Sample size for estimating a binomial proportion: Comparison of different methods

Article *in* Journal of Applied Statistics · November 2012

DOI: 10.1080/02664763.2012.713919

CITATIONS
21

READS
3,164

4 authors, including:

M. Rosário Oliveira
Technical University of Lisbon
36 PUBLICATIONS **665** CITATIONS
SEE PROFILE

Cláudia Pascoal
Technical University of Lisbon
5 PUBLICATIONS **158** CITATIONS
SEE PROFILE

Ana M Pires
Technical University of Lisbon
53 PUBLICATIONS **722** CITATIONS
SEE PROFILE

Journal of Applied Statistics
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/cjas20

Sample size for estimating a binomial proportion: comparison of different methods

Luzia Gonçalves [a] , M. Rosário de Oliveira [b] , Cláudia Pascoal [b] & Ana Pires [b]

[a] CEAUL and Unidade de Saúde Pública Internacional e Bioestatística, Instituto de Higiene e Medicina Tropical, Universidade Nova de Lisboa , Rua da Junqueira 100, 1349-008 , Lisboa , Portugal

[b] CEMAT and Departamento de Matemática, Instituto Superior Técnico , Universidade Técnica de Lisboa , Avenida Rovisco Pais, 1049-001 , Lisboa , Portugal

Published online: 10 Aug 2012.

To cite this article: Luzia Gonçalves , M. Rosário de Oliveira , Cláudia Pascoal & Ana Pires (2012) Sample size for estimating a binomial proportion: comparison of different methods, Journal of Applied Statistics, 39:11, 2453-2473, DOI: 10.1080/02664763.2012.713919

To link to this article: http://dx.doi.org/10.1080/02664763.2012.713919

Journal of Applied Statistics
Vol. 39, No. 11, November 2012, 2453–2473

Sample size for estimating a binomial proportion: comparison of different methods

Luzia Gonçalves[a]*, M. Rosário de Oliveira[b], Cláudia Pascoal[b] and Ana Pires[b]

[a]*CEAUL and Unidade de Saúde Pública Internacional e Bioestatística, Instituto de Higiene e Medicina Tropical, Universidade Nova de Lisboa, Rua da Junqueira 100, 1349-008 Lisboa, Portugal;* [b]*CEMAT and Departamento de Matemática, Instituto Superior Técnico, Universidade Técnica de Lisboa, Avenida Rovisco Pais, 1049-001 Lisboa, Portugal*

(*Received 7 July 2011; final version received 17 July 2012*)

The poor performance of the Wald method for constructing confidence intervals (CIs) for a binomial proportion has been demonstrated in a vast literature. The related problem of sample size determination needs to be updated and comparative studies are essential to understanding the performance of alternative methods. In this paper, the sample size is obtained for the Clopper–Pearson, Bayesian (Uniform and Jeffreys priors), Wilson, Agresti–Coull, Anscombe, and Wald methods. Two two-step procedures are used: one based on the expected length (EL) of the CI and another one on its first-order approximation. In the first step, all possible solutions that satisfy the optimal criterion are obtained. In the second step, a single solution is proposed according to a new criterion (e.g. highest coverage probability (CP)). In practice, it is expected a sample size reduction, therefore, we explore the behavior of the methods admitting 30% and 50% of losses. For all the methods, the ELs are inflated, as expected, but the coverage probabilities remain close to the original target (with few exceptions). It is not easy to suggest a method that is optimal throughout the range $(0, 1)$ for p. Depending on whether the goal is to achieve CP approximately or above the nominal level different recommendations are made.

Keywords: binomial proportion; confidence intervals; coverage probability; expected length; sample size

1. Introduction

Sample size is a critical aspect of the design of a study. In biomedical research, even for estimating a simple binomial proportion, using a confidence interval (CI), we find several difficulties to obtain the number of individuals suitable for a desired degree of precision. In the medical literature, some authors have emphasized that CIs should be used instead of hypothesis tests and corresponding p-values [15,31]. Newcombe reinforces [23] this idea, considering that CIs are presented on the same scale of measurement as opposed to the probabilistic abstraction of p-values. On the other hand, in the statistical literature, CIs for a proportion have been the subject of several recent

*Corresponding author. Email: luziag@ihmt.unl.pt

ISSN 0266-4763 print/ISSN 1360-0532 online

http://dx.doi.org/10.1080/02664763.2012.713919
http://www.tandfonline.com

studies, and the most widely taught Wald method for constructing a CI for a proportion, based on the asymptotic normality has been questioned. Other alternative methods are (re-)emerging and consequently the sample size calculation needs to be updated.

There is an extensive work about methods for constructing intervals for a binomial proportion. Only in the last two decades, several authors presented comparative studies and gave useful recommendations about the best methods [2,6,7,23,27,30,33]. However, it is not easy to identify a single method that is (uniformly) the best in a wide range of plausible settings since all of them have some deficiencies [19]. Most comparative studies assess the performance of the methods by the coverage probability (CP) and the expected length (EL). Vos and Hudson [34] added two other criteria (p-confidence and p-bias) and concluded that good methods, as measured by CP and EL, need not perform well in terms of p-confidence and p-bias. Newcombe [24] presented a new criterion based on the idea of interval location. Nevertheless, we compare the methods only by their CP and EL.

Krishnamoorthy and Peng [19] pointed out that the important related problem of sample size determination for computing CIs with a specified 'precision' has not been well addressed in literature. They refer to the lack of comparative studies, especially for sample sizes related to interval estimation and one-sided hypothesis testing. As a contribution to overcome this situation, these authors compare the sample sizes required to construct 95% CIs with several 'levels of precision' (note that those authors define 'precision' as half of the EL of the CIs) and present tables for p varying between 0.05 and 0.5 with an increment of 0.05 ($p = 0.05(0.05)0.5$) for Clopper–Pearson and Wilson (or score) methods.

Fosgate [14] proposes a sample size algorithm based on the exact mid-P method of CI estimation. Piegorsch [29] uses Agresti–Coull, Wilson and Jeffreys intervals to deduce the sample size. Liu and Bailey [21] explore the arcsine method and its modification due to Anscombe [4] to tabulate the sample size. Kron [20] presents sample size tables for a small binomial proportion, taking into account an upper Clopper–Pearson confidence bound. In the context of safety of a diagnostic test or therapeutic intervention, Jovanovic and Zalenski [18] also explore the one-sided exact CI to obtain the sample size to deal with rare adverse events.

The role of Bayesian Statistics for sample size determination is well described in several papers [12,16,17,22]. M'Lan *et al.* [22] examine several analytic and computational methods to deal with the sample size determination for a binomial proportion, using six different Bayesian criteria.

In this paper, we present two-step procedures to determine the sample size required for interval estimation of a binomial proportion with a given EL. A comparative study is also performed with seven distinct methods for interval estimation, providing guidelines for their application in practical studies. The paper is organized as follows. In Section 2, some notation and the seven selected methods (or variants) for constructing CIs are introduced. When necessary, the methods are modified to guarantee that the lower and upper limits are contained in [0,1]. This modification is important when dealing with proportions near 0 (e.g. rare prevalence) or 1 (e.g. high sensitivity or specificity of a diagnostic test) in epidemiological and clinical applications [13]. In Section 3, we present two-step procedures to determine the sample size, one based on the EL of the CI and another one on its first-order approximation. In the first step, all possible solutions that verify the optimal criterion are obtained. In the next step, the solution is chosen according to a new criterion. Among four criteria, the highest CP criterion is explored in more detail. For practical purposes, we add two tables with the sample sizes required for a given EL and confidence level. Additionally, tables with the coverage probabilities attained with those sample sizes are presented. The corresponding computer programme in R is available upon request. In many real situations, the actual sample size, achieved after the data collection, is often smaller than the planned optimal size, due to inevitable losses in the sampling and data collection processes. This practical problem is addressed in Section 3.5, where we assess the performance of the methods studied, considering a reduction in the optimal sample size. Finally, in Section 4, we discuss the results.

2. CIs for a binomial proportion

2.1 *Notation and selected methods*

Let us consider that a random sample of size n is drawn from a large or infinite population and that X is the number of successes ($0 \leq X \leq n$). Let p be the unknown proportion of the events of interest in the population. A two-sided CI with nominal confidence level $100 \times (1-\alpha)\%$, fixed in advance, and represented by $[L(X); U(X)]$ can be obtained by several methods.

In this paper, we consider seven methods (or variants):

(1) Clopper–Pearson exact method,
(2) Two Bayesian variants with non-informative priors – Uniform and Jeffreys,
(3) Wilson or score method,
(4) Agresti–Coull or Add-4 method,
(5) Arcsine method with Anscombe's continuity correction,
(6) Wald method,

whose lower, $L_i(X)$, and upper, $U_i(X)$, $i = 1, 2, \ldots, 7$, bounds are summarized in Table 1. To avoid problems on the boundary outcomes (especially for $X = 0$ and $X = n$), the lower and upper bounds are truncated, such that $L_i(0) = 0$ and $U_i(n) = 1$ [25,30,33].

These methods were chosen because they have been identified as having good properties by several authors [6,7,30]. According to Pires and Amado [30], the methods can be classified as strictly conservative (Clopper–Pearson and Anscombe), correct on average (Bayesian with uniform prior, Wilson and Agresti–Coull), and other (Wald). In [26], it is verified that the Bayesian method with Jeffreys prior is also correct on average. Further information is given in Table 1 and in the appendix.

2.2 *CP and EL*

Previous studies have compared the performance of the different methods using mainly two criteria: CP and EL. These quantities are functions of n and p, given, respectively, by

$$\mathrm{CP}_i(n,p) = \sum_{j=0}^{n} \binom{n}{j} p^j (1-p)^{n-j} I_{[L_i(j),U_i(j)]}(p) \tag{1}$$

(with $I_{[a,b]}(x) = 1$ if $x \in [a,b]$ and $I_{[a,b]}(x) = 0$, otherwise) and

$$\mathrm{EL}_i(n,p) = \sum_{j=0}^{n} \binom{n}{j} p^j (1-p)^{n-j} (U_i(j) - L_i(j)), \tag{2}$$

where i denotes the method (in our case, $i = 1, 2, \ldots, 7$).

A CI may or may not contain the actual value of the parameter, but the CP of a random interval $[L_i(X), U_i(X)]$, given by Equation (1), should be $1-\alpha$ or, at least, close to $1-\alpha$. And if two methods have similar coverage probabilities, the one with the smallest EL is preferred. The width or length of a CI computed for a particular sample with $X = x$, that is, the difference between the upper and lower limits, $U_i(x) - L_i(x)$, will be denoted by $\Delta_i(x)$.

Since several papers [2,6,7,30,33] have addressed exhaustively these and other criteria, providing graphical illustrations in several situations, we only add some notes that are important to understand the results of the next section.

Table 1. Lower – $L_i(X)$ – and upper – $U_i(X)$ – bounds of a $100 \times (1-\alpha)\%$ confidence level for a two-sided CI for p, using the seven selected methods, $i = 1, \ldots, 7$.

Method number of successes (X)	Lower limit $L_i(X)$	Upper limit $U_i(X)$
1. Clopper–Pearson		
$X = 0$	0	$1 - \left(\frac{\alpha}{2}\right)^{1/n}$
$0 < X < n$	$\text{Beta}_{\alpha/2}(X, n-X+1)$	$\text{Beta}_{1-\alpha/2}(X+1, n-X)$
$X = n$	$\left(\frac{\alpha}{2}\right)^{1/n}$	1
2. Bayesian-U		
$X = 0$	0	$1 - \alpha^{1/(n+1)}$
$0 < X < n$	$\text{Beta}_{\alpha/2}(X+1, n-X+1)$	$\text{Beta}_{1-\alpha/2}(X+1, n-X+1)$
$X = n$	$\alpha^{1/(n+1)}$	1
3. Jeffreys		
$X = 0$	0	$1 - \left(\frac{\alpha}{2}\right)^{1/n}$
$X = 1$	0	$\text{Beta}_{1-\alpha/2}(2, n)$
$1 < X < n-1$	$\text{Beta}_{\alpha/2}(X+\frac{1}{2}, n-X+\frac{1}{2})$	$\text{Beta}_{1-\alpha/2}(X+\frac{1}{2}, n-X+\frac{1}{2})$
$X = n-1$	$\text{Beta}_{\alpha/2}(n, 2)$	1
$X = n$	$\left(\frac{\alpha}{2}\right)^{1/n}$	1
4. Wilson	$\frac{2X + z^2_{1-\alpha/2} - z_{1-\alpha/2}\sqrt{z^2_{1-\alpha/2} + 4X(1-X/n)}}{2(n + z^2_{1-\alpha/2})}$	$\frac{2X + z^2_{1-\alpha/2} + z_{1-\alpha/2}\sqrt{z^2_{1-\alpha/2} + 4X(1-X/n)}}{2(n + z^2_{1-\alpha/2})}$

(*Continued*)

Table 1. Continued.

Method number of successes (X)	Lower limit $L_i(X)$	Upper limit $U_i(X)$
5. Agresti–Coull	$\max\left\{\frac{X+2}{n+4} - z_{1-\alpha/2}\sqrt{\frac{X+2}{(n+4)^2}\left(1-\frac{X+2}{n+4}\right)}; 0\right\}$	$\min\left\{\frac{X+2}{n+4} + z_{1-\alpha/2}\sqrt{\frac{X+2}{(n+4)^2}\left(1-\frac{X+2}{n+4}\right)}; 1\right\}$
6. Anscombe		
$X = 0$	0	$\sin^2\left(\min\left\{\arcsin\sqrt{\frac{3/8+1/2}{n+3/4}} + \frac{z_{1-\alpha/2}}{2\sqrt{n+1/2}}; \frac{\pi}{2}\right\}\right)$
$0 < X < n$	$\sin^2\left(\max\left\{\arcsin\sqrt{\frac{3/8+X-1/2}{n+3/4}} - \frac{z_{1-\alpha/2}}{2\sqrt{n+1/2}}; 0\right\}\right)$	$\sin^2\left(\min\left\{\arcsin\sqrt{\frac{3/8+X+1/2}{n+3/4}} + \frac{z_{1-\alpha/2}}{2\sqrt{n+1/2}}; \frac{\pi}{2}\right\}\right)$
$X = n$	$\sin^2\left(\max\left\{\arcsin\sqrt{\frac{3/8+n-1/2}{n+3/4}} - \frac{z_{1-\alpha/2}}{2\sqrt{n+1/2}}; 0\right\}\right)$	1
7. Wald	$\max\left\{\frac{X}{n} - z_{1-\alpha/2}\sqrt{\frac{X}{n^2}\left(1-\frac{X}{n}\right)}; 0\right\}$	$\min\left\{\frac{X}{n} + z_{1-\alpha/2}\sqrt{\frac{X}{n^2}\left(1-\frac{X}{n}\right)}; 1\right\}$

Notes: z_γ and $\text{Beta}_\gamma(a,b)$ represent the γ-quantiles of the $N(0,1)$ and the $\text{Beta}(a,b)$ distributions, respectively.

The oscillation in the coverage probabilities is due to the discreteness of the binomial distribution and affects all the methods. However, for certain values of p and n, some methods present unacceptable low CP. This is the case of the Wald CI, as shown in almost all the studies [5–7,10,23,30]. Brown *et al.* [6], for instance, point out that unsatisfactory coverage probabilities may occur for some (n, p) pairs, even with large n. These are called 'unlucky' pairs, as opposed to'lucky' pairs.

The main characteristic of the Clopper–Person CI, emphasized in almost all the literature, is the excessive conservativeness. Agresti and Coull [2] even say '*we believe it is inappropriate to treat this approach as optimal for statistical practice*'. Other authors [6,14,30] express a less restrictive position and say that, depending on the practical purposes (for instance when $\mathrm{CP}(n, p) \geq 1 - \alpha$ is mandatory), this method may be appropriate. However, we shall keep in mind that for this method, the actual CP can be much larger than the nominal confidence level (Pastor [28] proved that for every $p \in]0, 1[$ and every n less than $\ln(\alpha/2)/\ln(\max(p, 1 - p))$, the CP of the Clopper–Pearson CI is larger than or equal to $1 - \alpha/2$). According to Pires and Amado [30], the Anscombe method is almost equivalent to the exact Clopper–Pearson method in terms of EL and the degree of conservativeness.

Considering both CP and EL, it is possible to make some practical recommendations. Brown *et al.* [6] recommend the Wilson and Jeffreys intervals for small samples and the Agresti–Coull interval for larger n. In terms of EL of the set Wald, Agresti–Coull, Wilson, and Jeffreys methods, those authors have concluded that the Agresti–Coull interval is always the longest, the Wilson and the Wald have similar lengths, and the Jeffreys is always the shortest.

Here, we consider two different variants of the Bayesian method, one based on the Uniform prior and another one based on the Jeffreys prior, both with modified limits (Table 1).

In the next section, we discuss the determination of the sample size necessary for a given method to produce CI with a certain EL. Some CP results, which will be helpful when discussing the merits and the pitfalls of the methods under study, are also presented.

3. Sample size determination

3.1 *Closed-form solutions*

A closed-form formula for n is a desirable property when it is necessary to explain details about the CI and the sample size determination (for instance in teaching or in consulting). Böhning and Viwatwongkasem [5] explain very well the pedagogic value of simple methods in helping the presentation of a more complex method for constructing CI.

Suppose that the goal of a certain research study is to estimate p, using a $100(1 - \alpha)\%$ CI with width Δ, previously fixed by the researcher. Despite the poor behavior of the Wald method, it is still usual, in introductory courses and in consulting, to determine the required sample size by the following simple formula derived from the Wald CI and presented in the majority of elementary statistics and epidemiology textbooks,

$$n = \left\lceil \frac{4z^2_{1-\alpha/2}p(1-p)}{\Delta^2} \right\rceil, \tag{3}$$

where $\lceil a \rceil$ is the smallest integer larger than or equal to a. In practice and from a frequentist perspective, it is necessary to have a preliminary point estimate of p ($\hat{p}$) to use in formula (3). Such an estimate can be obtained from a pilot or previous study. If no such studies are available, the conservative choice $p = \frac{1}{2}$, which maximizes (3), can be considered.

As mentioned before, the modified expressions of the Wald limits (Table 1) should be used instead of the usual formula to avoid values outside [0, 1]. Accordingly, the sample size formula

(3) should also be updated. After some calculations, the new formula is given by

$$n_7 = \begin{cases} \left\lceil \dfrac{z^2_{1-\alpha/2}p(1-p)}{(\Delta - p)^2} \right\rceil, & 0 \le p < \dfrac{\Delta}{2} \\ \left\lceil \dfrac{4z^2_{1-\alpha/2}p(1-p)}{\Delta^2} \right\rceil, & \dfrac{\Delta}{2} \le p \le 1 - \dfrac{\Delta}{2} \\ \left\lceil \dfrac{z^2_{1-\alpha/2}p(1-p)}{(\Delta - (1-p))^2} \right\rceil, & 1 - \dfrac{\Delta}{2} < p \le 1. \end{cases} \tag{4}$$

This expression takes into account the correction of inconsistencies near the boundaries when computing the CI, but the resulting sample sizes are not appropriate when p is close to zero (or one), since the associated CP are very low. It is worth noting that, if we had used formula (3) instead of formula (4) those CP would have been even worse.

We cannot guarantee that the actual width of the CI computed after observing the sample (with $X = x$) is Δ, even if the true proportion is p. Assuming we are far from the boundaries, the width of the CI obtained using Wald formula is given by $\Delta_7(x) = 2z_{1-\alpha/2}\sqrt{x/n(1-x/n)/n}$, so $\Delta_7(x)$ will only be equal to Δ in the unlikely event of $x = np$ and no rounding is necessary in formula (3). Therefore, the goal of the sample size determination should be clearly stated as: determine the sample size, n, such that, for a given p, the expected length, $\mathrm{EL}_i(n,p) = E(\Delta_i(X))$ is equal to Δ.

It is easy to verify that Equation (3) gives only an approximate solution to this problem, for $i = 7$, by considering two crude approximations,

$$\begin{aligned} \mathrm{EL}_7(n,p) = E(\Delta_7(X)) &\simeq E\left(2z_{1-\alpha/2}\sqrt{\frac{(X/n)(1-X/n)}{n}}\right) \\ &\simeq 2z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}}. \end{aligned}$$

The first approximation ignores the values of x leading to confidence limits outside $[0,1]$, in which case the length is $\min(x/n, 1-x/n) + z_{1-\alpha/2}\sqrt{x/n(1-x/n)/n}$, not $2z_{1-\alpha/2}\sqrt{x/n(1-x/n)/n}$. The second approximation is a first-order approximation to the expected value, $E(g(X)) \simeq g(E(X))$, which can be misleading for small values of n and/or p.

It is possible to obtain sample size results for the remaining methods, $i = 1, \ldots, 6$, using the first-order approximation of the expected value. Note that $E(U_i(X) - L_i(X)) \simeq U_i(E(X)) - L_i(E(X)) = U_i(np) - L_i(np)$. Thus, n is such that $U_i(np) - L_i(np) = \Delta$. In the case of the Wilson interval, solving the previous equation also leads to a closed-formula for n, given by

$$n_4 = \left\lceil \frac{-z^2_{1-\alpha/2}(\Delta^2 - 2p(1-p)) + z^2_{1-\alpha/2}\sqrt{(\Delta^2 - 2p(1-p))^2 - \Delta^2(\Delta^2 - 1)}}{\Delta^2} \right\rceil. \tag{5}$$

Other authors [19,29] presented this formula with $\Delta = 2d$, naming d 'the precision of the CI'. It is important to remark at this point that asymmetric CIs cannot be interpreted as $\hat{p} \pm d$, where $\hat{p}$ is a point estimate of p [24]. This interpretation is valid only for symmetric intervals. All the methods enumerated in Section 2.1 produce, in general, asymmetric intervals (even the Wald and Agresti–Coull, if p is close enough to the boundaries). If we want to measure the precision of a generic interval $[L, U]$, not necessarily symmetric, relatively to a point estimate $\hat{p}$, we can define

the upper and lower precisions as

$$d_U = U - \hat{p} \quad \text{and} \quad d_L = \hat{p} - L, \tag{6}$$

respectively. The overall precision is then $d = \Delta/2 = (d_L + d_U)/2$. Obviously, in the case of symmetric intervals, we have that $d_U = d_L = d$.

For the remaining methods, inversion of the CI expressions does not produce closed-form formulae. Therefore, the determination of the sample size necessary to achieve the desired EL requires more sophisticated algorithms and some computer programming. In the next section we describe two types of procedures (denominated 'exact' and 'first-order approximation') which can be used to find n such that $E(\Delta_i(X)) = \Delta$. The proposed algorithms have been implemented in *Mathematica 6.0* (Wolfram Research, Inc., 2008) and `R` (R Development Core Team, 2007).

3.2 *Procedure based on the exact EL*

Our goal is to obtain n such that

$$\mathrm{EL}_i(n, p) = \Delta \quad (i = 1, \ldots, 7), \tag{7}$$

where p is an anticipated value of the proportion under study. These equations do not have neither a closed-form nor an integer solution. However, it is always possible to find an integer minimizing $|\mathrm{EL}_i(n, p) - \Delta|$, and in most cases to find n such that $\mathrm{EL}_i(n, p) \simeq \Delta$.

First step: Given α, Δ and a tolerance ξ, determine the integer values verifying

$$|\mathrm{EL}_i(n, p) - \Delta| \leq \xi \quad (i = 1, \ldots, 7). \tag{8}$$

The ξ can be seen as a tolerance value (we have used $\xi = 10^{-4}$) fixed by the investigator. The choice of this extra parameter does not impose additional difficulties since it is just an admissible discrepancy between the target, Δ, and a possible value for the EL, $\mathrm{EL}_i(np)$. If there is no integer solution for Equation (8), we increase ξ.

Second step: If there is more than one solution ($k > 1$) to Equation (8), denoted by $S_\xi = \{n_1, n_2, \ldots n_k\}$, choose n according to one of the following criteria:

(i) $n = \mathrm{argmax}_{l \in S_\xi}\{\mathrm{CP}_i(l, p)\}$
(ii) $n = \mathrm{argmin}_{l \in S_\xi}\{\mathrm{CP}_i(l, p)\}$
(iii) $n = \min\{S_\xi\}$
(iv) $n = \mathrm{argmin}_{l \in S_\xi}\{|\mathrm{CP}_i(l, p) - (1 - \alpha)|\}$.

The overall solution to this optimization problem can always be found by exhaustive search, i.e. trying successively $n = 1, 2, 3, \ldots$, but this can be very time-consuming. We have implemented a fast algorithm which uses the numerical routine `uniroot` of the R software, and the result of Equation (3) as the initial solution. Krishnamoorthy and Peng [19] also study the sample size determination based on the exact EL but consider only the Clopper–Pearson and Wilson methods, for which they find the smallest n such that $\mathrm{EL}(n, p) \leq \Delta$ and $\mathrm{CP}(n, p) \geq 1 - \alpha$.

3.3 *Procedure based on a first-order approximation*

The sample size determination based on the first-order approximation of the EL (which leads to the closed-form formulae related to the Wald and Wilson CI methods, given in the previous

subsection) was also used by other authors [14,19,29] to obtain approximate sample size solutions for CI methods other than the Wald or the Wilson. Recall that the aim is to find n such that

$$U_i(np) - L_i(np) = \Delta \quad (i = 1, 2, 3, 5, 6). \tag{9}$$

Note that the lower, $L_i(x)$, and upper, $U_i(x)$, limits associated to ith method are still well defined even if x is not an integer, the usual case when $x = np$. The exception is the Anscombe method ($i = 6$) when $0 < x < 1/8$. In this case, we consider $L_6(x) = 0$ and keep $U_6(x)$ unchanged, that is, as defined in Table 1 for $0 < x < n$.

Given that n must be an integer, Equation (9) does not usually have a solution. Therefore, a procedure similar to the one described in the previous subsection can also be implemented in this case. The algorithm starts to obtain the integer value (or values), $S_\xi = \{n_1, n_2, \ldots, n_k\}$, verifying

$$\mid U_i(np) - L_i(np) - \Delta \mid \leq \xi \quad (i = 1, 2, 3, 5, 6). \tag{10}$$

If condition (10) has no integer solution, the value of ξ has to be increased. When there are several solutions ($k > 1$), in a second step, we choose $n \in S_\xi$ according to one of the criteria described in the previous subsection. Note that for the strictly conservative methods, criteria (ii) and (iv) lead to the same solution since $\mathrm{CP}(n, p) \geq (1 - \alpha)$. Other authors [19,29] choose n as the first integer such that

$$U_i(np) - L_i(np) \leq \Delta. \tag{11}$$

3.4 *Results*

In this section, we will illustrate the two-step procedures to determine the sample sizes needed to attain an EL of 0.05 for a 95% CI provided by each of the seven methods under study. As pointed out by Pires and Amado [30], the results presented in this paper are theoretical (apart from numerical errors) and were not obtained by simulation.

Considering the exact procedure with each criteria listed before, in the second step, the results are similar or even equal when $p < 0.01$. The larger differences (11 or 12 units) in sample sizes occur when p tends to 0.50. For the seven methods and $p = 0.01(0.01)0.5$, the differences in sample sizes between each pair of criteria is shown in Figure 1 using boxplots. The corresponding coverage probabilities and ELs (data not shown) were also compared.

The highest CP criterion –(i)– almost always produces smaller sample size than criteria (ii) and (iv) for all methods. When compared with the third criterion, the results are very similar, particularly for Jeffreys, Wilson and Anscombe methods. For all methods, a difference of more than five units is highlighted as an outlier in Figure 1(i)–(iii). The third criterion is a valid choice for the strictly conservative methods (Clopper–Pearson and Anscombe) because the coverage probabilities are always above the nominal confidence level. However, for the Wald and the correct on average methods, some cautions are needed if the corresponding coverage probabilities are unsatisfactory. This issue also applies to criterion (ii) for the previous five methods. Criterion (iv) requires a larger sample size than criterion (i) and (iii), particularly for the strictly conservative methods. As mentioned before, by definition, (ii) and (iv) produce equal sample size for Clopper–Pearson and Anscombe methods.

From now on, we will focus on results using the highest CP criterion –(i)– in the second step to determine the sample sizes using the exact and the first-order approximation procedures. The results are summarized in Table 2 for $p = 0.001(0.001)0.009$ and in Table 3 for $p = 0.01(0.01)0.5$, considering a nominal confidence level of 0.95 and an EL of 0.05. It should be emphasized that the results of the various methods cannot be compared without taking into account the corresponding CP given in Tables 4 and 5.

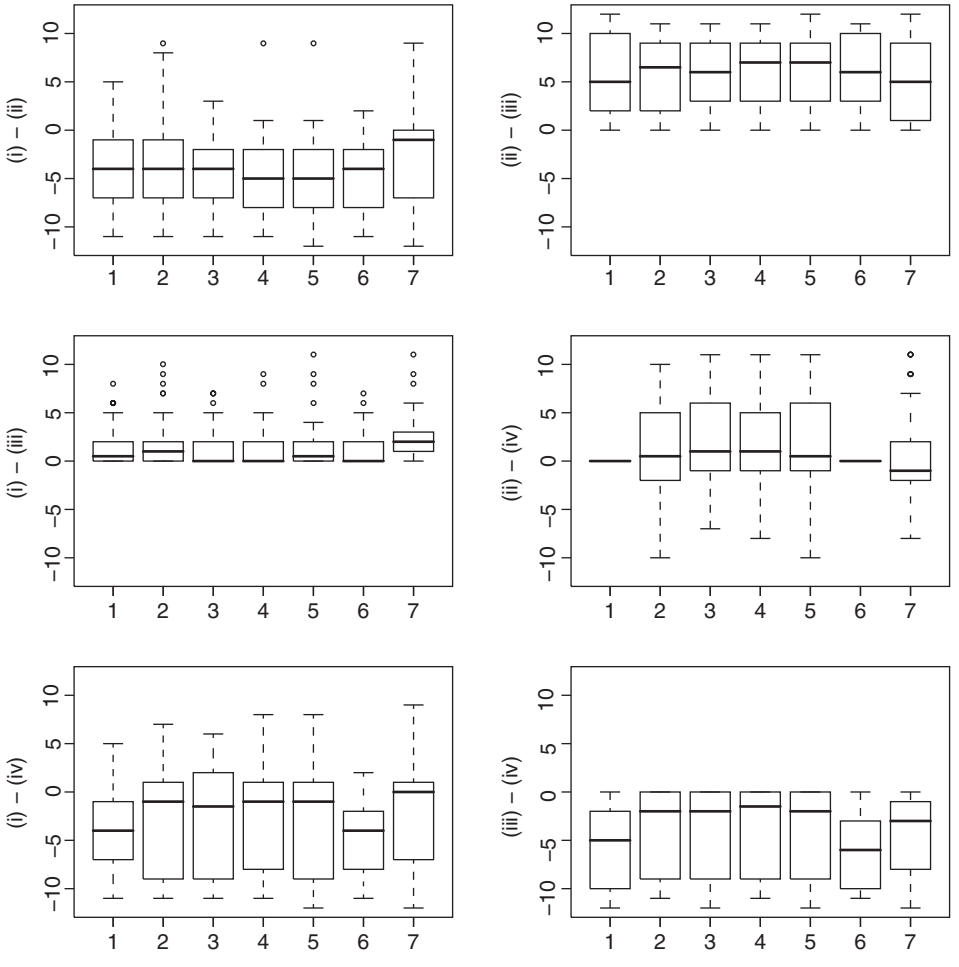

Figure 1. Boxplots with the differences in sample sizes between each pair of criteria – (i), (ii), (iii) and (iv) – needed to attain an EL of 0.05 by a 95% CI for each method (1. Clopper–Pearson, 2. Bayesian-U, 3. Jeffreys, 4. Wilson, 5. Agresti–Coull, 6. Anscombe, and 7. Wald), using the exact procedure for $p \geq 0.01$.

The figures reported in parentheses, in Tables 2–5, are the differences between the results obtained with the first-order approximation and the results obtained with the exact procedure (multiplied by 1000, in the cases of Tables 4 and 5). Although the first-order approximation may provide a crude estimate of the exact value, we decided to show these results since the traditional formulas presented in the literature follow this approach and there are cases where the exact procedure does not have a solution (e.g. Wald method in Table 2).

The differences between the two procedures (exact and first-order approximation) for p less than 0.01 are considerable for the Bayesian-U and the modified Jeffreys variants. But, as we can see in Table 3, these differences tend to decrease, for all the methods, as the value of p increases. An exception occurs for the Wilson and Wald methods which retain a difference of up to seven units even for $p > 0.2$. However, it is important to note that these correspond precisely to the first-order approximation solutions given in the closed-form, and which are not subject to any refinement based on the choice with the best CP, as for the other methods. In the rest of this paper, all the comments refer to the exact procedure.

Table 2. Sample sizes needed to compute a nominal 95% CI with $\Delta = 0.05$, obtained using the exact EL, for seven CI methods and $p = 0.001(0.001)0.009$ (in parentheses: difference between the values obtained using the first-order approximation to the EL and the exact one) and considering the highest CP criterion in the second step.

p	Clopper–Pearson	Bayesian-U	Jeffreys	Wilson	Agresti–Coull	Anscombe	Wald[a]
0.001	75 (0)	60 (14)	73 (−21)	75 (1)	94 (0)	74 (1)	NA (2)
0.002	77 (2)	63 (14)	75 (−20)	78 (2)	97 (1)	77 (1)	NA (4)
0.003	80 (2)	66 (14)	76 (−18)	81 (2)	100 (1)	80 (2)	NA (6)
0.004	84 (2)	69 (14)	78 (−17)	83 (3)	104 (1)	83 (2)	NA (8)
0.005	87 (3)	73 (13)	79 (−15)	86 (4)	107 (2)	87 (2)	NA (10)
0.006	91 (3)	76 (14)	81 (−13)	89 (4)	111 (2)	90 (3)	NA (12)
0.007	94 (4)	80 (13)	83 (−11)	93 (4)	115 (3)	94 (3)	NA (15)
0.008	98 (5)	84 (13)	85 (−9)	96 (4)	119 (4)	98 (4)	NA (18)
0.009	102 (6)	88 (13)	88 (−8)	99 (5)	123 (5)	102 (5)	NA (21)

Notes: [a]Exceptionally in this column, the figures in parentheses are the sample sizes (not the differences) based on the first-order approximation. NA, not available.

Analyzing more detail in Tables 2 and 3, it is clear that the Agresti–Coull method requires more experimental units until $p \leq 0.03$. After this point, the Clopper–Pearson and Anscombe methods always give similar results and sample sizes larger than the remaining methods. For $p \leq 0.009$, the Wilson, the strictly conservative methods lead to very similar sample sizes. For larger p, especially for $p > 0.1$, the values for the Wilson method become distant from those provided by the last two methods and approach the results for the two Bayesian methods. Comparing these latter two methods, we can see that, for small values of p, the sample sizes required by the modified Jeffreys CI are higher than those required by the Bayesian-U CI. The difference decreases until $p = 0.01$, vanishing for this value of p. When $0.02 \leq p \leq 0.05$, the Bayesian-U interval needs 14, 13, 10, and 8 units more than the Jeffreys interval. For $p \geq 0.11$, those two methods often lead to the same or to very similar values of n (a maximum difference of 8 units occurs at $p = 0.45$).

For $p \geq 0.15$, the sample sizes suggested by the Agresti–Coull method are very similar to the ones required by the Bayesian-U and the modified Jeffreys methods. Especially for $0.22 \leq p \leq 0.48$ and except $p = 0.26$, the values of the sample size and the CP corresponding to the Agresti–Coull and to the Bayesian-U methods are exactly the same (Tables 3 and 5).

For the Wald method ($i = 7$) and small values of p ($p < 0.02$), the results of the exact procedure reported in the tables are not available (NA) because it is not possible to find an integer value verifying $|\mathrm{EL}_7(n, p) - 0.05| \leq \xi$ for any reasonable choice of ξ. The reason is that the EL of the Wald CI has a singular behavior. Unlike all the other six methods for which the EL, when holding p fixed, is a strictly decreasing function of $n = 1, 2, \ldots$ (for every $p \in]0, 1[$), in the case of the Wald CI, the EL starts at zero, then increases up to a certain value $n^* > 1$, and then decreases. In fact, it is easy to verify that (for every $p \in]0, 1[$) $\mathrm{EL}_7(1, p) \equiv 0$, $\mathrm{EL}_7(n, p) > 0$, for every $n > 1$, and $\lim_{n\to\infty} \mathrm{EL}_7(n, p) = 0$.

The discrepancies between the results provided by the Wald method ($\mathsf{n}_{\mathsf{Wld}}$) and the results for the remaining methods (Clopper–Pearson: n_{CP}; Bayesian-U: n_{BU}; modified Jeffreys: $\mathsf{n}_{\mathsf{Jef}}$; Wilson: $\mathsf{n}_{\mathsf{Wls}}$; Agresti–Coull: n_{AC}; and Anscombe: $\mathsf{n}_{\mathsf{Asc}}$) are evident when $p \leq 0.05$. However, for $p \geq 0.05$, the distinction between the Wald method and the four correct on average methods is not so clear.

Figure 2 shows boxplots of the differences between the corresponding columns in Table 3 ($\mathsf{n}_{\mathsf{Wld}}$–$\mathsf{n}_{\mathsf{BU}}$, $\mathsf{n}_{\mathsf{Wld}}$–$\mathsf{n}_{\mathsf{Jef}}$, $\mathsf{n}_{\mathsf{Wld}}$–$\mathsf{n}_{\mathsf{Wls}}$, and $\mathsf{n}_{\mathsf{Wld}}$–$\mathsf{n}_{\mathsf{AC}}$). We can see that the modified Jeffreys method gives the closest results to the Wald method ($\mathsf{n}_{\mathsf{Wld}}$–$\mathsf{n}_{\mathsf{Jef}}$ varies between −2 and 7 units, except

Table 3. Sample sizes needed to compute a nominal 95% CI with $\Delta = 0.05$, obtained using the exact EL, for seven CI methods and $p = 0.01(0.01)0.5$ (in parentheses: difference between the values obtained, using the first-order approximation to the EL and the exact one) and considering the highest CP criterion in the second step.

p	Clopper–Pearson	Bayesian-U	Jeffreys	Wilson	Agresti–Coull	Anscombe	Wald
0.01	107 (6)	92 (13)	90 (−6)	103 (5)	127 (6)	107 (6)	NA (24)[a]
0.02	158 (9)	140 (8)	126 (7)	145 (7)	172 (6)	159 (9)	48 (36)
0.03	215 (7)	191 (7)	178 (8)	194 (7)	218 (4)	217 (7)	162 (17)
0.04	272 (6)	244 (6)	234 (6)	246 (6)	266 (3)	273 (6)	229 (8)
0.05	328 (4)	297 (4)	289 (5)	298 (6)	314 (4)	329 (5)	287 (5)
0.06	384 (2)	349 (6)	343 (5)	351 (5)	364 (3)	383 (6)	343 (4)
0.07	435 (6)	401 (4)	399 (1)	402 (5)	413 (3)	439 (2)	397 (4)
0.08	487 (5)	452 (3)	449 (3)	453 (5)	462 (3)	491 (1)	450 (3)
0.09	538 (3)	505 (0)	501 (1)	505 (3)	511 (4)	539 (5)	500 (4)
0.10	589 (1)	553 (1)	549 (4)	552 (5)	561 (0)	589 (2)	552 (2)
0.11	636 (6)	599 (2)	600 (−1)	600 (4)	608 (0)	637 (2)	601 (1)
0.12	688 (0)	649 (0)	646 (4)	649 (2)	651 (5)	687 (0)	648 (2)
0.13	729 (5)	691 (2)	690 (2)	691 (5)	697 (1)	734 (0)	691 (5)
0.14	775 (1)	736 (2)	734 (3)	736 (5)	740 (6)	775 (2)	736 (5)
0.15	821 (0)	783 (0)	781 (3)	783 (1)	783 (7)	818 (2)	781 (3)
0.16	864 (0)	822 (6)	820 (3)	822 (4)	828 (0)	863 (0)	827 (0)
0.17	901 (6)	862 (5)	865 (−2)	862 (5)	867 (0)	901 (5)	867 (1)
0.18	940 (5)	903 (0)	901 (3)	903 (4)	904 (4)	945 (0)	902 (6)
0.19	980 (0)	945 (0)	943 (3)	945 (0)	945 (0)	980 (5)	944 (2)
0.20	1021 (0)	977 (1)	981 (2)	977 (5)	983 (0)	1018 (8)	981 (3)
0.21	1052 (3)	1013 (1)	1013 (1)	1013 (5)	1015 (1)	1055 (0)	1015 (5)
0.22	1089 (0)	1049 (0)	1048 (2)	1049 (4)	1049 (2)	1089 (0)	1051 (4)
0.23	1121 (4)	1083 (0)	1082 (1)	1083 (4)	1083 (4)	1125 (0)	1087 (2)
0.24	1154 (1)	1121 (0)	1114 (1)	1113 (6)	1121 (0)	1154 (1)	1116 (6)
0.25	1185 (4)	1146 (0)	1145 (2)	1146 (4)	1146 (1)	1185 (3)	1147 (6)
0.26	1215 (9)	1183 (0)	1182 (1)	1174 (6)	1179 (0)	1220 (0)	1183 (0)
0.27	1244 (3)	1204 (3)	1206 (−1)	1204 (5)	1204 (3)	1244 (3)	1208 (4)
0.28	1272 (0)	1231 (2)	1232 (2)	1231 (6)	1231 (2)	1272 (1)	1235 (5)
0.29	1301 (0)	1257 (7)	1263 (2)	1257 (6)	1257 (7)	1301 (0)	1261 (5)
0.30	1323 (3)	1283 (0)	1283 (1)	1283 (5)	1283 (1)	1323 (3)	1288 (3)
0.31	1346 (1)	1308 (0)	1307 (2)	1308 (4)	1308 (0)	1349 (0)	1309 (6)
0.32	1369 (2)	1329 (2)	1330 (0)	1328 (7)	1329 (2)	1371 (0)	1333 (5)
0.33	1391 (0)	1350 (1)	1352 (0)	1350 (6)	1350 (1)	1391 (1)	1355 (4)
0.34	1411 (1)	1370 (1)	1371 (1)	1370 (6)	1370 (1)	1411 (1)	1375 (5)
0.35	1429 (3)	1390 (0)	1390 (1)	1390 (5)	1390 (0)	1430 (2)	1392 (7)
0.36	1447 (2)	1408 (0)	1408 (1)	1406 (7)	1408 (0)	1449 (0)	1410 (7)
0.37	1466 (0)	1423 (3)	1425 (1)	1423 (7)	1423 (3)	1466 (0)	1427 (6)
0.38	1484 (0)	1442 (0)	1439 (4)	1442 (3)	1442 (0)	1481 (0)	1446 (3)
0.39	1501 (0)	1463 (0)	1460 (−5)	1452 (7)	1463 (0)	1501 (0)	1467 (−4)
0.40	1506 (2)	1466 (0)	1466 (1)	1466 (6)	1466 (0)	1508 (0)	1470 (6)
0.41	1524 (0)	1478 (0)	1483 (1)	1478 (6)	1478 (0)	1524 (0)	1482 (5)
0.42	1528 (1)	1488 (0)	1488 (1)	1488 (6)	1488 (0)	1529 (1)	1492 (6)
0.43	1543 (0)	1499 (0)	1501 (−1)	1501 (2)	1499 (0)	1543 (0)	1503 (4)
0.44	1546 (0)	1506 (0)	1506 (1)	1506 (5)	1506 (0)	1546 (1)	1510 (5)
0.45	1552 (1)	1520 (0)	1512 (1)	1520 (−2)	1520 (0)	1553 (1)	1524 (−2)
0.46	1563 (0)	1524 (0)	1524 (1)	1524 (−1)	1524 (0)	1563 (0)	1528 (−1)
0.47	1564 (0)	1521 (5)	1523 (1)	1521 (7)	1521 (0)	1564 (0)	1525 (7)
0.48	1565 (2)	1525 (0)	1525 (1)	1525 (6)	1525 (0)	1565 (2)	1529 (6)
0.49	1567 (2)	1526 (1)	1527 (1)	1526 (7)	1525 (1)	1567 (2)	1529 (7)
0.50	1568 (0)	1528 (0)	1528 (1)	1526 (7)	1526 (2)	1568 (2)	1530 (7)

Notes: [a]Sample size obtained with the first-order approximation. NA, not available.

Table 4. Actual coverage probabilities of the nominal 95% CIs computed with the sample sizes given in Table 1 (in parentheses: difference between the values obtained using the sample size based on the first-order approximation to the EL and the sample size based on the exact EL, multiplied by 1000).

p	Clopper–Pearson	Bayesian-U	Jeffreys	Wilson	Agresti–Coull	Anscombe	Wald[a]
0.001	0.997 (0)	0.942 (13)	0.930 (−19)	0.928 (1)	0.996 (0)	0.997 (0)	NA (0.002)
0.002	0.989 (0)	0.882 (25)	0.990 (−4)	0.855 (3)	0.999 (0)	0.999 (0)	NA (0.008)
0.003	0.976 (−22)	0.820 (34)	0.978 (−9)	0.975 (1)	0.996 (0)	0.998 (0)	NA (0.018)
0.004	0.995 (0)	0.969 (13)	0.961 (−14)	0.956 (3)	0.991 (0)	0.995 (0)	NA (0.032)
0.005	0.990 (1)	0.948 (17)	0.940 (−19)	0.931 (6)	0.983 (−1)	0.990 (0)	NA (0.049)
0.006	0.982 (1)	0.923 (25)	0.987 (50)	0.900 (−81)	0.970 (−1)	0.998 (0)	NA (0.070)
0.007	0.996 (1)	0.892 (−80)	0.979 (−7)	0.972 (3)	0.991 (−1)	0.996 (−1)	NA (0.100)
0.008	0.992 (2)	0.970 (13)	0.969 (−8)	0.958 (5)	0.984 (−2)	0.992 (−1)	NA (0.135)
0.009	0.986 (3)	0.954 (18)	0.954 (−10)	0.939 (7)	0.974 (−3)	0.986 (−2)	NA (0.173)

Notes: [a]Exceptionally in this column, the figures in parentheses are the actual coverage probabilities (not the differences) for the sample size based on the first-order approximation. NA, not available. $\Delta = 0.05$ and $p = 0.001(0.001)0.009$.

when $p = 0.45$, for which there is a difference of 12 units). As a curiosity note that, for $p \geq 0.25$, those differences are often (approximately 54% of the times) equal to 4 units.

Piegorsch [29] interprets this difference as the square of the 0.975 quantile of the standard normal distribution ($z^2_{0.975} = 3.84 \simeq 4$). A thorough comparison of the methods is only possible if the coverage probabilities for each optimal pair (n, p) and each method i ($i = 1, 2, \ldots, 7$) are known. Tables 4 and 5 give the coverage probabilities corresponding to the sample sizes presented in Tables 2 and 3.

Table 4 shows that the sample sizes presented in Table 2 ensure, for the Clopper–Pearson, Agresti–Coull and Anscombe methods, coverage probabilities always above the nominal level. The Bayesian-U, Jeffreys and Wilson methods show, in turn, some coverage probabilities above and some below the nominal level, but note that the Bayesian-U method may have considerable low CP for some values of p (e.g. 0.820 for $p = 0.003$). For the Wald method, since there is no solution with the exact procedure (NA), only the first-order approximation is used, but the associated coverage probabilities are very poor (last column of Table 4).

The results presented in Table 5 show some discrepancies between methods for $0.01 \leq p \leq 0.1$. The Clopper–Pearson, Agresti–Coull, and Anscombe methods always lead to CP larger than 0.95 but the Anscombe method seems better in terms of distance to the nominal confidence level. The Wald method gives particularly low values for $p = 0.02$ and 0.03. For $p > 0.1$, all the methods give coverage probabilities close to the target. Note, however, that for the strictly conservative methods, the difference $\mathrm{CP}_i(n_i, p) - 0.95$ is always positive, as expected.

As a general remark, we can say that, although the sample sizes required to achieve a fixed EL using the seven methods may be different, the corresponding CP are relatively similar.

Despite the fact that our bounds and algorithms are slightly different from those of [29], it is important to confront his findings with ours. This author compares the Wilson CI and the Bayesian CI with Jeffreys prior and points out that for low values of p, the Wilson method leads, in general, to smaller sample sizes but that the opposite occurs when p approaches 0.5. Our conclusion is slightly different, we verify that, at least for an EL of 0.05, the Wilson method leads almost always to higher values than the modified Jeffreys or Bayesian-U methods. The same author adds that, when comparing the Wald and the Agresti–Coull methods, the latter guarantees a reduction of the sample size required. In the situations explored in Tables 2 and 3, we only see this effect for p larger than 0.25. Piegorsch [29] also says that the Agresti–Coull CI tends to be slightly wider than most of the other competitors and that a smaller required sample size would be naturally associated with this effect. However, our study shows precisely the opposite, methods known to

Table 5. Actual coverage probabilities of the nominal 95% CIs computed with the sample sizes given in Table 3 (in parentheses: difference between the values obtained using the sample size based on the first-order approximation to the EL and the sample size based on the exact EL, multiplied by 1000).

p	Clopper–Pearson	Bayesian-U	Jeffreys	Wilson	Agresti–Coull	Anscombe	Wald
0.01	0.977 (17)	0.935 (−24)	0.987 (−5)	0.980 (−4)	0.961 (28)	0.995 (−1)	NA (0.214)[a]
0.02	0.985 (−5)	0.937 (−17)	0.958 (−7)	0.973 (−7)	0.977 (−5)	0.985 (−6)	0.620 (195)
0.03	0.975 (−2)	0.949 (−4)	0.952 (−1)	0.948 (−5)	0.957 (−3)	0.975 (−2)	0.863 (35)
0.04	0.971 (−1)	0.951 (15)	0.958 (−26)	0.949 (16)	0.960 (0)	0.971 (−1)	0.942 (−35)
0.05	0.970 (−1)	0.956 (−2)	0.958 (−1)	0.956 (−3)	0.962 (−2)	0.969 (0)	0.942 (3)
0.06	0.969 (0)	0.946 (11)	0.948 (−2)	0.945 (12)	0.953 (11)	0.969 (−12)	0.937 (10)
0.07	0.962 (−2)	0.950 (0)	0.950 (0)	0.950 (−1)	0.958 (−1)	0.960 (0)	0.940 (3)
0.08	0.955 (7)	0.954 (−10)	0.944 (0)	0.954 (−1)	0.952 (−1)	0.962 (0)	0.937 (8)
0.09	0.958 (0)	0.957 (0)	0.958 (−10)	0.957 (−9)	0.947 (8)	0.958 (6)	0.954 (−12)
0.10	0.961 (0)	0.953 (0)	0.954 (−1)	0.953 (−1)	0.952 (0)	0.961 (0)	0.950 (−11)
0.11	0.958 (5)	0.950 (0)	0.950 (0)	0.957 (−8)	0.956 (0)	0.957 (0)	0.947 (0)
0.12	0.960 (0)	0.954 (0)	0.954 (−7)	0.954 (0)	0.954 (−1)	0.960 (0)	0.952 (−9)
0.13	0.959 (−1)	0.953 (−1)	0.953 (−1)	0.953 (−8)	0.951 (0)	0.958 (0)	0.951 (−2)
0.14	0.957 (0)	0.951 (−1)	0.951 (−7)	0.951 (−1)	0.956 (−1)	0.957 (−1)	0.950 (−2)
0.15	0.960 (0)	0.955 (0)	0.955 (−6)	0.955 (−6)	0.955 (−1)	0.955 (0)	0.954 (−7)
0.16	0.959 (0)	0.954 (0)	0.955 (−6)	0.954 (−6)	0.954 (0)	0.959 (0)	0.953 (0)
0.17	0.959 (−1)	0.954 (−1)	0.954 (−5)	0.954 (−1)	0.953 (0)	0.959 (−1)	0.953 (−7)
0.18	0.959 (−1)	0.954 (0)	0.954 (−5)	0.954 (−6)	0.954 (−1)	0.958 (0)	0.953 (−1)
0.19	0.958 (0)	0.954 (0)	0.954 (−5)	0.954 (0)	0.954 (0)	0.958 (0)	0.953 (−5)
0.20	0.958 (0)	0.950 (0)	0.954 (−5)	0.950 (−1)	0.954 (0)	0.954 (4)	0.949 (−1)
0.21	0.955 (0)	0.951 (0)	0.951 (0)	0.951 (−1)	0.951 (−1)	0.955 (0)	0.950 (0)
0.22	0.956 (0)	0.952 (0)	0.952 (0)	0.952 (−1)	0.952 (0)	0.956 (0)	0.951 (0)
0.23	0.957 (0)	0.953 (0)	0.953 (−4)	0.953 (0)	0.953 (0)	0.957 (0)	0.952 (−4)
0.24	0.954 (0)	0.954 (0)	0.951 (−1)	0.951 (−1)	0.954 (0)	0.954 (0)	0.950 (−1)
0.25	0.956 (0)	0.952 (0)	0.952 (−4)	0.952 (0)	0.952 (0)	0.956 (0)	0.952 (−5)
0.26	0.954 (3)	0.953 (0)	0.953 (−4)	0.950 (0)	0.954 (0)	0.957 (0)	0.953 (0)
0.27	0.956 (0)	0.952 (0)	0.952 (−4)	0.952 (−4)	0.952 (0)	0.956 (0)	0.952 (−1)
0.28	0.954 (0)	0.951 (0)	0.951 (0)	0.951 (−1)	0.951 (0)	0.954 (0)	0.950 (−4)
0.29	0.956 (0)	0.953 (0)	0.953 (−4)	0.953 (−4)	0.953 (0)	0.956 (0)	0.953 (−4)
0.30	0.956 (−1)	0.952 (0)	0.952 (−3)	0.952 (−4)	0.952 (0)	0.956 (−1)	0.952 (−1)
0.31	0.955 (0)	0.952 (0)	0.952 (−4)	0.952 (−1)	0.952 (0)	0.955 (0)	0.951 (0)
0.32	0.954 (0)	0.951 (0)	0.951 (0)	0.951 (0)	0.951 (0)	0.954 (0)	0.951 (−1)
0.33	0.954 (0)	0.951 (0)	0.951 (0)	0.951 (−1)	0.951 (0)	0.954 (0)	0.951 (−1)
0.34	0.954 (0)	0.951 (0)	0.951 (0)	0.951 (−1)	0.951 (0)	0.954 (0)	0.950 (0)
0.35	0.954 (0)	0.951 (0)	0.951 (−3)	0.951 (0)	0.951 (0)	0.954 (0)	0.951 (−1)
0.36	0.954 (0)	0.951 (0)	0.951 (−3)	0.951 (0)	0.951 (0)	0.954 (0)	0.951 (−1)
0.37	0.955 (0)	0.952 (0)	0.952 (−4)	0.952 (−4)	0.952 (0)	0.955 (0)	0.952 (−1)
0.38	0.955 (0)	0.952 (0)	0.953 (−4)	0.952 (0)	0.952 (0)	0.955 (0)	0.952 (0)
0.39	0.956 (0)	0.953 (0)	0.953 (−3)	0.950 (0)	0.953 (0)	0.956 (0)	0.952 (−2)
0.40	0.954 (0)	0.951 (0)	0.951 (−3)	0.951 (0)	0.951 (0)	0.954 (0)	0.951 (0)
0.41	0.955 (0)	0.953 (0)	0.952 (−3)	0.953 (−4)	0.953 (0)	0.955 (0)	0.952 (0)
0.42	0.954 (0)	0.951 (0)	0.951 (−3)	0.951 (−3)	0.951 (0)	0.954 (0)	0.951 (−4)
0.43	0.955 (0)	0.953 (0)	0.952 (−2)	0.952 (−3)	0.953 (0)	0.955 (0)	0.952 (−3)
0.44	0.954 (0)	0.951 (0)	0.951 (−3)	0.951 (−3)	0.951 (0)	0.954 (0)	0.951 (−3)
0.45	0.953 (0)	0.953 (0)	0.951 (−1)	0.953 (−3)	0.953 (0)	0.953 (0)	0.953 (−3)
0.46	0.955 (0)	0.952 (0)	0.952 (−3)	0.952 (−3)	0.952 (0)	0.955 (0)	0.952 (−3)
0.47	0.955 (0)	0.952 (0)	0.952 (−3)	0.952 (0)	0.952 (0)	0.955 (0)	0.952 (−1)
0.48	0.954 (0)	0.952 (0)	0.952 (−3)	0.952 (−1)	0.952 (0)	0.954 (0)	0.951 (0)
0.49	0.954 (0)	0.951 (0)	0.951 (0)	0.951 (0)	0.951 (0)	0.954 (0)	0.951 (0)
0.50	0.954 (0)	0.951 (0)	0.951 (−3)	0.951 (−3)	0.951 (0)	0.954 (0)	0.951 (−4)

Notes: [a]CP with the sample size based on the first-order approximation. NA, not available. $\Delta = 0.05$ and $p = 0.01(0.01)0.5$.

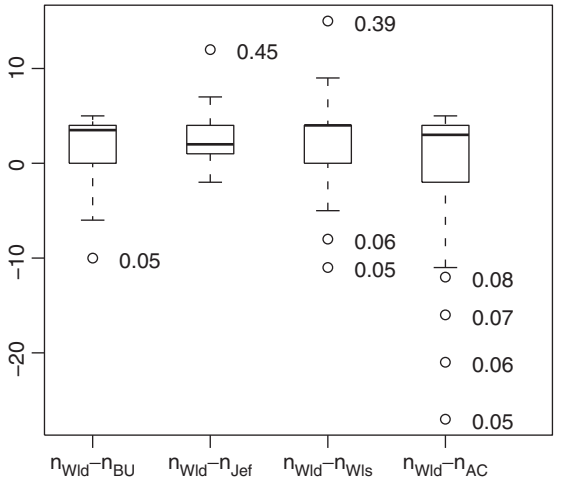

Figure 2. Boxplots of the differences between the sample size provided by the Wald method (n_{Wld}) and the sample sizes obtained for the four non-conservative methods (Bayesian-U: n_{BU}; modified Jeffreys: n_{Jef}; Wilson: n_{Wls}; Agresti–Coull: n_{AC}), with $\Delta = 0.05$, $1 - \alpha = 0.95$ and $p \geq 0.05$.

produce wider intervals, like the Anscombe and the Clopper–Pearson methods, also require larger samples to achieve a fixed EL. This happens because, as mentioned before and apart the Wald interval with small n, the EL is a decreasing function of n. So, to achieve an equal value, methods that produce wider intervals require larger samples than methods that give shorter intervals. The apparent paradox can be solved noting that the EL of the Agresti–Coull intervals tends to be larger than the length of the other intervals for extreme values of p, either close to 0 or to 1, but the opposite happens when p is not extreme [30]. The results given in Tables 2 and 3 are coherent with this: the Agresti–Coull method needs, for small values of p, larger sample sizes than the other methods, while the opposite happens as p approaches 0.5.

3.5 *Reduction of the optimal sample size*

In many real situations, sample units are lost in the data collection process. A common approach to compensate for those non-responses is to design the study with a sample size larger than the size required to obtain the desired EL and a specified nominal confidence level.

However, note that inflating the sample size does not necessarily address the potential bias due to non-response.

In this section, we analyze what happens to the CP and the EL of the CIs, computed by each of the seven methods, when a percentage of non-responses, ϕ, occurs. As an example, consider that the optimal n obtained by the exact procedure for each $p \geq 0.05$, presented in Table 3, is reduced to $n' = [(1 - \phi)n]$, (where $[x]$ is the nearest integer to x), with $\phi = 0.3$.

Figure 3 shows boxplots of the coverage probabilities of the CI computed with n and n'. It is easy to verify that the CP are not really affected by the reduction, except in the case of the Wald method. In general, all the other methods keep their main characteristics, already pointed out in Section 3.4. Analyzing the boxplots of the differences between the ELs of the CI computed with n' and n, displayed in Figure 4, we can conclude that the EL is increased by approximately 0.01, for all the methods. When $\phi = 0.5$ (results not shown), the conclusions are very similar, but the effect on the CP for the Wald method is stronger and the increase of the EL is approximately 0.02. Note that the approximate values of the EL observed for all the methods, 0.06 and 0.07, respectively,

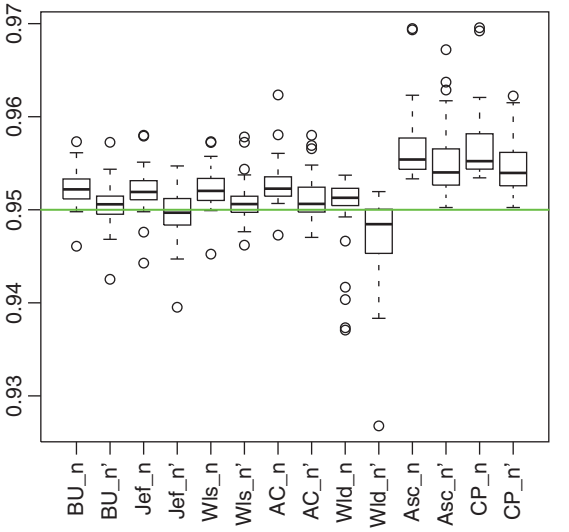

Figure 3. Coverage probabilities associated to n and n' for the Bayesian-U (BU), modified Jeffreys (Jef), Wilson (Wls), Agresti–Coull (AC), Wald (Wld), Anscombe (Asc), and Clopper–Pearson (CP) methods, with $\Delta = 0.05$, $1 - \alpha = 0.95$ and $p \geq 0.05$.

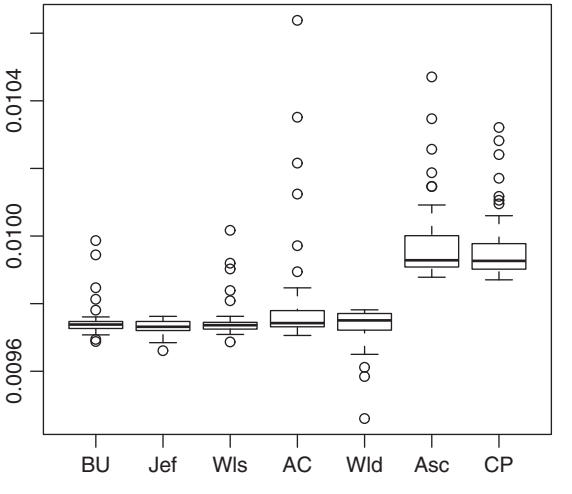

Figure 4. Difference between $\mathrm{EL}(n', p)$ and $\mathrm{EL}(n, p)$ for the Bayesian-U (BU), modified Jeffreys (Jef), Wilson (Wls), Agresti–Coull (AC), Wald (Wld), Anscombe (Asc), and Clopper–Pearson (CP) methods, with $\Delta = 0.05$, $1 - \alpha = 0.95$ and $p \geq 0.05$.

for $\phi = 0.3$ and $\phi = 0.5$, are very close to the values we would expect for $\sqrt{n}$-consistent methods, that is, $\mathrm{EL}(n', p) \simeq \mathrm{EL}(n, p)\sqrt{n}/\sqrt{n'} \simeq \mathrm{EL}(n, p)/\sqrt{1-\phi}$.

4. Discussion

Several aspects should be taken into account when analyzing the problem of sample size determination for interval estimation of a binomial proportion. Among others, we point out the degree of conservativeness, the mathematical complexity and the computational time (increasingly less

important). Moreover, when implementing a biomedical research project in the field, time, cost, human resources, ethical, and technical issues need to be addressed jointly with the statistical requirements. The nature of the event under study (e.g. pharmaceutical versus epidemiological context) is also important when choosing among methods that require distinct numbers of sample units to achieve the same EL.

Among the correct on average methods, Bayesian-U, Jeffreys and Wilson, by requiring a smaller sample size are advantageous in terms of economical, ethical, and technical aspects. On the other hand, the strictly conservative methods lead to larger sample sizes. The Agresti–Coull method shows an intermediate behavior. Larger sample sizes may be advantageous in some situations. For instance, in human populations, it is often possible, based on previous knowledge of the field, to have an idea of the difficulties in enrolling the number of individuals suitable for a desired degree of precision. In this case, it may be preferable to take into account the possibility of a low response rate, and to recommend an initial sample size larger than the one strictly necessary to achieve that degree of precision. Taking into account the practical importance of a possible reduction, in the field, of the optimal sample size, we also explored the behavior of all the methods admitting 30% and 50% of losses. We have concluded that, for all the methods, the EL are inflated, as expected, but the CP remain close to the original target (with some exceptions, mostly related to the Wald method).

From the main results presented in Section 3.4, it is possible to extract some practical recommendations. Since the behavior of the studied methods depends on the value of p, we will consider separately four intervals for p: $p < 0.01, 0.01 \leq p < 0.05, 0.05 \leq p < 0.1$ and $0.1 \leq p \leq 0.5$ (for $p > 0.5$ use $1 - p$).

(i) $p < 0.01$ – The sample size determination associated with a CI for a very low/high proportion (see [19,20,29]) is a sensitive statistical problem that has not been addressed in most of the published papers. Nevertheless, biomedical research often focuses on events which occur with a very low/high frequency (note, however, that what is a very low/high value for p is subjective and depends on the context).

As expected, the major discrepancies between methods occur for p in this range. Recall that the Wald method should never be used when p is so small. Based on the results of our study, we recommend the modified Jeffreys method (the Bayesian-U method leads to smaller sample sizes but its CP is lower than 0.9 for some p, see Table 4). Among the strictly conservative, Anscombe's method minimizes the sample size. On the other hand, the Agresti–Coull method gives the larger sample sizes with a CP always exceeding the nominal level. Note that, in an epidemiological context, a larger value of n may be preferred. Suppose that an important prevalence is unknown but assumed to be in this range. The epidemiologist usually wants to estimate the prevalence, but in addition aims to characterize the cases found. For this second objective, the sample sizes presented in Table 2 may not be appropriate. Instead, we should answer the question: for a given prevalence, how many individuals should be selected, such that, at least, x cases are found with a certain probability?

The problem of sample size determination associated with a very low/high proportion demands, however, further investigation. As explained by Newcombe [24], the degree of asymmetry of two-sided CI around a point estimate near 0 or 1, should be taken into account because it is reflected in the balance of the coverage probabilities in the two tails. In some applications, one kind of non-coverage could be of greater concern than the other. For the prevalence of a rare disease, it may prefer to err on the side of over-estimating, rather than under-estimating it. On the other hand, for a sensitivity or a specificity, it may prefer to err on the side of under-estimating rather than over-estimating. We advise the practitioner to explore the usefulness of the optimal value of n (n_{opt}), by performing the following extra steps: 1. Given n_{opt}, compute CIs, $[L(x), U(x)]$, varying x in an interesting range; 2. Calculate upper and lower precisions (given by Equation (6)), and judge how informative they are in his/her research context.

A related issue is the choice between one and two-sided CIs for p near the boundaries. Olivier and May [25] advise that the use of one-sided CIs should be decided before the data are collected and not post hoc. Those authors also recall that the more conservative two-sided approach should be preferred, to account for CP problems near the boundaries. Other authors [18] explore the sample size problem in the context of one-sided CI for rare events, to determine the safety of a diagnostic test or a therapeutic intervention.

(ii) $0.01 \leq p < 0.05$ – The use of the Wald method is also prohibited for p within this range. Among the strictly conservative methods, the Clopper–Pearson represents the best choice. In the other group, the modified Jeffreys method is appropriate because it guarantees a safe CP with the smallest sample size.

(iii) $0.05 \leq p < 0.1$ – In this situation, all the strictly conservative methods produce similar results. Again, the Jeffreys method yields very satisfactory results, according to the ethical and economical points of view.

(iv) $0.1 \leq p \leq 0.5$ – For p in this interval, and in particular when $p > 0.15$, the differences between the methods within a group (strictly conservatives and non-conservatives) are not substantial. The Agresti–Coull method becomes closer to the Bayesian and the Wilson methods, and even the Wald method produces values similar to those produced by the latter methods. The results presented, especially in Table 5, show that when the EL is fixed, the discrepancies between the CP of the different methods are smoothed (when compared to previous studies on the performance of CI methods). Even though all methods have a similar CP, the conservative methods have a CP always above the nominal level. Thus, to achieve the same EL, the conservative methods require larger samples sizes.

In global terms, the modified Jeffreys method is appropriate, particularly for $p \in [0.01, 0.2]$ (or $[0.8, 0.99]$). The Wilson method is also a good choice, though it requires some caution, in terms of CP, for p near 0 (or 1). However, the Wilson method has the advantage of yielding a closed-form formula for n that is tractable without requiring extra computational effort (an appealing property in teaching and consulting work). This advantage is only shared by the Wald method (but this popular method should only be used when $0.2 < p < 0.8$).

Another interesting feature that emerges from Table 3 is related to 'small' differences in the assumed proportion to estimate the sample size. For example, suppose we consider, based on a previous study, that $p = 0.01$. However, if the true value is $p = 0.02$, we will need an additional of 51 units for the Clopper–Pearson method and an additional of 52 units for Anscombe's method. The sample sizes associated to the Wilson and the modified Jeffreys methods would be increased by 42 and 36 units, respectively. In fact, for $0.01 < p < 0.1$ a positive difference of 0.01 requires, except for the Agresti–Coull method, an addition of more than 50 units. When $0.1 < p \leq 0.2$, the same difference can produce a variation of three or four dozens of units. For $0.2 < p \leq 0.3$, the variations in the sample size are of two or three dozens of units. For $0.3 < p \leq 0.4$, the differences are between 10 and 26 units. In the remaining block, $0.4 < p \leq 0.5$, we find some differences between 10 and 20. It is only for $0.45 < p \leq 0.5$ that the differences become insignificant. These findings suggest that a preliminary estimate of p, which is going to be used for sample size determination, must be chosen carefully, and in case of doubt, it may be better to use an overestimate, if $p < 0.5$ (or an underestimate, if $p > 0.5$).

Even though the continuous update of the statistical software for sample size determination, for example, in epidemiological and clinical practice, the Wald method is still popular. Thus, more emphasis should be given to alternative methods using different procedures and criteria to obtain the sample size. As in hypothesis testing, where the sample size is determined given that attention to Type-I error and power, when we compute the sample size based on a CI, the CP and the EL should be returned by the software in order to avoid misinterpretations.

Acknowledgements

The authors gratefully acknowledge the comments and suggestions of Prof. R. Newcombe on a previous version of the manuscript. This research was (partially) supported by FCT/OE and two projects: PTDC/SAU-ESA/81240/2006 and PTDC/MAT/64353/2006 (FCT – Portugal).

References

[1] A. Agresti and B. Caffo, *Simple and effective confidence intervals for proportions and differences of proportions result from adding 2 successes and 2 failures*, Amer. Statist. 54 (2000), pp. 280–288.
[2] A. Agresti and B. Coull, *Approximate is better than 'exact' for interval estimation of binomial proportions*, Amer. Statist. 52 (1998), pp. 119–126.
[3] A. Agresti and Y. Min, *Frequentist performance of Bayesian confidence intervals for comparing proportions in* 2×2 *contingency tables*, Biometrics 61 (2005), pp. 515–523.
[4] F. Anscombe, *Transformations of Poisson, binomial and negative-binomial data*, Biometrika 35 (1948), pp. 246–254.
[5] D. Böhning and C. Viwatwongkasem, *Revisiting proportion estimators*, Stat. Methods Med. Res. 14 (2005), pp. 147–169.
[6] L. Brown, T. Cai, and A. DasGupta, *Interval estimation for a binomial proportion*, Statist. Sci. 16 (2001), pp. 101–133.
[7] L. Brown, T. Cai, and A. DasGupta, *Confidence intervals for a binomial proportion and edgeworth expansions*, Ann. Statist. 30 (2002), pp. 160–201.
[8] B. Carlin and T. Louis, *Bayes and Empirical Bayes Methods for Data Analysis*, Chapman & Hall, London, 1996.
[9] Z. Chen and M. McGee, *A Bayesian approach to zero-numerator problems using hierarchical models*, J. Data Sci. 6 (2008), pp. 261–268.
[10] M. Chernick and C. Liu, *The saw-toothed behavior of power versus sample size and software solutions: Single binomial proportion using exact methods*, Amer. Statist. 56 (2002), pp. 149–155.
[11] C. Clopper and E. Pearson, *The use of confidence or fiducial limits illustrated in the case of the binomial*, Biometrika 26 (1934), pp. 404–413.
[12] N. Dendukuri, E. Rahme, P. Belisle, and L. Joseph, *Bayesian sample size determination for prevalence and diagnostic test studies in the absence of a gold standard test*, Biometrics 60 (2004), pp. 388–397.
[13] A. Flahault, M. Cadilhac, and G. Thomas, *Sample size calculation should be performed for design accuracy in diagnostic test studies*, J. Clin. Epidemiol. 58 (2005), pp. 859–863.
[14] G. Fosgate, *Modified exact sample size for a binomial proportion with special emphasis on diagnostic test parameter estimation*, Stat. Med. 24 (2005), pp. 2857–2866.
[15] M. Gardner and D. Altman, *Estimating with confidence*, in *Statistics with Confidence. Confidence Intervals and Statistical Guidelines*, D. Altman, D. Machin, T. Bryant, and M. Gardner, eds., BMJ Books, London, 2000, pp. 3–5.
[16] L. Joseph, R. Berger, and P. Belisle, *Bayesian and mixed Bayesian/likelihood criteria for sample size determination*, Stat. Med. 16 (1997), pp. 769–781.
[17] L. Joseph, D. Wolfson, and R. Berger, *Sample size calculations for binomial proportions via highest posterior density intervals*, Statistician 44 (1995), pp. 143–154.
[18] B. Jovanovic and R. Zalenski, *Safety evaluation and confidence intervals when the number of observed events is small or zero*, Ann. Emerg. Med. 30 (1997), pp. 301–306.
[19] K. Krishnamoorthy and J. Peng, *Some properties of the exact and score methods for binomial proportion and sample size calculation*, Comm. Statist. Simulation Comput. 36 (2007), pp. 1171–1186.
[20] E. Kron, *Sample size for bounding small proportions*, Biometrics 42 (1986), pp. 213–216.
[21] W. Liu and B. Bailey, *Sample size determination for constructing a constant width confidence interval for a binomial success probability*, Statist. Probab. Lett. 56 (2002), pp. 1–5.
[22] C. M'Lan, L. Joseph, and D. Wolfson, *Bayesian sample size determination for binomial proportions*, Bayesian Anal. 3 (2008), pp. 269–296.
[23] R. Newcombe, *Two-sided confidence intervals for the single proportion: Comparison of seven methods*, Stat. Med. 17 (1998), pp. 857–872.
[24] R. Newcombe, *Measures of location for confidence intervals for proportions*, Comm. Statist. Theory Methods 40 (2011), pp. 1743–1767.
[25] J. Olivier and W. May, *A discussion of binomial interval estimators on the boundary*, J. Miss. Acad. Sci. 52 (2007), pp. 178–183.
[26] M.R. Oliveira, A. Subtil, and L. Gonçalves, *Confidence intervals for sensitivity and specificity: A comparison study*, submitted.
[27] W. Pan, *Approximate confidence intervals for one proportion and difference of two proportions*, Comput. Statist. Data Anal. 40 (2002), pp. 143–157.

[28] D. Pastor, *On the coverage probability of the Clopper–Pearson confidence interval*, Tech. Rep., ENST Bretagne, France, 2007.
[29] W. Piegorsch, *Sample size for improved binomial confidence intervals*, Comput. Statist. Data Anal. 46 (2004), pp. 309–316.
[30] A. Pires and C. Amado, *Interval estimates for a binomial proportion: Comparison of twenty methods*, REVSTAT 6 (2008), pp. 165–197.
[31] J. Sim and N. Reid, *Statistical inference by confidence intervals: Issues of interpretation and utilization*, Phys. Ther. 79 (1999), pp. 186–195.
[32] F. Tuyl, R. Gerlach, and K. Mengersen, *A comparison of Bayes–Laplace, Jeffreys, and other priors: The case of zero events*, Amer. Statist. 62 (2008), pp. 40–44.
[33] S. Vollset, *Confidence intervals for a binomial proportion*, Stat. Med. 12 (1993), pp. 809–824.
[34] P. Vos and S. Hudson, *Evaluation criteria for discrete confidence intervals: Beyond coverage and length*, Amer. Statist. 59 (2005), pp. 137–142.
[35] H. Wang, *Exact average coverage probabilities and confidence coefficients of confidence intervals for discrete distributions*, Statist. Comput. 19 (2009), pp. 139–148.
[36] R. Winkler, J. Smith, and D. Frayback, *The role of informative priors in zero-numerator problems: Being conservative versus being candid*, Amer. Statist. 56 (2002), pp. 1–4.

Appendix

(1) *Clopper–Pearson exact method* – One of the oldest CI for a binomial proportion [11] is treated as the 'gold standard'. This method has been recommended in the most advanced statistics textbooks to avoid the approximation used by the historic Wald interval [2,14]. We can find different expressions associated with this interval, using the binomial distribution or its relationship with the beta and F distributions [2,30,35].

(2) *Bayesian method with non-informative priors* – In the context of Bayesian estimation for binary data, it is usual to choose the prior distribution for the binomial parameter, p, in the conjugate family of beta distributions. Given a prior distribution $\text{Beta}(a, b)$, $(a > 0, b > 0)$ and data (X), the posterior distribution of p is $\text{Beta}(X + a, n - X + b)$ and a $100(1 - \alpha)\%$ equal-tailed Bayesian interval is given by

$$[\text{Beta}_{\alpha/2}(X + a, n - X + b); \text{Beta}_{1-\alpha/2}(X + a, n - X + b)], \tag{A1}$$

where $\text{Beta}_{\gamma}(a, b)$ represents the γ-quantile of the $\text{Beta}(a, b)$ distribution [6]. In the remaining of the paper, these intervals are explored in a frequentist perspective, as justified by Carlin and Louis [8] (cited by Agresti and Caffo [1], Agresti and Min [3], and Newcombe [24]). We consider only non-informative priors [2,6,7,30], particularly, the well-known Jeffreys prior with $a = b = \frac{1}{2}$ [6,7] and the Uniform or Bayes–Laplace prior with $a = b = 1$ [30]. The resulting intervals are called Jeffreys and Bayesian-U, respectively. The limits shown in Table 1 already include slight modifications at the boundary values introduced previously by other authors (the Bayesian-U follows [30], and the Jeffreys is the *Modified Jeffreys interval* of [6]). In a recent work, Tuyl *et al*. [32] favor the Bayes–Laplace or Uniform prior as the choice of consensus over the Jeffreys distribution. Pires and Amado [30] also choose the Uniform, arguing that it is more intuitive. An open point is the assessment of other priors, namely informative priors, in terms of CP, EL and sample size determination. The delicate problem of choosing an informative prior has been addressed by some theoretical studies related to the case of zero events [32], also known as the zero-numerator problem [9,36].

(3) *Wilson or score method* – According to Brown and co-authors [6], this interval was introduced by Wilson in 1927. The performance and the simplicity of computation, at least in a classroom context, are the main advantages of this method [2,5–7,23].

(4) *Agresti–Coull or Add-4 method* – This method was introduced in 1998 by Agresti and Coull [2] and quickly gained popularity. Nowadays, it is recommended at least for large samples [6,7].

(5) *Arcsine method with Anscombe's continuity correction* – Some other methods are based on the approximation to the normal distribution after a variance stabilizing transformation

(see Table 2 in [30]). In this paper, we use only the modified arcsine interval suggested by Anscombe [4].

(6) *Wald method* – In spite of all the critics, the Wald method is still popular and it is presented in most of the elementary epidemiology and statistics textbooks, essentially due to its simple formula. Sometimes, the formula is accompanied by a warning stating that when $np < 5$ or $n(1-p) < 5$ another method should be chosen. The two main disadvantages of the Wald method are a breakdown of the CP, even for large sample sizes, and the possibility of limits outside [0, 1]. To avoid the latter drawback, we use the version of the Wald CI [30] presented in the last row of Table 1, which guarantees that $[L_7(X), U_7(X)] \subset [0, 1]$, for all X.