Electronic Journal of Statistics
Vol. 8 (2014) 817–840
ISSN: 1935-7524
DOI: 10.1214/14-EJS909

The cost of using exact confidence intervals for a binomial proportion

Måns Thulin

Department of Mathematics
Uppsala University
Box 480, 751 06 Uppsala, Sweden
e-mail: thulin@math.uu.se

Abstract: When computing a confidence interval for a binomial proportion p one must choose between using an exact interval, which has a coverage probability of at least $1-\alpha$ for all values of p, and a shorter approximate interval, which may have lower coverage for some p but that on average has coverage equal to $1-\alpha$. We investigate the cost of using the exact one and two-sided Clopper–Pearson confidence intervals rather than shorter approximate intervals, first in terms of increased expected length and then in terms of the increase in sample size required to obtain a desired expected length. Using asymptotic expansions, we also give a closed-form formula for determining the sample size for the exact Clopper–Pearson methods. For two-sided intervals, our investigation reveals an interesting connection between the frequentist Clopper–Pearson interval and Bayesian intervals based on noninformative priors.

AMS 2000 subject classifications: Primary 62F25; secondary 62F12.
Keywords and phrases: Asymptotic expansion, binomial distribution, confidence interval, expected length, sample size determination, proportion.

Received March 2013; revised May 2014.

Contents

1. Introduction

Inference for a binomial proportion p is one of the most commonly encountered statistical problems, with important applications in areas such as clinical trials, risk analysis and quality control. Consequently, a large number of two-sided confidence intervals and one-sided confidence bounds for p have been proposed by different authors. These are of two different types: *exact* methods, that have a coverage at least equal to $1 - \alpha$ for all $p \in (0, 1)$, and *approximate* methods, that may have coverage less than $1 - \alpha$ for some values of p, but that have a coverage that in some sense is approximately equal to $1 - \alpha$.

Research on confidence intervals and bounds for a binomial proportion has mostly focused on approximate intervals. In the methodological literature, exact intervals have often been deemed to be too conservative [2, 5, 21], as they tend to be quite wide and have actual coverage levels that often are noticeably greater than $1 - \alpha$. Nevertheless, the use of exact intervals for proportions is abundant among practitioners: see e.g. Abramson et al. [1], Ibrahim et al. [13], Ward et al. [31] and Sullivan et al. [26] for some recent examples. By far the most widely used exact interval is the Clopper–Pearson interval, introduced by Clopper & Pearson [9].

The benefit of using an exact interval is obvious: one does not risk that the actual coverage falls below $1 - \alpha$. For this reason, some regulatory authorities require that exact intervals be used. Moreover, the binomial distribution is unusual in that we often can be sure that it is an accurate description of that which we are modelling and not just an approximation to the true distribution, as is often the case when continuous distributions are used for modelling. In such a situation, using an exact method seems reasonable. But there are also costs associated with the use of such an interval. When choosing between approximate and exact confidence methods, there is a trade-off in that exact intervals and bounds by construction are wider than the best approximate intervals, or equivalently, require a larger sample size in order to obtain a certain expected length. If one is unwilling to accept intervals and bounds with undercoverage for some values of p, there is a cost to pay in terms of expected length or required sample size. This paper seeks to quantify these costs.

In planned experiments, it is always important to determine a suitable sample size. Sample size determination for binomial confidence intervals has received

much attention in recent years [15, 22, 17, 18, 12, 32], with different authors studying different intervals and methods for sample size calculations, the latter often of a computer-intensive nature. The first main contribution of this paper is closed-form formulas for computing the sample size required for the Clopper–Pearson methods to obtain a given expected length. This eliminates the need for computer-intensive methods for computing sample sizes and gives a better understanding of how the desired length and the parameters p and α affect the sample size.

The second main contribution is closed-form expressions for the excess length and increase in required sample size that comes from using the exact Clopper–Pearson methods instead of approximate methods. We obtain these expressions by deriving asymptotic expansions for the exact Clopper–Pearson methods, extending the work of Brown et al. [6], Cai [7] and Staicu [24] on the asymptotics of approximate binomial confidence methods to exact intervals and bounds.

The rest of the paper is organised as follows. In Section 2 we introduce the Clopper–Pearson methods along with other exact and approximate confidence methods. In Section 3 we give an asymptotic expression for the expected length of the Clopper–Pearson interval. This allows us to give a formula for computing the sample size, and to determine the cost of using an exact interval rather than an approximate interval, in terms of expected length and sample size. In Section 4 we discuss the one-sided Clopper–Pearson bound and give expressions for its expected distance to p and the cost of using an exact bound. In Section 5 we discuss costs associated with approximate intervals and state some conclusions. All proofs and technical details are deferred to an appendix.

2. Binomial confidence methods

2.1. The Clopper–Pearson interval and bounds

The two-sided Clopper–Pearson interval for a proportion p is an inversion of the equal-tailed binomial test: the interval contains all values of p that aren't rejected by the test at confidence level α. Given an observation X, the lower limit is thus given by the value of p_L such that

$$\sum_{k=X}^{n} \binom{n}{k} p_L^k (1-p_L)^{n-k} = \alpha/2 \tag{1}$$

and the upper limit is given by the p_U such that

$$\sum_{k=0}^{X} \binom{n}{k} p_U^k (1-p_U)^{n-k} = \alpha/2. \tag{2}$$

As is well-known, the computation of p_L and p_U is simplified by the following equality from Johnson et al. [14]. Let $f(t, a, b)$ be the density function of a

$Beta(a, b)$ random variable. Then

$$\sum_{k=X}^{n} \binom{n}{k} p^k (1-p)^{n-k} = \int_0^p f(t, X, n-X+1)dt. \tag{3}$$

When (3) is plugged into (1) and (2), the problem of finding p_L and p_U reduces to inverting the distribution functions of two beta distributions. Consequently, the endpoints of the Clopper–Pearson interval are given by quantiles of beta distributions:

$$(p_L, p_U) = \Big(B(\alpha/2, X, n-X+1), \quad B(1-\alpha/2, X+1, n-X)\Big). \tag{4}$$

When X is either 0 or n, closed-form expressions for the interval bounds are available. When $X = 0$ the interval is $(0, 1-(\alpha/2)^{1/n})$ and when $X = n$ it is $((\alpha/2)^{1/n}, 1)$. For other values of X, (4) must be evaluated numerically. The interval is implemented in most statistical software packages; it can for instance be found in the `PropCIs` package in R and computed using the `PROC FREQ` command in SAS.

Some authors [2, 5] have argued that when choosing between confidence intervals, it is often preferable to use an interval with a simple closed-form formula rather than one that requires numerical evaluation, as the former is easier to present and to interpret. Next, we give asymptotic expansions of p_L and p_U, that function as good approximations when $n \geq 40$, and can be used if a closed-form formula for the Clopper–Pearson interval is desired. As an example, when $n = 50$ the upper bound is accurate up to two decimal places for $X \notin \{0, 1, 2, n\}$.

Theorem 1. *Let $X \in \{1, 2, \ldots, n-1\}$ be fixed. Let $\hat{p} = X/n$, $\hat{q} = 1-\hat{p}$ and $z_{\alpha/2}$ be the upper $\alpha/2$ quantile of the standard normal distribution.*

The bounds of the Clopper–Pearson interval are, up to $O(n^{-3/2})$,

$$p_L = \hat{p} - n^{-1/2} z_{\alpha/2} (\hat{p}\hat{q})^{1/2} + (3n)^{-1}\Big(2(1/2-\hat{p})z_{\alpha/2}^2 - (1+\hat{p})\Big) \qquad \textit{and}$$
$$p_U = \hat{p} + n^{-1/2} z_{\alpha/2} (\hat{p}\hat{q})^{1/2} + (3n)^{-1}\Big(2(1/2-\hat{p})z_{\alpha/2}^2 + 1 + \hat{q}\Big).$$

Similar in construction to the two-sided interval, the one-sided Clopper–Pearson bounds are obtained by inverting one-sided binomial tests. Thus the $1-\alpha$ Clopper–Pearson upper bound p_U is given by the p_U such that

$$\sum_{k=0}^{X} \binom{n}{k} p_U^k (1-p_U)^{n-k} = \alpha. \tag{5}$$

In the following, we limit our study to upper bounds. For symmetry reasons, the results are however equally valid for lower bounds, as for the bounds under consideration, a lower bound p_L for p is equivalent to an upper bound for q, as $q_U = 1 - p_L$.

If a closed-form expression for p_U is desired, it can be obtained in the form of an asymptotic expansion by replacing $\alpha/2$ with α in Theorem 1 above.

2.2. Other exact intervals

In much of the medical literature, as well as the rest of the present paper, the Clopper–Pearson interval is refered to as *the* exact confidence interval for a binomial proportion. Despite this terminology, several other exact intervals have been proposed throughout the years. These alternative intervals do not admit closed-form expressions and are, to varying extents, computer-intensive.

There are several reasons as to why the Clopper–Pearson interval is the most widely used exact interval. One is simply tradition and availability: it has found its way in to classic statistical textbooks and has been implemented in almost all statistical software packages. Compared to the computer-intensive alternatives, the Clopper–Pearson interval is also considerably simpler computationally. Finally, it remains a natural choice in that it is the inversion of the well-known equal-tailed binomial test.

In the two-sided case, however, there is room for improvement, at least if one is willing to let go of some natural properties of confidence intervals. Other exact intervals have been designed to be shorter than the Clopper–Pearson interval, by inverting two-sided tests that need not be equal-tailed. Moreover, the coverage probabilities of these intervals often fluctuate less from $1 - \alpha$ than does the coverage of the Clopper–Pearson interval.

The Blyth–Still–Casella interval [4, 8] is guaranteed to be the shortest exact interval, but has the odd property that it is not nested, in the sense that the 90 % interval need not be contained in the 95 % interval [3, Theorem 2]. This is also true for the intervals of Crow [11].

The Sterne [25] procedure yields nested intervals that are shorter than the Clopper–Pearson interval, but will in some cases result in two separate intervals rather than one connected interval. Blaker [3] proposed a nested exact interval that, while wider than the Blyth–Still–Casella interval, always is contained in the Clopper–Pearson interval. It is however sometimes a union of disjoint intervals and its upper bound is decreasing but not strictly decreasing in α when n and X are fixed [30]. The interval based on the inverted exact likelihood ratio test suffers from similar problems [30].

The Clopper–Pearson interval, in contrast, is nested, is always a connected set and has bounds that are strictly monotone in α. While it is possible to obtain shorter exact confidence intervals for a binomial proportion, this seems to be associated with the loss of nestedness, connectedness or monotonicity. As we consider these properties to be of importance, we will only include the Clopper–Pearson interval and bounds in the following sections, and will out of convenience refer to them as *the* exact methods.

Implementations of some of the alternative exact intervals are readily available. The Blyth–Still–Casella interval has been implemented in StatXact and Blaker [3] gave an S-PLUS function for his interval. A more efficient implementation of Blaker's interval is found in the R package `BlakerCI` [16].

Finally, we mention that there also are exact randomized confidence intervals, but that we do not include such intervals in the study as they suffer from

ambiguity issues and rarely are used in practice; see Thulin [28] for a discussion of such intervals.

2.3. Approximate confidence intervals and bounds

Throughout the text, the Clopper–Pearson methods will be compared to several well-known approximate methods. These are described below, along with the commonly used Wald interval. For more thorough reviews of binomial confidence methods, see Newcombe [20], Cai [7] and Brown et al. [5, 6]. In the descriptions below, $\hat{p} = X/n$ is the sample proportion, $\hat{q} = 1-\hat{p}$ and $z_{\alpha/2}$ is the $100(1-\alpha/2)$th percentile of the standard normal distribution.

The Wald interval. Inversion of the large sample test $|(\hat{p}-p)(\hat{p}\hat{q}/n)^{-1/2}| \leq z_{\alpha/2}$ leads to the Wald interval, which is presented in virtually every introductory statistics course: $\hat{p} \pm z_{\alpha/2}\sqrt{\hat{p}\hat{q}/n}$. The Wald interval suffers from particularly erratic coverage properties, and cannot be recommended for general use [5, 20].

The Wilson score interval. Like the Wald interval, the Wilson [33] score interval is based on an inversion of the large sample normal test $|(\hat{p}-p)/d(\hat{p})| \leq z_{\alpha/2}$, where $d(\hat{p})$ is the standard error of $\hat{p}$. Unlike the Wald interval, however, the inversion is obtained using the null standard error $(p(1-p)/n)^{1/2}$ instead of the sample standard error. The solution of the resulting quadratic equation leads to the confidence interval

$$\frac{X + z_{\alpha/2}^2/2}{n + z_{\alpha/2}^2} \pm \frac{z_{\alpha/2}}{n + z_{\alpha/2}^2}\sqrt{\hat{p}\hat{q}n + z_{\alpha/2}^2/4}.$$

The Wilson score interval has favourable coverage and length properties and is often recommended for general use [5, 20].

The Agresti–Coull interval. For 95% nominal coverage, Agresti & Coull [2] proposed the use of the Wald interval with two successes and two failures added, i.e. with n replaced by $n+4$ and X replaced by $X+2$. More generally, let $\tilde{n} = n + z_{\alpha/2}^2$, $\tilde{X} = X + z_{\alpha/2}^2/2$, $\tilde{p} = \tilde{X}/\tilde{n}$ and $\tilde{q} = 1-\tilde{p}$. Brown et al. [5] dubbed the interval $\tilde{p} \pm z_{\alpha/2}\sqrt{\tilde{p}\tilde{q}/\tilde{n}}$ the Agresti-Coull interval. It has performance close to that of the Wilson interval, but is somewhat simpler to use.

Bayesian Beta intervals and bounds. Let $B(\alpha, a, b)$ denote the α-quantile of the $Beta(a, b)$ distribution. An equal-tailed Bayesian credible interval based on the $Beta(a, b)$ prior is given by $(B(\alpha/2, X+a, n-X+b),\ B(1-\alpha/2, X+a, n-X+b))$, where $B(\alpha, a, b)$ is the quantile function of the $Beta(a, b)$ distribution. Similarly, an upper bound is given by $B(1-\alpha, X+a, n-X+b)$. As these methods make use of beta quantiles, they are algebraically very similar to the Clopper–Pearson interval. This connection is discussed further in Section 3.4.

The Jeffreys interval and bound. A commonly used Bayesian interval for p is the Jeffreys interval $(B(\alpha/2, X+1/2, n-X+1/2),\ B(1-\alpha/2, X+1/2, n-X+1/2))$, which is the equal-tailed credible interval derived using the noninformative Jeffreys prior. Both the two-sided interval and the one-sided bound exhibit favourable frequentist properties [5, 20, 7].

The second-order correct bound. Cai [7] proposed a coverage-corrected version of the one-sided Wald bound, based on second-order asymptotic expansions. Cai [7] recommended it for general use and gave a closed-form expression for the bound.

The modified loglikelihood root bound. Staicu [24] studied the bound obtained by inverting the modified loglikelihood root test and found it to have very favourable coverage and length properties. It cannot be expressed in a closed form, but Staicu [24] gave asymptotic expansions that can be used as approximations.

3. Two-sided intervals

3.1. Expected length

Let $q = 1 - p$ and let $L_{CP} = p_U - p_L$ denote the length of the Clopper–Pearson interval. Next, we present an asymptotic expression for the expectation of L_{CP}.

Theorem 2. *As $n \to \infty$ the expected length of the $1 - \alpha$ Clopper–Pearson interval is*

$$E(L_{CP}) = 2z_{\alpha/2}n^{-1/2}(pq)^{1/2} + n^{-1} + n^{-3/2}(pq)^{-1/2}\frac{z_{\alpha/2}}{18}\left(z_{\alpha/2}^2 - \frac{5}{2} - 17pq - 13pqz_{\alpha/2}^2\right) + O(n^{-2}). \tag{6}$$

The expansion (6) is compared to the actual expected length in Figure 1. Even for small values of n, the approximation comes quite close to the actual expected length over the entire parameter space.

Having an expression for the expected length of the Clopper–Pearson interval allows us to evaluate its performance for different combinations of n, p and α. When planning an experiment, this is extremely useful as it can be used to determine what sample size we need in order to achieve a desired expected length. Methods for determining sample size are discussed next.

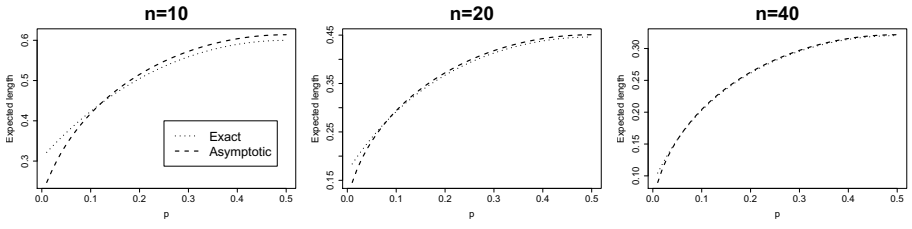

FIG 1. *Comparison between the actual expected length and the expansion (6) for the nominal 95 % Clopper–Pearson interval.*

3.2. Sample size determination

Several different criterions can be considered when determining sample size, as discussed e.g. by Gonçalves et al. [12]. We focus on a comparatively simple criterion: for a fixed confidence level $1-\alpha$ we wish to find the smallest sample size n such that the expected length of the confidence interval is some fixed value d. As the value of n will depend on p, we require that an initial guess p_0 for p is available.

Studying the Clopper–Pearson interval, Krishnamoorthy & Peng [17] gave a first-order approximation of $E(L_{CP})$ in the form of beta quantiles and used that to numerically calculate the sample size required to obtain a desired expected length d. Ignoring the higher terms of the expansion (6) we obtain the second-order approximation $E(L_{CP}) \approx 2z_{\alpha/2}n^{-1/2}(pq)^{1/2} + n^{-1}$, which can be evaluated analytically. Given an initial guess p_0 for p, the equation $2z_{\alpha/2}n^{-1/2}(p_0q_0)^{1/2} + n^{-1} = d$ has the solution

$$n = \left\lceil \frac{2z_{\alpha/2}^2 p_0 q_0 + 2z_{\alpha/2}\sqrt{z_{\alpha/2}^2 p_0^2 q_0^2 + d p_0 q_0} + d}{d^2} \right\rceil \tag{7}$$

when rounded up to the nearest integer. This is a good approximation of the actual required sample size, with a small positive bias. At the 95 % level it does typically not differ by more than 4 from the solution obtained by more complicated (and computer-intensive) exact numerical computations. For p close to $1/2$, the Krishnamoorthy–Peng method is slightly more accurate, whereas for p close to 0 or 1, (7) gives a better approximation. In either case, both approximations are accurate enough for most applications. As an example, when $p_0 = 0.05$ and $d = 0.05$, the actual required sample size is 329, while our approximation yields $n = 331$, corresponding to an actual expected length of 0.0498. In comparison with exact methods or the Krishnamoorthy–Peng procedure, (7) offers greater computational ease without sacrificing much accuracy.

It is likewise possible to solve the cubic equation that results from including the $n^{-3/2}$-term of (6), but the solution does not yield a simple formula and does not give substantially improved accuracy.

A downside to this approach to sample size determination is that the initial guess p_0 may be quite wrong. This is particularly problematic if p is closer to $1/2$ than is p_0, in which case the calculated required sample size will be too small. As a safety measure, it is sometimes recommended to use the conservative guess $p_0 = 1/2$, which maximizes the required sample size. More often than not, however, this choice is needlessly conservative.

An alternative approach, with a Bayesian flavour, is to use a prior distribution for p when determining the sample size. Beta distributions constitute a flexible and analytically tractable class of priors for p. For $p \sim Beta(a, b)$, we have

$$E\left(2z_{\alpha/2}n^{-1/2}(pq)^{1/2} + n^{-1}\right) = 2z_{\alpha/2}n^{-1/2}\frac{\Gamma(a+1/2)\Gamma(b+1/2)}{(a+b)\Gamma(a)\Gamma(b)} + n^{-1}.$$

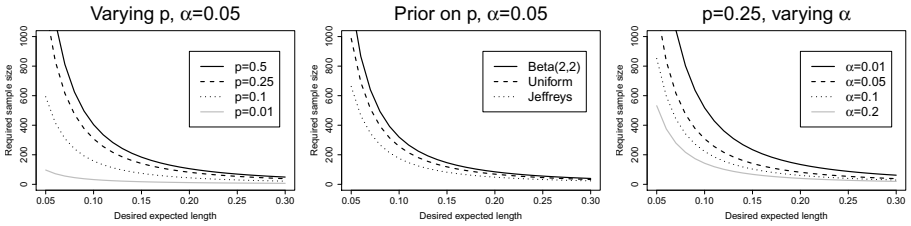

FIG 2. *The required sample size for the Clopper–Pearson interval for different combinations of p and α.*

With $R(a,b) = \Gamma(a+1/2)\Gamma(b+1/2)[(a+b)\Gamma(a)\Gamma(b)]^{-1}$, this gives the required sample size

$$n = \frac{2z_{\alpha/2}^2 R^2(a,b) + 2z_{\alpha/2}\sqrt{z_{\alpha/2}^2 R^4(a,b) + dR^2(a,b)} + d}{d^2}.$$

When applying a frequentist procedure, the prior information about p is typically diffuse, indicating that a low-informative prior should be used so as not to bias the sample size determination. One example is the Jeffreys prior $Beta(1/2, 1/2)$, which puts more probability mass close to 0 and 1 and yields $R(1/2, 1/2) = 1/\pi$. Other examples include the uniform $Beta(1,1)$ prior, for which we have $R(1,1) = \pi/8$ and the $Beta(2,2)$ prior, which puts more mass close to $1/2$, yielding $R(2,2) = 9\pi/64$.

The required sample size for different combinations of p and α is shown in Figure 2. It is decreasing in α, increasing in p when $p < 0.5$ and decreasing in p when $p > 0.5$.

Remark. In formulas similar to those above, some authors use d to denote the expected half-length, or error tolerance, of a confidence interval. This may be inappropriate in the binomial setting, since using the half-length might give the false impression that all confidence intervals are symmetric about the unbiased estimator $\hat{p} = X/n$. This is not the case for the Clopper–Pearson interval and most good approximate intervals, including those presented in Section 2.3. As an example, when $n = 50$ and $p = 0.01$, the expected length of the Clopper–Pearson interval is 0.044. Since the interval is boundary respecting, most of its length will be placed above p. The expected length is very much an interesting quantity when determining sample size, but for binomial proportions it should not be interpreted in terms of error tolerances.

3.3. The cost of using the exact interval

Next, we will study the cost of using the exact Clopper–Pearson interval instead of an approximate interval. We will do so by comparing the exact interval to three of the approximate intervals described in Section 2.3: the Wilson score, Jeffreys and Agresti–Coull intervals. These intervals have been recommended as default intervals for a single proportion by several authors [2, 5, 20].

First, we measure the cost in terms of increased expected length. By comparing the expansion in Theorem 2 to the expansions in Theorem 7 of Brown et al. [6], we get the following expressions for how much the expected length of the confidence interval increases when the Clopper–Pearson interval is used instead of an approximate interval.

Corollary 1. *The Clopper–Pearson interval is asymptotically wider than the approximate intervals described in Section 2.3. In particular, compared to the length L_J of the Jeffreys interval,*

$$E(L_{CP}) = E(L_J) + n^{-1} + O(n^{-2}), \tag{8}$$

and if L_A denotes the length of the Wilson or Agresti–Coull interval,

$$E(L_{CP}) = E(L_A) + n^{-1} + O(n^{-3/2}). \tag{9}$$

Expanded versions of (9) for the different intervals, including the $n^{-3/2}$-terms, are given in the proof in the appendix.

Up to $O(n^{-3/2})$, the increase in expected length is inversely proportional to n. Note that, up to $O(n^{-3/2})$, the increase does not depend on p or α. The cost of using an exact interval, in terms of expected length, is thus more or less constant for a fixed n. This is an interesting and somewhat unexpected fact, since the expected lengths of these confidence interval are highly dependent on both p and α.

Next, we consider required sample size. As the Clopper–Pearson interval is wider than the approximate intervals, it naturally requires larger sample sizes to obtain a particular expected length d. Let $n_{CP}(d, p, \alpha)$ be the minimum sample size for which $E_p(L_{CP}) \leq d$ at the $1 - \alpha$ level. Similarly, let $n_J(d, p, \alpha)$ be the minimum sample size for which the expected length of the Jeffreys interval is at most d under p at the $1 - \alpha$ level.

As noted by Piegorsch [22], the sample size for the Jeffreys interval is well approximated by $n_J(d, p_0, \alpha) = 4z^2_{\alpha/2}p_0q_0d^{-2}$. Comparing this to (7) without rounding, the increase in required sample size $n^+_J(d, p_0, \alpha) = n_{CP}(d, p_0, \alpha) - n_J(d, p_0, \alpha)$ can be approximated by

$$n^+_J(d, p_0, \alpha) \approx \frac{d - 2z_{\alpha/2}\left(z_{\alpha/2}p_0q_0 - \sqrt{(z_{\alpha/2}p_0q_0)^2 + dp_0q_0}\right)}{d^2}. \tag{10}$$

This approximation is quite accurate, generally differing by less than 1 when compared to the value for n^+_J obtained using substantially more computer-intensive exact computations.

(10) is plotted as a function of d for three choices of p_0 in Figure 3. When shorter intervals are desired, the increase in required sample size can be substantial. When $d = 0.05$, for instance, n^+_J is 40 for $0.05 \leq p_0 \leq 0.95$.

As was the case for the expected length, the increase n^+_J is remarkably insensitive to p and α: there is no concernable difference when $0.05 \leq p \leq 0.95$

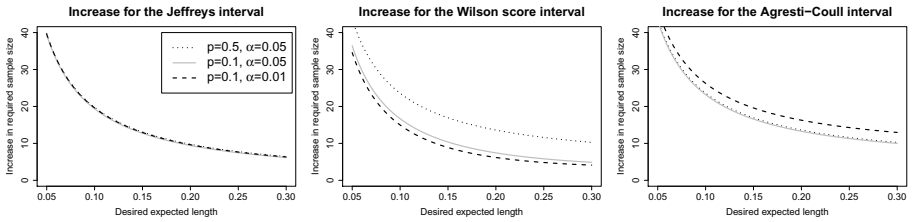

FIG 3. *The increase in required sample size when using the Clopper–Pearson interval instead of the Jeffreys, Wilson score and Agresti–Coull intervals, as approximated by (10)–(12).*

and $0.001 \leq \alpha \leq 0.2$. The cost of using an exact interval instead of the Jeffreys interval is, in terms of required sample size, constant for a fixed expected length d.

Moving on to the Wilson score interval, Piegorsch [22] gave the following formula for its sample size:

$$n_{WS}(d, p_0, \alpha) = z_{\alpha/2}^2[p_0q_0 + d^2/2 + \sqrt{p_0^2q_0^2 + d^2(p_0 - 1/2)^2}][d^2/2]^{-1}.$$

The increase $n_{WS}^+(d, p_0, \alpha) = n_{CP}(d, p_0, \alpha) - n_{WS}(d, p_0, \alpha)$ can thus be approximated by

$$n_{WS}^+(d, p_0, \alpha) \approx d^{-2}\Big[d(1 + dz_{\alpha/2}^2) + 2z_{\alpha/2}\Big(\sqrt{z_{\alpha/2}^2p_0^2q_0^2 + dp_0q_0} - \sqrt{z_{\alpha/2}^2p_0^2q_0^2 + d^2z_{\alpha/2}^2(p_0 - 1/2)^2}\Big)\Big]. \tag{11}$$

The approximation is good when p_0 is not very small, typically not differing by more than 2 from the exact value.

Similarly, Piegorsch [22] gave the formula $n_{AC}(d, p_0, \alpha) = 4z_{\alpha/2}^2p_0q_0d^{-2} - z_{\alpha/2}^2$ for the sample size of the Agresti–Coull interval. Consequently, the increase $n_{AC}^+(d, p_0, \alpha) = n_{CP}(d, p_0, \alpha) - n_{AC}(d, p_0, \alpha)$ is approximately

$$n_{AC}^+(d, p_0, \alpha) \approx \frac{d + z_{\alpha/2}^2(d^2 - 2p_0q_0) + 2z_{\alpha/2}\sqrt{(z_{\alpha/2}p_0q_0)^2 + dp_0q_0}}{d^2}. \tag{12}$$

The expressions (11) and (12) are plotted for some combinations of p and α in Figure 3. For the Agresti–Coull interval, the cost is more or less constant in p, but is sensitive to changes in α. For the Wilson score interval, the cost depends on both p and α.

3.4. The exact frequentist interval and Bayesian credible intervals with noninformative priors

Equation (8) in Corollary 1 and the fact that (10) is so insensitive to p and α reveal a strong connection between the frequentist Clopper–Pearson interval

and the Bayesian credible interval derived under the Jeffreys prior. In the light of these results, it seems natural to think of the Bayesian interval as a sort of continuity-correction of the Clopper–Pearson interval, in which conservativeness is sacrificed in order to get a short interval.

Attempts to connect the exact frequentist interval with Bayesian intervals have previously been made by Brown et al. [5], who argued that the Jeffreys interval can be thought of as a continuity-corrected version of the Clopper–Pearson interval. Their argument comes from a comparison between the Jeffreys interval and the mid-p interval, which generally is considered to be a continuity-corrected Clopper–Pearson interval. However, the key step in their argument is their equation (17), which is incorrect; it relies on the false assumption that for two continuous functions f_1 and f_2, $(f_1 + f_2)^{-1} = f_1^{-1} + f_2^{-1}$.

Another natural noninformative Bayesian interval is that based on the uniform prior, $Beta(1, 1)$. The Clopper–Pearson interval is essentially this interval *after half the prior information has been removed*, a fact which we have not seen mentioned before in the literature. To see this, note that for a central Bayesian interval with prior $Beta(a, b)$, $a, b > 0$, the lower bound is given by the beta quantile $p_{L,B}(a, b, X, n) = B(\alpha/2, X + a, n - X + b)$. The parameters a and b can be interpreted as additional successes and failures added to the data. For the uniform prior, $a = b = 1$. The lower bound of the Clopper–Pearson interval is similarly the beta quantile $B(\alpha/2, X, n - X + 1)$. When $X \notin \{0, n\}$ this can be written as $B(\alpha/2, (X-1)+1, (n-1)-(X-1)+1) = p_{L,B}(1, 1, X-1, n-1)$, the lower bound of the $Beta(1, 1)$ interval with one success removed. Similarly, the upper bound is $1 - p_{L,B}(1, 1, n - X, n - 1)$, i.e. 1 minus the lower bound for q under the uniform prior with one success removed. The $Beta(1, 1)$ interval can thus be thought of as a shrinkage Clopper–Pearson interval.

4. One-sided bounds

4.1. Expected distance to the true proportion

For one-sided confidence bounds, it is not the expected length that is of interest, but how close the bound is to p. Let $L_{U,CP} = p_U - p$ denote the distance from p_U to p. The next theorem gives an asymptotic expansion for the expectation of $L_{U,CP}$.

Theorem 3. *As $n \to \infty$ the expected distance to p for the $1 - \alpha$ one-sided Clopper–Pearson upper bound is*

$$E(L_{U,CP}) = n^{-1/2} z_\alpha (pq)^{1/2} + (3n)^{-1}\Big(2(1/2 - p)z_\alpha^2 + 1 + q\Big) + n^{-3/2} z_\alpha (pq)^{1/2}\Big(- \frac{53}{36} + \frac{\frac{1}{2} - p}{q} + \frac{z_\alpha^2 + \frac{13}{2}}{36pq} - \frac{13 z_\alpha^2}{36}\Big) + O(n^{-2}) \tag{13}$$

The expansion (13) is compared to the actual expected distance to p in Figure 4. Like the expansion for the expected length of the two-sided interval, (13) is close to the actual expected distance even for small n.

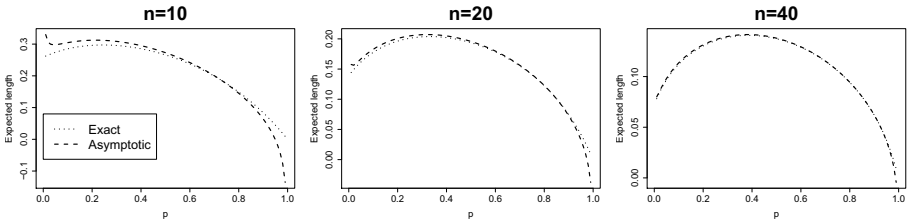

FIG 4. *Comparison between the actual expected distance and the expansion (13) for the nominal 95 % Clopper–Pearson upper bound.*

4.2. Sample size determination

The expressions we obtain in the one-sided case are not quite as simple as those in the two-sided case. Let d denote the desired expected distance to p and let p_0 be the initial guess for the value of p. Proceeding as before, using the second-order approximation

$$E(L_{U,CP}) \approx n^{-1/2} z_\alpha (pq)^{1/2} + (3n)^{-1}\Big(2(1/2-p)z_\alpha^2 + 1 + q\Big)$$

yields the required sample size

$$\Big\lceil n = (2d^2)^{-1}\Big(9z_\alpha^2 p_0 q_0 + 3z_\alpha\sqrt{3p_0q_0}\sqrt{3z_\alpha^2 p_0 q_0 + 4[dz_\alpha^2 - 2dz_\alpha^2 p_0 + d(1+q_0)]} + 6[2z_\alpha^2(1/2-p_0) + (1+q_0)]\Big)\Big\rceil.$$

This approximation is very good when d is not too small. For smaller d it has a small negative bias: when $\alpha = 0.05$ and $p_0 = 1/2$ the actual required sample size for $d = 0.02$ is $n = 1738$, whereas the above expression gives the approximation $n = 1721$, corresponding to a true expected distance of $d = 0.0201$. For most purposes, this will probably be a sufficiently accurate approximation.

As in the two-sided case, we may consider using a prior distribution of p, rather than a fixed p_0, to determine a reasonable sample size. The expectation of the second-order approximation with respect to a $Beta(a,b)$ prior for p is

$$\frac{(2+z_\alpha^2)\Gamma(2-a)\Gamma(2-b)}{3n\Gamma(4-a-b)} - \frac{(2z_\alpha^2+1)\Gamma(3-a)\Gamma(2-b)}{3n\Gamma(5-a-b)} + \frac{z_\alpha\Gamma(5/2-a)\Gamma(5/2-b)}{\sqrt{n}\Gamma(5-a-b)}. \tag{14}$$

Note that this expression is undefinied when $a, b \geq 2$, limiting which priors we can use. When (14) is well-defined, a general formula for the required sample size can be obtained by equating (14) to d and solving for n, but the resulting expression is rather complicated. It is however readily evaluated for particular values of a and b. For the Jeffreys prior for instance, the required sample size is

$$\Big\lceil n = \frac{6z_\alpha(z_\alpha + \sqrt{z_\alpha^2 + 9d\pi})}{d^2} + \frac{\pi}{16d}\Big\rceil.$$

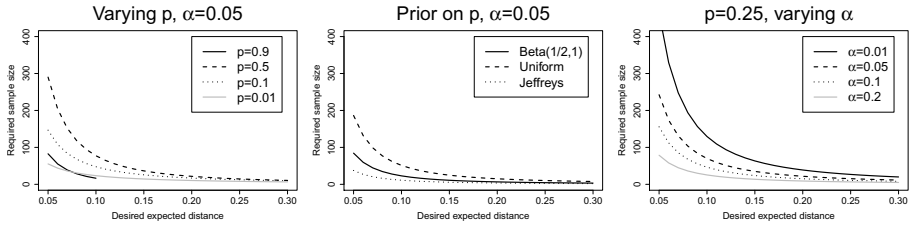

FIG 5. *The required sample size for the upper Clopper–Pearson bound for different combinations of p and α.*

The solutions for the Jeffreys and uniform priors as well as the low-informative asymmetric $Beta(1/2, 1)$ prior are shown in Figure 5, along with the solutions for fixed p_0 and different values of α.

In contrast to the two-sided case, d can in fact be interpreted as an error tolerance for the one-sided bound. This makes the interpretation of d easier in this case.

4.3. The cost of using the exact bound

The cost of using the exact bound will be evaluated in relation to three approximate bounds: The Jeffreys, second-order correct and modified loglikelihood root bounds, described in Section 2.3. Comparing (13) to the expansions in Corollary 1 of Cai [7] and Proposition 2.2 of Staicu [24], the following corollary is immediate.

Corollary 2. *When $L_{U,A}$ denotes the distance of the Jeffreys, second-order correct or modified loglikelihood root bounds,*

$$E(L_{U,CP}) = E(L_{U,A}) + (2n)^{-1} + O(n^{-3/2}).$$

It should be noted that there are one-sided versions of the Wald and Wilson score intervals, but since these have very poor performance [7] they are omitted from our comparison. They can however readily be compared to the Clopper–Pearson bound by comparing (13) to the corresponding expansions in Corollary 1 of Cai [7].

For one-sided bounds, the approximation of the increased sample size when the exact bound is used is more involved than it was for the two-sided cases. To keep the comparison brief, we simply use the naive first-order formula $n = z^2_{\alpha/2} pqd^{-2}$ to determine the sample sizes for the approximate bounds. This works reasonably well most of the time. Let $n^+(d, p, \alpha)$ be the increase in sample size when the Clopper–Pearson bound is used instead of an approximate bound. Then, with $\omega(z, d, p) = 9z^2pq + 12dz^2 - 24dz^2p$,

$$n^+(d, p_0, \alpha) \approx \frac{\sqrt{\omega(z_\alpha, d, p_0) + 12d(1 + q_0)} - \sqrt{\omega(z_\alpha, d, p_0) + 12d(1/2 - p_0)} + \frac{d}{2}}{d^2}. \tag{15}$$

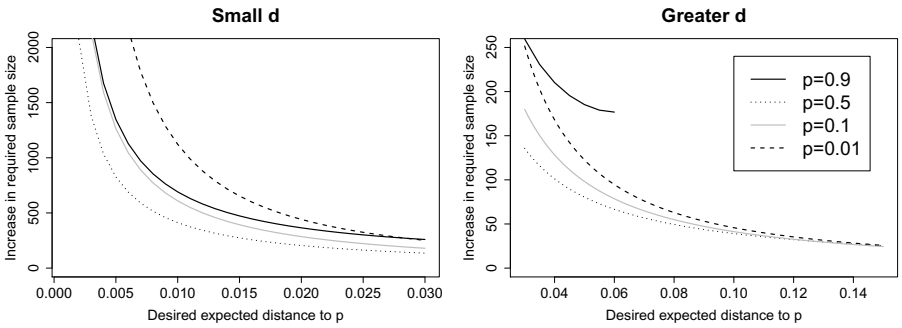

FIG 6. *The increase in required sample size when using the upper Clopper–Pearson bound instead of an approximate upper bound, as approximated by (15) for* $\alpha = 0.05$.

Compared to the increased sample size in the two-sided setting, (15) is more sensitive to changes in p and α. The cost is the smallest when $p = 0.5$. When evaluating the increased sample size $p_0 = 0.5$ is therefore not to be recommended as the default choice, as this can lead to a serious underestimation of the increase, especially for smaller d.

5. Discussion

5.1. Minimum coverage or mean coverage?

The Clopper–Pearson methods are exact in the sense that their minimum coverage over all p is at least $1 - \alpha$. An alternative measure of coverage is mean coverage, which typically is taken to be the expected coverage with respect to a uniform pseudo-prior of p. In recent papers on binomial confidence intervals, approximate methods have often been considered to be preferable to exact methods [2, 5, 7, 21], the argument being that it makes more sense to interpret the confidence level as the mean coverage probability rather than the minimum coverage probability, as this corresponds better to how many modern-day statisticians think of coverage levels. Reasoning along the lines of Newcombe & Nurminen [21], the minimum coverage can occur in an uninteresting part of the parameter space, typically close to the boundaries, possibly rendering it an uninteresting measure of coverage. This is discussed further in the next section.

As noted e.g. by Newcombe & Nurminen [21], using mean coverage is very much in line with current statistical practice in other problems. Widely used methods based on boostrapping and MCMC, for instance, typically only control confidence levels and type I error rates approximately, attaining the $1 - \alpha$ level only on average. This is particularly reasonable when the model is known to be an imperfect representation of the underlying process, in which case even minimum coverage criterions are approximate at best. Unlike in many other applications however, one can often be rather certain that a random variable truly

is binomial. This begs the question whether one should resort to approximations or use methods that really are guaranteed to be exact.

If the Bayesian credible intervals based on either the Jeffreys $Beta(1/2, 1/2)$ or the uniform $Beta(1, 1)$ priors are used, an additional argument for the mean coverage criterion is given by the Bayesian interpretation of these intervals. If we accept mean coverage as a criterion when choosing between confidence intervals, we can obtain intervals that simultaneously admit both frequentist and objective Bayesian interpretations.

The minimum coverage criterion underlying the Clopper–Pearson interval is in line with classical statistical theory. It asserts that overcoverage is a less serious problem than undercoverage, or, in other words, that it is better to be more confident than you think that you are than to be overconfident. Next, in order to evaluate this argument further, we will discuss just how overconfident one risks being when using approximate intervals.

5.2. The cost of using approximate methods

Just as there are costs associated with using exact methods, there are costs associated with using approximate methods: the actual coverage level may, even for large n, drop below the nominal $1-\alpha$. There is no guarantee that the true p is not in an unfortunate area with low coverage. However, these coverage anomalies usually occur close to the boundaries of the parameter space, so unless we are interested in inference for p close to 0 or 1, it may therefore be more relevant to investigate the minimum over a central subset, such as $[0.1, 0.9]$.

The problem of undercoverage is illustrated in Figure 7, in which the minimum coverages of the Jeffreys, Wilson and Agresti–Coull intervals are shown for different n when the minimum is taken over either $p \in [0.01, 0.99]$ or $p \in [0.1, 0.9]$. For $p \in [0.01, 0.99]$ and a moderately large sample size of $n = 250$, the minimum coverage of the Jeffreys interval is approximately 0.88, whereas the minimum coverage of the Wilson score interval is about 0.93. The Agresti–Coull interval fares somewhat better, with a minimum coverage of 0.94. In this setting neither the Jeffreys nor the Wilson score interval has a minimum coverage above 0.94 even for a sample size as large as $n = 2000$.

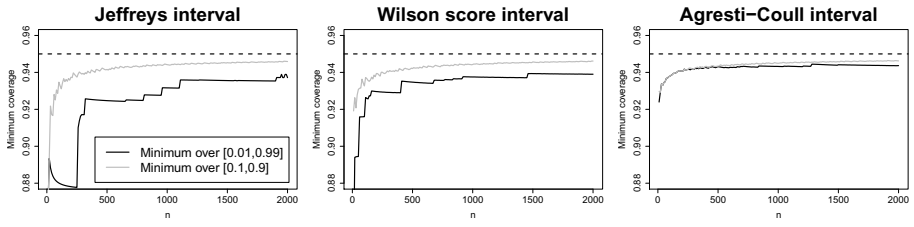

FIG 7. *Minimum coverage of two-sided approximate intervals over* $p \in [0.01, 0.99]$ *or* $p \in [0.1, 0.9]$ *when* $\alpha = 0.05$*, computed over a grid of 200,000 equidistant points.*

A coverage of 0.94 for a nominal 0.95 method is well below what one should expect for sample sizes as large as $n = 2000$. If undercoverage of this size is unacceptable, one may apply computer-intensive coverage-adjustment method similar to those discussed in Reiczigel [23], decreasing α to some γ for which the minimum coverage over some set of values of p is at least $1 - \alpha$, thus making the methods exact. Decreasing α will however *increase* the expected length of the intervals.

Comparing sample sizes of the $1-\gamma$ Jeffreys interval and the $1-\alpha$ Clopper–Pearson interval, we have:

$$n^+(d, p_0, \alpha, \gamma)) \approx \frac{d + 2p_0q_0(z^2_{\alpha/2} - 2z^2_{\gamma/2}) + 2z_{\alpha/2}\sqrt{z^2_{\alpha/2}p_0^2q_0^2 + dp_0q_0}}{d^2}.$$

For n between 1000 and 1500, computer-intensive adjustments of the Jeffreys interval lead to $\gamma \approx 0.04$ (the actual γ being somewhat larger than 0.04). For $p_0 = 1/2$ and $d = 0.04$, we get $n^+(0.04, 1/2, 0.05, 0.04)) \approx -186$, i.e. that the Clopper–Pearson interval requires 186 observations *fewer* to obtain the desired expected length. In general, approximate intervals that have been adjusted to be exact are outperformed by the Clopper–Pearson interval.

Similarly, if one is willing to use approximate intervals, it is possible to apply coverage-adjustments to the Clopper–Pearson interval in order to adjust its mean coverage to $1-\alpha$. The resulting $\gamma > \alpha$, meaning that the interval becomes shorter after the adjustment. Thulin [27] studied this problem in detail for $n \leq 100$, showing that the adjusted Clopper–Pearson intervals often outperformed its competitors.

It should be noted that other criterions than coverage and expected length can be used for comparing confidence intervals. Newcombe [19, 20] compared location properties, i.e. left and right non-coverage, of intervals and found the Clopper–Pearson interval to have good properties in comparison to some approximate intervals. Vos & Hudson [29] considered two criterions related to p-values, motivated by the interpretation of confidence intervals as inverted tests, and found the Clopper–Pearson interval to be better than its competitors.

5.3. On sample size determination

One of the main contributions of this paper is the formulas for sample size determination that are given in Sections 3.2 and 4.2. Bearing in mind the rapid increase of computational power, one might question whether there is a need for such formulas, or if computer-intensive sample size methods should be used instead. Some arguments in defence of formulas are presented next.

Despite the computational resources available, time can still be an issue when comparing formulas and computer-intensive methods. While computer-intensive sample size determination certainly is feasible on modern computers, comparing the sample sizes for different combinations of α and p can be time-consuming. Using a formula, such a comparison is a trivial task. In Sections 3.2 and 4.2, we propose using a prior for p (rather than a fixed guess p_0) when determining

the sample size. While this also can become very time-consuming if a computer-intensive method is used, the sample size under the prior is readily computed using the formulas in Sections 3.2 and 4.2.

Apart from the merits of computational simplicity, the benefit of having a formula is that it becomes clear *how* the parameters p, d and α affect the sample size n. Computer-intensive methods, in contrast, are black boxes that are useful for computing sample sizes, but not for much else. A formula is more useful when a statistician shows and explains sample size calculations to a client, and is likewise more useful in teaching.

Finally, the formulas in 3.2 and 4.2 are of interest even if one prefers to use computer-intensive methods, as they can be used to obtain an initial guess for the sample size n required. This can speed up computer-intensive determination substantially, particularly if d is small, in which case a large number of iterations tend to be needed to find the sample size if no good starting guess is available.

5.4. Conclusion

When choosing between exact and approximate confidence methods, it is important to be aware of the benefits and the costs associated with the two types of methods. The coverage fluctuations of approximate intervals have been compared in several studies, making it easy for practitioners to compare how costly these intervals can be in terms of undercoverage. We have attempted to make the costs of using exact methods explicit, by giving expressions for how much larger the expected length of the exact intervals are and for how much the sample size increases when a fixed expected length is to be attained.

For the two-sided Jeffreys interval, exactness comes at a fixed price: the cost of using the Clopper–Pearson interval instead of this intervals is, in terms of expected length and required sample size, insensitive to p and α. For the Agresti–Coull interval, the cost only depends on α. This stands in contrast to the Wilson score interval and one-sided bounds, for which p and α can greatly affect the cost. In either case the required sample sizes for the exact methods can be substantially larger than those of the approximate methods. That α can have a large impact on the cost is interesting since most numerical comparisons of binomial confidence intervals only consider $\alpha = 0.05$.

In our comparison of exact and approximate methods, the only exact methods considered were the Clopper–Pearson interval and bound. While other shorter exact two-sided intervals exist, they suffer from various problems that make them unsuitable for use. Moreover, the Clopper–Pearson interval is used far more often than the other exact intervals, which merits its role as the main subject of this study.

Acknowledgements

The author would like to thank the editor and an anonymous reviewer for constructive feedback, as well as Robert Newcombe and Silvelyn Zwanzig for their many thoughtful comments on an earlier version of this paper.

Appendix: Proofs

Theorem 1 follow directly from the following lemma, which is used in the proofs of Theorems 2 and 3.

Lemma 1. *With assumptions and notation as in Theorem 1, the bounds of the Clopper–Pearson interval are*

$$\begin{aligned}
p_L &= \hat{p} - n^{-1/2} z_{\alpha/2} (\hat{p}\hat{q})^{1/2} + (3n)^{-1}\Big(2(1/2-\hat{p}) z_{\alpha/2}^2 - (1+\hat{p})\Big) \\
&\quad - n^{-3/2} z_{\alpha/2} (\hat{p}\hat{q})^{1/2}\Big(- \frac{53}{36} - \frac{\frac{1}{2}-\hat{p}}{\hat{p}} + \frac{z_{\alpha/2}^2+11}{36\hat{p}\hat{q}} - \frac{13 z_{\alpha/2}^2}{36}\Big) + O(n^{-2}), \\
p_U &= \hat{p} + n^{-1/2} z_{\alpha/2} (\hat{p}\hat{q})^{1/2} + (3n)^{-1}\Big(2(1/2-\hat{p}) z_{\alpha/2}^2 + (1+\hat{q})\Big) \\
&\quad + n^{-3/2} z_{\alpha/2} (\hat{p}\hat{q})^{1/2}\Big(- \frac{53}{36} + \frac{\frac{1}{2}-\hat{p}}{\hat{q}} + \frac{z_{\alpha/2}^2+11}{36\hat{p}\hat{q}} - \frac{13 z_{\alpha/2}^2}{36}\Big) + O(n^{-2}).
\end{aligned}$$

The approximations are close to the actual bounds even for small sample sizes. When $n = 25$ and $\hat{p}$ is not too close to 0 or 1, the approximations are typically accurate up to at least least two decimal places.

Proof of Lemma 1. First, we note that the lower limit of the Bayesian interval with prior $Beta(a,b)$, $a,b>0$, is given by the beta quantile $p_B(a,b,X,n) = B(\alpha/2, X+a, n-X+b)$.

For the Clopper–Pearson interval p_L is the beta quantile $B(\alpha/2, X, n-X+1)$. When $X \notin \{0,n\}$ this can be written as $B(\alpha/2,(X-1)+1,(n-1)-(X-1)+1)$, i.e. $p_B(1,1,X-1,n-1)$, the lower limit of the $Beta(1,1)$ interval for $X-1$ and $n-1$.

Brown et al. [6] gave the following asymptotic expression for $p_B(1,1,X,n)$:

$$\begin{aligned}
p_B(a,b,X,n) &= \hat{p} + \frac{1}{3}\frac{(1-2\hat{p})(z_{\alpha/2}^2+2)}{n} - \Big[\frac{z_{\alpha/2}(\hat{p}\hat{q})^{1/2}}{\sqrt{n}} \\
&\quad + \frac{z_{\alpha/2}(\hat{p}\hat{q})^{1/2}}{(n)^{3/2}}\Big(\frac{z_{\alpha/2}^2+11}{36}(\hat{p}\hat{q})^{-1} - \frac{13 z_{\alpha/2}^2+71}{36}\Big)\Big] + O(n^{-2}).
\end{aligned}$$

In particular, for $X \in \{1,2,\ldots,n-1\}$ and $n>1$ we have

$$\begin{aligned}
p_L &= p_B(1,1,X-1,n-1) \\
&= \frac{X-1}{n-1} + \frac{1}{3}\frac{(1-2\cdot\frac{X-1}{n-1})(z_{\alpha/2}^2+2)}{(n-1)} - \Big[\frac{z_{\alpha/2}(\frac{X-1}{n-1}\cdot\frac{n-X}{n-1})^{1/2}}{(n-1)^{1/2}} \\
&\quad + \frac{z_{\alpha/2}(\frac{X-1}{n-1}\cdot\frac{n-X}{n-1})^{1/2}}{((n-1))^{3/2}}\Big(\frac{z_{\alpha/2}^2+11}{36}\Big(\frac{X-1}{n-1}\cdot\frac{n-X}{n-1}\Big)^{-1} \\
&\quad - \frac{13 z_{\alpha/2}^2+71}{36}\Big)\Big] + O((n-1)^{-2}).
\end{aligned} \tag{16}$$

Next, we obtain the asymptotic expression for the lower Clopper–Pearson bound by rewriting each part of (16) in terms of $\hat{p} = X/n$ and $\hat{q} = 1 - \hat{p}$. First of all,

$$\frac{X-1}{n-1} = \frac{X}{n} + \frac{X-1}{n-1} - \frac{X}{n} = \frac{X}{n} + \frac{X - 1 - \frac{X}{n}(n-1)}{n-1} = \hat{p} + \frac{\hat{p}-1}{n-1}.$$

The $(n-1)^{-1}$-part of (16) now becomes

$$\begin{aligned} &\frac{\hat{p}-1}{n-1} + \frac{1}{3}\frac{(1 - 2\cdot\frac{X-1}{n-1})(z_{\alpha/2}^2+2)}{(n-1)} \\ &= \frac{3\hat{p} - 3 + z_{\alpha/2}^2 + 2 - 2\left(\hat{p} + \frac{\hat{p}-1}{n-1}\right)(z_{\alpha/2}^2+2)}{3(n-1)} \\ &= \frac{2(1/2-\hat{p})z_{\alpha/2}^2 - (1+\hat{p})}{3(n-1)} + O((n-1)^{-2}) \\ &= \frac{2(1/2-\hat{p})z_{\alpha/2}^2 - (1+\hat{p})}{3n} + O(n^{-2}). \end{aligned} \tag{17}$$

Next, we have

$$\frac{n-X}{n-1} = 1 - \frac{X-1}{n-1} = \hat{q} - \frac{\hat{p}-1}{n-1},$$

whence it follows

$$\begin{aligned} \frac{X-1}{n-1}\cdot\frac{X-n}{n-1} &= \left(\hat{p} + \frac{\hat{p}-1}{n-1}\right)\left(\hat{q} - \frac{\hat{p}-1}{n-1}\right) \\ &= \hat{p}\hat{q} + (\hat{q}-\hat{p})\frac{\hat{p}-1}{n-1} - \left(\frac{\hat{p}-1}{n-1}\right)^2 \\ &= \hat{p}\hat{q} + 2(1/2-\hat{p})\frac{\hat{p}-1}{n-1} - \left(\frac{\hat{p}-1}{n-1}\right)^2 \\ &= \hat{p}\hat{q} - 2(1/2-\hat{p})\frac{\hat{q}}{n-1} - \left(\frac{\hat{q}}{n-1}\right)^2. \end{aligned}$$

Since

$$\left(pq - 2(1/2-p)\frac{q}{n-1} - \left(\frac{q}{n-1}\right)^2\right)^{1/2} = \sqrt{pq} - \sqrt{pq}(1/2-p)(pn)^{-1} + O(n^{-2}),$$

$$(n-1)^{-1/2} = n^{-1/2} + \frac{1}{2}n^{-3/2} + O(n^{-5/2})$$

and

$$\left(pq - 2(1/2-p)\frac{q}{n-1} - \left(\frac{q}{n-1}\right)^2\right)^{-1} = (pq)^{-1} + O(n^{-1})$$

the $(n-1)^{-1/2}$ and $(n-1)^{-3/2}$-parts of (16) can be written as

$$-\Big[\frac{z_{\alpha/2}(\frac{X-1}{n-1}\cdot\frac{n-X}{n-1})^{1/2}}{(n-1)^{1/2}} + \frac{z_{\alpha/2}(\frac{X-1}{n-1}\cdot\frac{n-X}{n-1})^{1/2}}{((n-1))^{3/2}}\times$$

$$
\begin{aligned}
&\times \Big(\frac{z_{\alpha/2}^2+11}{36}\Big(\frac{X-1}{n-1}\cdot\frac{n-X}{n-1}\Big)^{-1}-\frac{13z_{\alpha/2}^2+71}{36}\Big)\Big]+O((n-1)^{-2})\\
&=-\Big[z_{\alpha/2}\sqrt{\hat{p}\hat{q}}n^{-1/2}+z_{\alpha/2}\sqrt{\hat{p}\hat{q}}n^{-3/2}\times\\
&\times\Big(1/2-\frac{1/2-\hat{p}}{\hat{p}}+\frac{z_{\alpha/2}^2+11}{36}(\hat{p}\hat{q})^{-1}-\frac{13z_{\alpha/2}^2+71}{36}\Big)\Big]+O(n^{-2})\\
&=-\Big[z_{\alpha/2}\sqrt{\hat{p}\hat{q}}n^{-1/2}+z_{\alpha/2}\sqrt{\hat{p}\hat{q}}n^{-3/2}\times\\
&\times\Big(-\frac{1/2-\hat{p}}{\hat{p}}+\frac{z_{\alpha/2}^2+11}{36}(\hat{p}\hat{q})^{-1}-\frac{13z_{\alpha/2}^2+53}{36}\Big)\Big]+O(n^{-2}).
\end{aligned}
\tag{18}
$$

The expansion for p_L is now obtained as $\hat{p}+(17)+(18)$. The expansion for p_U is derived analogously. □

Proof of Theorem 2. Using the expansion in Lemma 1, when $X \notin \{0,n\}$

$$L_{CP}=p_U-p_L=2n^{-1/2}z_{\alpha/2}(\hat{p}\hat{q})^{1/2}+n^{-1}+n^{-3/2}m(\hat{p})+R_n,$$

where

$$m(\hat{p})=(\hat{p}\hat{q})^{-1/2}\frac{z_{\alpha/2}}{18}\Big(z_{\alpha/2}^2+2-17\hat{p}\hat{q}-13\hat{p}\hat{q}z_{\alpha/2}^2\Big)$$

and $E(R_n)=O(n^{-2})$ by the mean value theorem. As the contribution to expected length given by $X\in\{0,n\}$ is $P(X\in\{0,n\})\cdot(1-(\alpha/2)^{1/n})=O((1/2)^n)$, when computing $E(L_{CP})$ we can disregard the fact that the above expansion is invalid for $X\in\{0,n\}$.

The $n^{-1/2}$-term is the length of the Wald interval, the expectation of which was given in Brown et al. [6]:

$$E\Big(2z_{\alpha/2}n^{-1/2}(\hat{p}\hat{q})^{1/2}\Big)=2z_{\alpha/2}n^{-1/2}(pq)^{1/2}\Big(1-(8npq)^{-1}\Big)+O(n^{-2}).$$

$m(\hat{p})$ is bounded when $X\neq\{0,n\}$ and $m(p)$ is twice differentiable for $0<p<1$. Thus, by the theorem in Section 27.7 of Cramér [10],

$$E(m(\hat{p}))=(pq)^{-1/2}\frac{z_{\alpha/2}}{18}\Big(z_{\alpha/2}^2+2-17pq-13pqz_{\alpha/2}^2\Big)+O(n^{-1})$$

and (6) follows after all terms of the same order are collected. □

Proof of Corollary 1. (8) and (9) are obtained by comparing (6) to the expansions in Theorem 7 of Brown et al. [6]. In particular, compared to the length L_{WS} of the Wilson score interval,

$$
\begin{aligned}
E(L_{CP})=E(L_{WS})&+n^{-1}\\
&-n^{-3/2}\frac{z_{\alpha/2}}{36(pq)^{1/2}}\Big[9z_{\alpha/2}\Big(z_{\alpha/2}+\Big(\frac{26}{9}pq-\frac{2}{9}\Big)^2\Big)\\
&\qquad+34pq(1-2z_{\alpha/2}^2)-4\Big]+O(n^{-2}),
\end{aligned}
$$

compared to the length L_{AC} of the Agresti–Coull interval,

$$E(L_{CP}) = E(L_{AC}) + n^{-1} - n^{-3/2}\frac{z_{\alpha/2}}{36(pq)^{1/2}}\Big[9z_{\alpha/2}\Big(2z_{\alpha/2} + \Big(\frac{26}{9}pq - \frac{2}{9}\Big)^2\Big) + pq(34 - 108z_{\alpha/2}^2) - 4\Big] + O(n^{-2}). \qquad \square$$

The proof of Theorem 3 is in complete analogue with the proof of Theorem 2 and is therefore omitted. As in the proof of Theorem 2, the expected distance of the one-sided Wald bound must be computed in an intermediate step: this expectation can be found in Corollary 1 of Cai [7].

References

[1] Abramson, J.S., Takvorian, R.W., Fisher, D.C., Feng, Y., Jacobsen, E.D., et al. (2013). Oral clofarabine for relapsed/refractory non-Hodgkin lymphomas: results of a phase 1 study. *Leukemia & Lymphoma*, **54**, 1915–1920.

[2] Agresti, A., Coull, B.A. (1998). Approximate is better than "exact" for interval estimation of a binomial proportion. *The American Statistician*, **52**, 119–126. MR1628435

[3] Blaker, H. (2000). Confidence curves and improved exact confidence intervals for discrete distributions. *The Canadian Journal of Statistics*, **28**, 783–798. MR1821434

[4] Blyth, C.R., Still, H.A. (1983). Binomial confidence intervals. *Journal of the American Statistical Association*, **78** 108–116. MR0696854

[5] Brown, L.D., Cai, T.T., DasGupta, A. (2001). Interval estimation for a binomial proportion. *Statistical Science*, **16**, 101–133. MR1861069

[6] Brown, L.D., Cai, T.T., DasGupta, A. (2002). Confidence intervals for a binomial proportion and asymptotic expansions. *The Annals of Statistics*, **30**, 160–201. MR1892660

[7] Cai, T.T. (2005). One-sided confidence intervals in discrete distributions. *Journal of Statistical Planning and Inference*, **131**, 63–88. MR2136006

[8] Casella, G. (1986). Refining binomial confidence intervals. *The Canadian Journal of Statistics*, **14**, 113–129. MR0849867

[9] Clopper, C.J., Pearson, E.S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, **26**, 404–413.

[10] Cramér, H. (1946). *Mathematical Methods of Statistics*, Princeton University Press. MR0016588

[11] Crow, E.L. (1956). Confidence intervals for a proportion. *Biometrika*, **43**, 423–435. MR0093077

[12] Gonçalves, L., de Oliviera, M.R., Pascoal, C., Pires, A. (2012). Sample size for estimating a binomial proportion: comparison of different methods. *Journal of Applied Statistics*, **39**, 2453–2473. MR2993297

[13] Ibrahim, T., Farolfi, A., Scarpi, E., Mercatali, L., Medri, L., et al. (2013). Hormonal receptor, human epidermal growth factor receptor-2, and Ki67 discordance between primary breast cancer and paired metastases: clinical impact. *Oncology*, **84**, 150–157.

[14] Johnson, N.L., Kemp, A.W., Kotz, S. (2005). *Univariate Discrete Distributions*, 3rd edition, Wiley. MR2163227

[15] Katsis, A. (2001). Calculating the optimal sample size for binomial populations. *Communications in Statistics – Theory and Methods*, **30**, 665–678. MR1863030

[16] Klaschka, J. (2010). On calculation of Blaker's binomial confidence limits. COMPSTAT'10.

[17] Krishnamoorthy, K., Peng, J. (2007). Some properties of the exact and score methods for binomial proportion and sample size calculation. *Communications in Statistics – Simulation and Computation*, **36**, 1171–1186. MR2415711

[18] M'Lan, C.E., Joseph, L., Wolfson, D.B. (2008). Bayesian sample size determination for binomial proportions. *Bayesian Analysis*, **3**, 269–296. MR2407427

[19] Newcombe, R.G. (2011). Measures of location for confidence intervals for proportions. *Communications in Statistics – Theory and Methods*, **40**, 1743–1767. MR2781501

[20] Newcombe, R.G. (2012). *Confidence Intervals for Proportions and Related Measures of Effect Size*. Chapman & Hall. MR2986377

[21] Newcombe, R.G., Nurminen, M.M. (2011). In defence of score intervals for proportions and their differences. *Communications in Statistics – Theory and Methods*, **40**, 1271–1282. MR2771780

[22] Piegorsch, W.W. (2004). Sample sizes for improved binomial confidence intervals. *Computational Statistics & Data Analysis*, **46**, 309–316. MR2062050

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

[24] Staicu, A.-M. (2009). Higher-order approximations for interval estimation in binomial settings. *Journal of Statistical Planning and Inference*, **139**, 3393–3404. MR2549089

[25] Sterne, T.H. (1954). Some remarks on confidence or fiducial limits. *Biometrika*, **41**, 275–278. MR0062387

[26] Sullivan, A.K., Raben, D., Reekie, J., Rayment M., Mocroft, A., et al. (2013). Feasibility and effectiveness of indicator condition-guided testing for HIV: results from HIDES I (HIV Indicator Diseases across Europe Study). *PLoS ONE*, **8**, e52845.

[27] Thulin, M. (2014). Coverage-adjusted confidence intervals for a binomial proportion. *Scandinavian Journal of Statistics*, **41**, 291–300.

[28] Thulin, M. (2014). On split sample and randomized confidence intervals for binomial proportions. *Statistics and Probability Letters*, **92**, 65–71.

[29] Vos, P.W., Hudson, S. (2005). Evaluation criteria for discrete confidence intervals. *The American Statistician*, **59**, 137–142. MR2133560

[30] Vos, P.W., Hudson, S. (2008). Problems with binomial two-sided tests and the associated confidence intervals. *Australian & New Zealand Journal of Statistics*, **50**, 81–89. MR2414657

[31] Ward, L.G., Heckman, M.G., Warren, A.I., Tran, K. (2013). Dosing accuracy of insulin aspart FlexPens after transport through the pneumatic tube system. *Hospital Pharmacy*, **48**, 33–38.

[32] Wei, L., Hutson, A.D. (2013). A comment on sample size calculations for binomial confidence intervals. *Journal of Applied Statistics*, **40**, 311–319. MR3047150

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