Received: 29 August 2018 | Revised: 18 March 2019 | Accepted: 10 April 2019

DOI: 10.1002/sim.8189

RESEARCH ARTICLE

WILEY Statistics in Medicine

Ensemble confidence intervals for binomial proportions

Hayeon Park | Lawrence M. Leemis

Department of Mathematics, The College of William & Mary, Williamsburg, Virginia

Correspondence
Lawrence M. Leemis, Department of Mathematics, The College of William & Mary, Williamsburg, VA 23187.
Email: leemis@math.wm.edu

We propose two measures of performance for a confidence interval for a binomial proportion p: the root mean squared error and the mean absolute deviation. We also devise a confidence interval for p based on the actual coverage function that combines several existing approximate confidence intervals. This "Ensemble" confidence interval has improved statistical properties over the constituent confidence intervals. Software in an R package, which can be used in devising and assessing these confidence intervals, is available on CRAN.

KEYWORDS

binomial distribution, confidence interval, coverage, statistical computing

1 | INTRODUCTION

Calculating point and interval estimators for the binomial proportion occurs in many applications, including Monte Carlo simulation, survey sampling, and survival analysis. Assume that n independent binary responses have been collected, with x successes, where success is broadly defined, and the sample size n is a fixed, positive integer constant. The maximum likelihood estimator for the binomial proportion p is the fraction of successes in the sample, $\hat{p} = x/n$, which is an intuitive, unbiased, and consistent estimator of p.

The existing approximate confidence interval procedures for p typically have two significant shortcomings: they do not perform well in terms of coverage (a) for small sample sizes and (b) near the extremes, that is, near $p = 0$ and $p = 1$.

Two measures of performance are developed here that can be used to assess the effectiveness of these confidence intervals. We also suggest combining the following five approximate confidence intervals based on their actual coverage in order to achieve an Ensemble approximate confidence interval for p whose coverage is closer to the stated coverage than the constituent confidence intervals: Clopper-Pearson, Wilson-score, Jeffreys, Agresti-Coull, and arcsine transformation. Meeker et al[1] devote a chapter to overviewing methods for constructing confidence intervals for binomial proportions. The five confidence intervals are presented in the next section. Section 3 contains plots of the actual coverage for these confidence interval procedures and defines the two measures of performance. Section 4 presents some graphics associated with the Clopper-Pearson confidence interval and compares this conservative confidence interval with the Blaker confidence interval. Section 5 contains an algorithm for calculating an Ensemble confidence interval from several constituent confidence intervals and evaluates its statistical properties. Section 6 illustrates the use of the Ensemble confidence interval in an application, and Section 7 contains conclusions.

2 | CONFIDENCE INTERVALS

Five confidence intervals have been selected in this section for use in the calculation of an Ensemble confidence interval whose statistical properties are better than the constituent confidence intervals. In all cases, we consider a two-sided confidence interval for p with stated coverage $1 - \alpha$. One-sided confidence intervals can be constructed from the two-sided confidence intervals in the usual fashion. For most of the existing two-sided confidence interval procedures for p, the

probability that the confidence interval misses the true parameter high does not equal the probability that the confidence interval misses low. Hence, a separate construction of the Ensemble confidence intervals described here must be developed for one-sided confidence intervals (see the work of Pradhan et al[2] for details). The criteria for selecting the constituent confidence interval procedures considered here are as follows:

- For a fixed sample size n, the confidence interval should be complementary for any particular x and $n - x$ successes.
- The confidence interval should be asymptotically exact (which is defined subsequently) for $0 < p < 1$.
- The confidence interval should not degenerate to a confidence interval of width zero for $x = 0$ or $x = n$.

The third criterion eliminates the well-known Wald confidence interval for consideration. The Wald confidence interval is based on the normal approximation to the binomial distribution

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},$$

where $z_{\alpha/2}$ is the $1 - \alpha/2$ percentile of the standard normal distribution. When $x = 0$ ($x = n$), the point estimate for p is $\hat{p} = 0$ ($\hat{p} = 1$), so the Wald confidence interval degenerates to a zero-width confidence interval at the extremes. When a confidence interval bound falls outside of $(0, 1)$, the bound is typically set to 0 or 1. The next five paragraphs briefly describe the five confidence intervals for p that will be combined subsequently to establish an Ensemble confidence interval.

The $100(1 - \alpha)\%$ Clopper-Pearson confidence interval for the binomial proportion p given by Clopper and Pearson[3] can be expressed as the quantiles of beta distributions

$$B_{x,n-x+1,1-\alpha/2} < p < B_{x+1,n-x,\alpha/2},$$

for $x = 0, 1, 2, \ldots, n$, where the first two values in the subscripts are the parameters of the beta distribution and the third value in the subscript is a right-hand tail probability. The Clopper-Pearson confidence interval bounds can also be written as functions of percentiles of the F distribution as shown by Leemis and Trivedi.[4]

The bounds on the Wilson-score $100(1 - \alpha)\%$ confidence interval for p are[5]

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

where $z_{\alpha/2}$ is the $1 - \alpha/2$ percentile of the standard normal distribution. The center of the Wilson-score confidence interval is

$$\frac{\hat{p} + z_{\alpha/2}^2/(2n)}{1+z_{\alpha/2}^2/n},$$

which is a weighted average of the point estimator $\hat{p} = x/n$ and $1/2$, with more weight on $\hat{p}$ as n increases.

The Jeffreys $100(1 - \alpha)\%$ confidence interval for p is a Bayesian credible interval that uses a Jeffreys noninformative prior distribution for p. As was the case with the Clopper-Pearson confidence interval, the bounds of the Jeffreys confidence interval for p are percentiles of a beta random variable[6]

$$B_{x+1/2,n-x+1/2,1-\alpha/2} < p < B_{x+1/2,n-x+1/2,\alpha/2},$$

for $x = 1, 2, \ldots, n - 1$. When $x = 0$, the lower bound is set to zero and the upper bound calculated using the formula above; when $x = n$, the upper bound is set to one and the lower bound calculated using the formula above.

The bounds of the Agresti-Coull $100(1 - \alpha)\%$ confidence interval, which was originally developed to approximate the Wilson-score confidence interval, are[7]

$$\tilde{p} \pm z_{\alpha/2}\sqrt{\frac{\tilde{p}(1-\tilde{p})}{\tilde{n}}},$$

where $\tilde{n} = n + z_{\alpha/2}^2$ and $\tilde{p} = (x + z_{\alpha/2}^2/2)/\tilde{n}$. In the special case of $\alpha = 0.05$, if one is willing to round $z_{\alpha/2} = 1.96$ to 2, this interval can be interpreted as "add two successes and add two failures and use the Wald confidence interval formula."

The arcsine transformation uses a variance-stabilizing transformation when constructing a confidence interval for p. Using a modification suggested by Anscombe,[8] the bounds on a $100(1 - \alpha)\%$ confidence interval for p are[9]

$$\sin^2\left(\arcsin\left(\sqrt{\tilde{p}}\right) \pm \frac{z_{\alpha/2}}{2\sqrt{n}}\right),$$

TABLE 1 Approximate 95% confidence intervals for p for $n = 10$ and $x = 3$

Interval Name	Confidence Interval
Wald	$0.0160 < p < 0.584$
Clopper-Pearson	$0.0667 < p < 0.652$
Wilson-score	$0.108 < p < 0.603$
Jeffreys	$0.0927 < p < 0.606$
Agresti-Coull	$0.103 < p < 0.608$
Arcsine	$0.0790 < p < 0.618$

where $\tilde{p} = (x + 3/8)(n + 3/4)$. In the rare cases in which a confidence interval does not include the point estimator, one of the bounds is adjusted to include the point estimator.

These are not the only confidence intervals for p. Some of these confidence intervals have variations that include a continuity correction. In addition, there are other intervals, such as the logit interval, which could have been included in our Ensemble confidence interval procedure. We consider only the five intervals described above when constructing the Ensemble confidence interval procedure developed subsequently.

Example 1. When $n = 10, x = 3$ and $\alpha = 0.05$, the point estimate for p is $\hat{p} = 0.3$. The Wald, Clopper-Pearson, Wilson-score, Jeffreys, Agresti-Coull, and arcsine transformation 95% confidence intervals can be calculated with the following R commands after installing and loading the `conf` package.

```
binomTest(n = 10, x = 3, intervalType = "Wald")
binomTest(n = 10, x = 3, intervalType = "Clopper-Pearson")
binomTest(n = 10, x = 3, intervalType = "Wilson-Score")
binomTest(n = 10, x = 3, intervalType = "Jeffreys")
binomTest(n = 10, x = 3, intervalType = "Agresti-Coull")
binomTest(n = 10, x = 3, intervalType = "Arcsine")
```

The 95% confidence intervals for p are given in Table 1. The confidence interval bounds vary significantly between confidence interval procedures. The Clopper-Pearson confidence interval is the widest of the six; the Wilson-score confidence interval is the narrowest of the six. Only the Wald interval is symmetric about the maximum likelihood estimator $\hat{p} = 0.3$.

3 | COMPUTING ACTUAL COVERAGE

The actual coverage $c(p)$ of a confidence interval for the binomial proportion is[10]

$$c(p) = \sum_{x=0}^{n} I(x, p) \binom{n}{x} p^x (1-p)^{n-x} \qquad 0 < p < 1,$$

where $I(x, p)$ is an indicator function that denotes whether a confidence interval includes the binomial proportion p when the number of successes $X = x$. In this paper, we use the following terms to describe the performance, in terms of actual coverage, of a confidence interval for a binomial proportion p.

- A confidence interval is *exact* if its actual coverage equals its stated (or nominal) coverage $1 - \alpha$ for all values of n and p, that is, $c(p) = 1 - \alpha$ for $n = 1, 2, \ldots$, and $0 < p < 1$.
- A confidence interval is *asymptotically exact* if $\lim_{n\to\infty} c(p) = 1 - \alpha$ for all $0 < p < 1$.
- A confidence interval is *approximate* if it is not exact.
- A confidence interval is *conservative* if $c(p) \geq 1 - \alpha$ for all values of n and all $0 < p < 1$.

There are no exact confidence interval procedures for the binomial proportion p from a random sample of binary data values. Section 4 will show why this must be the case.

The formula for $c(p)$ is applied to the six confidence intervals described in the previous section with $n = 10$ and $\alpha = 0.05$ with axis ranges $0 < p < 1$ and $0.9 < c(p) < 1.0$. The actual coverage for the Wald interval, for example, is plotted with the R statement (using the `conf` package)

```
binomTestCoveragePlot(n = 10, intervalType = "Wald")
```

which results in the upper-left-hand plot given in Figure 1. In all six plots, a horizontal line is drawn at the stated coverage of 0.95.

Some general conclusions that can be drawn from the six plots in Figure 1 are given below.

- The actual coverage function, $c(p)$, is symmetric about $p = 1/2$ for all six confidence interval procedures.
- The actual coverage degrades on all confidence interval procedures near the extremes at $p = 0$ and $p = 1$.

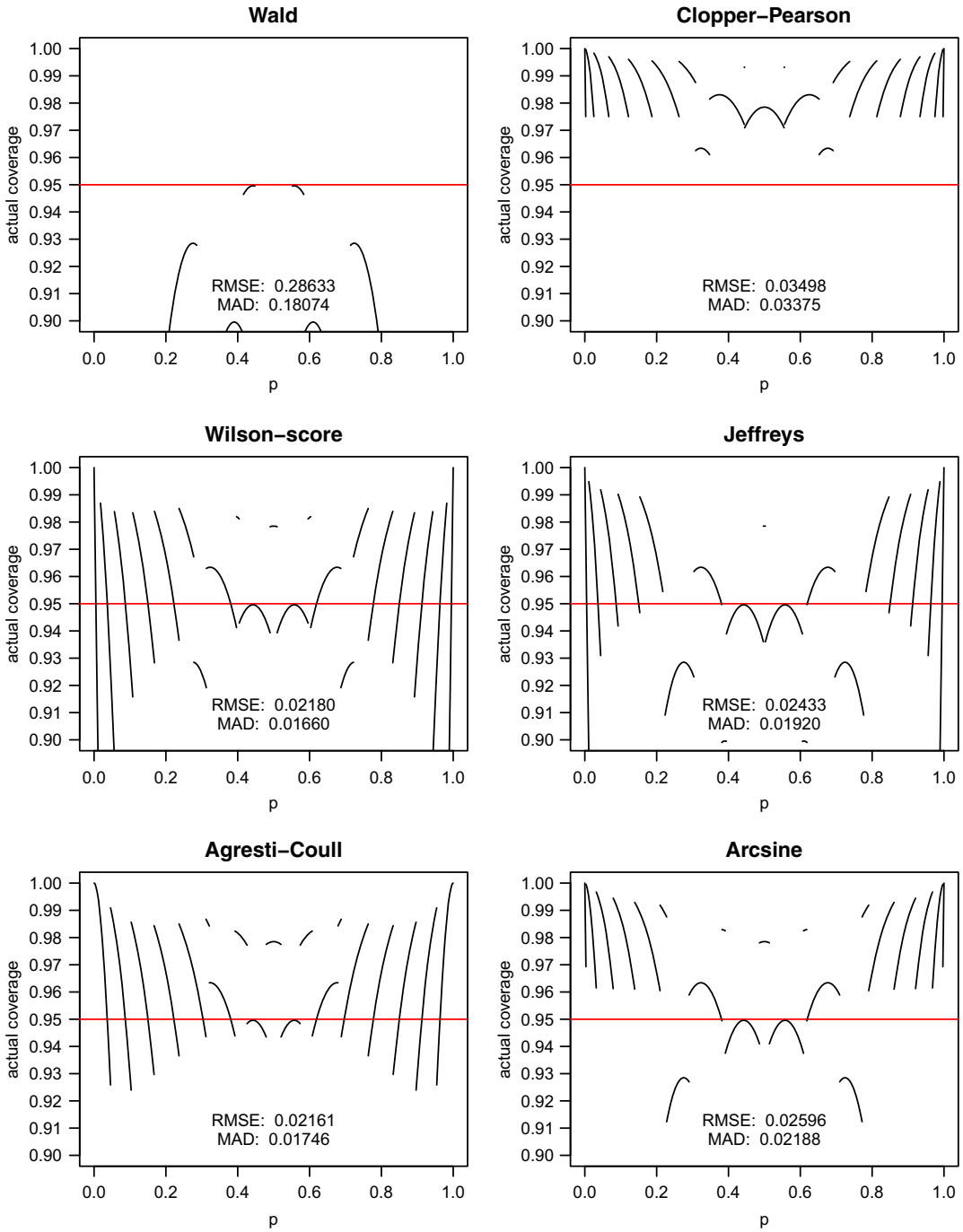

FIGURE 1 Actual coverage for six approximate 95% confidence intervals for $n = 10$. MAD, mean absolute deviation; RMSE, root mean squared error [Colour figure can be viewed at wileyonlinelibrary.com]

- The actual coverage of the Wald confidence interval differs most from the associated stated coverage among the six confidence interval procedures. See Brown et al[6] for a thorough analysis of this poor performance for other values of n.
- The Clopper-Pearson confidence interval is popular with statisticians because its actual coverage is always greater than or equal to its stated coverage. In other words, this confidence interval is always wider than it should be. It will never claim more precision on p than it should. Figure 1 shows that the Clopper-Pearson confidence interval is clearly not an exact confidence interval, but is rather a conservative confidence interval.

When the six actual coverage functions $c(p)$ in Figure 1 are plotted for larger values of n, it is apparent that the actual coverage functions approach the stated coverage as the sample size n increases, that is, these confidence intervals are asymptotically exact.

It is difficult to compare these plots of the actual coverage visually for the various confidence interval procedures. We define two measures here: the root mean squared error (RMSE) and the mean absolute deviation (MAD). The mean actual coverage m for a confidence interval procedure is the average value of the actual coverage for a fixed sample size n:

$$m = \int_0^1 c(p)\,dp.$$

The ideal case, of course, is $m = 1 - \alpha$. It is also important to measure how far the actual coverage strays from $1 - \alpha$. The variance of the actual coverage v is defined as

$$v = \int_0^1 c^2(p)\,dp - m^2.$$

The ideal case, assuming that $m = 1 - \alpha$, is $v = 0$. The two measures of performance can be combined into a single measure by devising a measure that is similar to the root mean squared error (which is the square root of the variance plus the squared bias):

$$\text{RMSE} = \sqrt{v + (m - (1 - \alpha))^2}.$$

The ideal case for the RMSE measure is zero. Such an RMSE would correspond to an exact confidence interval. A second measure of the performance of a confidence interval procedure for p based on its actual coverage function is the mean absolute deviation, which is defined as

$$\text{MAD} = \int_0^1 |c(p) - (1 - \alpha)|\,dp.$$

The ideal case for the MAD measure is also zero. Such a MAD would correspond to an exact confidence interval. The RMSE and MAD measures are given in Figure 1 for each of the confidence interval procedures. The Agresti-Coull confidence interval has the smallest RMSE, and the Wilson-score confidence interval has the smallest MAD for $n = 10$.

The RMSE and MAD are plotted for the five confidence intervals for sample sizes $n = 1, 2, \ldots, 10$ in Figures 2 (RMSE) and 3 (MAD). These measures put the Clopper-Pearson confidence interval at a disadvantage because it is a conservative interval in the sense that $c(p) \geq 1 - \alpha$ for all values of p, which inflates both RMSE and MAD. A more appropriate measure of performance for a conservative confidence interval is to minimize the average value of the actual coverage m. When the constituent confidence interval procedures were evaluated individually for larger values of n, we found that the Wilson-score interval was superior to the others in terms of both measures of performance.

The six confidence interval procedures are not the only options for a confidence interval for p. Pires and Amado[11] compare twenty confidence interval procedures. Blaker[12] defines a conservative confidence interval for p that has an improved $c(p)$ function over the Clopper-Pearson confidence interval. This confidence interval procedure is based on the acceptability function given in Spjøtvoll[13] which possess a nesting property: if $\alpha < \alpha'$, then the $1 - \alpha'$ confidence interval is contained in the associated $1 - \alpha$ confidence interval for fixed values of x and n. Schilling and Doi[14] developed confidence intervals that minimize the discontinuities in the $c(p)$ functions shown in Figure 1. Lang[15] has devised mean-minimum confidence intervals that guarantee that the mean and minimum coverage never drop below prescribed values. Copas[16] considered a Bayesian confidence interval for p. Wang and Zhang[17] use the asymptotic infimum actual coverage probability as a criteria for constructing confidence intervals. Blyth and Still[18] and Blyth[19] also consider the actual coverage in evaluating confidence intervals for p.

The Clopper-Pearson confidence interval is often chosen because it is a well-known conservative confidence interval. As indicated earlier, conservative confidence intervals perform poorly in terms of the RMSE and MAD criteria because $c(p) \geq 1 - \alpha$ for $0 < p < 1$. Because of this, we present some graphics associated with the Clopper-Pearson confidence interval and compare it to the Blaker confidence interval in the next section before developing the Ensemble confidence interval. Many of the graphics presented in the next section also apply to any of the other confidence intervals.

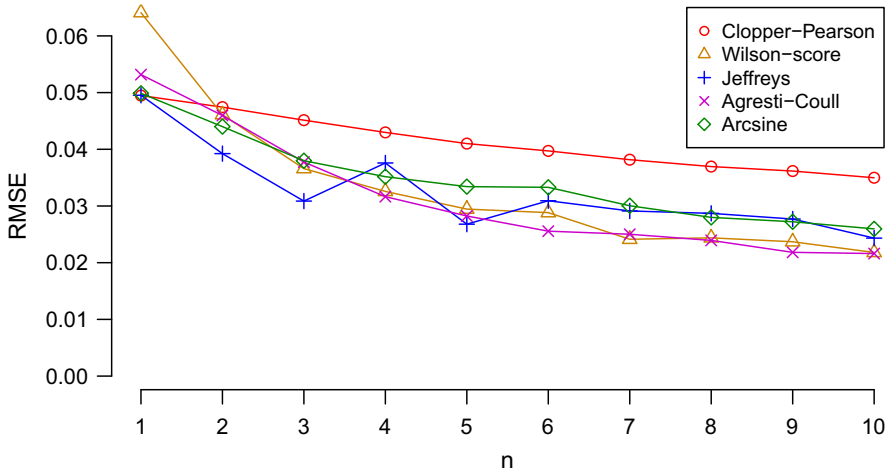

FIGURE 2 Root mean squared error (RMSE) for five 95% approximate confidence intervals [Colour figure can be viewed at wileyonlinelibrary.com]

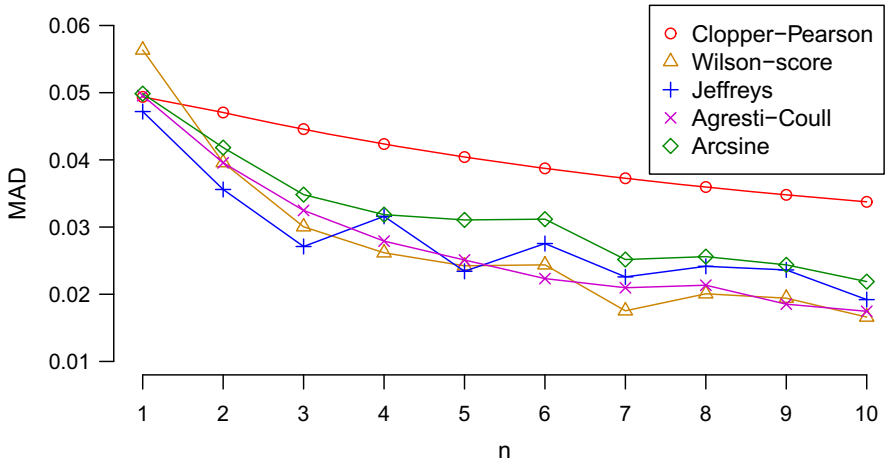

FIGURE 3 Mean absolute deviation (MAD) for five 95% approximate confidence intervals [Colour figure can be viewed at wileyonlinelibrary.com]

4 | CLOPPER-PEARSON CONFIDENCE INTERVAL GRAPHICS

The defining formula for the actual coverage function $c(p)$ and the fact that the lower bounds and upper bounds on *any* confidence interval procedure for the binomial proportion p are nondecreasing functions of x means that the actual coverage $c(p)$ must lie on one of the *acceptance curves* defined as

$$b(p, x_0, x_1) = \sum_{x=x_0}^{x_1} \binom{n}{x} p^x (1-p)^{n-x}$$

for a prescribed value of p, for $0 < p < 1$ and for integers x_0 and x_1 satisfying $0 \le x_0 \le x_1 \le n$. (Formulas for m and v based on this observation which avoid numerical integration are given in the appendix.) These acceptance curves are graphed in Figure 4 for $n = 10$. The acceptance curves were also given by Schilling and Doi[14] and Blaker,[12] where they are called "shortest acceptance regions." Wang[20] showed that $b(p, x_0, x_1)$ on $0 < p < 1$ is (a) a decreasing function of p when $x_0 = 0$ and $0 \le x_1 < n$, (b) an increasing function of p when $x_1 = n$ and $0 < x_0 \le n$, (c) equal to 1 when $x_0 = 0$ and $x_n = n$ since all binomial probability mass functions sum to 1, and (d) a unimodal function that achieves a maximum at

$$p = \left[1 + \left(\frac{\binom{n}{x_1}(n - x_1)}{\binom{n}{x_0} x_0} \right)^{1/(x_1 - x_0 - 1)} \right]^{-1}$$

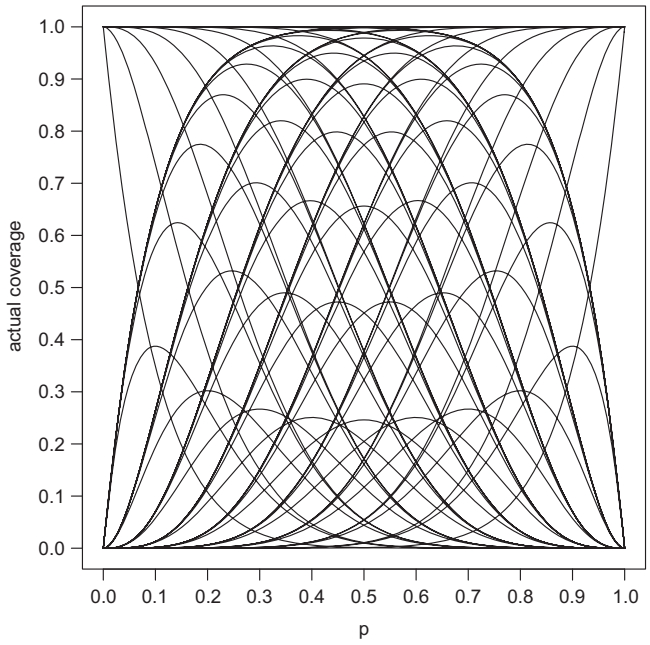

FIGURE 4 Actual coverage acceptance curves for $n = 10$

when $0 < x_0 \leq x_1 < n$. Since the actual coverage function for all confidence interval procedures must lie on one of these curves, there will never be an exact confidence interval for p.

The scale on the vertical axis in Figure 5 is altered to range from 0.94 to 1. The acceptance curves from Figure 4 for $n = 10$ are plotted in gray. The actual coverage function $c(p)$ for a 95% Clopper-Pearson confidence interval is given by black lines. A solid horizontal line at 0.95 marks the stated coverage of a 95% confidence interval. As indicated in the previous paragraph, the actual coverage function $c(p)$ for the unknown binomial proportion p must lie on one of these acceptance curves for one particular value of p. The Clopper-Pearson confidence interval is conservative because $c(p) \geq 1 - \alpha$ for all p. Unlike the Clopper-Pearson approach, the other confidence intervals considered thus far have

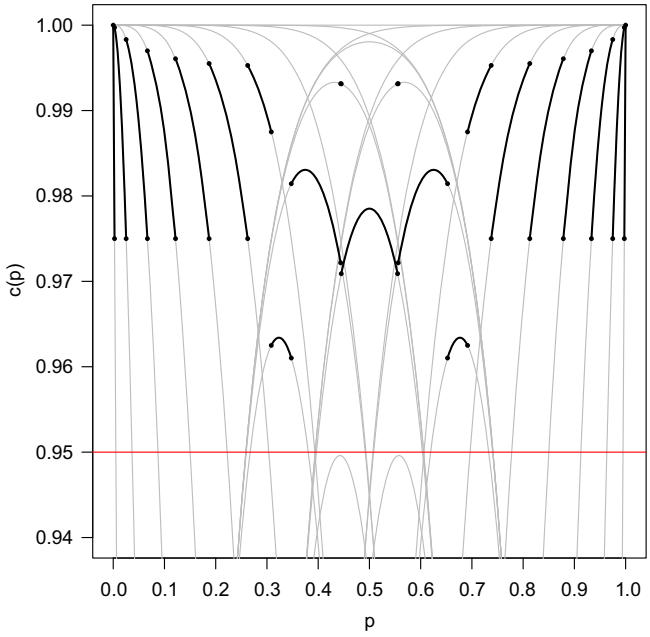

FIGURE 5 Clopper-Pearson actual coverage and acceptance curves for $n = 10$ and $\alpha = 0.05$ [Colour figure can be viewed at wileyonlinelibrary.com]

an actual coverage that falls below $1 - \alpha$ for some values of p, which means that they could potentially give narrow confidence intervals that claim more precision than they should.

One key observation from Figure 5 is that the values of p associated with the discontinuities in the actual coverage function are the confidence interval bounds. In general, there are $2n + 1$ segments in the actual coverage for the Clopper-Pearson confidence interval, so there are $(2)(10) + 1 = 21$ segments in Figure 5. This means that there are $2n + 2 = (2)(10) + 2 = 22$ endpoints of these segments, and these 22 endpoints correspond to the lower and upper bounds associated with the $n + 1 = 11$ confidence intervals for $x = 0, 1, \ldots, 10$ successes in the binomial random experiment. A second key observation from Figure 5 is that the discontinuities in $c(p)$ are a result of either an increase in either x_0 or x_1 in $b(p, x_0, x_1)$. If x_0 is increased, the discontinuity is associated with an upper confidence interval bound; if x_1 is increased, the discontinuity is associated with a lower confidence interval bound.

Table 2 illustrates the two observations from the previous paragraph for the 95% Clopper-Pearson confidence interval for $n = 10$. The first two rows give (x_0, x_1) pairs corresponding to the appropriate acceptance curves in Figure 5. The third row indicates whether the value of p associated with the *leftmost* endpoint of a segment in Figure 5 corresponds to a lower bound p_L or an upper bound p_U. The fourth row gives the value of p associated with the leftmost endpoint of the segment in Figure 5. For example, the first segment of the actual coverage of the 95% Clopper-Pearson confidence interval corresponds to $x_0 = 0$ and $x_1 = 0$, which corresponds to the acceptance curve

$$b(p, 0, 0) = \sum_{x=0}^{0} \binom{10}{x} p^x (1-p)^{10-x} = (1-p)^{10}$$

for $0 < p < 1$. The leftmost endpoint of the first segment is 0, which corresponds to a lower confidence interval limit. Likewise, the second segment of the actual coverage of the 95% Clopper-Pearson confidence interval corresponds to $x_0 = 0$ and $x_1 = 1$, which corresponds to the acceptance curve

$$b(p, 0, 1) = \sum_{x=0}^{1} \binom{10}{x} p^x (1-p)^{10-x} = (1-p)^{10} + 10p(1-p)^9$$

for $0 < p < 1$. The leftmost endpoint of the second segment is 0.0025, which also corresponds to a lower confidence interval limit. Table 2 only considers confidence interval bounds between 0 and 0.5 because the rest of the confidence interval bounds are symmetric about $p = 0.5$, as illustrated in Figure 5.

Continuing in this fashion yields the 11 two-sided Clopper-Pearson 95% confidence intervals for p, the first three of which are given in Table 3.

Another way of visualizing the bounds of a Clopper-Pearson confidence interval was suggested by Kang and Schmeiser.[21] Their approach is applied for $n = 10$, $p = 0.3$, and $\alpha = 0.05$ in Figure 6. All (p_L, p_U) pairs must fall above the line $p_L = p_U$. This confidence interval scatterplot plots the eleven (p_L, p_U) pairs for $n = 10$ at the center of each circle, with the area associated with the circles surrounding each of the points proportional to the associated binomial probability. Any point that falls to the northwest of $(p, p) = (0.3, 0.3)$ covers the true parameter $p = 0.3$. Any point falling below

TABLE 2 Discontinuity points on 95% Clopper-Pearson actual coverage function for $n = 10$

x_0	0	0	0	0	0	0	0	1	1	1	2
x_1	0	1	2	3	4	5	6	6	7	8	8
p_L or p_U	p_L	p_L	p_L	p_L	p_L	p_L	p_L	p_U	p_L	p_L	p_U
Confidence limit	0	.0025	.0252	.0667	.1216	.1871	.2624	.3085	.3475	.4439	.4450

TABLE 3 The first three 95% Clopper-Pearson confidence interval bounds for $n = 10$

x	p_L	p_U
0	0.0000	0.3085
1	0.0025	0.4450
2	0.0252	0.5561

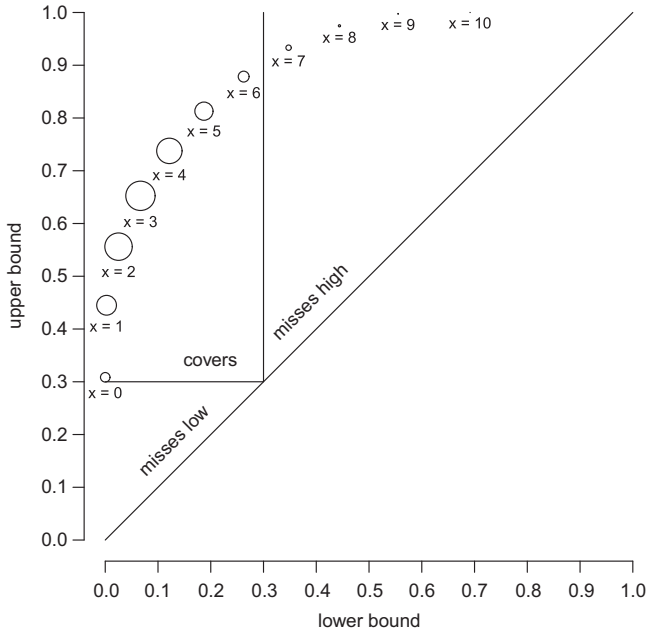

FIGURE 6 Lower and upper Clopper-Pearson confidence interval bounds for $n = 10$, $p = 0.3$, and $\alpha = 0.05$

$p_U = 0.3$ misses low, and any point falling to the right of $p_L = 0.3$ misses high. In this particular plot, the probabilities associated with these three outcomes are

$$P(\text{missing low}) = 0, \quad P(\text{covering } p) = 0.9894, \quad P(\text{missing high}) = 0.0106.$$

The probability that the Clopper-Pearson confidence interval covers $p = 0.3$ for this plot, $P(\text{covering } p) = 0.9894$, is the point $(0.3, c(0.3))$ in Figure 5. For the fixed values of $n = 10$ and $\alpha = 0.05$, Figures 5 and 6 relate to one another in the following fashion. As p varies in Figure 6, (a) the centers of the points remain in the same position, (b) the sizes of the points change according the binomial distribution probability mass function, and (c) the point (p, p) shifts along the line connecting $(0, 0)$ and $(1, 1)$. Changes along the continuous portions in Figure 5 correspond to (b); discontinuities in Figure 5 correspond to changes in (c), which result in points entering and exiting the region in Figure 6 labeled "covers." As p increases, points exiting the "misses high" region correspond to an increase in x_1, and points entering the "misses low" region correspond to an increase in x_0 in the $b(p, x_0, x_1)$ function.

The expected confidence interval width for a prescribed sample size n and stated coverage $1 - \alpha$ for any confidence interval for p can be written as[22]

$$E[W] = \sum_{x=0}^{n} (p_U - p_L) \binom{n}{x} p^x (1-p)^{n-x},$$

for $0 < p < 1$. A plot of the expected confidence interval width for 95% Clopper-Pearson confidence intervals for various sample sizes n is given in Figure 7. Not surprisingly, the confidence intervals narrow as n increases. Also, the confidence intervals narrow around the extremes.

The actual coverage of the Clopper-Pearson confidence interval performs worst when the sample size n is small or the value of p is near the extremes. A reasonable question to ask is: what p and n combinations provide an actual coverage that stays within some bounds of the stated coverage? Arbitrarily choosing the bound $0.95 \leq c(p) \leq 0.96$ for $\alpha = 0.05$, the values of p and n falling above the concave scatter in Figure 8 satisfy the constraint. The boundary of the region is not smooth because the pattern of the actual coverage function $c(p)$ varies significantly as n increases. The message of Figure 8 is that very large values of n and an assurance that p does not fall at an extreme value are necessary in order to use the Clopper-Pearson confidence interval procedure to obtain a confidence interval for p whose coverage is close to the stated coverage.

One of the shortcomings of the Clopper-Pearson confidence interval for p is that it can be unnecessarily wide, particularly for small n and extreme values of p. Blaker[12] devised a conservative confidence interval for p based on the acceptability function as a basis. Wang[23] devised a similar confidence interval by adjusting lower and upper confidence bounds to achieve a conservative interval. The conservative confidence interval given by Blyth and Still[18] is not compared

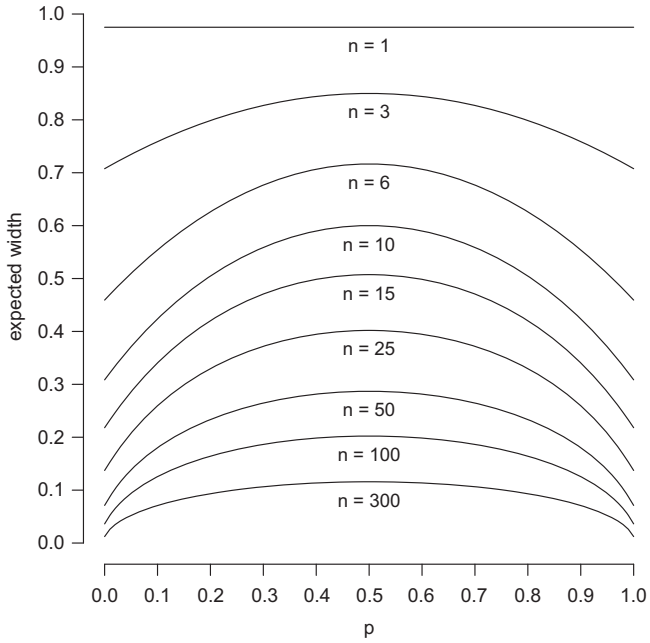

FIGURE 7 Expected Clopper-Pearson 95% confidence interval widths for various n

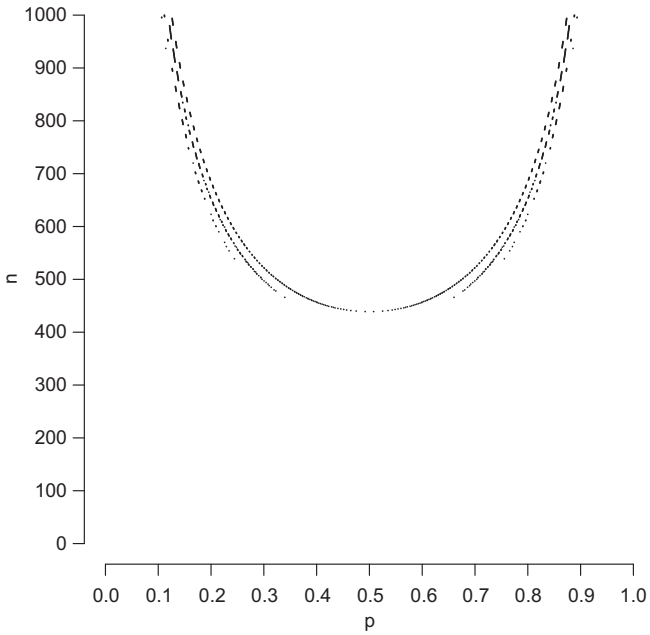

FIGURE 8 Values of n and p such that $0.95 \leq c(p) \leq 0.96$ for Clopper-Pearson 95% confidence intervals for p

to the Blaker confidence interval because the latter possesses the nesting property defined in Section 3. Agresti and Min[24] also compare conservative confidence intervals.

A direct comparison between the Clopper-Pearson and the Blaker confidence intervals is best made in terms of m, the mean actual coverage. Figure 9 shows the mean actual coverage m for $n = 1, 2, \ldots, 10$. The Blaker confidence interval outperforms the Clopper-Pearson for all values of n given in Figure 9, and continues to do so for larger values of n as well. Thus, we suggest that the Blaker confidence interval be used in place of the Clopper-Pearson confidence interval for all values of n when a confidence interval whose actual coverage cannot be less than $1 - \alpha$ is desired. It will produce narrower confidence intervals that keep the actual coverage above the stated coverage for all values of n.

Using m as a metric in Figure 9 assigns equal weight to all values of p in the interval $(0, 1)$. Applications will arise in which there are sound reasons to examine restricted values of p or to weight some values of p more heavily. Examples

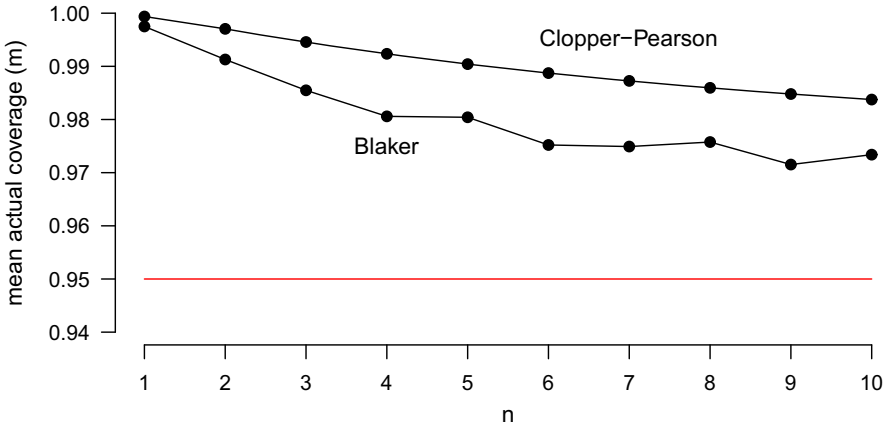

FIGURE 9 Values of m for Clopper-Pearson and Blaker 95% confidence intervals [Colour figure can be viewed at wileyonlinelibrary.com]

include restricting p to a subinterval (a, b) or weighting the values of p on $(0, 1)$ by a beta probability density function. In either case, the values plotted in Figure 9 will adjust to account for the modified metric.

In general, the goal of a confidence interval procedure is to come as close to the stated coverage as possible. One possible way to improve on the RMSE and MAD measures of performance is to combine the five confidence intervals presented in Section 2 in order to form an "Ensemble" confidence interval with potentially superior statistical properties to the constituent confidence intervals. The next section addresses one way to devise an Ensemble confidence interval.

5 | ENSEMBLE CONFIDENCE INTERVAL

In some applications, the goal of constructing a confidence interval for a binomial proportion is to attain an actual coverage that is as close as possible to the stated coverage. We propose a technique here that combines the five confidence interval procedures described in Section 2 (the Wald confidence interval is omitted because it degenerates for $x = 0$ and $x = n$) based on their actual coverage of the associated confidence interval at $\hat{p}$. The five constituent confidence interval procedures were chosen because they are likely to be a part of most statistical packages. We return to the example of $n = 10$, $x = 3$, and $\alpha = 0.05$. Table 4 augments Table 1 to include an extra column that gives the actual coverage of the confidence interval procedures at $\hat{p} = 3/10$, that is, $c(\hat{p})$. Three of the actual coverages fall above the stated coverage of 0.95, and two of the actual coverages fall below the stated coverage. The actual coverage associated with the Clopper-Pearson confidence interval for $n = 10$ and $x = 3$ at $\hat{p} = 3/10$, for example, is calculated with the R command

```
binomTestCoverage(n = 10, p = 0.3, intervalType = "Clopper-Pearson"),
```

which uses the `binomTestCoverage` function from the `conf` package.

Figure 10 contains a plot of the five lower bounds on the left and the five upper bounds on the right associated with $n = 10, x = 3$, and $p = 3/10$. The horizontal axis contains the (p_L, p_U) pairs, and the heights of the points plotted are the associated $c(\hat{p})$ values. The points are labeled CP (Clopper-Pearson), WS (Wilson-score), JF (Jeffreys), AC (Agresti-Coull), and AR (Arcsine). Based on the actual coverage, we know that the widest of the confidence intervals, the Clopper-Pearson confidence interval, is too wide, and the narrowest of the confidence intervals, the Wilson-score confidence interval, is

TABLE 4 Approximate 95% confidence intervals for p and actual coverage at $\hat{p}$ for $n = 10$ and $x = 3$

Interval Name	Confidence Interval	$c(\hat{p})$
Clopper-Pearson	$0.0667 < p < 0.652$	0.9894
Wilson-score	$0.108 < p < 0.603$	0.9244
Jeffreys	$0.0927 < p < 0.606$	0.9244
Agresti-Coull	$0.103 < p < 0.608$	0.9527
Arcsine	$0.0790 < p < 0.618$	0.9611

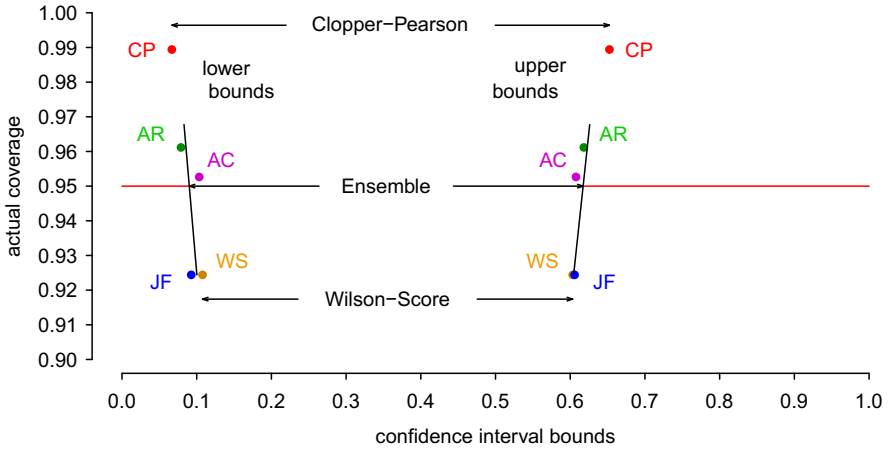

FIGURE 10 Geometry behind an Ensemble 95% confidence interval for p for $n = 10, x = 3$ [Colour figure can be viewed at wileyonlinelibrary.com]

too narrow at $p = 0.3$. Our procedure is to connect the centroid of the lower limits whose actual coverage falls above the stated coverage with the centroid of the lower limits whose actual coverage falls below the stated coverage. In the case of the lower bounds, the centroid of the lower bounds falling below the stated coverage is at (0.100, 0.924), and the centroid of the lower bounds falling above the stated coverage is at (0.0830, 0.968). A similar procedure is applied to the upper bounds of the confidence intervals. In the case of the upper bounds, the centroid of the upper bounds falling below the stated coverage is at (0.605, 0.924), and the centroid of the upper bounds falling above the stated coverage is at (0.626, 0.968). The centroids are connected by the line segments shown in Figure 10. We considered using linear regression rather than the method illustrated in Figure 10 involving centroids, but the centroid approach proved to be more stable. When actual coverages are similar for the confidence interval procedures, we found that simple linear regression could potentially give confidence interval bounds that strayed significantly from the bounds of the constituent confidence intervals.

The five confidence intervals can be combined to form a new Ensemble confidence interval, which might perform better than each of the five constituent confidence intervals alone. We determined the intersection points of the two line segments connecting the two centroids with the stated coverage and used the associated p values as the confidence interval limits. For this particular case, the Ensemble 95% confidence interval is

$$0.0901 < p < 0.617.$$

This confidence interval can be determined with the R command

```
binomTestEnsemble(n = 10, x = 3)
```

which uses the `binomTestEnsemble` function from the `conf` package.

Connecting the centroids of the points below and above $1 - \alpha$ fails when all five of the points lie above $1 - \alpha$. In this case, the Ensemble estimator uses the maximum of the constituent p_L values and the minimum of the constituent p_U values.

All permutations of the five confidence intervals were tested for the RMSE and MAD measures of performance. All $2^5 - 1 = 31$ combinations of the five constituent confidence intervals were employed to determine the best confidence interval in terms of RMSE (Table 5) and MAD (Table 6). Bullet(s) in a particular row of these tables indicate the particular constituent confidence intervals that were combined to achieve the smallest value of RMSE or MAD. The pattern of choice between the constituent confidence intervals is consistent for the two measures of performance. Multiple methods are required in all cases to minimize the measures of performance. The RMSE and MAD values in Tables 5 and 6 generally decrease as n increases. This is consistent with $c(p)$ having more pieces, which means that $c(p)$ will generally lie closer to the stated coverage on $0 < p < 1$. The discrete nature of the binomial distribution accounts for the observation that the decrease is not monotone.

For larger sample sizes, the Wilson-score interval performs unusually well. It appears as part of the Ensemble confidence interval for every combination for $n \geq 5$. In addition, it outperforms all of the potential Ensemble confidence intervals alone for $n \geq 49$ (using the RMSE criterion) and $n \geq 27$ (using the MAD criterion).

TABLE 5 Best 95% Ensemble confidence intervals using root mean squared error (RMSE) as a measure of performance

n	CP	WS	JF	AC	AR	*m*	RMSE
1	•	•	•		•	0.991	0.0466
2		•	•		•	0.975	0.0389
3			•			0.963	0.0309
4			•	•		0.964	0.0280
5		•	•			0.957	0.0266
6		•		•		0.961	0.0243
7		•	•	•		0.961	0.0216
8		•	•	•		0.960	0.0222
9		•		•		0.961	0.0205
10		•		•		0.961	0.0201

Abbreviations: AC, Agresti-Coull; AR, Arcsine; CP, Clopper-Pearson; JF, Jeffreys; WS, Wilson-score.

TABLE 6 Best 95% Ensemble confidence intervals using mean absolute deviation (MAD) as a measure of performance

n	CP	WS	JF	AC	AR	*m*	MAD
1	•	•	•		•	0.991	0.0450
2		•	•		•	0.975	0.0354
3			•			0.963	0.0271
4			•	•		0.964	0.0240
5		•	•			0.957	0.0230
6		•		•		0.961	0.0212
7		•	•	•		0.961	0.0169
8		•	•	•		0.960	0.0194
9		•		•		0.961	0.0177
10		•		•		0.961	0.0163

Abbreviations: AC, Agresti-Coull; AR, Arcsine; CP, Clopper-Pearson; JF, Jeffreys; WS, Wilson-score.

6 | APPLICATION

Consider the nonparametric estimation of the survivor function associated with the $n = 7$ rat survival times (in days) from Efron and Tibshirani[25]:

$$16 \qquad 23 \qquad 38 \qquad 94 \qquad 99 \qquad 141 \qquad 197.$$

The empirical survivor function, which takes a downward step of $1/n = 1/7$ at each data value, is given by the solid lines in Figure 11. The dashed lines that denote the 95% confidence intervals associated with the survival probability at any time are calculated using the Ensemble confidence interval for the associated binomial probability. As seen in Tables 5 and 6, this will correspond to an Ensemble 95% confidence interval consisting of the Wilson-score, Jeffreys, and Agresti-Coull constituent confidence intervals. These intervals are superior to the usual intervals for the survival probability associated with Greenwood's formula, which collapses to Wald intervals in the case of uncensored data.

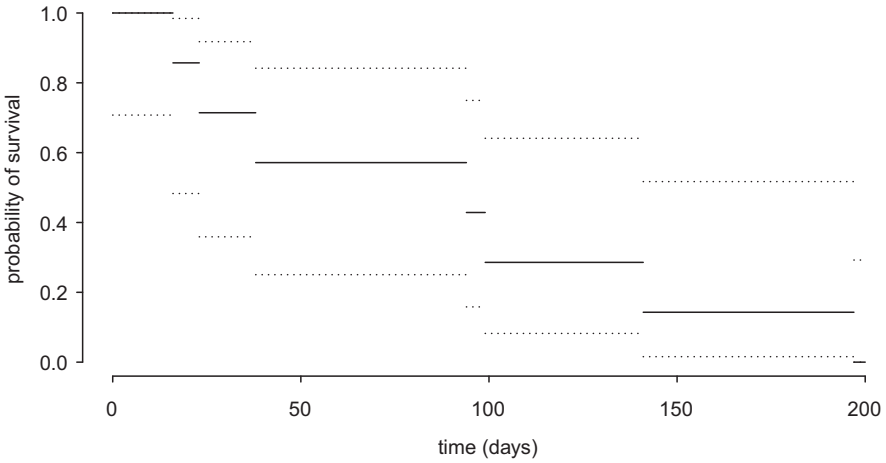

FIGURE 11 Survivor function estimate for rat survival data

7 | CONCLUSIONS

Two measures of performance for a confidence interval for a binomial proportion p have been developed here: the RMSE and MAD. Based on these two measures of performance, we draw the following conclusions in terms of which confidence interval to use for a binomial proportion.

- When the purpose of establishing a confidence interval is to select the best conservative confidence interval, we recommend using the Blaker confidence interval over the Clopper-Pearson confidence interval for all values of n (see Figure 9).
- When the purpose of establishing a confidence interval is to minimize the absolute difference between the stated and actual coverage, the preferred confidence interval depends on the sample size n.
  - The Wilson-score confidence interval should be used alone for $n \geq 49$ because this confidence interval alone minimizes both the RMSE and MAD over all other permutations of constituent confidence intervals in the Ensemble estimator. Although Brown et al[6] recommend the Agresti-Coull because it is easiest to "describe, remember, and compute" widely available software such as R make computing a Wilson-score confidence interval trivial. The Wilson-score, Jeffreys, Agresti-Coull, and arcsine transformation are all roughly comparable in terms of their actual coverage.
  - For smaller values of n, an Ensemble confidence interval that combines several existing confidence intervals should be considered. Specific optimal permutations of the constituent confidence intervals are given in Tables 5 and 6 for $n = 1, 2, \ldots, 10$.

R code is available in the `conf` package[26] for calculating the lower and upper limits of the confidence interval (`binomTest`), calculating the actual coverage (`binomTestCoverage`), plotting the actual coverage (`binomTestCoveragePlot`), and calculating Ensemble confidence intervals using different combinations of constituent distributions (`binomTestEnsemble`).

ACKNOWLEDGEMENTS

The authors thank the referees, Prof Roger Berger, Brennan Dolson, and Prof Chris Weld, for their helpful suggestions, which have improved this paper.

ORCID

Lawrence M. Leemis https://orcid.org/0000-0001-9071-985X

REFERENCES

1. Meeker WQ, Hahn GJ, Escobar LA. *Statistical Intervals: A Guide for Practitioners and Researchers*. 2nd ed. Hoboken, NJ: Wiley; 2017.
2. Pradhan V, Evans JC, Banerjee T. Binomial confidence intervals for testing non-inferiority or superiority: a practitioner's dilemma. *Stat Methods Med Res*. 2016;25(4):1707-1717.
3. Clopper CJ, Pearson ES. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*. 1934;26(4):404-413.
4. Leemis LM, Trivedi KS. A comparison of approximate interval estimators for the Bernoulli parameter. *Am Stat*. 1996;50(1):63-68.
5. Wilson EB. Probable inference, the law of succession, and statistical inference. *J Am Stat Assoc*. 1927;22(158):209-212.
6. Brown LD, Cai TT, DasGupta A. Interval estimation for a binomial proportion. *Statistical Science*. 2001;16(2):101-133.
7. Agresti A, Coull BA. Approximate is better than 'exact' for interval estimation of binomial proportions. *Am Stat*. 1998;52(2):119-126.
8. Anscombe FJ. On estimating binomial response relations. *Biometrika*. 1956;43(3/4):461-464.
9. Chen H. The accuracy of approximate intervals for a binomial parameter. *J Am Stat Assoc*. 1990;85(410):514-518.
10. Vollset SE. Confidence intervals for a binomial proportion. *Statist Med*. 1993;12(9):809-824.
11. Pires AM, Amado C. Interval estimators for a binomial proportion: comparison of twenty methods. *REVSTAT Stat J*. 2008;6(2):165-197.
12. Blaker H. Confidence curves and improved exact confidence intervals for discrete distributions. *Can J Stat*. 2000;28(4):783-798.
13. Spjøtvoll E. Preference functions. In: Bickel PJ, Doksum KA, Hodges Jr JL, eds. *A Festschrift for Erich L. Lehmann*. Belmont, CA: Wadsworth; 1983:409-432.
14. Schilling MF, Doi JA. A coverage probability approach to finding an optimal binomial confidence procedure. *Am Stat*. 2014;68(3):133-145.
15. Lang JB. Mean-minimum exact confidence intervals. *Am Stat*. 2017;71(4):354-368.
16. Copas JB. Exact confidence limits for binomial proportions—Brenner & Quan revisited. *J R Stat Soc Ser D Stat*. 1992;41(5):569-572.
17. Wang W, Zhang Z. Asymptotic infimum coverage probability for interval estimation of proportions. *Metrika*. 2014;77(5):635-646.
18. Blyth CR, Still HA. Binomial confidence intervals. *J Am Stat Assoc*. 1983;78(381):108-116.
19. Blyth CR. Approximate binomial confidence limits. *J Am Stat Assoc*. 1986;81(395):843-855.
20. Wang H. Exact confidence coefficients of confidence intervals for a binomial proportion. *Statistica Sinica*. 2007;17(1):361-368.
21. Kang K, Schmeiser B. Graphical methods for evaluating and comparing confidence-interval procedures. *Operations Research*. 1990;38(3):546-553.
22. Feng C. *On Confidence Intervals for Proportions With Focus on the U.S. National Health and Nutrition Examination Surveys* [master's thesis]. Burnaby, Canada: Simon Fraser University; 2006.
23. Wang W. An iterative construction of confidence intervals for a proportion. *Statistica Sinica*. 2014;24(3):1389-1410.
24. Agresti A, Min Y. On small-sample confidence intervals for parameters in discrete distributions. *Biometrics*. 2001;57(3):963-971.
25. Efron B, Tibshirani RJ. *An Introduction to the Bootstrap*. Boca Raton, FL: Chapman & Hall/CRC; 1993.
26. Weld C, Park H, Leemis L. conf: visualization and analysis of statistical measures of confidence. R package version 1.6. 2019. https://CRAN.R-project.org/package=conf

How to cite this article: Park H, Leemis LM. Ensemble confidence intervals for binomial proportions. *Statistics in Medicine*. 2019;38:3460–3475. https://doi.org/10.1002/sim.8189

APPENDIX

For a fixed sample size n, a confidence interval procedure for the binomial proportion p associated with $x = 0, 1, 2, \ldots, n$ successes results in $n + 1$ confidence intervals. Thus, there are $2n + 2$ associated confidence interval bounds. Let $p_1, p_2, \ldots, p_{2n+2}$ denote these ordered confidence interval bounds. These bounds correspond to the endpoints of the piecewise actual coverage function $c(p)$ defined in Section 3. Each of the $2n + 1$ pieces of $c(p)$ corresponds to a piece of one of the acceptance curves

$$b(p, x_0, x_1) = \sum_{x=x_0}^{x_1} \binom{n}{x} p^x (1-p)^{n-x}$$

defined in Section 4. Let x_{0i} and x_{1i} denote the lower and upper summation limits associated with the ith piece of $c(p)$, for $i = 1, 2, \ldots, 2n + 1$. Using this notation and the binomial theorem, an expression for the mean actual coverage that

avoids numerical integration is

$$\begin{aligned} m &= \int_0^1 c(p)dp \\ &= \sum_{i=1}^{2n+1} \int_{p_i}^{p_{i+1}} \sum_{x=x_{0i}}^{x_{1i}} \binom{n}{x} p^x (1-p)^{n-x} dp \\ &= \sum_{i=1}^{2n+1} \int_{p_i}^{p_{i+1}} \sum_{x=x_{0i}}^{x_{1i}} \left[\binom{n}{x} p^x \sum_{k=0}^{n-x} \binom{n-x}{k} (-p)^k \right] dp \\ &= \sum_{i=1}^{2n+1} \sum_{x=x_{0i}}^{x_{1i}} \int_{p_i}^{p_{i+1}} \binom{n}{x} \sum_{k=0}^{n-x} \binom{n-x}{k} (-1)^k p^{k+x} dp \\ &= \sum_{i=1}^{2n+1} \sum_{x=x_{0i}}^{x_{1i}} \binom{n}{x} \sum_{k=0}^{n-x} \binom{n-x}{k} (-1)^k \left[\frac{p_{i+1}^{k+x+1} - p_i^{k+x+1}}{k+x+1} \right]. \end{aligned}$$

Using a similar approach and again applying the binomial theorem, an expression for the variance of the actual coverage $v = \int_0^1 c^2(p)dp - m^2$ which avoids numerical integration is

$$v = \left\{ \sum_{i=1}^{2n+1} \sum_{x=x_{0i}}^{x_{1i}} \sum_{y=x_{0i}}^{x_{1i}} \binom{n}{x} \binom{n}{y} \sum_{k=0}^{2n-x-y} \binom{2n-x-y}{k} (-1)^k \left[\frac{p_{i+1}^{k+x+y+1} - p_i^{k+x+y+1}}{k+x+y+1} \right] \right\} - m^2.$$