Recommended Confidence Intervals for a Binomial Proportion

Authored by: Kyle Kolsti, PhD

6 August 2021

HSCoBP@afit.edu
937-255-3636 x 4768

The HS CoBP core mission involves leading a consortium of government, industry, and academic experts to assess future homeland threats to inform strategic plans across nine DHS T&E capability areas.

Homeland Security Community of Best Practices
2950 Hobson Way – Wright-Patterson AFB, OH 45433

Table of Contents

Executive Summary

This paper applies to a test or experiment where a proportion is to be estimated, such as probability of detection. The response must be binomial, meaning each trial must result in one of two outcomes. A confidence interval is an effective way to quantify the uncertainty when estimating the true proportion. Unfortunately, the literature provides many methods for computing these intervals and it can be difficult to discern which to use. This paper proposes a simple decision tool for selecting a method based on test objectives and risk tolerance. The paper also provides formulas for the recommended methods, analysis of the performance of the methods, and justifications for the recommendations put forth.

Keywords: binomial, proportion, confidence interval, Bayesian

Introduction

The problem at hand is how to report the uncertainty of an estimated binomial proportion in system Test and Evaluation (T&E). The word binomial indicates there are only two possible outcomes for each trial – generically, success or failure in the judgment of the tester. For example, testing a sensor may result in X detections out of n trials for an estimated probability of detection $\hat{p} = X/n$. The uncertainty in that estimate may be reported using a confidence interval. A confidence interval is defined to be a range of probabilities from a lower bound to an upper bound, $[l, u]$ where the probability of the interval covering the true probability p is equal to the user's desired confidence – typically 80%, 90%, or 95% – which is a risk-based choice.

The task of computing this interval is more complicated than it seems because of the discrete response. A test with n trials can produce only one of $n + 1$ outcomes: $X = 0, 1, \ldots, n - 1$, or n successes. Any selected method will therefore produce $n + 1$ unique intervals. The intervals are known in advance and the test results merely indicate which of the intervals to report. Changing X by one means the interval jumps from one location to another rather than smoothly adjusting. This phenomenon is the source of the difficulties; it makes it impossible for any method to consistently deliver on its stated confidence.

The problematic nature of intervals for a binomial proportion has led to substantial research and a wide variety of methods. The community agrees on some points, such as the inferiority of the most commonly seen method, but disagrees on other substantial matters. Appendix D contains a compilation of 23 methods & variants as well as significant articles on this subject. For in-depth background, the reader is referred to (Brown, Cai, & Dasgupta, Interval Estimation for a Binomial Proportion, 2001), who describe the history of these intervals, analyze many methods in detail, and include interesting responses from other researchers. That paper is a good launching point for learning more about the topic.

The objective of this paper is to recommend which of these methods to use in the context of government T&E limited to small sample sizes due to resource constraints. The recommendations are based on an extensive review of the literature, independent analysis of the methods, and experience

supporting Department of Defense (DOD) and Department of Homeland Security (DHS) acquisition programs. It is acknowledged that these recommendations are opinions open to debate. Testers are always encouraged to make their own informed judgments on which method is optimal for their purposes.

This paper is organized in an unconventional way to accommodate readers with different motivations. The recommendations are listed first with little context to immediately present the results of this investigation and save time for readers who are familiar with the subject. Readers who wish to dig deeper into the methods or understand the reasoning behind the recommendations may continue on to the subsequent sections.

Recommendations

Flowchart

Figure 1 structures the decision of picking a confidence interval into a series of four questions. The formulas and justification will follow in subsequent sections.

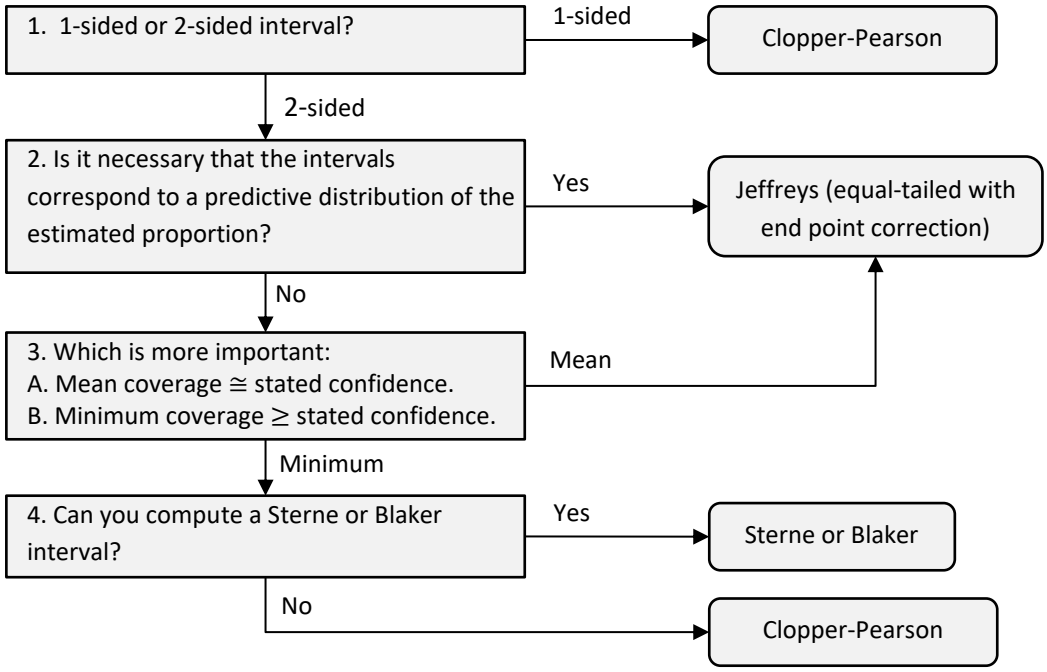

Figure 1: Recommendation Flowchart

Question #1 addresses the nature of the test objective or requirement: is the goal to estimate the range of values for the proportion (two-sided), or to estimate only an upper or a lower bound (one-sided)? Question #2 involves future uses of the interval. If the true proportion needs to be described by a statistical distribution – for example, to perform Monte Carlo simulations or to perform inference about

the proportion in addition to calculating an interval – the Jeffreys method provides that statistical distribution. Question #3 relates to risk: if the potential risk of making an incorrect conclusion based on the interval must be strictly controlled, as is often the case in government T&E, seek a method that guarantees that the minimum coverage meets or exceeds the stated confidence (i.e., Sterne, Blaker, or Clopper Pearson intervals). On the other hand, if the risk level may be managed on average (meaning the risk will be higher than anticipated in some tests), methods that tend to provide smaller intervals may be used (i.e., Jeffreys interval). Question #4 acknowledges that some methods are easier to compute than others. If unable to compute a Sterne or Blaker level, contact the HS CoBP for assistance before resorting to the Clopper-Pearson method.

Explanation and Justification

The flow chart begins with the assumption that there is not relevant information that can be incorporated into the uncertainty, such as previous test results. If this information does exist, a Bayesian approach using an informative prior may be preferred. That topic is outside the scope of this paper, but the reader is referred to (Albers, Kiers, & van Ravenzwaaij, 2018) for more information.

One-sided Interval

The first question is whether the interval must be one-sided or two-sided. A one-sided interval may be most appropriate when evaluating the performance against a less than or greater than requirement; for example, a situation where the lower limit l must exceed 90% for the system to pass. For one-sided intervals, the Clopper-Pearson interval is consistent with p-values generated from the exact binomial hypothesis test (Reiczigel, 2003; Wang, 2006).

The literature (and the remainder of this paper) is concerned almost entirely with two-sided intervals because they are more problematic.

Jeffreys Interval and Bayesian Inference

The Jeffreys interval is unique relative to the other methods addressed in this paper, as it is derived through a Bayesian formulation of the problem (Albers, Kiers, & van Ravenzwaaij, 2018). Bayesian methods produce a statistical distribution, called the posterior distribution or just "the posterior", instead of a traditional confidence interval. Sampling from the resulting statistical distribution can be useful for further inference and for conducting simulations. In addition, the interpretation of Bayesian results allows for more natural-sounding claims like "there is an 86% chance the system's probability of detection exceeds 90%", a statement which would be technically incorrect with the "frequentist" (meaning non-Bayesian) interval methods.

Bayesian methods start with what is known as a prior distribution (a.k.a "the prior"), representing the belief regarding the distribution of $\hat{p}$ before consideration of the data. Bayesian methods for intervals almost always use a beta distribution as a prior because they are mathematically convenient and flexible, in the sense that they can describe a wide range of beliefs regarding $\hat{p}$. In particular, the Jeffreys prior is a $\mathrm{Beta}(0.5,0.5)$ and the uniform prior is a $\mathrm{Beta}(1,1)$. Both of these priors are called "non-

informative" meaning they represent a belief that we have no particular expectations of the system performance. The Jeffreys prior is more prevalent in the literature and generally accepted as the preferred option. This paper will echo the literature by using the informal name "Jeffreys interval" as short for the Bayesian method using a $\mathrm{Beta}(0.5,0.5)$ conjugate prior to the binomial likelihood.

The relevance here is that the posterior distribution can be used to calculate the Bayesian counterpart of a confidence interval, called a credible interval. Two techniques are prevalent: the equal-tails method and the highest probability density "HPD" method (M'Lan, Joseph, & Wolfson, 2008). The equal-tail method ensures the area of the posterior distribution below the lower limit is the same as the area above the upper limit. The HPD method ensures that the probabilities of p are greater within the interval than anywhere outside the interval. Analysis for this paper supports previous conclusions that the HPD method does not improve performance, especially considering the additional computational effort required (Brown, Cai, & Dasgupta, Interval Estimation for a Binomial Proportion, 2001). Therefore, the equal-tail method is recommended.

End point corrections have been developed in the literature to fix undesirable interval behavior for extreme outcomes. These corrections are exceptions to the method's general procedure. These exceptions do improve the interval coverage, but they do not fix the posterior accordingly. The common one for the Jeffreys method is to apply different rules when $X = 0$ or $X = n$ as shown in Appendix A. The resulting inconsistency with additional inference from the posterior will probably be negligible: for $X = 0$ and $n = 5$, the lower limit of Eq. 3 is $l = 9.34e - 5$. In the unlikely event that p is in this gap, the coverage will be zero (no interval contains p) without the correction. The correction is also small enough that it should probably not be significantly inconsistent with inference using the posterior.

Brown et al (2001) apply an additional boundary correction to the Jeffreys method for $X = 1$ or $X = n - 1$. This correction is not recommended because it makes the mean coverage diverge from the confidence and further impinges on use of the posterior for other inference.

Optimal Coverage Objectives

The third question deals with optimal coverage. This is the most contentious issue seen in the reviewed literature. As stated in a rejoinder other authors' comments, "It seems that the primary source of disagreement is based on differences in interpretation of the coverage goals for confidence intervals." (Brown, Cai, & Dasgupta, Interval Estimation for a Binomial Proportion, 2001).

To illustrate the problem, consider a situation where a sample of size $n = 5$ is used to create a 95% confidence interval for p. There are six possible outcomes for this experiment ($x = 0, 1, 2, 3, 4, 5$), and thus six possible confidence intervals for each method, as shown in Figure 2. Suppose for the moment that the true (unknown) value of p is 0.5. Using the Jeffreys method as an example, outcomes $x = 1, 2, 3, 4$ produce confidence intervals that include $p = 0.5$. From the binomial distribution the probability of seeing any one of these four outcomes is 0.9375, which is notably below the nominal confidence of 0.95. In other words, if $n = 5$ and $p = 0.5$, the probability of $p = 0.5$ being within a 95% confidence interval is actually only 93.75%. This calculation can be carried out for all possible values of p (for fixed n and α) resulting in a coverage plot, as shown in Figure 2. Each vertical rise or drop of

coverage coincides with the lower or upper bound of an interval – at these points, the coverage takes a large step up or down with an infinitesimal change in p into or out of the interval. The shaded areas highlight the difference between the coverage and the intended confidence. These plots are symmetric about $p = 0.5$ but the full range of p is shown for illustrative purposes. Coverage plots for all recommended methods are shown in Appendix C for $n = 10$ and 80% and 95% confidence.

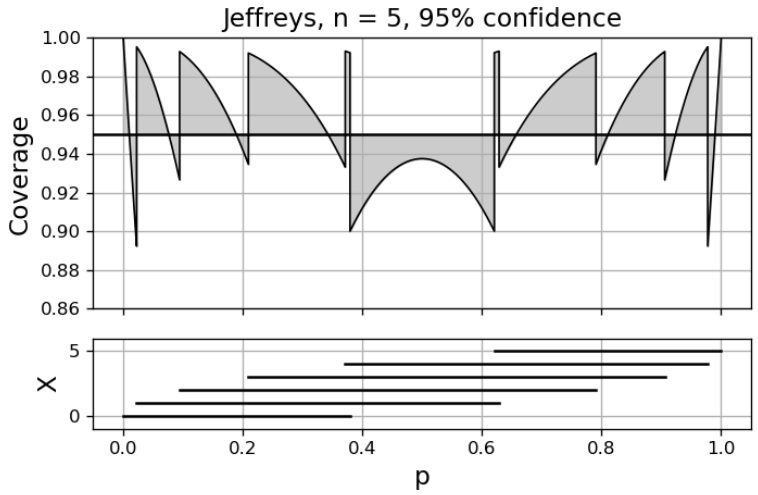

Figure 2: Example of a coverage plot (upper plot) along with the n+1 intervals (lower plot).

The problem is immediately apparent: is it accurate to claim that any of these intervals contains the true p with precisely 95% confidence? Technically, no, since coverage is a function of p. Two camps have evolved in the literature to deal with this quandary. Rather than pick a side, this paper's recommendations will accommodate both schools of thought by asking how much increased risk the decision maker is willing to tolerate.

The competing objectives of the two camps are minimum coverage versus mean coverage. Newcomb wrote "Choice of method must depend on an explicit decision whether to align minimum or mean coverage with $1 - a$" (Newcombe, 1998). In Figure 2 the mean coverage (the integral of the coverage from $p = 0$ to $p = 1$) is 95.8%, close to the confidence; however, for many potential values of p the coverage is below the confidence, with a minimum below 90% in two places. This method is suitable in one sense but not the other. The two constraints are mutually exclusive and are best addressed by two corresponding families of methods: exact and approximate.

Exact Methods

If it is important to the decision makers that the risk be managed at or below what is advertised, regardless of p, an exact method should be used. Exact methods are not exact in that they provide perfect results; rather, it means they are derived by inverting the equal-tailed binomial hypothesis test (Thulin, 2014). These methods are guaranteed to keep the coverage at or above a lower bound, namely the confidence, for all p (Newcombe, 1998) (Agresti & Min, On Small-Sample Confidence Intervals for Parameters in Discrete Distributions, 2001). This approach prevents unintended elevation of risk beyond

what was intended when selecting the confidence. This conservative coverage is gained at the price of increased confidence interval width (Thulin, 2014). The classic benchmark is the Clopper-Pearson method, which is well known and easy to calculate. It is also sometimes referred to as the "exact" method for obvious reasons, even though other exact methods exist.

There are exact methods that are less conservative than the Clopper-Pearson overall while still guaranteeing the minimum coverage; however, they are more difficult to compute and therefore less commonly encountered in practice. The method of Sterne assembles a subset of the intervals to be "active" for every given value of p (Sterne, 1954). Conceptually, for a given value of p, the binomial distribution provides the probability to observe each of the $n + 1$ intervals. The intervals are included one at a time, in order of decreasing probability, until the sum or their probabilities exceeds the confidence. This procedure is repeated for all values of p to define a set of included points for each interval. The procedure guarantees the minimum coverage in this way, but unfortunately sometimes results in intervals with gaps in them. In practice these gaps are small so one contiguous interval can be made using the farthest end points with little impact. It therefore does not guarantee it makes the smallest possible intervals, but it nearly does so. The algorithm can be computationally expensive if performed at many values of p as described above to precisely locate the limits of an interval. A faster algorithm has been proposed along with an improvement to align the method more with its corresponding hypothesis tests (Klaschka & Reiczigel, 2020).

The Blyth-Still interval (Blyth & Still, 1983), built upon by (Casella, 1986), is built similarly to Sterne in that for each value of p it builds a set of intervals to obtain the required confidence, but in contrast to Sterne it selects them to give the smallest possible interval length. It succeeds in guaranteeing contiguous intervals that are also the shortest exact interval, but it is not nested, meaning the interval may not always get shorter as confidence is decreased (Klaschka & Reiczigel, 2020) (Thulin, 2014). In fact, it has been proven that exact intervals cannot be both the shortest possible and nested simultaneously (Blaker, 2000). Perhaps this behavior is why it has not generally gained favor.

The Blaker interval follows the same construction concept as the Sterne but uses yet another set of criteria for selecting intervals (Blaker, 2000). Both the Sterne and Blaker methods guarantee nested intervals unlike the Blyth & Still method. They guarantee the minimum coverage is at least equal to the confidence while providing shorter intervals than the Clopper-Pearson. Therefore, if an exact method is needed, either the Sterne or Blaker methods should be utilized (the Sterne and Blaker intervals are highly similar in practice, so either can be used).

The primary barrier to greater utilization of these methods seems to be computation. If such a tool is available, it should be used. If one is not, the second-best option by a wide margin is to resort to the more conservative but easy to compute Clopper-Pearson method.

Ideally, computational difficulty should no longer be a barrier to using the optimal method (Reed, 2007). Unfortunately, "It is generally true in statistical practice that only those methods that are easy to describe, remember and compute are widely used" (Brown, Cai, & Dasgupta, Interval Estimation for a Binomial Proportion, 2001). This phenomenon is probably why the Wald method is still pervasive even

though researchers have been highlighting its inferiority for decades. But in this age of apps and immediate internet access, computational convenience should no longer be as high a priority as statistical performance, particularly for multi-million-dollar programs. The burden is probably on the statisticians to supply these tools rather to expect T&E practitioners to develop code from academic articles. The optimal target audience for this plea may indeed be software developers to provide the tools to the T&E community as suggested by (Klaschka & Reiczigel, 2020).

Approximate Methods

Managing risk with the conservatism of the exact methods is a more traditional point of view. Anecdotally it has been favored in government T&E which tends to be risk averse. However, the necessarily high coverage is considered excessive by many authors. Brown (2001) called the Clopper-Pearson "wastefully conservative" and Agresti and Coull (1998) went so far as to title their article "Approximate is Better than 'Exact' for Interval Estimation of Binomial Proportions." They further conclude for the Clopper-Pearson method that "it is inappropriate to treat this approach as optimal for statistical practice."

The response has been the creation of methods that use approximations rather than directly deriving the intervals from the binomial hypothesis test; for example, assuming a normal distribution for p. This class includes the ubiquitous Wald method and the many variants attempting to cure its performance that is "persistently chaotic and unacceptably poor." (Brown, Cai, & Dasgupta, Interval Estimation for a Binomial Proportion, 2001) Its inadequacies are present even at sample sizes deemed sufficient by many sources (Brown, Cai, & DasGupta, Confidence Intervals for a Binomial Proportion and Asymptotic Expansions, 2002). It is difficult to imagine any T&E setting where the Wald method should be used given the better alternatives available.

The Wilson method performs well and is generally viewed favorably in the reviewed literature. The mean coverage is near the confidence. Unfortunately, the minimum coverage is significantly below the confidence as can be seen in the plots in Appendix C. The modified Wilson improves the minimum coverage by removing those spikes with only a small penalty to the mean coverage. One additional variant of the Wilson method in the literature has a continuity correction to improve its coverage. This variant is not recommended as the mean coverage rises to the point where the Sterne or Blaker methods may as well be used.

A close competitor among the approximate methods is the Agresti-Coull method (Agresti & Coull, Approximate is Better than "Exact" for Interval Estimation of Binomial Proportions, 1998). The Agresti-Coull method is well regarded but its mean coverage is more conservative, making it less attractive than the Wilson or Jeffreys in the framework of this paper.

Despite standing apart as Bayesian rather than exact or approximate, the reviewed literature generally considers the Jeffreys interval to have good frequentist behavior so many recommend it for general use (Warfield & Roberts, 2015) (Agresti & Coull, Approximate is Better than "Exact" for Interval Estimation of Binomial Proportions, 1998) (Albers, Kiers, & van Ravenzwaaij, 2018) (Brown, Cai, & Dasgupta, Interval Estimation for a Binomial Proportion, 2001). The plots in Appendix C show its advantageous behavior –

its minimum coverage is better than the Wilson interval (except for 80% confidence and larger n), while its mean coverage tracks the confidence better than the Modified Wilson interval. It also offers the fringe benefits of more natural interpretation and opportunities for further inference. It is interesting to note that coverage is not a relevant metric from a strictly Bayesian perspective (Newcombe, 1998); however, it seems reasonable to compare the Jeffreys credible intervals to competing confidence intervals from the frequentist viewpoint, as has been done extensively in the literature.

Based on this analysis, both the Jeffreys method and the modified Wilson are competitive when the goal is for the mean coverage to be near the confidence. The Jeffreys is recommended due to the added benefits of the Bayesian approach. However, the modified Wilson is a defensible alternative.

Coverage Comparison

The appendices provide products for comparing the recommended methods. Appendix A has a table for the intervals when $n = 10$ at 80% and 95% confidence. It shows that the Jeffreys intervals are the shortest while the exact intervals are longer, with the Clopper-Pearson being the longest. Appendix B has plots of the coverage for the methods corresponding to the intervals of Appendix A. Appendix C provides plots that depict the mean and minimum coverage for each method from $n = 5$ to $n = 30$. These products confirm that the exact methods ensure the minimum coverage meets or exceeds the confidence, while the approximate methods have mean coverage closer to the confidence. Note that the Sterne and Blaker performance is similar, and both perform better than the Clopper-Pearson method.

Conclusion

This paper distills the results of a literature review and independent analysis into a simple decision tool to help T&E practitioners select an appropriate binomial proportion confidence interval method. The flow chart suggests which confidence interval method would be most optimal given the objectives and risk tolerance of the reader to accommodate the wide variety of R&D efforts and acquisition programs in DOD and DHS. This paper also considers the performance of the methods down to 80% confidence, which was not addressed in the reviewed literature, and limits attention to small samples ($n < 30$). It is acknowledged that these recommendations are subjective; testers are of course encouraged to make their own informed decisions on which method is most appropriate for their needs.

Works Cited

Agresti, A., & Coull, B. A. (1998, May). Approximate is Better than "Exact" for Interval Estimation of Binomial Proportions. *The American Statistician, 52*(2).

Agresti, A., & Min, Y. (2001, Sep.). On Small-Sample Confidence Intervals for Parameters in Discrete Distributions. *Biometrics, 57*(3), 963-971.

Albers, C. J., Kiers, H. A., & van Ravenzwaaij, D. (2018). Credible Confidence: A Pragmatic View on the Frequentist vs Bayesian Debate. *Collabra: Psychology, 4*(1). doi:https://doi.org/10.1525/collabra.149

Blaker, H. (2000). Confidence curves and improved exact confidence intervals for discrete distributions. *The Canadian Journal of Statistics, 28*(4), 783-798.

Blyth, C. R., & Still, H. A. (1983, Mar.). Binomial Confidence Intervals. *Journal of the American Statistical Association, 78*(381), 108-116.

Brown, L. D., Cai, T. T., & Dasgupta, A. (2001). Interval Estimation for a Binomial Proportion. *Statistical Science, 16*(2), 101-133. Retrieved from http://dx.doi.org/10.1214/ss/1009213286

Brown, L. D., Cai, T. T., & DasGupta, A. (2002). Confidence Intervals for a Binomial Proportion and Asymptotic Expansions. *The Annals of Statistics, 30*(1), 160-201.

Burke, S., Divis, E., Guldin, S., Harman, M., Kolsti, K., McBride, A., . . . Welker, T. (2019). *Guide to Developing an Effective STAT Test Strategy V7.0.* Dayton: Scientific Test and Analysis Techniques Center of Excellence (STAT COE).

Casella, G. (1986). Refining binomial confidence intervals. *The Canadian Journal of Statistics, 14*(2), 113-129.

Fay, M. P. (2010, June). Two-sided Exact Tests and Matching Confidence Intervals for Discrete Data. *The R Journal, 2*(1).

Klaschka, J., & Reiczigel, J. (2020). On matching confidence intervals and tests for some discrete distributions: methodological and computational aspects. *Computational Statistics*. Retrieved from http://creativecommons.org/licenses/by/4.0/

Lecoutre, B., & Poitevineau, J. (2014). New Results for Computing Blaker's Exact Confidence Interval for One Parameter Discrete Distributions. *Communication in Statistics- Simulation and Computation, 45*(3). doi:10.1080/03610918.2014.91 1900

M'Lan, C. E., Joseph, L., & Wolfson, D. B. (2008). Bayesian Sample Size Determination for Binomial Proportions. *Bayesian Analysis, 3*(2), 269-296. doi:10.1214/08-BA310

Newcombe, R. G. (1998). Two-sided Confidence Intervals for the Single Proportion: Comparison of Seven Methods. *Statistics in Medicine, 17*, 857-872.

Reed, J. F. (2007). Better Binomial Confidence Intervals. *Journal of Modern Applied Statistical Methods, 6*(1). doi:10.22237/jmasm/1177992840

Reiczigel, J. (2003). Confidence intervals for the binomial parameter: some new considerations. *Statistics in Medicine, 22*, 611-621.

Sterne, T. E. (1954). Some Remarks on Confidence or Fiducial Limits. *Biometrika, 41*(1/2), 275-278. Retrieved from https://www.jstor.org/stable/2333026

Thulin, M. (2014). The cost of using exact confidence intervals for a binomial proportion. *Electronic Journal of Statistics, 8*, 817-840. Retrieved from arXiv:1303.1288 [math.ST]

Wang, W. (2006). Smallest confidence intervals for one binomial proportion. *Journal of Statistical Planning and Inference, 136*, 4293-4306. doi:10.1016/j.jspi.2005.08.044

Warfield, J., & Roberts, S. E. (2015). Comparison of Test Sizing Approaches for Initial and Follow-On Evaluation of Strategic Weapon Systems. *Quality Engineering, 27*(2), 230-252. doi:10.1080/08982112.2014.957 402

Appendix A: Formulas

General notes on the formulas

The formulas provide the confidence interval defined by the lower and upper confidence limits, l and u, for X successes observed in n trials. By definition, $\alpha = 1 - \text{confidence}/100$, so $\alpha = 0.20, 0.10, 0.05$ for 80%, 90%, and 95% confidence, respectively. Tables 2 and 3 in Appendix A provided calculated intervals for verifying calculations.

Clopper-Pearson Interval (One- and Two-Sided)

There are multiple ways to calculate the Clopper-Pearson interval. The form provided here uses the Beta distribution, chosen to highlight the similarity with the Jeffreys method (Thulin, 2014). The first argument is the quantile and the second and third arguments are the two parameters. The result may be calculated using Excel using the `BETA.INV` function using the three arguments as shown here. If $X = 0$, then $l = 0$. If $X = n$, then $u = 1$. Otherwise, apply Eq. 1 or Eq. 2.

The one-sided Clopper-Pearson limits are

$$l = \text{Beta}(\alpha, X, n - X + 1) \quad \text{or} \quad u = \text{Beta}(1 - \alpha, X + 1, n - X) \tag{1}$$

The two-sided Clopper-Pearson interval is

$$\begin{aligned} l &= \text{Beta}\left(\frac{\alpha}{2}, X, n - X + 1\right) \\ u &= \text{Beta}\left(1 - \frac{\alpha}{2}, X + 1, n - X\right) \end{aligned} \tag{2}$$

Jeffreys Interval

The lower and upper limits of the Jeffreys interval are the $\alpha/2$ and $1 - \alpha/2$ quantiles of the Beta posterior distribution, respectively. This approach is called the equal-tail method since the lower and upper tails beyond the limits both have an area of $\alpha/2$. The lower and upper confidence limits are

$$\begin{aligned} l &= \text{Beta}\left(\frac{\alpha}{2}, X + \frac{1}{2}, n - X + \frac{1}{2}\right) \\ u &= \text{Beta}\left(1 - \frac{\alpha}{2}, X + \frac{1}{2}, n - X + \frac{1}{2}\right) \end{aligned} \tag{3}$$

The end point corrections fill in gaps outside the set of intervals: If $X = 0$, then $l = 0$; or if $X = n$, then $u = 1$ (Same exception as in the Clopper-Pearson method).

Wilson Interval

The Wilson interval will be described here because it is the basis for the Modified Wilson Interval, which is described next. Define $z_{\alpha/2}$ as the $1-\alpha/2$ quantile of the standard normal distribution. It may be calculated in Excel as `NORM.S.INV(1-alpha/2)`. For 80%, 90%, and 95% confidence, $z_{\alpha/2} = 1.282, 1.645$, and 1.960 respectively. The point prediction is $\hat{p} = X/n$. The lower and upper limits of the Wilson interval are

$$l, u = \frac{\hat{p} + \frac{z_{\alpha/2}^2}{2n} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p}) + \frac{z_{\alpha/2}^2}{4n}}{n}}}{1 + \frac{z_{\alpha/2}^2}{n}} \tag{4}$$

An alternative form is available as a matter of personal preference, where $\hat{q} = 1 - \hat{p}$ and $\kappa = z_{\alpha/2}$ (Brown, Cai, & Dasgupta, Interval Estimation for a Binomial Proportion, 2001).

$$l, u = \frac{X + \kappa^2/2}{n + \kappa^2} \pm \frac{\kappa\sqrt{n}}{n + \kappa^2}\sqrt{\hat{p}\hat{q} + \frac{\kappa^2}{4n}} \tag{5}$$

Modified Wilson Interval

The modified Wilson interval is obtained by first calculating the Wilson interval using Eq. 4 or Eq. 5, then applying corrections if X is near 0 or n (Brown, Cai, & Dasgupta, Interval Estimation for a Binomial Proportion, 2001). The new limit will replace either the lower limit or the upper limit from the Wilson formula, but not both. Table 1 provides the formulas. The constant value in each formula can be precisely calculated as $\lambda_x = \frac{1}{2}\chi^2_{2X,\alpha}$ which is half the α^{th} quantile of the Chi-Square distribution with $2X$ degrees of freedom. The upper bounds for $X = n-3, n-2$, or $n-1$ are the "mirror image" ($u = 1 - l$) of the lower bounds for $X = 1, 2, 3$. To use the table, enter the row for the observed value of X, then go to the column for the confidence, and apply the cell's formula to replace the applicable Wilson limit.

Table 1: Boundary correction for the Modified Wilson interval

If X = …	80% confidence	90% confidence	95% confidence
1	$l = 0.2231/n$	$l = 0.1054/n$	$l = 0.0513/n$
2	$l = 0.8244/n$	$l = 0.5318/n$	$l = 0.3554/n$
3 ($n > 50$ only)	$l = 1.5350/n$	$l = 1.1021/n$	$l = 0.8177/n$
$n-3$ ($n > 50$ only)	$u = 1 - 1.5350/n$	$u = 1 - 1.1021/n$	$u = 1 - 0.8177/n$
$n-2$	$u = 1 - 0.8244/n$	$u = 1 - 0.5318/n$	$u = 1 - 0.3554/n$
$n-1$	$u = 1 - 0.2231/n$	$u = 1 - 0.1054/n$	$u = 1 - 0.0513/n$

Sterne and Blaker Intervals

The Sterne and Blaker intervals are not amenable to calculation by hand or by Excel. Both of these intervals are available in the R scripting language within the package **exactci** (Fay, 2010). An algorithm for the Blaker method with R code included in the article is provided in (Lecoutre & Poitevineau, 2014). A computationally efficient algorithm for both the Stene and the Blaker intervals has recently been published with R code available through a link within the article (Klaschka & Reiczigel, 2020).

Contact the HS CoBP if assistance is needed in utilizing the Sterne or Blaker intervals.

Appendix B: Tables of intervals for n=10

Tables 2 and 3 provide intervals so the reader can verify formulas and compare interval widths.

Table 2: Intervals for $n = 10$, 80% Confidence

X	Jeffreys	Modified Wilson	Clopper-Pearson	Sterne	Blaker
0	[0.0000,0.1236]	[0.0000,0.1411]	[0.0000,0.2057]	[0.0000,0.2080]	[0.0000,0.1957]
1	[0.0295,0.2746]	[0.0223,0.2824]	[0.0105,0.3368]	[0.0221,0.3086]	[0.0221,0.2997]
2	[0.0836,0.3948]	[0.0824,0.3984]	[0.0545,0.4496]	[0.0833,0.4511]	[0.0833,0.4489]
3	[0.1506,0.5018]	[0.1538,0.5026]	[0.1158,0.5517]	[0.1576,0.5489]	[0.1535,0.5511]
4	[0.2265,0.5997]	[0.2296,0.5986]	[0.1876,0.6458]	[0.2080,0.6131]	[0.1957,0.6131]
5	[0.3099,0.6901]	[0.3122,0.6878]	[0.2673,0.7327]	[0.3086,0.6914]	[0.2997,0.7003]
6	[0.4003,0.7735]	[0.4014,0.7704]	[0.3542,0.8124]	[0.3869,0.7920]	[0.3869,0.8043]
7	[0.4982,0.8494]	[0.4974,0.8462]	[0.4483,0.8842]	[0.4511,0.8424]	[0.4489,0.8465]
8	[0.6052,0.9164]	[0.6016,0.9176]	[0.5504,0.9455]	[0.5489,0.9167]	[0.5511,0.9167]
9	[0.7254,0.9705]	[0.7176,0.9777]	[0.6632,0.9895]	[0.6914,0.9779]	[0.7003,0.9779]
10	[0.8764,1.0000]	[0.8589,1.0000]	[0.7943,1.0000]	[0.7920,1.0000]	[0.8043,1.0000]

Table 3: Intervals for $n = 10$, 95% Confidence

X	Jeffreys	Modified Wilson	Clopper-Pearson	Sterne	Blaker
0	[0.0000,0.2172]	[0.0000,0.2775]	[0.0000,0.3085]	[0.0000,0.2909]	[0.0000,0.2829]
1	[0.0110,0.3813]	[0.0051,0.4042]	[0.0025,0.4450]	[0.0051,0.4465]	[0.0051,0.4444]
2	[0.0441,0.5028]	[0.0355,0.5098]	[0.0252,0.5561]	[0.0368,0.5535]	[0.0368,0.5556]
3	[0.0927,0.6058]	[0.1078,0.6032]	[0.0667,0.6525]	[0.0873,0.6194]	[0.0873,0.6194]
4	[0.1531,0.6963]	[0.1682,0.6873]	[0.1216,0.7376]	[0.1500,0.7091]	[0.1500,0.7171]
5	[0.2235,0.7765]	[0.2366,0.7634]	[0.1871,0.8129]	[0.2224,0.7776]	[0.2224,0.7776]
6	[0.3037,0.8469]	[0.3127,0.8318]	[0.2624,0.8784]	[0.2909,0.8500]	[0.2829,0.8500]
7	[0.3942,0.9073]	[0.3968,0.8922]	[0.3475,0.9333]	[0.3806,0.9127]	[0.3806,0.9127]
8	[0.4972,0.9559]	[0.4902,0.9645]	[0.4439,0.9748]	[0.4465,0.9632]	[0.4444,0.9632]
9	[0.6187,0.9890]	[0.5958,0.9949]	[0.5550,0.9975]	[0.5535,0.9949]	[0.5556,0.9949]
10	[0.7828,1.0000]	[0.7225,1.0000]	[0.6915,1.0000]	[0.7091,1.0000]	[0.7171,1.0000]

Notes on Tables 2 and 3

The following sources provide calculated confidence intervals for verifying calculations:

Intervals	**Source**
95% Clopper-Pearson	Agresti (2001), Table 1
95% Blaker	Agresti (2001), Table 1 R package **exactci** (Fay, 2010)
95% Blyth & Still	All intervals from n=1 to 30 (Blyth and Still, 1983)
95% Jeffreys	Brown (2001), Table 5
50% and 90% Sterne	Sterne (1954)
95% Wilson	Newcomb (1998), Table 1

Appendix C: Coverage plots for n=10

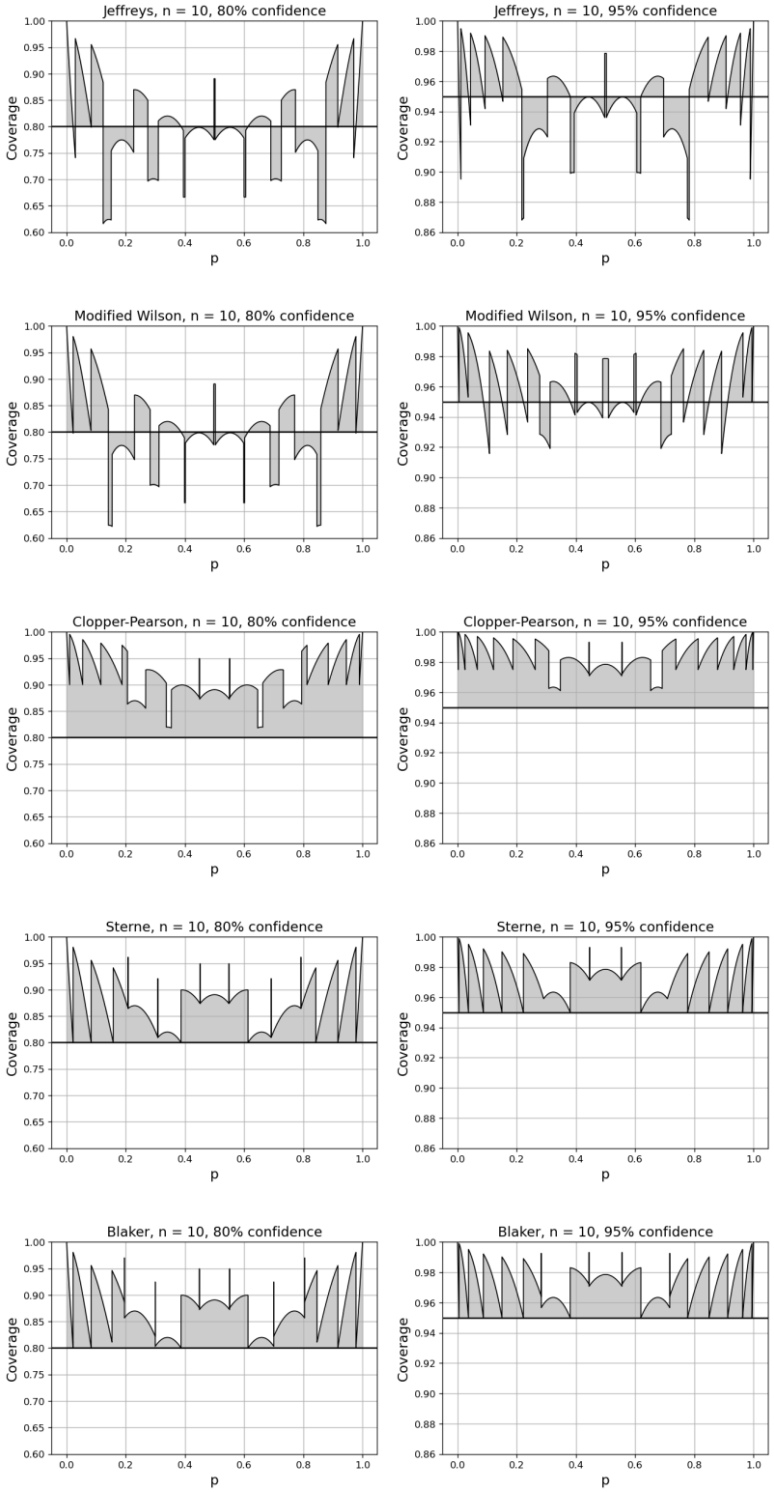

Figure 3: Coverage plots for n=10 with 80% (left column) and 95% (right column) confidence.

Appendix D: Plots of coverage versus sample size

Figure 4 presents plots of coverage metrics as a function of sample size from $n = 5$ to $n = 30$. The Sterne and Blaker curves are essentially coincident with each other. They are also equal to the confidence for all p in the minimum coverage plots.

The minimum coverage of the Wilson and the Modified Wilson intervals are coincident for 95% confidence so only one curve is visible on the plot.

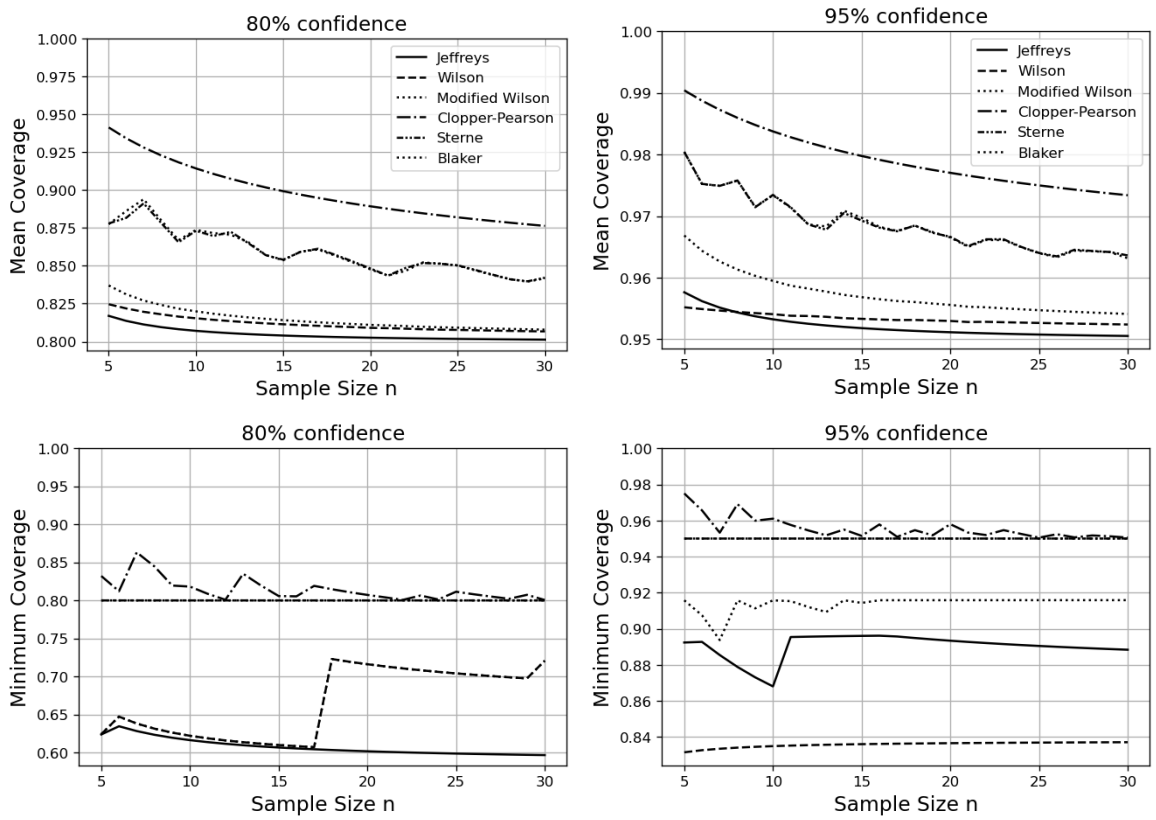

Figure 4: Mean (top row) and minimum coverage (bottom row) as a function of sample size and confidence (80% left column, 95% right column)

Appendix E: Table of Methods

The following table is a guide for further research. It is not intended to be exhaustive. Some of the references given are not in the works cited as they were not directly cited in this paper.

"X" indicates a more thorough discussion; "M" indicates a briefer mention.

Confidence Interval Method	Primary Reference	Albers 2018	Agresti 1998	Agresti 2001	Brown 2001	Brown 2002	Fay 2010	Klaschka 2020	Newcombe 1998	Reed 2007	Reiczigel 2003	Thulin 2013	Wang2004
Agresti & Coull ("Adjusted Wald") ("Plus-Four Method")	Agresti and Coull (1998)	X	X		X	X				X	X	X	
Arc Sine	Shao (1998)	X			X					M			
Bayesian - Jeffreys prior, equal-tailed ("Jeffreys Interval"		X	M		X	X						X	
Bayesian - Jeffreys prior, equal-tailed, with boundary modification	Brown (2001), "Modified Jeffreys"				X								
Bayesian - Jeffreys prior, highest-probability density (HPD)					X								
Bayesian - uniform prior, equal tailed		X	M									X	
Blaker	Blaker (2000)			X			X	X				X	
Blyth & Still (& Casella)	Blyth and Still (1983) Casella et al (1994)		M	X			M	X	M	M	X	X	M
Clopper-Pearson ("Exact method")	Clopper and Pearson (1934)	X	X	X	X		M	M	X	X	X	X	X
Likelihood Ratio	Rao (1973)		M		X	X			X				
Logit	Stone (1995)				X								
Mean Pratt modification	Pratt (1968)								M	M			
Mid-P ("Mid-P ClopperPearson")			M	M		X			X	M	X		
Probability-Based method	Hirji (2006)						X						
SAIFS-z	Borkowf (2005)									X			
Sterne	Sterne (1954)			M			M	X		M	X	X	
Sterne-Klaschka	Klaschka (2020)							X					
Wald			X		X	X		M	X	X		M	M
Wald (Blyth-Still modification)	Blyth and Still (1983)									X			
Wald with contintuity correction	Blyth (1986)								X	X			
Wilson ("Wilson-Score"; "score")	Wilson (1927)		X	X	X	X			X	X	X		
Wilson with boundary modification	Brown (2001), "Modified Wilson"				X								
Wilson with continuity correction	Casella (1986)				X				X	X			