c). Interval estimators that satisfy the criterion will be said to yield *locally correct confidence* (LCC) *intervals* and the criterion is called the LCC criterion. In Section 3, we present a new interval estimator that yields such intervals and examine its properties. We call the new estimator the *local average coverage* (LAC) *method*. In Section 4, it is compared with a number of methods that have been recommended for forming equal-tail confidence intervals. These include the mid-p method and we identify a property of this method that seems to have been overlooked in the past. While intervals satisfying the LCC criterion can meet requirements (a)–(c) there are, no doubt, alternative criteria that give such intervals. However, some seemingly plausible criteria give unreasonable intervals and, in Section 5, we describe two such criteria that we briefly considered. Concluding comments are given in Section 6.

To facilitate use of the LAC method, in an appendix we provide tables of its two-sided 95% and 99% intervals for sample sizes up to 30. Intervals for other sample sizes and nominal levels can be obtained through a Shiny R application available at https://olcbinomialci.shinyapps.io/binomial/.

2 | LOCALLY CORRECT CONFIDENCE INTERVALS

We first consider upper-tail (upper one-sided) intervals. Let X denote a binomial variate based on n trials with success probability p and suppose an interval estimator gives $(0, u_x)$ as its upper-tail estimate for p when x is the observed value of X. For a sensible estimator,

$$0 < u_0 < u_1 < \cdots < u_n \leq 1, \tag{1}$$

and we assume that (1) holds. The coverage probability of the interval estimator depends upon the value of p and is the probability that the random interval $(0, u_X)$ contains p. We denote this coverage probability by $C_u(p)$. When $u_{i-1} < p \leq u_i$,

$$C_u(p) = P(X \geq i \mid p) = \sum_{x=i}^{n} \binom{n}{x} p^x (1-p)^{n-x}. \tag{2}$$

For $x = 0, 1, \ldots, n$, the difference between $C_u(u_x)$ and $C_u(u_x + \delta)$ does not tend to 0 as $\delta \to 0$, but equals $\binom{n}{x} u_x^x (1-u_x)^{n-x}$. This is the reason that the coverages in Figure 1 have a saw-tooth appearance. The points of the teeth (the spikes) occur where p equals $u_0, \ldots, u_n$. The coverage increases monotonically between the spikes and drops by $\binom{n}{x} u_x^x (1-u_x)^{n-x}$ at $p = u_x$.

For a good interval estimator, how should $C_u(p)$ vary with p? If the nominal confidence level is $(1-\alpha)$, then $C_u(p)$ should exceed $(1-\alpha)$ when p is just before a spike, as $C_u(p)$ follows a cycle with its largest values just before spikes. If the estimator is not to be very conservative, then $C_u(p)$ should be less than $(1-\alpha)$ when p is just after a spike, as then $C_u(p)$ is at the lowest part of its cycle. However, the extent to which the estimator is liberal should be restricted. It seems reasonable to require the *average coverage within each cycle* to exceed or equal the target confidence level. The LCC criterion is that, for one-sided intervals, the average coverage between any pair of consecutive spikes must never be less than the nominal level. This reflects the saw-tooth pattern of coverage that is illustrated in Figure 1. An interval estimator gives LCC intervals if it meets the criterion.

Definition 1. Suppose an interval estimator gives $(0, u_x)$ as its upper-tail interval for p when $X = x$, that $u_0, \ldots, u_n$ satisfy (1) and that $u_n = 1$. If, for $i = 1, \ldots, n$,

$$\frac{1}{u_i - u_{i-1}} \int_{p=u_{i-1}}^{u_i} C_u(p)\, dp \geq 1 - \alpha \tag{3}$$

then the interval estimator gives *upper-tail LCC intervals* with confidence level $1-\alpha$.

In Definition 1 it is assumed that $u_n = 1$; otherwise the average coverage over the interval $(u_n, 1)$ would be 0, which is inconsistent with the required coverage in other intervals.

A cycle—the interval between two spikes—is quite small. Hence, while the coverage need not equal (or exceed) the nominal confidence level at individual values of p, it must do so on average over quite narrow ranges of p. In Figure 1, short horizontal (red) lines show the average coverage between pairs of consecutive spikes. The Clopper–Pearson intervals are LCC intervals while those given by Wilson's method are not.

Definition 2 is the corresponding definition for lower-tail intervals.

Definition 2. Suppose an interval estimator gives $(l_x, 1)$ as its lower-tail interval for p when $X = x$ and that $0 = l_0 < l_1 < \cdots < l_n < 1$. Define the coverage probability, $C_l(p)$, by

$$C_l(p) = P(X \leq i \mid p) = \sum_{x=0}^{i} \binom{n}{x} p^x (1-p)^{n-x} \tag{4}$$

for $l_i < p \leq l_{i+1}$. If, for $i = 0, \ldots, n-1$,

$$\frac{1}{l_{i+1} - l_i} \int_{p=l_i}^{l_{i+1}} C_l(p)\, dp \geq 1 - \alpha \tag{5}$$

then the interval estimator gives *lower-tail LCC intervals* with confidence level $1 - \alpha$.

Two-sided equal-tail LCC interval estimators are defined in terms of one-sided LCC intervals.

Definition 3. Suppose that an interval estimator gives (l_x, u_x) as its two-sided equal-tail interval for p when $X = x$, for $x = 0, \ldots, n$. Then it gives *equal-tail LCC intervals* with confidence level $(1 - 2\alpha)$ if and only if $\{(l_x, 1), x = 0, \ldots, n\}$ and $\{(0, u_x), x = 0, \ldots, n\}$ are sets of one-sided lower-tail and upper-tail LCC intervals, respectively, each with nominal confidence level $(1 - \alpha)$.

An interval estimator that gives equal-tail LCC intervals will be referred to as an *LCC interval estimator*. We use "$1 - \alpha$ intervals" as a generic term for $1 - \alpha$ LCC intervals and $1 - \alpha$ confidence intervals.

Confidence intervals have an invariance property under monotonic transformations. If $f(.)$ is a monotonic function and (l, u) is a $(1 - 2\alpha)$ confidence interval for p, then $(f(l), f(u))$ or $(f(u), f(l))$ is a $(1 - 2\alpha)$ confidence interval for $f(p)$, depending upon whether $f(.)$ is an increasing or decreasing function. A minor drawback of LCC intervals is that they do not have this property because they are defined in terms of *average* coverage. While average coverage is invariant under linear transformations, in general, with other transformations the average coverage between $f(a)$ and $f(b)$ will not be the same as the average coverage between a and b.

3 | A NEW INTERVAL ESTIMATOR

In this section, we present a novel interval estimator that yields LCC intervals with a smaller average length than any other LCC interval estimator. The proofs of most results require separate consideration of each binomial sample size that is of interest. The conservatism of the Clopper–Pearson method dissipates as the sample size increases, becoming negligible, so there is little reason to seek an alternative to the gold-standard Clopper–Pearson method for large sample sizes. Hence, asymptotic results are of limited relevance and instead we prove results for all sample sizes up to 200.

The new interval estimator, the *local average coverage* (LAC) *method*, uses a straightforward iterative algorithm to obtain one-sided interval estimates. The following proposition underpins the algorithm. Proofs of propositions are given in supplementary material.

Proposition 1. *Suppose* $1 \leq n \leq 200$ *and* $0.0001 < \alpha < 0.27$. *Suppose also that*

$$\frac{1}{u_i - u_{i-1}} \int_{p=u_{i-1}}^{u_i} \sum_{x=i}^{n} \binom{n}{x} p^x (1-p)^{n-x}\, dp = 1 - \alpha \tag{6}$$

and $u_i > u_{i-1}$ *for* $i = j+1, j+2, \ldots, n$ *(where* j *is a non-negative integer less than* n*). Then there exists* u_{j-1} *such that Equation (6) also holds for* $i = j$. *This* u_{j-1} *is unique and* $u_j > u_{j-1} > 0$. *Also, each* u_j *is a monotonic decreasing function of* α *for* $0.0001 < \alpha < 0.27$.

A necessary condition in Proposition 1 is that $\alpha < 0.27$. For larger values of α, sometimes $P(X \geq i | p = u_i)$ is less than $1 - \alpha$ and then Equation (6) cannot be satisfied with $u_{i-1} \leq u_i$. However, α is generally set equal to 0.05, 0.025 or 0.005 for a confidence interval, and for an inter-quartile range $\alpha = 0.25$. Hence, Proposition 1 covers the values of n and α of practical interest. The monotonicity property given in Proposition 1 is used in proofs of further propositions.

For an upper-tail LCC interval with confidence level $1 - \alpha$ and sample size n, the algorithm sets u_n equal to 1 and then sequentially determines $u_{n-1}, u_{n-2}, \ldots, u_0$. Given u_i, a simple numerical search is used to find u_{i-1} that satisfies Equation (6). From Proposition 1, there is always a unique u_{i-1} for which Equation (6) holds. The LAC method sets $(0, u_i)$ as the upper-tail interval when $X = i$. From Proposition 1, we also have that $0 < u_0 < \cdots < u_n = 1$. Consequently, under Definition 1 the LAC method gives upper-tail LCC intervals.

The LAC method uses similar steps to form lower-tail intervals. The iterative search starts by putting $l_0 = 0$ and then $l_1, \ldots, l_n$ are determined sequentially. Given l_i, the value l_{i+1} is found that satisfies

$$\frac{1}{l_{i+1} - l_i} \int_{p=l_i}^{l_{i+1}} \sum_{x=0}^{i} \binom{n}{x} p^x (1-p)^{n-x}\, dp = 1 - \alpha \tag{7}$$

for $i = 0, \ldots, n-1$. Then $(l_i, 1)$ is the $1 - \alpha$ lower-tail interval when $X = i$ and, under Definition 2, the method gives lower-tail LCC intervals. Two-sided intervals are obtained by combining the endpoints of one-sided intervals. Thus, when $X = i$ the method gives (l_i, u_i) as the two-sided equal-tails interval for a confidence level of $1 - 2\alpha$, where $(l_i, 1)$ and $(0, u_i)$ are its $1 - \alpha$ lower-tail and upper-tail LCC intervals, respectively. From its construction, the LAC method is an LCC interval estimator.

Examples showing the coverage of the estimator are given in Figures 2 and 3. Figure 2 plots coverage against p for upper-tail LCC intervals with nominal confidence levels of 97.5% and 99.5% for sample sizes 8, 20, and 50. The positions of spikes have similar characteristics to their positions in Figure 1: apart from the gap from 0 to the first spike, the spikes seem fairly evenly spaced and any trend in the size of the inter-spike interval is smooth. However, the coverages in Figure 2 are evenly spread around the nominal level, unlike the conservative coverage of the Clopper–Pearson method (left-hand diagram in Figure 1). Also, the coverages in Figure 2 have no long-term trend,

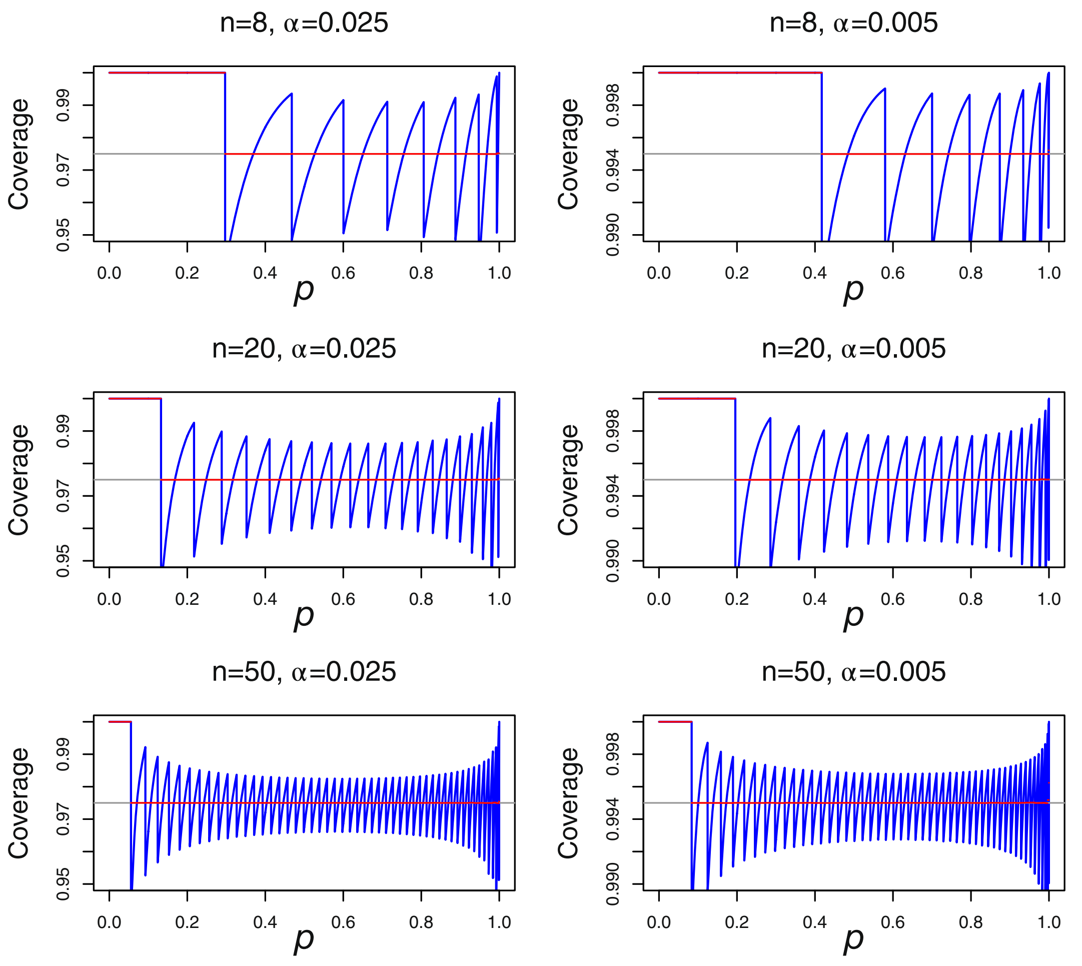

FIGURE 2 Coverage of upper-tail LCC intervals given by the new estimator for sample sizes of 8, 20 and 50, and nominal confidence levels of 97.5% and 99.5%.

unlike the coverages with Wilson's method (right-hand diagram in Figure 1), which are mostly above the nominal level for small values of p and below the nominal level for large values of p. Consequently, Figure 2 suggests that the LAC method gives sensible one-tail interval estimates. By design, with the method the average coverage between two spikes will always equal the nominal level.

In the top four panels in Figure 3, coverage of the LAC method is plotted against p for two-tail LCC intervals with nominal confidence levels of 95% and 99% for sample sizes 8 and 20. For comparison, the bottom two panels plot the coverage of 95% two-tail intervals given by the Clopper–Pearson method and Wilson's method for a sample size of 20. The plots have a more complicated pattern than the one-tail plots of Figures 1 and 2. There is a discontinuity in the $(1-2\alpha)\%$ two-tail coverage wherever there is a spike for a $(1-\alpha)\%$ upper-tail or lower-tail interval. In each plot, the average two-tail coverage between consecutive spikes for the upper-tail are plotted by short red lines and that between spikes from the lower-tail are plotted by short purple lines. With the LAC method, the lines are never far below the nominal level and, for each value

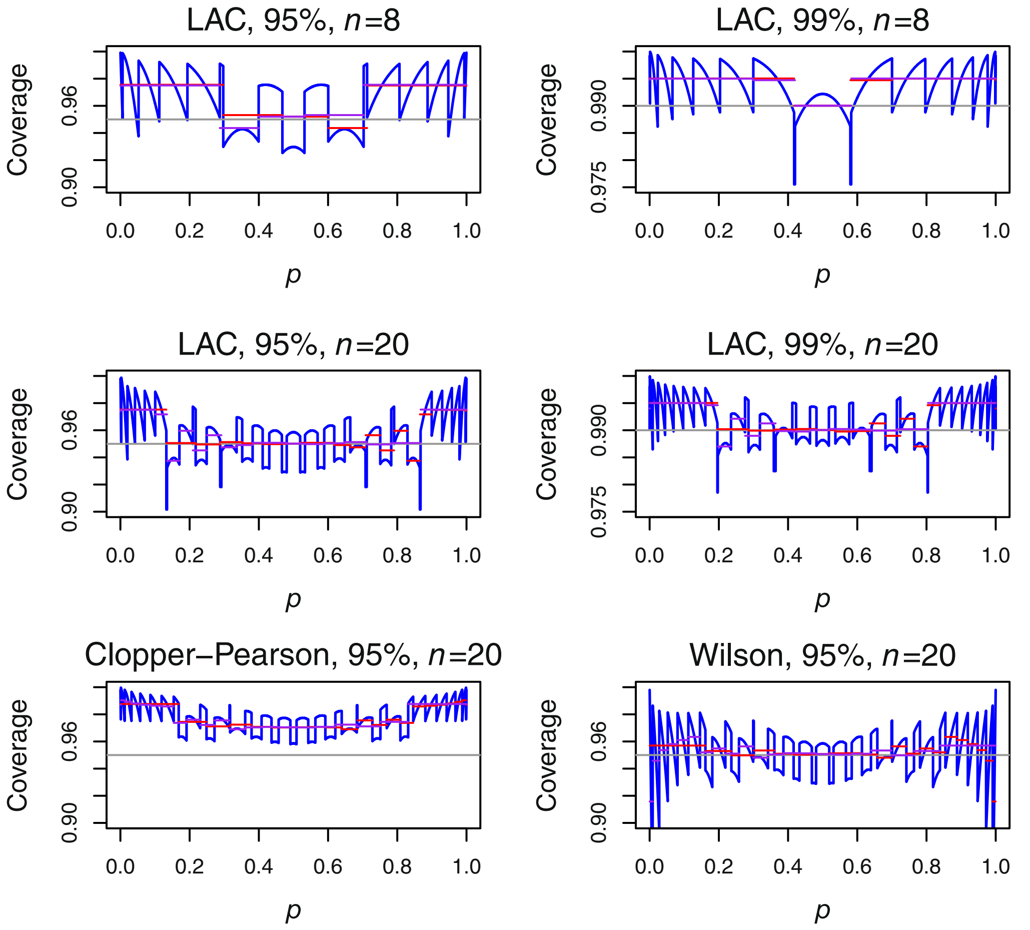

FIGURE 3 Top four panels show coverage of two-tail LCC intervals given by the LAC method for sample sizes of 8 and 20, and nominal confidence levels of 95% and 99%. Lowest two panels show coverage given by the Clopper–Pearson method and Wilson's method for two-tail 95% intervals with sample size 20.

of p, at least one of the lines is above or equal to the nominal level. Coverage is noticeably conservative for $p < u_0$ and $p > l_n$, which could perhaps be alleviated by a method that gives unequal tails, but that is outside the scope of this paper.

The lower-left panel in Figure 3 illustrates the marked conservatism of the Clopper–Pearson method, with a two-tail coverage noticeably above the nominal value for all values of p. The two-tail coverage of Wilson's method is slightly conservative for most values of p but, when p is near 0 or 1, it varies substantially and can be well below the nominal level. Overall, Figure 3 suggests that the two-tail coverage of the LAC method is less conservative than that of the Clopper–Pearson method and less variable than that of Wilson's method.

An important feature of an interval estimator is the length of its intervals. If an estimator gives (l_x, u_x) as its interval estimate when $X = x$, then the expected length of its interval, $L_n(p)$ say, is given by

$$L_n(p) = \sum_{x=0}^{n} (u_x - l_x) \binom{n}{x} p^x (1-p)^{n-x} \tag{8}$$

and, averaging over p, its average expected length (AEL) is

$$E[L_n(p)] = \int_{p=0}^{1} L_n(p)\, dp. \tag{9}$$

This definition holds for two-tail intervals that do not necessarily have equal tails and also for one-tail intervals—in Equation (8) l_x is set equal to 0 for upper-tail intervals and u_x is set equal to 1 for lower-tail intervals. When p is equally likely to take any value in the interval (0, 1), $\mathrm{P}(X = x) = 1/(n+1)$ for $x = 0, \ldots, n$. Hence, another expression for the AEL is that $\mathrm{AEL} = \sum_{x=0}^{n}(u_x - l_x)/(n+1)$.

The LAC method has the optimality properties given in Proposition 2, whose proof is given in the supplementary material.

Proposition 2. *Suppose* $1 \le n \le 200$ *and* $0.0001 < \alpha < 0.27$. *Then*

(i) for a one-tail LCC interval with confidence level $1-\alpha$, *the LAC method has the smallest AEL of any interval estimator that gives one-tail LCC intervals*;
(ii) for an equal-tail LCC interval with confidence level $1-2\alpha$, *the LAC method has the smallest AEL of any LCC estimator*.

The following are desirable properties in a method of forming confidence regions (see, e.g., Blyth and Still (1983) and Schilling and Doi (2014)).

1. *Interval valued.* A confidence region should be an interval and not a collection of disjoint intervals.

The remaining properties assume that the confidence region is a two-tail interval. When $X = x$, the sample size is n and the nominal confidence level is $1-2\alpha$, denote this interval as $(L(x, n, \alpha),\ U(x, n, \alpha))$.

2. *Equivariance.* As the binomial distribution is invariant under the transformation $X \to n - X$; $p \to 1-p$, confidence intervals should also be invariant under these transformations. That is, $L(x, n, \alpha)$ should equal $1 - U(n-x, n, \alpha)$ for $x = 0, \ldots, n$.
3. *Nesting.* If two confidence intervals have different nominal confidence levels then, for any given x and n, the interval for the higher nominal level should contain the interval for the lower nominal level. If the nominal levels are $1-\alpha_1$ and $1-\alpha_2$ with $\alpha_1 < \alpha_2$, this requires $(L(x, n, \alpha_2),\ U(x, n, \alpha_2)) \in (L(x, n, \alpha_1),\ U(x, n, \alpha_1))$.
4. *Monotonicity in* x. For fixed n and α, the endpoints should be increasing in x. This requires $L(x+1, n, \alpha) > L(x, n, \alpha)$ and $U(x+1, n, \alpha) > U(x, n, \alpha)$.
5. *Monotonicity in* n. For fixed x and α, the lower endpoint should be non-increasing in n and the upper endpoint should be decreasing in n. This requires $L(x, n+1, \alpha) \le L(x, n, \alpha)$ and $U(x, n+1, \alpha) < U(x, n, \alpha)$.

For the property of monotonicity in n, the "greater or equal" inequality cannot be a "strictly greater" inequality for the lower limit because, regardless of n, the lower limit should be 0 when $x = 0$. Instead, the property implies that, when an additional trial results in a failure, the limits of the confidence interval should not increase. In conjunction with the invariance property it also implies that, when an additional trial results in a success, the limits of the confidence interval should not decrease.

Suppose $1 \leq n \leq 200$ and $0.0001 < \alpha < 0.27$, so that Proposition 1 applies. Then clearly the LAC method will have the first four properties, as (a) its estimate is an interval and never a set of disjoint intervals, (b) equivalent procedures are used to construct lower and upper tails, (c) the nesting property follows from the monotonicity property given by Proposition 1, and (d) interval endpoints increase as x increases. Proposition 3 shows that the LAC method also gives interval endpoints that have the 'monotonicity in n' property. Hence it seems clear that the LAC method meets the above requirements for being a well-behaved interval estimator.

Proposition 3. *The endpoints of intervals given by the LAC method satisfy* $L(x, n+1, \alpha) \leq L(x, n, \alpha)$ *and* $U(x, n+1, \alpha) < U(x, n, \alpha)$ *for* $1 \leq n \leq 200$ *and* $0.0001 < \alpha < 0.27$.

4 | COMPARISON WITH OTHER METHODS

In this section, we compare the LAC method with the following six methods of forming interval estimates: Clopper–Pearson, mid-p, Wald, Agresti–Coull, Wilson, and Jeffreys methods. An optimality property of the mid-p method is also given.

4.1 | Clopper–Pearson and mid-p methods

To simplify notation, we let (l_i, u_i) denote a method's $1 - 2\alpha$ confidence interval when i successes occur in a sample of size n. The Clopper–Pearson method determines (l_i, u_i) by inverting equal-tail tests of the hypothesis $H_0 : p = p_0$. Thus l_i satisfies $P(X \geq i | p = l_i) = \alpha$ and u_i satisfies $P(X \leq i | p = u_i) = \alpha$, except that $l_0 = 0$ and $u_n = 1$. The endpoints of the interval may also be obtained as Buehler limits (Buehler, 1957): u_i is the smallest possible upper limit under an ordering restriction and a coverage restriction, and l_i is the largest possible lower limit under similar restrictions. Of the methods we consider, the Clopper–Pearson method is the only one that strictly meets the definition of a method for forming confidence intervals. Since it gives equal-tail confidence intervals, it also gives LCC intervals. As noted earlier, its conservative coverage (cf. Figures 1 and 3) leads to intervals that are frequently considered unnecessarily long.

The mid-p method reduces the conservatism of the Clopper–Pearson method by halving the probability of the observed result (Newcombe, 1998). That is, l_i satisfies

$$P(X > i \mid p = l_i) + \frac{1}{2}P(X = i \mid p = l_i) = \alpha \tag{10}$$

and u_i satisfies

$$P(X < i \mid p = u_i) + \frac{1}{2}P(X = i \mid p = u_i) = \alpha \tag{11}$$

except, by definition, $l_0 = 0$ and $u_n = 1$, as otherwise the coverage is 0 for $p < l_0$ and $p > u_n$ (Agresti & Gottard, 2005). It is a well-recommended method of forming confidence intervals (Agresti & Gottard, 2005; Berry & Armitage, 1995; Mehta & Walsh, 1992; Vollset, 1993), suggesting that its intervals are acceptably short.

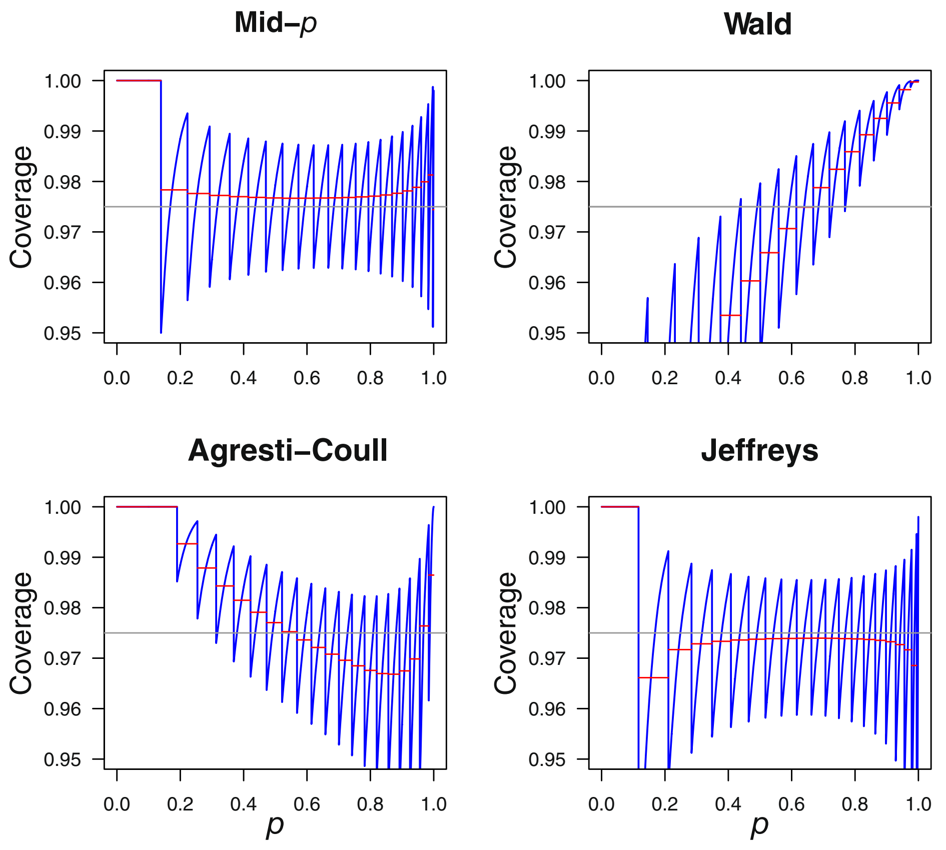

FIGURE 4 Coverage of upper one-sided 97.5% confidence intervals for the mid-p, Wald, Agresti–Coull and Jeffreys methods for a fixed sample size of 20 and variable p.

While the method does not give interval estimates that meet the definition of confidence intervals, it does give LCC intervals for $n < 200$ and $0.0001 < \alpha < 0.1$. The range of α $(0.0001 < \alpha < 0.1)$ is more restrictive than with the LAC method and, in particular, does not include interquartile ranges, but it does include all levels of confidence that are commonly of interest when forming confidence intervals. The result is given in the following proposition.

Proposition 4. *For* $1 \le n \le 200$ *and* $\alpha \in (0.0001,\ 0.1)$, *the mid-p method gives LCC intervals.*

The coverage of 97.5% upper-tail intervals for the mid-p method for a bin$(20, p)$ distribution is plotted against p in the upper-left panel of Figure 4. The spikes in the plot are spaced fairly regularly and the actual coverage always crosses the nominal coverage level between consecutive spikes.

Agresti and Gottard (2007) compare the coverages of the mid-p and Clopper–Pearson methods as p varies uniformly between 0 and 1. They conclude that "the mid-p approach is an excellent one to adopt if one hopes to achieve close to the nominal level in using a method repeatedly for various studies in which p itself varies" (Agresti & Gottard, 2007, p. 6455). In fact, among an important class of methods of forming confidence intervals, the mid-p method gives one-tail confidence intervals whose coverage is optimally close to the nominal confidence level for any value of p. This property is given in Proposition 5. It seems unlikely that the property has not

TABLE 1 Endpoints of two-sided 90% confidence intervals of the Clopper–Pearson method and 95% intervals of the LAC and mid-p methods for $X \sim \text{bin}(12, p)$.

	Clopper–Pearson		LAC method		Mid-p method	
x	Lower	Upper	Lower	Upper	Lower	Upper
0	0	0.221	0	0.210	0	0.221
1	0.004	0.339	0.004	0.339	0.004	0.347
2	0.030	0.438	0.034	0.443	0.029	0.451
3	0.072	0.527	0.072	0.535	0.068	0.541
4	0.123	0.609	0.122	0.617	0.116	0.623
5	0.181	0.685	0.177	0.692	0.172	0.698
6	0.245	0.755	0.240	0.760	0.234	0.766

been noted before, but we have been unable to find a reference to it in the literature, so a proof of the proposition is given in the supplementary material. We only give the property for upper-tail intervals but the corresponding result obviously holds for lower-tail intervals.

Proposition 5. *Consider the class of methods of forming upper-tail confidence intervals that (i) do not involve randomisation (i.e., confidence interval are determined by* x *and* n*, and do not involve the value of a further hypothetical random variable), and (ii) satisfy* $0 \le u_0 \le \cdots \le u_n = 1$*. Then the absolute error in the coverage probability,* $|C_u(p) - (1-\alpha)|$*, is as small or smaller for the mid-p method as for any method in the class, for any value of* p*.*

The class of methods in Proposition 5 includes all sensible methods of forming equal-tail confidence intervals that do not involve randomisation so, in particular, it includes all methods discussed in this section. The property is quite strong because it relates to every value of p, and hence gives other properties. For example, in comparing methods of interval estimation, it is common to examine the average absolute error in coverage or the root mean-square error in coverage, where averaging is over $p \sim U(0,1)$. Under either of these measures, Proposition 5 implies that the mid-p method would be the optimal method of forming one-tail confidence intervals.

As the mid-p method gives LCC intervals, its intervals tend to be longer than those of the LAC method (in accordance with Proposition 2). However, differences are small and the two methods generally give intervals with similar endpoints. These intervals are often similar to the intervals given by the Clopper–Pearson intervals for a much smaller nominal coverage. Indeed, $1-4\alpha$ confidence intervals given by the Clopper–Pearson method can be similar to the $1-2\alpha$ intervals given by the LAC and mid-p methods. To illustrate, Table 1 gives 90% two-sided confidence intervals for the Clopper–Pearson method and 95% two-sided intervals given by the LAC and mid-p methods for samples from a $\text{bin}(12, p)$ distribution. [Only intervals for $x = 0, \ldots, 6$ are shown. From symmetry, intervals for $x = 7, \ldots, 12$ have identical features.] In each row, the intervals are very similar to each other and the LAC interval is consistently within the mid-p interval.

Table 1 raises the question of how a mid-p interval is best described. Usually, its intervals in Table 1 would be referred to as 95% confidence intervals. However, they are similar to the 90% confidence intervals given by the gold-standard Clopper–Pearson method, and the difference between a 95% nominal coverage and a 90% nominal coverage is substantial. Rather, it seems better to describe the mid-p intervals as 95% LCC intervals, as they meet the criteria of 95% LCC intervals, while they do not strictly meet the criteria of 95% confidence intervals.

4.2 | Wald, Agresti–Coull, Wilson, and Jeffreys methods

The remaining methods we consider do not give LCC intervals. The Wald method is the simple method that is commonly taught in introductory statistics courses. Its interval endpoints are given by

$$l_i = \hat{p} - z\{\hat{p}(1-\hat{p})/n\}^{1/2} \quad \text{and} \quad u_i = \hat{p} + z\{\hat{p}(1-\hat{p})/n\}^{1/2}, \tag{12}$$

where $\hat{p} = i/n$ and z is is the $1-\alpha$ quantile of the standard normal distribution. Wald confidence intervals have been criticized heavily in, for example, Brown et al. (2001) and Vos and Hudson (2008). One fault is that their intervals tend to have low coverage and Agresti and Coull (1998) suggested a simple adjustment that appreciably improves the coverage of the 95% confidence intervals. The adjustment, which gives the Agresti–Coull method, is to add two "successes" and two "failures" to the sample, so the endpoints of its confidence interval are $\tilde{p} \pm z\{\tilde{p}(1-\tilde{p})/(n+4)\}^{1/2}$, where $\tilde{p} = (i+2)/(n+4)$. The coverages of Wald and Agresti–Coull 97.5% upper-tail intervals for the bin$(20, p)$ distribution are plotted against p in the upper-right and lower-left panels of Figure 4. In contrast to the mid-p coverage, the coverages show clear trends. The coverage of Wald intervals tends to be liberal for small values of p and conservative for large values, while the trend is in the opposite direction for Agresti–Coull intervals, and much less pronounced. Agresti–Coull intervals are recommended by Brown et al. (2001) as a simple method of forming confidence intervals for a binomial proportion when the sample size exceeds 40.

When the length of confidence intervals is of paramount concern, perhaps the most well-recommended methods are the Wilson method and Jeffreys method (see, e.g., Brown et al., 2001). Wilson intervals are based on inversion of the score test, so they are also known as score intervals. The interval endpoints are

$$\frac{n\hat{p} + z^2/2}{n+z^2} \pm \frac{z}{n+z^2}\left\{n\hat{p}(1-\hat{p}) + \frac{z^2}{4}\right\}^{1/2}.$$

A plot of the coverage of 97.5% upper-tail intervals against p for bin$(20, p)$ distributions was given in the right panel of Figure 1. Coverage is a little conservative for small values of p and quite liberal for values of p near 1. These characteristics are reversed for lower-tail intervals, and so the conservatism and liberalism partly balance out when the coverage of two-tail intervals is plotted against p. This can be seen in the lowest right-hand panel of Figure 3, where the coverage of two-tail intervals for Wilson's method shows much less trend than its coverage of one-tail intervals (cf. Figure 1).

The Jeffreys method is a Bayesian approach that takes a Beta$\left(\frac{1}{2}, \frac{1}{2}\right)$ distribution as the prior distribution, which is Jeffreys' choice of noninformative prior distribution for sampling from a binomial model. The sample consists of x successes in n trials and leads to Beta$\left(x+\frac{1}{2}, n-x+\frac{1}{2}\right)$ as the posterior distribution for p. The Jeffreys interval estimate is the $1-2\alpha$ equal-tail credible interval given by this posterior distribution, except for setting $l_0 = 0$ and $u_n = 1$ (Brown et al., 2001). Thus l_i and u_i are set equal to the α and $1-\alpha$ quantiles of the Beta$\left(i+\frac{1}{2}, n-i+\frac{1}{2}\right)$ distribution for $i = 1, \ldots, n-1$. Brown et al. (2001) note that Jeffreys intervals are always within Clopper–Pearson intervals and are approximately equal to mid-p intervals. The coverage of 97.5% upper-tail intervals for the Jeffreys method is given in the lower-right panel of Figure 4. Its coverage resembles that of the mid-p method and does not display the trend found with Wilson's method for one-tail intervals.

4.3 | Coverage and length of intervals

We will restrict attention to upper-tail intervals and two-tail intervals. As defined in Section 2, $C_u(p)$ is the coverage of an upper-tail interval estimator and is the probability that the random interval $(0, u_X)$ contains p. We define the quantity T_u as

$$T_u = \frac{1}{1-u_0}\int_{p=u_0}^{1} C_u(p)\, dp, \tag{13}$$

and refer to it as the *truncated average coverage*. In calculating T_u, values of p in the range $(0, u_0)$ are excluded because within that range the coverage is 1, so the coverage for $p \leq u_0$ differs radically from the coverage for $p > u_0$. Consequently, it is more informative to give the values of both T_u and u_0. Other average coverages may be calculated from these: the average coverage of a one-tail interval over the full range $(0, 1)$ equals $\{(1-u_0)T_u + u_0\}$ and the average coverage for two-tail intervals over the range $(0, 1)$ equals $2\{(1-u_0)T_u + u_0\} - 1$.

Table 2 gives both the values T_u and u_0 for the methods described earlier, for each of $\alpha = 0.05$, 0.025, and 0.005, and $n = 8$, 20, and 50. The results indicate the merits of the LAC method. From its construction, the LAC method necessarily has a truncated average cover that equals the nominal α (apart from rounding error). In contrast, the Clopper–Pearson method is very conservative, mid-p is slightly conservative, and Agresti–Coull is generally conservative, while the Wilson, Wald and Jeffreys methods are consistently liberal. (The Wald method is particularly liberal with values of T_u that are almost always far below the nominal values.)

A small value of u_0 is desirable, as then the range over which the coverage equals 1 is small. On that basis the Wald method does exceptionally well, as u_0 always equals 0 for that method. However, the coverage of the Wald method is too liberal for it to be the preferred method of forming interval estimates. Based on u_0, the LAC method is a little poorer than Jeffreys (a consistently liberal method), but a little better than mid-p, and much better than Clopper–Pearson, Agresti–Coull and Wilson.

As noted in the introduction, it has been argued that a good interval estimator should (i) give short intervals, and (ii) give coverage probabilities that are usually quite close to the nominal level (see, e.g., Agresti and Coull (1998)). To examine the latter criterion, the root mean-square error (RMSE) of each method's coverage was determined over the truncated range $(u_0, 1)$. This RMSE is given by

$$\text{RMSE} = \left[\frac{1}{1-u_0}\int_{u_0}^{1} \{C_u(p) - (1-\alpha)\}^2\, dp\right]^{1/2}, \tag{14}$$

where $1-\alpha$ is the nominal confidence level. The RMSE for each method is given in Table 3 for $\alpha = 0.05$, 0.025, and 0.005 and $n = 8$, 20, and 50. From Proposition 5, the mid-p method has the minimum possible RMSE of any method of forming interval estimates that does not use randomisation. Consequently, the mid-p method has the smallest RMSE in every row of Table 3. The new method, LAC, has the second smallest RMSE in every row and is always only a little poorer than the mid-p (its RMSE is never more than 20% bigger). The RMSE of the Clopper–Pearson was sometimes more than 45% bigger than the RMSE of the mid-p method and each of the other methods has an RMSE that is at least 80% bigger for some combination of α and n, with the Wald method often doing extremely badly. Hence LAC, while not the optimal method, has a very respectable RMSE.

TABLE 2 Truncated average coverage (T_u) of upper-tail $1-\alpha$ intervals and smallest upper limit (u_0) of seven methods of forming interval estimates, for $\alpha = 0.05, 0.025, 0.005$ and sample sizes (n) of 8, 20, and 50.

α	n	Statistic	Clopper–Pearson	Mid-p	Agresti–Coull	Wilson	Wald	Jeff.	LAC
0.05	8	T_u	0.976	0.956	0.949	0.941	0.852	0.941	0.950
0.05	8	u_0	0.312	0.250	0.293	0.253	0.000	0.208	0.239
0.05	20	T_u	0.971	0.954	0.953	0.947	0.903	0.946	0.950
0.05	20	u_0	0.139	0.109	0.141	0.119	0.000	0.091	0.105
0.05	50	T_u	0.966	0.952	0.953	0.949	0.928	0.948	0.950
0.05	50	u_0	0.058	0.045	0.062	0.051	0.000	0.038	0.043
0.025	8	T_u	0.989	0.979	0.972	0.966	0.867	0.969	0.975
0.025	8	u_0	0.369	0.312	0.372	0.324	0.000	0.262	0.297
0.025	20	T_u	0.986	0.977	0.976	0.972	0.923	0.972	0.975
0.025	20	u_0	0.168	0.139	0.190	0.161	0.000	0.117	0.133
0.025	50	T_u	0.983	0.976	0.977	0.974	0.950	0.974	0.975
0.025	50	u_0	0.071	0.058	0.085	0.071	0.000	0.049	0.056
0.005	8	T_u	0.998	0.996	0.991	0.988	0.882	0.993	0.995
0.005	8	u_0	0.484	0.438	0.509	0.453	0.000	0.379	0.417
0.005	20	T_u	0.998	0.996	0.994	0.992	0.941	0.994	0.995
0.005	20	u_0	0.233	0.206	0.289	0.249	0.000	0.177	0.196
0.005	50	T_u	0.997	0.995	0.995	0.994	0.970	0.995	0.995
0.005	50	u_0	0.101	0.088	0.139	0.117	0.000	0.075	0.084

TABLE 3 Root mean-square error (RMSE) of coverage of upper-tail $1-\alpha$ intervals for seven methods of forming interval estimates, for $\alpha = 0.05, 0.025$, and 0.005 and sample sizes (n) of 8, 20, and 50.

α	n	Clopper–Pearson	Mid-p	Agresti–Coull	Wilson	Wald	Jeff.	LAC
0.05	8	0.0290	0.0224	0.0267	0.0313	0.2249	0.0314	0.0242
0.05	20	0.0235	0.0166	0.0188	0.0213	0.1481	0.0212	0.0175
0.05	50	0.0180	0.0119	0.0139	0.0151	0.0972	0.0144	0.0124
0.025	8	0.0153	0.0121	0.0177	0.0237	0.2334	0.0186	0.0134
0.025	20	0.0125	0.0091	0.0118	0.0157	0.1522	0.0123	0.0097
0.025	50	0.0097	0.0066	0.0088	0.0110	0.0983	0.0083	0.0069
0.005	8	0.0033	0.0027	0.0085	0.0147	0.2407	0.0051	0.0032
0.005	20	0.0028	0.0021	0.0041	0.0088	0.1557	0.0032	0.0023
0.005	50	0.0022	0.0015	0.0028	0.0057	0.0997	0.0021	0.0017

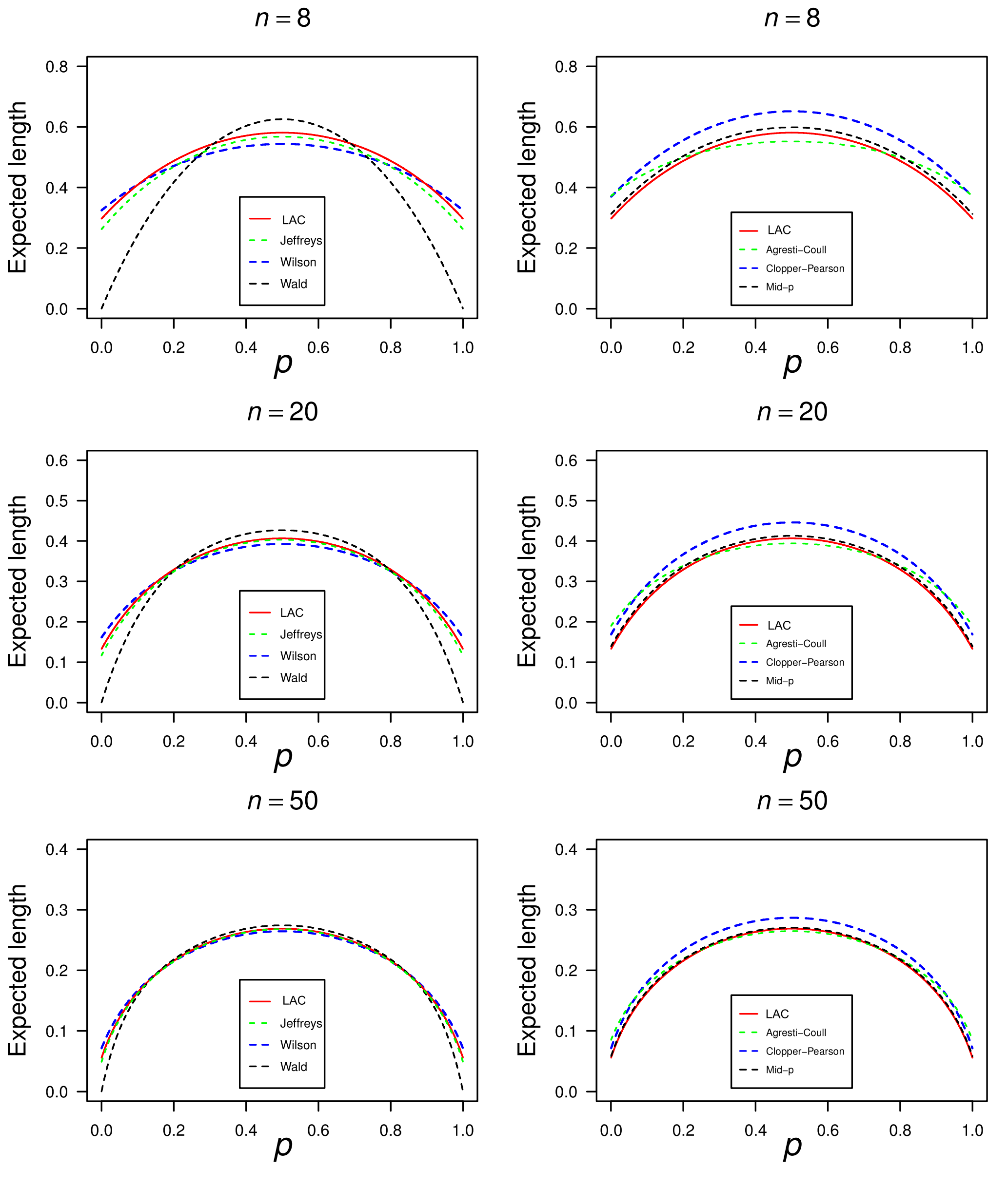

FIGURE 5 Expected lengths of two-sided 95% interval estimates for the LAC, Jeffreys, Wilson and Wald methods (left-hand panels) and the LAC, Agresti–Coull, Clopper–Pearson and mid-p methods (right-hand panels), plotted against p for sample sizes of 8, 20, and 50.

Turning to the length of intervals, two-tail intervals are examined as the length of one-tail intervals varies too much with p: the length of one-tail intervals is approximately proportional to p for upper-tail intervals and to $1 - p$ for lower-tail intervals. In Figure 5, the expected lengths of 95% two-tail intervals are plotted against p for sample sizes 8, 20, and 50. To aid comparison, the expected length for the LAC method is included in all plots. For all values of p and each combination of n and α, the expected lengths of the LAC, mid-p, Agresti–Coull, Wilson and Jeffreys intervals are all very similar, and a little smaller than the expected lengths of the Clopper–Pearson intervals. Wald intervals have a much smaller expected length than other methods when p is quite

TABLE 4 Average expected length (AEL) of two-tail $1-2\alpha$ intervals for seven methods of forming interval estimates, for $\alpha = 0.05$, 0.025, and 0.005 and $n = 8$, 20, and 50.

α	n	Clopper–Pearson	Mid-p	Agresti–Coull	Wilson	Wald	Jeff.	LAC
0.050	8	0.497	0.435	0.427	0.407	0.372	0.402	0.421
0.050	20	0.317	0.283	0.284	0.275	0.268	0.273	0.278
0.050	50	0.197	0.181	0.182	0.179	0.178	0.178	0.179
0.025	8	0.561	0.508	0.499	0.474	0.427	0.472	0.492
0.025	20	0.366	0.335	0.337	0.325	0.316	0.323	0.328
0.025	50	0.231	0.215	0.218	0.213	0.211	0.212	0.213
0.005	8	0.673	0.634	0.614	0.586	0.520	0.597	0.617
0.005	20	0.457	0.431	0.435	0.417	0.403	0.417	0.423
0.005	50	0.295	0.281	0.286	0.278	0.275	0.276	0.278

large or quite small, but it only achieves this by giving coverages that are well-short of the nominal confidence level (cf. Table 2).

Equation (9) defines the average expected length (AEL) of intervals given by a method. Important questions about the AEL of the new method are:

1. The LAC method gives intervals that have a longer AEL than some of the other methods considered here. But those methods do not give intervals with locally correct coverage. The question is whether the difference in AEL is small, so that there is little cost in requiring intervals to have locally correct coverage.
2. From Proposition 2, the LAC method gives intervals with a shorter AEL than the AEL of the Clopper–Pearson method. However, the gold-standard Clopper–Pearson method gives intervals that meet the definition of a confidence interval. The question is whether the LAC method reduces the AEL by enough to compensate for failing to meet that definition.

To answer those questions requires a means of deciding when a difference is "sufficiently large" and when it is "sufficiently small." Now, in the literature it has been variously argued that the Wilson, Agresti–Coull, mid-p and Jeffreys methods should be favoured over the Clopper–Pearson method because their intervals are shorter. Hence, the differences in AEL between these methods and the Clopper–Pearson method are considered sufficiently large to compensate for failing to meet the definition of a confidence interval. This provides a measure of a "sufficiently large" difference.

Table 4 gives the AEL of the methods considered earlier for each combination of $\alpha = 0.05$, 0.025, and 0.005 and $n = 8$, 20, and 50. The Wald intervals have the smallest or equally smallest AEL for each combination but, as noted earlier (cf. Table 2), this is obtained by having an actual coverage that is well below the nominal coverage. Apart from the Wald method and Clopper–Pearson, the AEL of the LAC method is usually similar in size to that of the other methods, and is always shorter than the mid-p method. Hence, in response to question 1, it is reasonable to conclude that the AEL of the LAC method compares satisfactorily with that of other methods, so there is little cost in requiring intervals to be locally correct. Also, the LAC method improves upon the AEL of Clopper–Pearson by amounts of similar size to the improvements given by the other methods. As these other methods are commonly favoured over the

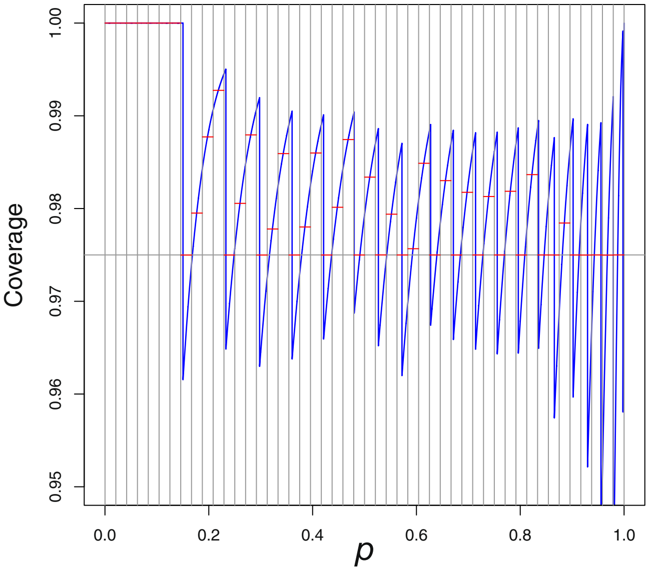

FIGURE 6 Coverage of upper one-sided 97.5% confidence intervals for $n = 20$ and 47 subintervals when the average coverage in each subinterval must be no less than the nominal value. Short horizontal (red) lines show the average coverage in each subinterval.

Clopper–Pearson method, the response to the second question is also positive—it is fair to say that the LAC method improves on the AEL of Clopper–Pearson by enough to compensate for failing to meet the definition of a confidence interval.

5 | OTHER CRITERIA FOR DEFINING AN INTERVAL ESTIMATOR

Requiring intervals to be locally correct is just one criterion that might be used in an alternative definition of a "confidence interval." In this section we briefly examine two other criteria that seem plausible alternatives but find that they lead to optimal estimators that are often unsatisfactory. We continue to assume that equal-tail intervals are required.

A simple criterion is to require the average coverage of a one-tail interval to be no less than the nominal level. In the notation of Equations (2) and (4), this criterion requires $\int_0^1 C_u(p)\, dp \geq 1 - \alpha$ for upper-tail intervals and $\int_0^1 C_l(p)\, dp \geq 1 - \alpha$ for lower-tail intervals. The optimal estimator would minimize the average expected length of two-tail intervals under this criterion. Unfortunately, examples show that it will often give endpoints that are identical for a number of different x values. To illustrate, when $n = 10$ the optimal estimator gives (0.3445, 0.6555) as the 90% equal-tail interval for p if x equals 3, and it gives exactly the same interval if x equals, 4, 5, 6, or 7. This is clearly unsatisfactory. (Intervals given by the optimal estimator were calculated using the Rsolnp package in R (Ghalanos & Theussl, 2015).)

A more stringent criterion is to partition the [0, 1] interval into a number of subintervals of equal length and require the average coverage of a one-tail interval to equal or exceed the nominal coverage in each sub-interval. Again, the optimal estimator would minimize the average expected length of two-tail intervals. This criteria offers some flexibility in its application, as the number of subintervals given by the partition must be chosen. However, if the number of subintervals is small, the same problem can arise as with the previous criterion—different values of x will sometimes lead to identical upper endpoints or identical lower endpoints. Also, if the number is large, then the coverage suffers from the form of conservatism that besets the Clopper–Pearson method. Indeed, as the number of subintervals increases, the optimal intervals under this criterion become indistinguishable from those given by the Clopper–Pearson method.

Moreover, there is not always an in-between ground that suffers from neither of these problems. As an example, when $n = 20$ and $\alpha = 0.025$, at least one pair of x values give the same upper-endpoint to the interval estimate unless the number of subintervals is at least 48. Figure 6 shows the coverage of the optimal estimator for an upper-tail 97.5% interval when the number of subintervals is 48. The short horizontal lines give the average coverage as p varies across each subinterval. The average coverage is equal to the nominal level in fewer than half the subintervals and is often substantially above it, indicating marked conservatism.

6 | CONCLUDING COMMENTS

This paper aimed to find a satisfactory criterion for choosing an interval estimator for a binomial proportion. As noted in Section 2, with an appropriate criterion there should be some interval estimators that (a) satisfy the new criterion; (b) give intuitively sensible intervals; and (c) give intervals whose average length is acceptably short.

As a criterion we proposed that an interval estimator should yield locally correct confidence intervals, meaning that, for one-tail intervals, the average coverage between any pair of consecutive spikes should be no smaller than the nominal confidence level. Three of the methods that were examined met this criterion: the Clopper–Pearson method, the mid-p method and the new LAC method. The Clopper–Pearson method arguably does not satisfy point (c), as over the years the conservative length of its intervals has motivated the construction of many other methods of forming confidence intervals for a binomial proportion. One of these other methods is the mid-p method, which has been advocated because of its shorter intervals (Agresti and Gottard, 2005; Vollset, 1993). The mid-p method gives intuitively sensible intervals so it meets points (a)–(c) above.

Turning to the LAC method, in the examples given in Figure 2 the method gave end-points that are fairly evenly spaced with coverages that are balanced around the nominal confidence level. This has also been the case in every other example we have examined and, in particular, there has never been a hint of the defects found with the methods considered in Section 5. Blyth and Still (1983) and Schilling and Doi (2014) list some properties that are desirable in an interval estimator, such as equivariance and monotonicity. We showed that the LAC method has these for the many combinations of n and α that were examined through extensive computation [$1 \leq n \leq 200$ and α in $(0.0001, 0.27)$]. Hence it is reasonable to conclude that the new method gives sensible intervals and meets point (b).

Regarding the third point, length of intervals, six methods of constructing equal-tail confidence intervals were compared in Section 4.3. The six methods include the mid-p, Wilson, Agresti–Coull and Jeffreys methods, which have each been recommended in preference to the Clopper–Pearson method because of the lengths of their intervals. The intervals given by the LAC method had an average expected length that was shorter than Clopper–Pearson intervals and comparable to those given by other methods for all values of n and α that were examined, except when a method gave intervals whose coverage was decidedly smaller than the nominal confidence level. Hence the LAC appears to give intervals that are acceptably short, and meets point (c).

As there are at least two methods that meet points (a)–(c), it can be concluded that the LCC criterion is a reasonable criterion to place on an interval estimator. (The results in Section 5 show that finding a suitable criterion is a non-trivial task.) Choosing between the mid-p method and the LAC method is tricky because each has an optimality property. On the one hand, for any value of p, the coverage of one-tail intervals is as close to the nominal level as possible when intervals are determined using the mid-p method (cf. Proposition 5). On the other hand, the average expected length of intervals is smaller with the LAC method than with any other method that gives locally correct intervals (cf. Proposition 2). However, in choosing between estimators, both coverage and length of intervals are important. Hence either (a) a restriction should be placed on coverage and methods should be differentiated on the basis of interval width, or (b) a restriction should be placed on interval width and methods differentiated on the basis of coverage. Requiring an interval to be locally correct places a restriction on coverage, so differentiating between methods on the basis of interval length seems appropriate, in which case the LAC method is the preferred interval estimator.

It is also the case that people are more conscious of the width of intervals than the absolute difference between nominal and actual coverage, which again favours the LAC method. However, the mid-p method is a simpler method of computing intervals since it does not require an iterative algorithm, unlike the LAC method, so properties of the method are more readily determined. The method is also well-known, gives acceptably short intervals, and is implemented in a number of statistical packages. When $\alpha < 0.1$ the mid-p method gives LCC intervals and is a reasonable alternative to the LAC method. A strong recommendation of this paper is that the mid-p intervals should be referred to as $1 - 2\alpha$ LCC intervals (provided $\alpha < 0.1$), rather than $1 - 2\alpha$ confidence intervals, since they strictly meet the criterion to be LCC intervals.

In further work we will extend the use of LCC intervals to other discrete one-parameter sampling models, such as the Poisson and negative binomial models. Interval estimates from two binomial samples is another task where coverage can exhibit large spikes, so that LCC intervals may be noticeably shorter than confidence intervals. Forming confidence intervals for the product of two proportions has been considered (Buehler, 1957) and much work has focused on the difference between two binomial proportions (e.g., Berger and Boos (1994) and Farrington and Manning (1990)). The latter task has been shown to pose unexpected problems (Röhmel, 2005), so the construction of good LCC intervals is likely to be challenging.

PACKAGE AND TABLES

A Shiny R application is available at https://olcbinomialci.shinyapps.io/binomial/ that determines one-tail or two-tail intervals given by the LAC method for any values of n, x and $\alpha < 0.5$. Tables of its two-sided 95% and 99% intervals are given in the appendix for sample sizes up to 30.

ACKNOWLEDGEMENTS

We are grateful to three reviewers and an associate editor for their insightful and constructive comments. Table 1 in the paper is based on a table given by one of the reviewers. The second author was supported by a studentship from The Open University, UK.

ORCID

Fadlalla G. Elfadaly https://orcid.org/0000-0002-7221-285X

REFERENCES

Agresti, A., & Coull, B. A. (1998). Approximate is better than “exact” for interval estimation of binomial proportions. *The American Statistician*, *52*, 119–126.

Agresti, A., & Gottard, A. (2005). Comment: Randomized confidence intervals and the mid-*P* approach. *Statistical Science*, *20*, 367–371.

Agresti, A., & Gottard, A. (2007). Nonconservative exact small-sample inference for discrete data. *Computational Statistics and Data Analysis*, *51*, 6447–6458.

Berger, R. L., & Boos, D. D. (1994). *P*-values maximized over a confidence set for the nuisance parameter. *Journal of the American Statistical Association*, *89*, 1012–1016.

Berry, G., & Armitage, P. (1995). Mid-*p* confidence intervals: A brief review. *The Statistician*, *44*, 417–423.

Blyth, C. R., & Still, H. A. (1983). Binomial confidence intervals. *Journal of the American Statistical Association*, *78*, 209–212.

Brown, L. D., Cai, T. T., & DasGupta, A. (2001). Interval estimation for a binomial proportion (with discussion). *Statistical Science*, *16*, 108–116.

Buehler, R. J. (1957). Confidence intervals for the product of two binomial parameters. *Journal of the American Statistical Association*, *52*, 482–493.

Casella, G. (1986). Refining binomial confidence intervals. *Canadian Journal of Statistics*, *14*, 113–129.

Crow, E. L. (1956). Confidence intervals for a proportion. *Biometrika*, *43*, 423–435.

Farrington, C. P., & Manning, G. (1990). Test statistics and sample size formulae for comparative binomial trials with null hypothesis of non-zero risk difference or non-unit relative risk. *Statistics in Medicine*, *9*, 1447–1454.

Ghalanos, A., & Theussl, S. (2015). *Package Rsolnp*. CRAN Repository.

Jovanovic, B. D., & Levy, P. S. (1997). A look at the rule of three. *The American Statistician*, *51*, 137–139.

Leemis, L. M., & Trivedi, K. S. (1996). A comparison of approximate interval estimators for the Bernoulli Parameter. *The American Statistician*, *50*, 63–68.

Mehta, c. R., & Walsh, S. J. (1992). Comparison of exact, Mid-*p*, and Mantel-Háenszel confidence intervals for the common odds ratio across several 2×2 contingency tables. *The American Statistician*, *46*, 146–150.

Newcombe, R. G. (1998). Two-sided confidence intervals for the single proportion: Comparison of seven methods. *Statistics in Medicine*, *17*, 857–872.

Röhmel, J. (2005). Problems with existing procedures to calculate exact unconditional *p* values for non-inferiority/superiority and confidence intervals for two binomials and how to resolve them. *Biometrical Journal*, *47*, 37–47.

Schilling, M. F., & Doi, J. A. (2014). A coverage probability approach to finding an optimal binomial confidence procedure. *The American Statistician*, *68*, 133–145.

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

Vos, P. W., & Hudson, S. (2008). Problems with binomial two-sided tests and associated confidence intervals. *Australian and New Zealand Journal of Statistics*, *50*, 81–89.

Wilson, E. B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, *22*, 209–212.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Garthwaite, P. H., Moustafa, M. W., & Elfadaly, F. G. (2023). Locally correct confidence intervals for a binomial proportion: A new criteria for an interval estimator. *Scandinavian Journal of Statistics*, 1–25. https://doi.org/10.1111/sjos.12672

APPENDIX . TABLES OF LOWER AND UPPER ENDPOINTS OF 95% AND 99% EQUAL-TAIL CONFIDENCE INTERVALS GIVEN BY THE LAC METHOD

Tables A1 and A2.

TABLE A1 Limits of 95% equal-tail LAC interval when x successes are observed in a sample of size n.

x	$n = 1$		$n = 2$		$n = 3$		$n = 4$		$n = 5$		$n = 6$	
0	0.0000	0.9500	0.0000	0.7396	0.0000	0.6038	0.0000	0.5003	0.0000	0.4291	0.0000	0.3735
1	0.0500	1.0000	0.0252	0.9748	0.0169	0.8444	0.0127	0.7400	0.0101	0.6470	0.0085	0.5759
2					0.1556	0.9831	0.1115	0.8885	0.0870	0.8049	0.0714	0.7239
3									0.1951	0.9130	0.1566	0.8434
x	$n = 7$		$n = 8$		$n = 9$		$n = 10$		$n = 11$		$n = 12$	
0	0.0000	0.3314	0.0000	0.2971	0.0000	0.2696	0.0000	0.2465	0.0000	0.2271	0.0000	0.2105
1	0.0072	0.5160	0.0063	0.4680	0.0056	0.4272	0.0051	0.3933	0.0046	0.3639	0.0042	0.3388
2	0.0606	0.6584	0.0526	0.6006	0.0464	0.5525	0.0416	0.5105	0.0377	0.4746	0.0344	0.4430
3	0.1309	0.7726	0.1125	0.7129	0.0986	0.6585	0.0878	0.6121	0.0792	0.5706	0.0721	0.5345
4	0.2274	0.8691	0.1937	0.8063	0.1688	0.7519	0.1496	0.7011	0.1344	0.6569	0.1220	0.6167
5					0.2481	0.8312	0.2186	0.7814	0.1954	0.7340	0.1768	0.6921
6									0.2660	0.8046	0.2398	0.7602
x	$n = 13$		$n = 14$		$n = 15$		$n = 16$		$n = 17$		$n = 18$	
0	0.0000	0.1962	0.0000	0.1836	0.0000	0.1726	0.0000	0.1628	0.0000	0.1541	0.0000	0.1462
1	0.0039	0.3167	0.0036	0.2974	0.0034	0.2802	0.0032	0.2650	0.0030	0.2512	0.0028	0.2389
2	0.0317	0.4154	0.0294	0.3908	0.0273	0.3691	0.0256	0.3495	0.0241	0.3319	0.0227	0.3159
3	0.0662	0.5022	0.0611	0.4737	0.0568	0.4479	0.0531	0.4249	0.0498	0.4040	0.0469	0.3851
4	0.1117	0.5812	0.1030	0.5490	0.0956	0.5203	0.0892	0.4941	0.0836	0.4705	0.0786	0.4489
5	0.1614	0.6534	0.1486	0.6189	0.1376	0.5873	0.1281	0.5588	0.1199	0.5326	0.1127	0.5087
6	0.2184	0.7206	0.2005	0.6835	0.1854	0.6501	0.1724	0.6192	0.1611	0.5912	0.1512	0.5652
7	0.2794	0.7816	0.2559	0.7441	0.2361	0.7086	0.2192	0.6765	0.2046	0.6464	0.1918	0.6189
8					0.2914	0.7639	0.2700	0.7300	0.2516	0.6990	0.2356	0.6698
9									0.3010	0.7484	0.2815	0.7185

TABLE A1 Continued

x	n = 19		n = 20		n = 21		n = 22		n = 23		n = 24	
0	0.0000	0.1391	0.0000	0.1327	0.0000	0.1268	0.0000	0.1215	0.0000	0.1165	0.0000	0.1120
1	0.0027	0.2277	0.0025	0.2174	0.0024	0.2081	0.0023	0.1995	0.0022	0.1916	0.0021	0.1843
2	0.0215	0.3015	0.0204	0.2882	0.0194	0.2761	0.0185	0.2649	0.0177	0.2547	0.0169	0.2451
3	0.0443	0.3678	0.0420	0.3520	0.0399	0.3374	0.0380	0.3240	0.0363	0.3116	0.0348	0.3002
4	0.0742	0.4292	0.0703	0.4110	0.0668	0.3944	0.0636	0.3790	0.0607	0.3647	0.0580	0.3515
5	0.1063	0.4868	0.1006	0.4667	0.0954	0.4480	0.0908	0.4309	0.0866	0.4149	0.0828	0.4000
6	0.1425	0.5414	0.1348	0.5194	0.1278	0.4991	0.1215	0.4802	0.1158	0.4627	0.1107	0.4464
7	0.1805	0.5934	0.1706	0.5698	0.1616	0.5479	0.1536	0.5276	0.1463	0.5086	0.1397	0.4910
8	0.2216	0.6430	0.2091	0.6179	0.1980	0.5947	0.1880	0.5730	0.1790	0.5528	0.1708	0.5338
9	0.2644	0.6903	0.2493	0.6642	0.2358	0.6396	0.2238	0.6168	0.2129	0.5953	0.2031	0.5753
10	0.3097	0.7356	0.2917	0.7083	0.2758	0.6829	0.2615	0.6589	0.2486	0.6365	0.2370	0.6154
11					0.3171	0.7242	0.3004	0.6996	0.2854	0.6761	0.2719	0.6542
12									0.3239	0.7146	0.3083	0.6917
x	n = 25		n = 26		n = 27		n = 28		n = 29		n = 30	
0	0.0000	0.1078	0.0000	0.1039	0.0000	0.1002	0.0000	0.0969	0.0000	0.0937	0.0000	0.0907
1	0.0020	0.1775	0.0020	0.1713	0.0019	0.1654	0.0018	0.1599	0.0018	0.1548	0.0017	0.1500
2	0.0162	0.2363	0.0156	0.2280	0.0150	0.2204	0.0145	0.2132	0.0140	0.2064	0.0135	0.2001
3	0.0333	0.2895	0.0320	0.2795	0.0308	0.2703	0.0296	0.2616	0.0286	0.2534	0.0276	0.2457
4	0.0556	0.3392	0.0534	0.3277	0.0513	0.3169	0.0494	0.3068	0.0476	0.2974	0.0460	0.2885
5	0.0793	0.3862	0.0761	0.3733	0.0731	0.3612	0.0704	0.3498	0.0678	0.3392	0.0654	0.3291
6	0.1059	0.4312	0.1016	0.4169	0.0976	0.4036	0.0939	0.3910	0.0905	0.3792	0.0873	0.3681
7	0.1336	0.4744	0.1281	0.4589	0.1230	0.4444	0.1183	0.4308	0.1140	0.4179	0.1099	0.4058
8	0.1634	0.5161	0.1565	0.4995	0.1503	0.4839	0.1445	0.4692	0.1391	0.4554	0.1341	0.4423
9	0.1941	0.5565	0.1859	0.5388	0.1784	0.5222	0.1714	0.5065	0.1650	0.4917	0.1590	0.4777
10	0.2264	0.5956	0.2167	0.5769	0.2078	0.5594	0.1997	0.5428	0.1921	0.5271	0.1851	0.5123
11	0.2596	0.6335	0.2484	0.6140	0.2381	0.5955	0.2287	0.5781	0.2200	0.5616	0.2119	0.5460
12	0.2942	0.6703	0.2813	0.6499	0.2696	0.6307	0.2588	0.6124	0.2488	0.5952	0.2396	0.5788
13	0.3297	0.7058	0.3152	0.6848	0.3018	0.6648	0.2896	0.6460	0.2784	0.6280	0.2680	0.6110
14					0.3352	0.6982	0.3214	0.6786	0.3088	0.6600	0.2972	0.6423
15									0.3400	0.6912	0.3271	0.6729

TABLE A2 Limits of 99% equal-tail LAC interval when x successes are observed in a sample of size n.

x	$n = 1$		$n = 2$		$n = 3$		$n = 4$		$n = 5$		$n = 6$	
0	0.0000	0.9900	0.0000	0.8801	0.0000	0.7572	0.0000	0.6565	0.0000	0.5761	0.0000	0.5121
1	0.0100	1.0000	0.0050	0.9950	0.0033	0.9297	0.0025	0.8441	0.0020	0.7636	0.0017	0.6928
2					0.0703	0.9967	0.0500	0.9500	0.0388	0.8841	0.0318	0.8177
3									0.1159	0.9612	0.0925	0.9075

x	$n = 7$		$n = 8$		$n = 9$		$n = 10$		$n = 11$		$n = 12$	
0	0.0000	0.4602	0.0000	0.4175	0.0000	0.3819	0.0000	0.3518	0.0000	0.3260	0.0000	0.3037
1	0.0014	0.6320	0.0013	0.5800	0.0011	0.5353	0.0010	0.4966	0.0009	0.4629	0.0008	0.4334
2	0.0269	0.7558	0.0233	0.7005	0.0206	0.6513	0.0184	0.6078	0.0167	0.5693	0.0152	0.5351
3	0.0770	0.8510	0.0660	0.7965	0.0577	0.7462	0.0513	0.7004	0.0462	0.6590	0.0420	0.6217
4	0.1490	0.9230	0.1261	0.8739	0.1094	0.8253	0.0966	0.7794	0.0866	0.7368	0.0784	0.6976
5					0.1747	0.8906	0.1532	0.8468	0.1365	0.8046	0.1232	0.7649
6									0.1954	0.8635	0.1754	0.8246

x	$n = 13$		$n = 14$		$n = 15$		$n = 16$		$n = 17$		$n = 18$	
0	0.0000	0.2842	0.0000	0.2670	0.0000	0.2518	0.0000	0.2382	0.0000	0.2260	0.0000	0.2150
1	0.0008	0.4072	0.0007	0.3840	0.0007	0.3633	0.0006	0.3446	0.0006	0.3277	0.0006	0.3124
2	0.0140	0.5045	0.0130	0.4771	0.0121	0.4524	0.0113	0.4300	0.0106	0.4097	0.0100	0.3912
3	0.0385	0.5879	0.0356	0.5574	0.0331	0.5297	0.0309	0.5045	0.0289	0.4815	0.0272	0.4604
4	0.0717	0.6618	0.0660	0.6291	0.0612	0.5991	0.0570	0.5717	0.0534	0.5464	0.0502	0.5233
5	0.1122	0.7280	0.1030	0.6938	0.0953	0.6622	0.0886	0.6331	0.0828	0.6061	0.0777	0.5812
6	0.1592	0.7875	0.1458	0.7527	0.1345	0.7200	0.1248	0.6897	0.1165	0.6614	0.1092	0.6351
7	0.2125	0.8408	0.1939	0.8061	0.1785	0.7731	0.1653	0.7420	0.1539	0.7128	0.1441	0.6855
8					0.2269	0.8215	0.2097	0.7903	0.1949	0.7606	0.1822	0.7326
9									0.2394	0.8051	0.2233	0.7767

x	$n = 19$		$n = 20$		$n = 21$		$n = 22$		$n = 23$		$n = 24$	
0	0.0000	0.2049	0.0000	0.1958	0.0000	0.1875	0.0000	0.1798	0.0000	0.1728	0.0000	0.1662
1	0.0005	0.2984	0.0005	0.2857	0.0005	0.2739	0.0005	0.2631	0.0004	0.2531	0.0004	0.2438
2	0.0095	0.3743	0.0090	0.3587	0.0086	0.3444	0.0082	0.3311	0.0078	0.3188	0.0075	0.3074
3	0.0257	0.4411	0.0244	0.4232	0.0232	0.4067	0.0221	0.3915	0.0211	0.3773	0.0202	0.3640
4	0.0473	0.5019	0.0448	0.4821	0.0425	0.4637	0.0405	0.4467	0.0386	0.4308	0.0369	0.4160
5	0.0733	0.5581	0.0693	0.5367	0.0657	0.5167	0.0624	0.4982	0.0595	0.4809	0.0569	0.4647
6	0.1028	0.6106	0.0970	0.5878	0.0919	0.5665	0.0873	0.5466	0.0832	0.5280	0.0794	0.5106
7	0.1354	0.6599	0.1278	0.6359	0.1209	0.6135	0.1148	0.5925	0.1092	0.5727	0.1042	0.5542
8	0.1710	0.7062	0.1611	0.6814	0.1523	0.6580	0.1444	0.6360	0.1374	0.6153	0.1309	0.5958
9	0.2093	0.7498	0.1969	0.7243	0.1860	0.7002	0.1762	0.6774	0.1674	0.6559	0.1595	0.6356
10	0.2502	0.7907	0.2351	0.7649	0.2218	0.7403	0.2100	0.7169	0.1993	0.6947	0.1897	0.6737
11					0.2597	0.7782	0.2456	0.7544	0.2329	0.7317	0.2215	0.7102
12									0.2683	0.7671	0.2549	0.7451

TABLE A2 Continued

x	$n = 25$		$n = 26$		$n = 27$		$n = 28$		$n = 29$		$n = 30$	
0	0.0000	0.1602	0.0000	0.1545	0.0000	0.1493	0.0000	0.1444	0.0000	0.1398	0.0000	0.1355
1	0.0004	0.2352	0.0004	0.2271	0.0004	0.2196	0.0004	0.2126	0.0003	0.2060	0.0003	0.1998
2	0.0072	0.2968	0.0069	0.2869	0.0066	0.2776	0.0064	0.2689	0.0062	0.2607	0.0059	0.2530
3	0.0193	0.3517	0.0185	0.3401	0.0178	0.3293	0.0172	0.3192	0.0166	0.3096	0.0160	0.3006
4	0.0353	0.4022	0.0339	0.3892	0.0326	0.3771	0.0314	0.3656	0.0302	0.3548	0.0292	0.3446
5	0.0544	0.4495	0.0522	0.4352	0.0501	0.4218	0.0482	0.4092	0.0465	0.3973	0.0448	0.3861
6	0.0760	0.4942	0.0728	0.4788	0.0699	0.4643	0.0672	0.4506	0.0647	0.4377	0.0624	0.4255
7	0.0996	0.5367	0.0954	0.5203	0.0916	0.5048	0.0880	0.4902	0.0847	0.4763	0.0817	0.4632
8	0.1251	0.5774	0.1198	0.5600	0.1149	0.5436	0.1104	0.5281	0.1062	0.5134	0.1023	0.4995
9	0.1523	0.6163	0.1457	0.5982	0.1397	0.5810	0.1341	0.5646	0.1290	0.5492	0.1242	0.5345
10	0.1810	0.6537	0.1731	0.6348	0.1658	0.6169	0.1592	0.5999	0.1530	0.5837	0.1473	0.5683
11	0.2112	0.6896	0.2018	0.6701	0.1933	0.6515	0.1854	0.6339	0.1782	0.6171	0.1715	0.6010
12	0.2429	0.7241	0.2319	0.7040	0.2219	0.6849	0.2128	0.6667	0.2044	0.6493	0.1966	0.6328
13	0.2759	0.7571	0.2633	0.7367	0.2518	0.7172	0.2413	0.6985	0.2316	0.6806	0.2227	0.6635
14					0.2828	0.7482	0.2709	0.7291	0.2599	0.7108	0.2498	0.6933
15									0.2892	0.7401	0.2778	0.7222