Published in final edited form as:

J Stat Comput Simul. 2020 ; 90(1): 75–89. doi:10.1080/00949655.2019.1672695.

Calibrated Bayesian Credible Intervals for Binomial Proportions

Robert H. Lyles¹, Paul Weiss¹, Lance A. Waller¹

¹Department of Biostatistics and Bioinformatics, The Rollins School of Public Health of Emory University, 1518 Clifton Rd. N.E., Mailstop 1518-002-3AA, Atlanta, GA 30322

Abstract

Drawbacks of traditional approximate (Wald test-based) and exact (Clopper-Pearson) confidence intervals for a binomial proportion are well-recognized. Alternatives include an interval based on inverting the score test, adaptations of exact testing, and Bayesian credible intervals derived from uniform or Jeffreys beta priors. We recommend a new interval intermediate between the Clopper-Pearson and Jeffreys in terms of both width and coverage. Our strategy selects a value κ between 0 and 0.5 based on stipulated coverage criteria over a grid of regions comprising the parameter space, and bases lower and upper limits of a credible interval on $Beta(\kappa, 1 - \kappa)$ and $Beta(1 - \kappa, \kappa)$ priors, respectively. The result tends toward the Jeffreys interval if the criterion is to ensure an average overall coverage rate $(1 - \alpha)$ across a single region of width 1, and toward the Clopper-Pearson if the goal is to constrain both lower and upper lack of coverage rates at $\alpha/2$ with region widths approaching zero. We suggest an intermediate target that ensures all average lower and upper lack of coverage rates over a specified set of regions are $\alpha/2$. Interval width subject to these criteria is readily optimized computationally, and we demonstrate particular benefits in terms of coverage balance.

Keywords

Approximate inference; Confidence interval; Exact inference; Lower bound; Upper bound

1. INTRODUCTION

Interval estimation for a binomial proportion (p) is one of the most common and fundamental goals in statistics, often tending to be filed in the methodologically “solved” category by practitioners and statisticians. Interest in its nuances remains prominent, however, even in literature more recent and more compelling than one might expect. Much of this renewed interest stems from well-recognized drawbacks attributed to the most common approaches to the problem, i.e., the standard Wald interval typically taught in elementary statistics courses and the most common of the “exact” intervals (Clopper and Pearson, 1934). Among others, Agresti and Coull (1998) provide a clear picture of the coverage failures of the former, together with the ultra-conservatism and undesirably high interval widths associated with the latter. The fact that this and similar references are extremely

*phone: 404-727-1310; fax: 404-727-1370; rlyles@sph.emory.edu.

commonly cited is a testament both to the insight provided and to the continued every day relevance of this fundamental problem to biometrical practice.

While numerous alternative methods with more favorable properties have been proposed, we can loosely categorize them into three groups. The first consists of better-behaved (relative to the Wald) approximate confidence intervals (CIs) for p that are simple to calculate, among which the interval derived from inverting the score test is arguably most prominent (e.g., Wilson, 1927; Vollset, 1993; Agresti and Coull, 1998; Agresti and Min, 2001; Brown, Cai and DasGupta, 2001; Thulin, 2014). The second category may be viewed as a set of “exact” methods essentially akin to the Clopper-Pearson that are designed to ensure no less than a specified overall coverage rate $(1-\alpha)$ for any value of p , but which buy improvements in terms of interval width by discarding the notion that both lower and upper lack of coverage rates need be controlled at the level $\alpha/2$ (e.g., Sterne, 1954; Crow, 1956; Blyth and Still, 1983; Casella, 1986; Blaker, 2000). These innovative methods reflect improvements over time, but remain conservative and require less transparent calculations when compared with the score test-based and similar approximate CIs. They can also exhibit unusual properties in terms of a lack of nestedness, connectivity, and/or monotonicity in α (Thulin, 2014). Finally, a third set of procedures targets Bayesian credible intervals, generally based on beta priors. Such procedures often exhibit favorable frequentist properties (e.g., Carlin and Louis, 2009), and the interval derived from a non-informative Jeffreys beta prior is particularly well-accepted and recommended (e.g., Brown, Cai and DasGupta, 2001; 2002).

In this article, we propose a new variant on the Bayesian credible interval strategy, rooted in the well-known observation that the Clopper-Pearson limits are equivalent to percentiles of specific beta posterior distributions. Like the approximate (e.g., score-based) CIs and common (e.g., uniform or Jeffreys prior-based) credible intervals, our approach sacrifices the assurance of at least nominal coverage for some values of p . However, it retains significant homage for the balanced coverage property of the conservative Clopper-Pearson interval by targeting an average upper and lower coverage rate no worse than $\alpha/2$ within each of a set of specified grid regions that encompass the parameter space of p . For a given sample size n , this is accomplished by determining the value of a tuning parameter (κ) between 0 and 0.5. Lower and upper limits of the interval are based on the posterior distributions corresponding to two separate beta priors determined from the selected value of κ . Setting $\kappa=0$ yields the Clopper-Pearson approach, while setting $\kappa=0.5$ yields the Jeffreys interval. Intervals obtained in the proposed way fall between those two extremes with respect to both average width and coverage levels for any true value of p . One can search for an optimal value of κ , in terms of minimizing average interval width subject to specified coverage criteria. Such a search is facilitated because the average upper and lower coverage rates for any grid interval can get no larger as κ increases, while the average interval width is monotonically decreasing with κ .

In the following sections, we briefly review specific existing interval approaches considered here that fall into the three aforementioned categories. We then describe the proposed method, with graphical illustrations of its coverage characteristics relative to other intervals. For brevity, we restrict these illustrations to the case of $n=10$ Bernoulli trials; however,

tabular results provide the essentials of the comparisons across a wider variety of sample sizes.

2. METHODS

We consider the standard scenario of n Bernoulli trials yielding y “successes” and $n-y$ “failures”, with Y denoting the binomial random variable counting successes and $\hat{p} = y/n$. We denote the $100(\alpha/2)$ th percentile of the standard normal distribution as $Z_{\alpha/2}$, and the $100(\alpha/2)$ th percentile of a beta distribution with shape parameters ϕ_1 and ϕ_2 as $(Beta(\alpha/2; \phi_1, \phi_2))$. When discussing Bayesian credible intervals, we make use of the well-known conjugacy property that a $Beta(\phi_1, \phi_2)$ prior for p yields a $Beta(\phi_1 + y, \phi_2 + n-y)$ posterior. For all such intervals considered here, we stipulate a lower posterior limit of 0 when $y=0$ and an upper posterior limit of 1 when $y=n$. This logical and minimal adjustment avoids the possibility of 0% coverage for extremely large or small true values of p , and is in alignment with the frequentist alternatives considered here in those two extreme cases.

2.1 Approximate frequentist intervals

The common CI based on inverting the Wald test for a binomial proportion needs no review. It is well known that its coverage properties can be extremely poor in small samples; they can remain deficient even with moderate to large n , particularly for relatively small or large values of p . Most textbook rules of thumb designed to defend its use tend to be unsatisfactory (e.g., Agresti and Coull, 1998; Brown et al., 2001). However, we agree with others (e.g., Agresti and Min, 2001) that a sizable share of its coverage deficiencies stem from the degradation of the CI to a single point (0 or 1) when 0 or n successes are observed. Nevertheless, most attempts to refine the Wald CI are less effective than alternatives, such as the CI obtained from inverting the score test (Wilson, 1927):

$$\left(\hat{p} + \frac{z_{\alpha/2}^2}{2n} \pm z_{\alpha/2} \sqrt{\left[\hat{p}(1 - \hat{p}) + z_{\alpha/2}^2/4n \right] / n} \right) / (1 + z_{\alpha/2}^2/n) \quad (1)$$

Agresti and Coull (1998) and Agresti and Caffo (2000) noted that $z_{\alpha/2}^2 \approx 4$ in most hypothesis testing scenarios and showed that (1) could be approximated by adding in 4 artificial trials (2 successes and 2 failures) and then calculating the standard Wald CI. They noted the appeal of such an adjusted Wald CI for use in introductory statistics texts or courses. The adjusted Wald interval is slightly more conservative (see Discussion), but we focus primarily on (1) as one of the best representatives among the set of proposed approximate frequentist CIs for p .

2.2 “Exact” intervals

The best known among this class is the Clopper-Pearson (C-P) interval, obtained by inverting the two one-sided tests of size $\alpha/2$ based directly on the binomial distribution. To our knowledge, this is the only viable CI to maintain strict control of both lower and upper lack of coverage rates at no more than $\alpha/2$ for all values of p . For our purposes, it is useful to frame the C-P interval in a well-known way, where the lower and upper C-P limits are the following beta distribution percentiles:

$$Beta(\alpha/2; y, n - y + 1) \text{ and } Beta(1 - \alpha/2; y + 1, n - y). \quad (2)$$

That is, one may view the lower limit as derived from a $Beta(0,1)$ prior, and the upper limit from a $Beta(0,1)$ prior. The exceptions occur when $y=0$ (in which case the lower limit is 0), and when $y=n$ (in which case the upper limit is 1).

Alternative exact intervals to the C-P include those proposed by Sterne (1954), Crow (1956), Blyth and Still (1983), Casella (1986), and Blaker (2000). The essential advantage of these methods is their basis in a single two-sided test targeting an overall error rate of $100(1-\alpha)\%$ for any p . Lower and upper limits so obtained cannot both be relied upon as exact lower and upper bounds, however, as balance in coverage is sacrificed in order to obtain narrower average interval widths relative to the C-P approach. In what follows, we illustrate some comparative properties of the most recent of this class of intervals (Blaker, 2000).

2.3 Credible intervals and proposed coverage criteria

Arguably, the two most popular Bayesian credible intervals for p are those based on the non-informative uniform [$Beta(1,1)$] and Jeffreys [$Beta(0.5,0.5)$] priors. As both lead to beta posteriors, there is no need for sampling and the intervals can easily be computed using any statistical software program (e.g., SAS or R) that gives access to percentiles of the beta distribution. Aside from the minor modifications when $y=0$ or n , the two intervals are:

$$\text{Uniform: } [Beta(\alpha/2; y + 1, n - y + 1), Beta(1 - \alpha/2; y + 1, n - y + 1)]$$

$$\text{Jeffreys: } [Beta(\alpha/2; y + 0.5, n - y + 0.5), Beta(1 - \alpha/2; y + 0.5, n - y + 0.5)] \quad (3)$$

The Jeffreys interval in particular comes highly recommended for its satisfactory average coverage properties, including for low or high values of p .

In developing the proposed approach, we first observed that most articles recommending particular approximate (e.g., score-based) or Bayesian (e.g., Jeffreys prior-based) intervals do so on the basis of average overall coverage and favorable average width over the entire range (0 to 1) of p (see, e.g., Agresti and Coull, 1998; Brown et al., 2001). These methods are quite defensible on that basis, although they involve no direct effort to control coverage within any particular subset of the parameter space or to ensure balancedness of coverage. Similarly, proposed “exact” alternatives to the C-P interval focus on control of overall coverage at all values of p without specific control of lower and upper coverage error rates. An appealing alternative to the existing methods might thus be one that respects coverage balance and control of coverage within subsets of the parameter space as criteria, yet affords notable improvements in average width relative to the C-P interval.

Our proposal stems from simple observations about the connection between the C-P interval in (2) and the Jeffreys interval in (3). Both methods base the lower and upper limits on beta priors with shape parameters summing to 1. While the Jeffreys interval derives both limits from the same prior, the C-P interval achieves formal control of both the lower and upper coverage rates by concentrating all weight on the first and second shape parameters in the

beta prior used to compute the upper and lower limits, respectively. The Jeffreys and C-P intervals can be seen as the two extremes to a rule in which one calculates the lower limit based on a $Beta(\kappa, 1 - \kappa)$ prior and the upper limit based on a $Beta(1 - \kappa, \kappa)$ prior for p , where $\kappa \in [0, 0.5]$ is a tuning parameter. Equivalently, they are the extremes of a class of intervals with lower and upper limits equal to the following beta distribution percentiles:

$$Beta(\alpha/2; y + \kappa, n - y + 1 - \kappa) \text{ and } Beta(1 - \alpha/2; y + 1 - \kappa, n - y + \kappa). \quad (4)$$

Clearly, $\kappa=0$ corresponds to the C-P and $\kappa=0.5$ yields the Jeffreys interval.

To select the value of κ , we propose the following strategy. First the analyst segments the range of p into a set of regions, across which he/she stipulates the coverage property that the maximum average probabilities of missing high and missing low are both to be controlled at a level no larger than $\alpha/2$. Because the lower and upper limits in (4) are monotonically increasing and decreasing in κ , respectively, it follows that the optimal κ (in terms of interval width) is the largest value between 0 and 0.5 such that this property holds. This monotonicity facilitates the search for the optimal κ ; a unique solution is guaranteed regardless of how fine the chosen grid regions are, due to the fact that $\kappa=0$ yields the C-P.

While the specified grid is at the discretion of the analyst, we suggest defining the regions to be of equal width (w) and equally spaced as long as there is no a priori reason for deference to a particular subset of the parameter space (see Discussion). For all interval estimation methods considered here (including the proposed approach), the lower limit for y successes will be one minus the upper limit for $(n-y)$ successes. Thus, to ensure coverage properties are met for all regions of width w over the entire range (0 to 1) of p , it is sufficient to define the regions over the interval 0 to 0.5. The overall coverage within any region to the left of $p=0.5$ is the same for the corresponding region to the right of 0.5, with the only difference being that the lower and upper lack of coverage rates are interchanged for those two regions. For most applications, we therefore recommend defining regions over $p \in [0, 0.5]$ as follows:

$$[0, w], (w, 2w], (2w, 3w], \dots, ((R-1)w, Rw],$$

where $w=1/(2R)$ and R is the number of regions.

Note that as R becomes large, the optimal κ must go to 0 for any n and the resulting interval tends toward the C-P. On the other hand, if the analyst chooses $R=1$, then the resulting interval will be equivalent to the Jeffreys ($\kappa=0.5$) unless the latter fails to yield overall average lower and upper lack of coverage rates that are both $\alpha/2$ for the specified value of n . Otherwise it will be a slightly more conservative interval (with κ likely near but less than 0.5), i.e., a minor variant on the Jeffreys that achieves such balanced control of coverage on average over the entire (0,1) range of p .

Our preference is to designate a scenario between the two extremes, such that $w=0.1$ and $R=5$. That is, we recommend specifying the grid regions (0, 0.1), (0.1, 0.2), (0.2, 0.3), (0.3, 0.4), (0.4, 0.5) and determining the largest value of $\kappa \in [0, 0.5]$ such that the maximal average lower and upper lack of coverage rates across those 5 regions are both no more than

$\alpha/2$. A SAS macro that searches to find the optimal κ under this grid definition for any specified n is available from the authors by request; the macro optionally produces plots similar to those shown in Section 3.

3. RESULTS

To explore coverage properties, we consider the probabilities of an interval missing low (p_{low}), missing high (p_{high}), and missing at all (p_{miss}), over any particular region ($\mathfrak{R}$) of the parameter space. For example, one can define the probability of an interval missing on the low side for a particular value of p as

$$p_{low|p} = \sum_{y=0}^n I_{low}(y, p) \Pr(Y = y | P = p),$$

where $I_{low}(y, p)$ is a (0,1) indicator for whether the upper limit of the interval is less than p when $Y=y$. The average probability of the interval missing low over the region $\mathfrak{R}$ is then given by

$$p_{low|p \in \mathfrak{R}} = \frac{1}{\Pr(p \in \mathfrak{R})} \int_{\mathfrak{R}} \sum_{y=0}^n I_{low}(y, p) \Pr(Y = y | P = p) f_P(p) dp. \quad (5)$$

Prior authors (e.g., Agresti and Coull, 1998) examined a version of (5) with p_{miss} as the target and the entire parameter space as the region of interest. Here, we consider the three separate targets (p_{low} , p_{high} , p_{miss}) with grid regions in keeping with the specifications of the proposed interval approach. For computations, we made accurate Monte Carlo approximations to (5) by generating large numbers of random draws for p from a uniform distribution over (0, 0.5) and averaging $p_{low|p}$, $p_{high|p}$ and $p_{miss|p}$ across values of p within region $\mathfrak{R}$.

Figure 1 plots the overall coverage rates and the probabilities of missing low and missing high vs. p for the 95% score and uniform intervals, when $n=10$. As with similar plots in prior references (e.g., Agresti and Coull, 1998), one sees in panels A and C that overall coverage is close to nominal on average but fluctuates markedly about the 95% level, especially when p is close to 0 or 1. Panels B and D provide particularly useful added information about coverage balance, however, with dashed horizontal lines drawn to identify excursions beyond the level $\alpha/2=0.025$. For values of $p < 0.5$, both the score and uniform intervals show a clear tendency to miss predominantly on the high side; this is particularly so for lower true values of p . This pattern is mirrored exactly for the corresponding values of $p > 0.5$, so that the tendency switches toward misses on the low side. Coverage properties of these two intervals are remarkably similar, except high and low side excursions for values of p near the boundaries are somewhat worse for the uniform interval. Comparisons across a variety of sample sizes (not shown) yield the same conclusions; thus, we henceforth drop the uniform in favor of the score interval after noting their similarity.

Figure 2 plots probabilities of missing low and missing high vs. p for the case of $n=10$ for the C-P, score, Jeffreys, and the proposed interval in (4). For the latter, specifying $w=0.1$ and $R=5$ yields an optimal value of $\kappa=0.378$. Having noted the symmetry on either side of $p=0.5$ in Figure 1 (which holds for all intervals considered), we restrict the plots in Figure 2 to the range $p \in (0, 0.5)$ to allow a more detailed visual comparison of the four intervals. As expected, the C-P interval admits no excursions beyond the $\alpha/2=0.025$ level. Again the tendency for the score interval to miss predominantly on the high side is clear, especially for lower true values of p . In contrast, the Jeffreys and the proposed interval are less subject to missing high, and somewhat more subject to missing on the low side. All four intervals provide overall average coverage rates across the full range of p that are at least nominal (0.984, 0.954, 0.953, and 0.964 for the C-P, score, Jeffreys, and proposed intervals, respectively).

Interestingly, Figure 1 revealed that the score interval performs very much like a more conservative version of the uniform, while Figure 2 confirms that the proposed interval behaves like a more conservative version of the Jeffreys (as designed). Again, although not plotted in Figure 2, it is important to note that the performance of each interval over $p \in (0.5, 1)$ mirrors that seen over $p \in (0, 0.5)$. However, the average widths and average overall coverages of each interval over the full $(0, 1)$ range are the same as observed over the $(0, 0.5)$ range. As an aside, note that the “jumps” in Figures 1 and 2 relate to the fact that there are $(n+1)$ distinct intervals that correspond to the observable values of y .

Figure 3 summarizes coverage properties for the same four CI methods for $n=10$, by plotting the average upper and lower excursion rates over the $R=5$ regions of width $w=0.1$ that were specified in conjunction with the proposed approach. As expected, the ultra-conservativeness of the C-P interval is apparent. Note that the score interval yields an average upper excursion rate markedly above nominal for all five regions except for $(0.4, 0.5)$, with all lower excursion rates below nominal. The Jeffreys interval yields average upper excursion rates minimally above the nominal level for each region, but surrenders a rate of low-side errors well above nominal in the region $p \in (0.2, 0.3)$. The proposed interval with $\kappa=0.378$ performs as designed, constraining both the average lower and upper lack of coverage rates at no higher than 2.5% across all five regions.

Table 1 provides summary statistics for 95% interval width and coverage properties, over a range of sample sizes. Coverage is summarized in terms of averages over the 5 regions as well as maximal excursion probabilities over the entire range of p . As expected, the C-P interval is always highly conservative on average and excessively wide relative to the others, given that it admits upper and lower error rates no higher than 0.025 for any value of p .

Summary statistics in Table 1 for the score interval for each sample size (n) reveal similar patterns to those seen in Figures 2 and 3. In particular, it is noticeably anticonservative with respect to high-side errors when $p < 0.5$ (these translate to the same rates of low-side errors for $p > 0.5$), both in terms of maximal overall and regional-specific average excursion probabilities. Overall average coverage over the full p range remains near nominal for the score interval, as high-side excursions are countered by conservativeness with respect to low-side errors when $p < 0.5$ and the converse occurs for $p > 0.5$. Summary statistics on the

Jeffreys and proposed intervals are similarly in keeping with patterns seen in the figures, with the proposed method more conservative than the Jeffreys in terms of both types of errors. As expected, average widths and coverage control properties of the proposed interval are always intermediate between those of the C-P and Jeffreys, and the method controls maximal average lower and upper excursion rates at no more than 2.5% across all 5 regions (see bolded numbers; Table 1).

For $n=10$, Table 2 shows how the optimal value of κ decreases and the proposed 95% interval becomes wider and more conservative as the chosen number of regions (R) increases. For example, with $R=1$ the approach achieves a small conservative adjustment to the Jeffreys in which the overall average probabilities of high- and low-side errors across the full range of p are both no larger than 0.025. As such, the optimal κ (0.427) is close to 0.5 and the average width of the interval is only slightly larger than that of the Jeffreys (0.444 vs. 0.433; Tables 1 and 2). On the other extreme, the optimal κ approaches 0 and the proposed interval approaches the C-P as R becomes large. With $R=40$ and correspondingly small interval widths of 0.0125, the average width of the proposed interval is just slightly narrower than that of the C-P (0.501 vs. 0.508) and the method admits maximal lower and upper errors only slightly above nominal over the full range of p .

Figure 4 compares the proposed 95% interval (for $n=10$) against one of a class of “exact” intervals that is also guaranteed to be contained within the C-P (Blaker, 2000). We note that the Blaker interval, like others (e.g., the C-P and Jeffreys) considered here, is available in common commercial software (SAS Institute Inc., 2015). For this comparison, we specified a somewhat finer grid in which $R=10$ and $w=0.05$; the optimal κ value under this grid definition is 0.254 (see Table 2). As in Figure 1, we present this plot over the entire (0, 1) range of p in order to highlight the symmetry in terms of lower and upper excursion probabilities for corresponding regions to the right and left of $p=0.5$. As designed, the Blaker interval never allows a total error rate larger than 0.05 for any value of p (Fig. 4A). Interestingly, this is narrowly avoided near the point $p=0.283$, where a dip in the upper error rate is quickly followed by a jump in the lower error rate; the analogous event occurs at $p=1-0.283=0.717$. In contrast, the proposed approach allows a maximal total error rate of approximately 0.072, which occurs roughly over the ranges $p \in (0.266, 0.283)$ and $p \in (0.717, 0.734)$; see Fig. 4C. However, because the Blaker method strictly controls only overall error, Figure 4B shows that it is subject to average upper error rates higher than $\alpha/2=0.025$ for two of the 10 intervals defined in the grid when $p < 0.5$ and to average lower error rates higher than 0.025 for the two corresponding intervals in the grid when $p > 0.5$.

Fig. 4D confirms that the proposed approach controls average lower and upper error rates over all 20 intervals at no higher than 0.025, as expected. For comparison, the average widths (over all p) for the Blaker and the proposed interval are 0.476 and 0.471, respectively. In general, the Blaker CI achieves its specified strict error control for any p by admitting a greater level of coverage imbalance in favor of high-side errors over the range (0, 0.5) and low-side errors over the range (0.5, 1). While we chose the Blaker method for comparison here, the same general conclusion would be expected to hold for other members of the class of “exact” intervals that sacrifice simultaneous control of lower and upper coverage rates (e.g., Sterne, 1954; Crow, 1956; Blyth and Still, 1983; Casella, 1986).

4. VARIATIONS ON THE PROPOSED APPROACH

While we favor the proposed criteria for balanced control (at $\alpha/2$) of average lower and upper lack of coverage rates across a specified set of grid regions, one could consider alternative criteria as well. For example, note in Table 1 that for $n=10$ and an overall target coverage of $100(1-\alpha)\% = 95\%$, the proposed interval with $\kappa=0.378$ admits maximal lower and upper excursion probabilities of 0.062 and 0.081. A similar search for the optimal κ to control both of these, for example at 0.050 or less, yields $\kappa=0.197$. The resulting interval would be better balanced but slightly wider than the Blaker interval (overall average width 0.480 vs. 0.476), where the latter clearly also meets the new specified criterion by design and as seen in Fig. 4A.

A potential generalization of the proposed prior specifications could be to allow the value of κ to vary both with the observed value of y and according to whether one is deriving the lower or upper limit. That is, one could base the lower limit on a $Beta(\kappa_{yL}, 1-\kappa_{yL})$ prior and the upper limit on a $Beta(1-\kappa_{yU}, \kappa_{yU})$ prior for p , for $y=0, 1, \dots, n$. The resulting generalized version of the proposed interval then takes on something of an empirical Bayes flavor (e.g., Carlin and Louis, 2009) and yields lower and upper limits equal to the following percentiles:

$$Beta(\alpha/2; y + \kappa_{yL}, n - y + 1 - \kappa_{yL}) \text{ and } Beta(1 - \alpha/2; y + 1 - \kappa_{yU}, n - y + \kappa_{yU}). \quad (6)$$

In this expanded framework, one would make the following restrictions on the κ 's to ensure the typical property of lower limits (LL) and upper limits (UL) for a given value of y and $y^* = n-y$, namely, that $LL_{y^*} = 1-UL_y$:

$$\kappa_{yL} = \kappa_{y^*U}, \quad y = 0, 1, \dots, n.$$

Table 3 shows an example of how the general specification in (6) can reproduce other proposed intervals designed to meet specific coverage targets. Specifically, we show the unique κ_{yL} and κ_{yU} values that correspond to the Blaker (2000) interval for $n=10$. While intriguing, we note that the value of the generalization in (6) may be modest given that any effort to optimize with regard to specific coverage criteria is complicated by introducing multiple κ 's (see Discussion).

5. DISCUSSION

All alternatives to the C-P approach seek less conservative (i.e., narrower) intervals. In general, this is achieved either by accepting approximate rather than “exact” coverage properties (as in the case of the score or Bayesian credible intervals), or by maintaining the notion of exactness while admitting unbalancedness in lower and upper coverage rates (e.g., Blyth and Still, 1983; Casella, 1986; Blaker, 2000). Our approach uses the Bayesian credible interval machinery in order to define an approximate interval that targets a level of balance as well as strict control of average error rates within each of a set of defined regions of the parameter space. It allows the analyst the option of a calibrated refinement to approximate intervals such as the score or Jeffreys, that tend to be advocated in the literature on the basis

of average overall width and coverage rates over the entire (0 to 1) range of p (e.g., Agresti and Coull, 1998; Brown et al., 2001).

For a given value of κ between 0 and 0.5, the proposed interval is trivial to calculate given access to statistical software (eqn. 4), and is guaranteed both to contain the Jeffreys interval and to be contained by the C-P interval. Thus, one could simply choose an intermediate value of κ such as 0.25 or 0.33 to be assured of an interval more conservative than the Jeffreys and less conservative than the C-P. Optimizing κ to minimize average interval width subject to specified coverage criteria requires a search that is greatly facilitated by monotonicity, in the sense that the lower and upper limits in (4) increase and decrease, respectively, as κ increases. While other criteria can be considered, we prefer the proposed control of maximal average lower and upper excursion probabilities over a specified set of R regions of width $w=1/(2R)$ defined over $p \in [0, 0.5]$. As noted, the criteria met pertaining to lower (upper) excursions over those R regions naturally matches the criteria met for upper (lower) excursions over the corresponding R regions over $p \in [0.5, 1]$. A SAS macro that performs the required search for the case $R=5$ and $w=0.1$ is available from the authors by request.

The improved balance in coverage achievable using (4) relative to common alternatives like the score and Jeffreys intervals brings it some inherent appeal. Figure 1 and Table 1 reveal that the score interval shows a clear tendency to miss on the high side for true values of $p < 0.5$, with a corresponding tendency to miss low for $p > 0.5$. This is particularly true for more extreme values of p , and reveals that the near-nominal overall coverage rate of the score interval across the full (0, 1) range of p noted in the literature comes at the expense of coverage balance. While we find it similar but preferable to the uniform interval and agree with prior authors (e.g., Agresti and Coull, 1998) that the score is much to be preferred over the traditional Wald interval, we view this tendency toward coverage imbalance as a drawback. An approximation to the score interval proposed and dubbed by Agresti and Coull as the “adjusted Wald” or “add two successes and two failures” approach makes an easily calculated and readily taught alternative (Agresti and Coull, 1998; Agresti and Caffo, 2000; Thulin, 2014). Although we did not include this interval in the comparisons discussed in Section 3, we found it to behave like a more conservative version of the score interval

with similar tendencies toward unbalancedness. The values of $\text{Max}(\overline{p_{\text{high}}})$ for this interval under the grid definition with $R=5$ and $w=0.1$ for $p \in (0, 0.5)$ (see Table 1) were 0.046, 0.036, 0.037, 0.036, 0.031, and 0.029 for $n=5, 10, 15, 25, 50$, and 100, respectively. While lower than the corresponding values for the score interval, these all remain considerably above 0.025. The adjusted Wald interval was always wider on average than the score, and also wider than the proposed optimized interval for all sample sizes in Table 1 except for $n=5$.

The Jeffreys is a generally preferred Bayesian credible interval option (e.g., Brown et al., 2001). We find (Table 1; Figures 2 and 3) that for $p < 0.5$ the Jeffreys mitigates the tendency toward high-side errors relative to the score and uniform, but at the expense of low-side errors. The proposed approach is designed to achieve better coverage balance, while assuring a calibrated level of control over average region-specific error rates. We believe it also offers

considerable benefits in cases where the goal is only to find a lower (or upper) bound on p . Among those considered here, it is the only method aside from the C-P that is expressly designed to achieve specified nominal coverage criteria for both high- and low-side errors. It thus provides a criteria-driven method for seeking reduced conservatism relative to the C-P for obtaining lower or upper bounds.

For future work, we note that it may be possible to produce narrower CIs subject to specified coverage criteria using the generalization in (6). Limited initial experimentation, however, suggests there may be little room to improve overall average interval width via (6) as opposed to (4) when the goal is to constrain maximal region-specific average lower and upper lack of coverage rates. We prefer the simplicity of (4) in general, along with the advantage that a single κ offers in terms of the potential to optimize subject to well-defined coverage criteria. A potentially more intriguing angle that we are currently considering is to adapt the approach in (4) to cases in which it can be assumed a priori with full or very high confidence that p is within a specified subset of the parameter space (e.g., when the goal is to estimate prevalence of a rare disease or condition).

ACKNOWLEDGEMENTS

This research was enabled in part by the Bill and Melinda Gates Foundation's support of the Child Health and Mortality Prevention Surveillance Network (CHAMPS), and by grant R01HD092580 from the US Eunice Kennedy Shriver National Institute of Child Health & Human Development. The views represent those of the authors, and do not necessarily represent those of the Gates Foundation or the US NICHD. We are grateful to Dr. Helge Blaker for a helpful correspondence related to characteristics of existing interval methodologies.

REFERENCES

- Agresti A, Caffo B. Simple and effective confidence intervals for proportions and differences of proportions result from adding two successes and two failures. *The American Statistician* 2000; 54:280–288.
- Agresti A, Coull BA. Approximate is better than 'exact' for interval estimation of binomial proportions. *The American Statistician* 1998; 52:119–126.
- Agresti A, Min Y. On small-sample confidence intervals for parameters in discrete distributions. *Biometrics* 2001; 57:963–971. [PubMed: 11550951]
- Blaker H. Confidence curves and improved exact confidence intervals for discrete distributions. *The Canadian Journal of Statistics* 2000; 28:783–798.
- Blyth CR, Still HA. Binomial confidence intervals. *Journal of the American Statistical Association* 1983; 78:108–116.
- Brown LD, Cai TT, DasGupta A. Interval estimation for a binomial proportion. *Statistical Science* 2001; 16:101–133.
- Brown LD, Cai TT, DasGupta A. Confidence intervals for a binomial proportion and asymptotic expansions. *The Annals of Statistics* 2002; 30:160–201.
- Carlin BP, Louis TA. *Bayesian Methods for Data Analysis* (3rd ed). Chapman & Hall/CRC: Boca Raton, FL, 2009.
- Casella G. Refining binomial confidence intervals. *The Canadian Journal of Statistics* 1986; 14:113–129.
- Clopper CJ, Pearson ES. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika* 1934; 26:404–413.
- Crow EL. Confidence intervals for a proportion. *Biometrika* 1956; 43:423–435.
- SAS Institute Inc. *SAS/STAT®14.1 User's Guide: High-Performance Procedures*. SAS Institute Inc.: Cary, NC, 2015.

- Sterne TE. Some remarks on confidence or fiducial limits. *Biometrika* 1954; 41:275–278.
- Thulin M. The cost of using exact confidence intervals for a binomial proportion. *Electronic Journal of Statistics* 2014; 8:817–840.
- Vollset SE. Confidence intervals for a binomial proportion. *Statistics in Medicine* 1993; 12:809–824. [PubMed: 8327801]
- Wilson EB. Probable inference, the law of succession and statistical inference. *Journal of the American Statistical Association* 1927; 22:209–212.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Figure 1.

Overall coverage rates of 95% score (panel A) and uniform (panel C) intervals plotted over the full range of the true proportion (p) for $n=10$, together with upper and lower lack of coverage rates for these intervals (panels B and D). Positive y-axis values in panels B and D represent upper excursion probabilities and negative y-axis values represent lower excursion probabilities (e.g., a value at 0.05 means the interval misses high 5% of the time at that value of p ; a value at -0.05 means the interval misses low 5% of the time at that value of p). Dashed lines are drawn at ± 0.025 .

Figure 2.

Upper and lower lack of coverage rates plotted over the range $p \in (0, 0.5)$ for $n=10$ for the 95% Clopper-Pearson (panel A), score (panel B), Jeffreys (panel C) and the proposed (optimized at $\kappa=0.378$ for $R=5$ regions of width $w=0.10$; panel D) intervals.

Figure 3.

Average upper and lower lack of coverage rates across the $R=5$ regions of width $w=0.1$ over the range $p \in (0, 0.5)$ for $n=10$ for the 95% Clopper-Pearson (panel A), score (panel B), Jeffreys (panel C) and the proposed (optimized at $\kappa=0.378$; panel D) intervals.

Figure 4.

Upper and lower lack of coverage rates of the 95% Blaker (panel A) and proposed (optimized at $\kappa=0.254$ for $R=10$ regions of width $w=0.05$; panel C) intervals plotted over the full range of p for $n=10$, together with average upper and lower lack of coverage rates for these intervals across the specified regions (panels B and D).

Table 1.

Performance properties of alternative intervals for p for various sample sizes ($\alpha=0.05$; Grid definition: $R=5$ regions, $w=.10$)

Interval	n	Avg. Width ^a	Avg. Coverage ^a	Max($\overline{p_{low}}$), Max($\underline{p_{low}}$) ^b	Max($\overline{p_{high}}$), Max($\underline{p_{high}}$) ^b	Max($\overline{p_{miss}}$), Max($\underline{p_{miss}}$) ^c
Score	5	.558	.955	.029, .058	.058, .168	.058, .168
Clopper-Pearson	5	.678	.990	0, 0	.013, .025	.013, .025
Optimal ($\kappa=0.308$) ^d	5	.610	.976	.025 , .055	.021 , .069	.044, .071
Jeffreys	5	.560	.958	.051, .092	.038, .108	.070, .108
Score	10	.435	.954	.024, .044	.054, .165	.054, .165
Clopper-Pearson	10	.508	.984	.013, .025	.012, .025	.025, .039
Optimal ($\kappa=0.378$) ^d	10	.452	.964	.025 , .062	.024 , .081	.049, .081
Jeffreys	10	.433	.953	.043, .086	.032, .105	.073, .132
Score	15	.369	.953	.023, .041	.050, .164	.050, .164
Clopper-Pearson	15	.421	.980	.014, .025	.015, .025	.028, .049
Optimal ($\kappa=0.412$) ^d	15	.376	.959	.024 , .067	.025 , .086	.048, .093
Jeffreys	15	.366	.952	.028, .085	.031, .104	.053, .104
Score	25	.294	.953	.024, .038	.046, .163	.049, .163
Clopper-Pearson	25	.328	.975	.016, .025	.016, .025	.032, .050
Optimal ($\kappa=0.369$) ^d	25	.302	.959	.025 , .059	.023 , .078	.047, .078
Jeffreys	25	.292	.951	.033, .083	.028, .103	.059, .110
Score	50	.213	.952	.024, .034	.041, .162	.050, .162
Clopper-Pearson	50	.231	.969	.018, .025	.018, .025	.036, .049
Optimal ($\kappa=0.446$) ^d	50	.214	.953	.025 , .071	.025 , .092	.050, .092
Jeffreys	50	.212	.950	.026, .082	.027, .102	.052, .116
Score	100	.152	.951	.024, .031	.037, .162	.050, .162
Clopper-Pearson	100	.161	.965	.020, .025	.020, .025	.040, .050
Optimal ($\kappa=0.456$) ^d	100	.153	.952	.025 , .073	.025 , .094	.050, .094
Jeffreys	100	.152	.950	.026, .082	.026, .103	.051, .120

^a Average interval width and coverage across all p uniformly distributed on (0,1)

^b $\overline{\text{Max}(p_{low})}$ and $\overline{\text{Max}(p_{high})}$ are maximal region-specific average probabilities of interval missing low and missing high; $\text{Max}(p_{low})$ and $\text{Max}(p_{high})$ are maximal probabilities of missing low and high over all p in (0, 0.5), where “low” and “high” interchange for p in (0.5, 1)

^c $\overline{\text{Max}(p_{miss})}$ is maximal region-specific average lack of coverage rate; $\text{Max}(p_{miss})$ is maximal lack of coverage rate over all p in (0, 1)

^d Bold type illustrates meeting of specified coverage criteria for proposed approach

Table 2.Properties of proposed optimal interval under different grid definitions ($n=10$)^{a,b}

Grid definition	Optimal κ	Avg. Width ^c	Avg. Coverage ^c	Max($\overline{p_{low}}$), Max(p_{low}) ^d	Max($\overline{p_{high}}$), Max(p_{high}) ^d	Max($\overline{p_{miss}}$), Max(p_{miss}) ^e
$R=1, w=.5$	.427	.444	.960	.015 , .071	.025 , .090	.040, .090
$R=2, w=.25$	.400	.449	.962	.025 , .066	.024 , .085	.049, .085
$R=5, w=.10$	.378	.452	.964	.025 , .062	.024 , .081	.049, .081
$R=10, w=.05$	.254	.471	.972	.025 , .046	.021 , .059	.046, .072
$R=20, w=.025$	.142	.488	.978	.025 , .035	.024 , .042	.048, .052
$R=40, w=.0125$	.050	.501	.982	.025 , .028	.024 , .030	.046, .051

^a R = # of regions; w = width of each region^b Bold type illustrates meeting of specified coverage criteria for proposed approach^c Average interval width and coverage across all p uniformly distributed on (0,1)^d $\overline{p_{low}}$ and $\overline{p_{high}}$ are maximal region-specific average probabilities of interval missing low and missing high; p_{low} and p_{high} are maximal probabilities of missing low and high over all p in (0, 0.5), where “low” and “high” interchange for p in (0.5, 1)^e $\overline{p_{miss}}$ is maximal region-specific average lack of coverage rate; p_{miss} is maximal lack of coverage rate over all p in (0, 1)

Table 3.

Values of κ_{yL} and κ_{yU} in (6) to match 95% Blaker (2000) interval for $n=10$

y	Blaker Interval ^a		Generalized credible interval in (6)	
	LL_y	UL_y	κ_{yL}	κ_{yU}
0	0	0.283	$_{--}b$	0.156
1	0.005	0.444	0.20	0.012
2	0.037	0.556	0.32	0.005
3	0.087	0.619	0.40	0.36
4	0.150	0.717	0.45	0.25
5	0.222	0.778	0.48	0.48
6	$1 - UL_4$	$1 - LL_4$	0.25	0.45
7	$1 - UL_3$	$1 - LL_3$	0.36	0.40
8	$1 - UL_2$	$1 - LL_2$	0.005	0.32
9	$1 - UL_1$	$1 - LL_1$	0.012	0.20
10	$1 - UL_0$	$1 - LL_0$	0.156	$_{--}b$

^aInterval limits as reported in Agresti and Min (2001); also available in SAS *Freq* Procedure (SAS Institute Inc., 2015)

^b $_{LL}$ set to 0 if $y=0$; $_{UL}$ set to 1 if $y=10$