REVSTAT – Statistical Journal
Volume 6, Number 2, June 2008, 165–197

INTERVAL ESTIMATORS FOR A BINOMIAL PROPORTION: COMPARISON OF TWENTY METHODS

Authors: Ana M. Pires
– Department of Mathematics and CEMAT, Instituto Superior Técnico, Technical University of Lisbon (TULisbon), Portugal
apires@math.ist.utl.pt

Conceição Amado
– Department of Mathematics and CEMAT, Instituto Superior Técnico, Technical University of Lisbon (TULisbon), Portugal
camado@math.ist.utl.pt

Received: March 2007 Revised: March 2008 Accepted: March 2008

Abstract:

- In applied statistics it is often necessary to obtain an interval estimate for an unknown proportion (p) based on binomial sampling. This topic is considered in almost every introductory course. However, the usual approximation is known to be poor when the true p is close to zero or to one. To identify alternative procedures with better properties twenty non-iterative methods for computing a (central) two-sided interval estimate for p were selected and compared in terms of coverage probability and expected length. From this study a clear classification of those methods has emerged. An important conclusion is that the interval based on asymptotic normality, but after the arcsine transformation and a continuity correction, and the Add 4 method of Agresti and Coull (1998) yield very reliable results, the choice between the two depending on the desired degree of conservativeness.

Key-Words:

- *confidence interval; binomial distribution; proportion test; normal approximation; arcsine transformation; continuity correction; bootstrap; HPD credibility intervals.*

AMS Subject Classification:

- 62F25, 62F40.

1. INTRODUCTION

In many practical situations it is important to compute a two-sided interval estimate for a population proportion (e.g. acceptance sampling by attributes, marketing research, survey sampling). The interval estimate may be either a confidence interval (in the frequentist framework) or a credibility interval (in the Bayesian framework). This is a well known topic considered in almost every introductory course on statistics. However, most of the standard methods rely on asymptotic approximations and the validity of the approximations is not always stated or differs from author to author (Leemis and Trivedi, 1996, give a survey of these "rules of thumb"). Moreover, comparisons between methods are usually based on single cases. We found no recent text book listing the most common methods and making a general comparison, not even in Fleiss, Levin and Paik (2003), dedicated exclusively to rates and proportions. A good discussion but somehow out of date can be found in Santner and Duffy (1989).

Nevertheless, several authors have addressed this subject in the last thirty years: Ghosh (1979); Fujino (1980); Angus and Schafer (1983); Blyth and Still (1983); Blyth (1986); Chen (1990); Copas (1992); Vollset (1993); Cohen and Yang (1994); Newcombe (1994 and 1998); Agresti and Coull (1998); Agresti and Caffo (2000); Brown, Cai and DasGupta (2001 and 2002); Edwardes (1998); Pan (2002); Reiczigel (2003); García-Pérez (2005); Geyer and Meeden (2005), Puza and O'neill (2006) and Lee (2006). Large comparative studies were presented by Vollset (1993), Newcombe (1998), Brown, Cai and DasGupta (2001) and, to a smaller extent, by Agresti and Coull (1998) and Pan (2002).

The present paper considers twenty simple non-iterative methods. Eleven of these have been included in the aforementioned comparisons: eight of the seventeen considered by Vollset (1993); five of the seven considered by Newcombe (1998); the four considered by Agresti and Coull (1998); five of the twelve considered by Brown, Cai and DasGupta (2001) and the four considered by Pan (2002). The nine (also simple) methods considered here and which were not included in previous comparisons are: a Bayesian interval with uniform prior; two simple corrections to the usual interval; four variants of bootstrap intervals not needing Monte Carlo; and two intervals based on the arcsine transformation followed by a correction. As with most of the cited recent studies it was decided not to include methods without explicit solutions and for which there is an explicit almost equivalent method (this is the case namely of the interval obtained by inverting the likelihood ratio test).

The results of this work are important for the applied statistician, who in a particular situation usually wants to use the best method and that this method is available or can easily be implemented in common statistical software, and for teachers of statistics who have to decide which methods to include in a given course.

Let $[L;U]$ be an interval estimator of a certain parameter θ and attach to it a level (confidence or credibility), $\eta \in (0;1)$. $[L;U]$ is a good interval estimator of θ with level η if the probability of containing the unknown θ (the coverage probability) is in fact η and its length is "small" (usually in a stochastic sense, for instance, on average). Note, however, that in discrete situations, like the one considered here, it is not possible to achieve the target coverage probability for all possible values of the parameter. We will therefore consider two classes of acceptable methods, those strictly conservative (for which the coverage probability is at least η) and those correct on average (for which the mean coverage probability is at least η) and look for small mean expected length within each class. Note also that it is reasonable to apply the same criteria to both confidence and credibility intervals.

Attention will be focused primarily on central intervals, that is, with approximately equal uncertainty associated to each side. This is how practitioners usually interpret two-sided intervals and matches better with one-sided intervals (the two-sided being the intersection of upper and lower one-sided intervals with the appropriate precision). However non-central intervals are also considered for comparison.

The paper is organized as follows: in Section 2 the twenty selected methods are described. In Section 3 results regarding coverage probability and expected length of the different intervals are presented and analyzed, first considering only the central intervals and at second stage including two optimal non-central intervals. Section 4 is devoted to concluding remarks.

2. DESCRIPTION OF THE METHODS

In order to establish the notation suppose that a random sample of size n is observed on a large (possibly infinite) population and that X observations ($0 \leq X \leq n$) belong to a certain category of interest. Let p be the unknown proportion of the category of interest in the population and suppose that a two-sided central interval estimate for p is wanted. Note that in order to use the methods based on the binomial distribution the total sample size must be fixed a priori and the variable to be observed is the number of successes. If the sampling plan fixes the number of successes and the total sample size is variable (inverse or negative binomial sampling) most of the methods can not be applied directly (see e.g. Lui, 1995, or Cai and Krishnamoorthy, 2005).

For all the intervals the nominal confidence level is fixed in advance as $100 \times (1-\alpha)\%$, meaning that the coverage probability of the random interval $[L; U]$ should be $1-\alpha$. The random variables L and U depend on X, number of successes in the random sample, on n and on the method. Twenty methods are described in the following five subsections and the corresponding solutions are represented as $[L_i(X); U_i(X)]$, where $i = I, ..., XX$ denotes the method. The final expressions are given in Tables 1 and 2. As in Vollset (1993) it was imposed that, for all i, $0 \leq L_i(X)$ and $U_i(X) \leq 1$, for all X, and that $L_i(0) = 0$ and $U_i(n) = 1$ (this means that for the boundary cases the centrality property is dropped but it is a natural choice). To be theoretically correct, but otherwise with no practical effect, the confidence intervals do not include the left (right) end point if $L_i(X) = 0$ $(U_i(X) = 1)$ but $X \neq 0$ $(X \neq n)$.

2.1. Exact results

Under the previous conditions X has a Binomial(n, p) distribution. Because this is a discrete distribution it is not possible to have a confidence interval with exactly the specified confidence level. But an interval with a coverage probability of at least $1-\alpha$ can be obtained by solving

$$\sum_{j=X}^{n} \binom{n}{j} L_I^j (1-L_I)^{n-j} = \alpha' \tag{2.1}$$

and

$$\sum_{j=0}^{X} \binom{n}{j} U_I^j (1-U_I)^{n-j} = \alpha'' , \tag{2.2}$$

where α' and α'', such that $\alpha' + \alpha'' = \alpha$, are fixed in advance and do not depend on X. If different values of α' and α'' are chosen for each X, for instance those minimizing the length of the interval, the exactness, meaning a coverage probability of at least $1-\alpha$ can no longer be guaranteed for every p. $\alpha' = \alpha'' = \alpha/2$ corresponds to the inversion of the two sided exact binomial test and leads to the central exact interval, usually called Clopper–Pearson interval (Clopper and Pearson, 1934). For $X = 0$ and $X = n$ the solutions of (2.1) and (2.2) are explicit:

$$L_I(0) = 0, \quad U_I(0) = 1 - (\alpha/2)^{1/n}, \quad L_I(n) = (\alpha/2)^{1/n}, \quad U_I(n) = 1 . \tag{2.3}$$

Otherwise, the solution of (2.1) and (2.2) is easy to obtain by using the relation (see for instance Johnson and Kotz 1969, pp. 58–59, or Stevens, 1950)

$$\sum_{j=k}^{n} \binom{n}{j} p^j (1-p)^{n-j} = \int_0^p f_B(t)\, dt ,$$

where f_B denotes the p.d.f. of a Beta$(k, n-k-1)$ random variable. In this way the extremes of the interval are appropriate quantiles of that Beta distribution (see Table 1) and can easily be obtained in most statistical packages. For instance, in S-plus or R the appropriate commands are `LI<-qbeta(alfa/2,X,n-X+1)` and `UI<-qbeta(1-alfa/2,X+1,n-X)`. If the percentiles of the Beta distribution are not available then the relation of those with the percentiles of the F distribution can be used, eventually in tables (this is again mentioned in Johnson and Kotz, 1969, and e.g. in Armitage and Berry, 1987).

It is worth noting that from the papers referred in the Introduction and addressing the same issue only Blyth (1986), Agresti and Coull (1998), Newcombe (1998), Brown, Cai and DasGupta (2001) and García-Pérez (2005) mention the relation with the Beta distribution while some of the others consider approximations to the percentiles of the F distribution (e.g. Fujino, 1980). Vollset (1993) proposes a very sophisticated numerical method (the "Pratt" approximation) which turns out to be completely unnecessary.

Other exact intervals, in the sense of having coverage probability of at least $1-\alpha$, based on the binomial probabilities have been considered in the literature but are not central and do not have explicit solutions (Sterne, 1954, Crow, 1956, Clunies-Ross, 1958, see also Blyth and Still, 1983, and Reiczigel, 2003). As mentioned in the Introduction, an interval of this type will be considered in Subsection 3.2 for comparative purposes.

As a second alternative in the class of exact methods we consider a Bayesian credibility interval. This method is exact in the second sense because it guarantees a mean coverage probability of $1-\alpha$ under the specified prior distribution for p. If this prior is the uniform distribution in $(0,1)$ or Beta$(1,1)$, which is non-informative, we have that *a posteriori* p follows a Beta$(X+1, n-X+1)$ distribution. Brown, Cai and DasGupta (2001) have chosen the Jeffreys prior, Beta$(1/2, 1/2)$, which is also non-informative. The results shall not differ considerably, but the uniform prior seems more intuitive.

In order to obtain a central interval, equal credibility tails $(\alpha/2)$ are considered, except for the boundary cases, $X=0$ and $X=n$. The explicit results are again shown is Table 1 (Method II). The similarity between these results and the results for the Clopper–Pearson interval is a consequence of the chosen *a priori* distribution.

In the Bayesian framework the optimal exact interval (that is of minimal length) is the HPD interval, but this again is non-central and has no explicit solution, and will be considered only in Subsection 3.2.

Table 1: Explicit limits of the intervals for Methods I to X.

i (Method)	$L_i(X)$ (Lower limit)	$U_i(X)$ (Upper limit)
I [(a)]	0 if $X=0$, $(\alpha/2)^{1/n}$ if $X=n$, $B_{X,n-X+1;\alpha/2}$ if $0<X<n$	$1-(\alpha/2)^{1/n}$ if $X=0$, 1 if $X=n$, $B_{X+1,n-X;1-\alpha/2}$ if $0<X<n$
II [(a)]	0 if $X=0$, $\alpha^{1/(n+1)}$ if $X=n$, $B_{X+1,n-X+1;\alpha/2}$ if $0<X<n$	$1-\alpha^{1/(n+1)}$ if $X=0$, 1 if $X=n$, $B_{X+1,n-X+1;1-\alpha/2}$ if $0<X<n$
III [(b)]	$\dfrac{2X+c^2-c\sqrt{c^2+4X(1-X/n)}}{2(n+c^2)}$	$\dfrac{2X+c^2+c\sqrt{c^2+4X(1-X/n)}}{2(n+c^2)}$
IV [(b)]	0 if $X=0$, otherwise $\dfrac{2X+c^2-1-c\sqrt{c^2-(2+1/n)+4X(1-X/n+1/n)}}{2(n+c^2)}$	1 if $X=n$, otherwise $\dfrac{2X+c^2+1+c\sqrt{c^2+(2-1/n)+4X(1-X/n-1/n)}}{2(n+c^2)}$
V [(b)]	$\max\left\{\frac{X}{n}-c\sqrt{\frac{X}{n^2}\left(1-\frac{X}{n}\right)};\ 0\right\}$	$\min\left\{\frac{X}{n}+c\sqrt{\frac{X}{n^2}\left(1-\frac{X}{n}\right)};\ 1\right\}$
VI [(b)]	$\max\left\{\frac{X}{n}-c\sqrt{\frac{X}{n^2}\left(1-\frac{X}{n}\right)}-\frac{1}{2n};\ 0\right\}$	$\min\left\{\frac{X}{n}+c\sqrt{\frac{X}{n^2}\left(1-\frac{X}{n}\right)}+\frac{1}{2n};\ 1\right\}$
VII [(b)]	0 if $X=0$, $(\alpha/2)^{1/n}$ if $X=n$, otherwise $\max\left\{\frac{X}{n}-c\sqrt{\frac{X}{n^2}\left(1-\frac{X}{n}\right)};\ 0\right\}$	$1-(\alpha/2)^{1/n}$ if $X=0$, 1 if $X=n$, otherwise $\min\left\{\frac{X}{n}+c\sqrt{\frac{X}{n^2}\left(1-\frac{X}{n}\right)};\ 1\right\}$
VIII [(b)]	0 if $X=0$, $(\alpha/2)^{1/n}$ if $X=n$, otherwise $\max\left\{\frac{X}{n}-c\sqrt{\frac{X}{n^2}\left(1-\frac{X}{n}\right)}-\frac{1}{2n};\ 0\right\}$	$1-(\alpha/2)^{1/n}$ if $X=0$, 1 if $X=n$, otherwise $\min\left\{\frac{X}{n}+c\sqrt{\frac{X}{n^2}\left(1-\frac{X}{n}\right)}+\frac{1}{2n};\ 1\right\}$
IX [(b)]	0 if $X=0$, $(\alpha/2)^{1/n}$ if $X=n$, otherwise $\max\left\{\frac{X+c^2/2}{n+c^2}-c\sqrt{\frac{X}{n^2}\left(1-\frac{X}{n}\right)};\ 0\right\}$	$1-(\alpha/2)^{1/n}$ if $X=0$, 1 if $X=n$, otherwise $\min\left\{\frac{X+c^2/2}{n+c^2}+c\sqrt{\frac{X}{n^2}\left(1-\frac{X}{n}\right)};\ 1\right\}$
X [(b)]	0 if $X=0$, $(\alpha/2)^{1/n}$ if $X=n$, otherwise $\max\left\{\frac{X+c^2/2}{n+c^2}-c\sqrt{\frac{X}{n^2}\left(1-\frac{X}{n}\right)}-\frac{1}{2n};\ 0\right\}$	$1-(\alpha/2)^{1/n}$ if $X=0$, 1 if $X=n$, otherwise $\min\left\{\frac{X+c^2/2}{n+c^2}+c\sqrt{\frac{X}{n^2}\left(1-\frac{X}{n}\right)}+\frac{1}{2n};\ 1\right\}$

[(a)] $B_{\theta_1,\theta_2;\gamma}$ is the γ percentile of the Beta(θ_1,θ_2) distribution.

[(b)] $c=z_{1-\alpha/2}$ where z_γ is the γ percentile of the $\mathcal{N}(0,1)$ distribution.

Table 2: Explicit limits of the intervals for Methods XI to XX.

i (Method)	$L_i(X)$ (Lower limit)	$U_i(X)$ (Upper limit)
XI [(a)]	0 if $X=0$, $(\alpha/2)^{1/n}$ if $X=n$, otherwise $\dfrac{\mathrm{Bin}_{n,X/n;\alpha/2}}{n}$	$1-(\alpha/2)^{1/n}$ if $X=0$, 1 if $X=n$, otherwise $\dfrac{\mathrm{Bin}_{n,X/n;1-\alpha/2}}{n}$
XII [(a)]	0 if $X=0$, $(\alpha/2)^{1/n}$ if $X=n$, otherwise $\max\left\{\frac{\mathrm{Bin}_{n,X/n;\alpha/2}}{n}-\frac{1}{2n};\,0\right\}$	$1-(\alpha/2)^{1/n}$ if $X=0$, 1 if $X=n$, otherwise $\min\left\{\frac{\mathrm{Bin}_{n,X/n;1-\alpha/2}}{n}+\frac{1}{2n};\,1\right\}$
XIII [(a)(b)]	0 if $X=0$, $(\alpha/2)^{1/n}$ if $X=n$, otherwise $\dfrac{\mathrm{Bin}_{n,X/n;\alpha'}}{n}$	$1-(\alpha/2)^{1/n}$ if $X=0$, 1 if $X=n$, otherwise $\dfrac{\mathrm{Bin}_{n,X/n;\alpha''}}{n}$
XIV [(a)(b)]	0 if $X=0$, $(\alpha/2)^{1/n}$ if $X=n$, otherwise $\max\left\{\frac{\mathrm{Bin}_{n,X/n;\alpha'}}{n}-\frac{1}{2n};\,0\right\}$	$1-(\alpha/2)^{1/n}$ if $X=0$, 1 if $X=n$, otherwise $\min\left\{\frac{\mathrm{Bin}_{n,X/n;\alpha''}}{n}+\frac{1}{2n};\,1\right\}$
XV [(c)]	0 if $X=0$, otherwise $\sin^2\left(\arcsin\sqrt{\frac{X}{n}}-\frac{c}{2\sqrt{n}}\right)$	1 if $X=n$, otherwise $\sin^2\left(\arcsin\sqrt{\frac{X}{n}}+\frac{c}{2\sqrt{n}}\right)$
XVI [(c)]	0 if $X=0$, otherwise otherwise $\sin^2\left(\arcsin\sqrt{\frac{X-0.5}{n}}-\frac{c}{2\sqrt{n}}\right)$	1 if $X=n$, otherwise otherwise $\sin^2\left(\arcsin\sqrt{\frac{X+0.5}{n}}+\frac{c}{2\sqrt{n}}\right)$
XVII [(c)]	0 if $X=0$, otherwise $\sin^2\left(\arcsin\sqrt{\frac{3/8+X-0.5}{n+3/4}}-\frac{c}{2\sqrt{n+1/2}}\right)$	1 if $X=n$, otherwise $\sin^2\left(\arcsin\sqrt{\frac{3/8+X+0.5}{n+3/4}}+\frac{c}{2\sqrt{n+1/2}}\right)$
XVIII [(c)]	$\max\left\{\frac{X+2}{n+4}-c\sqrt{\frac{X+2}{(n+4)^2}\left(1-\frac{X+2}{n+4}\right)};\,0\right\}$	$\min\left\{\frac{X+2}{n+4}+c\sqrt{\frac{X+2}{(n+4)^2}\left(1-\frac{X+2}{n+4}\right)};\,1\right\}$
XIX [(d)]	$\max\left\{\frac{X}{n}-t'\sqrt{\frac{X}{n^2}\left(1-\frac{X}{n}\right)};\,0\right\}$	$\min\left\{\frac{X}{n}+t'\sqrt{\frac{X}{n^2}\left(1-\frac{X}{n}\right)};\,1\right\}$
XX [(d)]	$\max\left\{\frac{X+2}{n+4}-t''\sqrt{\frac{X+2}{(n+4)^2}\left(1-\frac{X+2}{n+4}\right)};\,0\right\}$	$\min\left\{\frac{X+2}{n+4}+t''\sqrt{\frac{X+2}{(n+4)^2}\left(1-\frac{X+2}{n+4}\right)};\,1\right\}$

[(a)] $\mathrm{Bin}_{n,\theta;\gamma}$ is the γ percentile of the $\mathrm{Bin}(n,\theta)$ distribution.

[(b)] α' and α'' are given by equations (2.5) and (2.6) in the text, respectively.

[(c)] $c=z_{1-\alpha/2}$ where z_γ is the γ percentile of the $\mathcal{N}(0,1)$ distribution.

[(d)] t' and t'' are percentiles $(1-\alpha/2)$ of t-distributions with degrees of freedom given by equations (2.8) and (2.9) in the text, respectively.

2.2. Normal approximations

The most referred methods are based on the approximation of the Binomial(n,p) by the $\mathcal{N}\big(np, np(1-p)\big)$ distribution, that is on

$$\frac{X-np}{\sqrt{np\,(1-p)}} = \frac{\frac{X}{n}-p}{\sqrt{\frac{p(1-p)}{n}}} \xrightarrow{d} \mathcal{N}(0,1)$$

and

$$P\left(-z_{1-\alpha/2} \le \frac{\frac{X}{n}-p}{\sqrt{\frac{p(1-p)}{n}}} \le z_{1-\alpha/2}\right) \simeq 1-\alpha\ , \tag{2.4}$$

where z_γ denotes the γ percentile of the $\mathcal{N}(0,1)$ distribution.

From (2.4) one can obtain, solving a second degree equation, the score or Wilson (1927) interval (Method III in Table 1). Note that $L_{III}(X) > 0$ and $U_{III}(X) < 1$ for every $0 < X < n$ and that $L_{III}(0) = 0$ and $U_{III}(n) = 1$.

A modification of the score method is obtained by introducing a continuity correction (cc) of $\pm 1/(2n)$ in the numerator of the central expression of (2.4), which is expected to improve the approximation (Method IV in Table 1). For the boundary cases one obtains $L_{IV}(0) > 0$ and $U_{IV}(n) < 1$, which are corrected in the obvious way.

Most of the elementary text books do not present the previous methods, considering instead a second approximation in expression (2.4),

$$P\left(-c \le \frac{\frac{X}{n}-p}{\sqrt{\frac{X}{n^2}\left(1-\frac{X}{n}\right)}} \le c\right) \simeq 1-\alpha\ ,$$

leading to the Wald interval (Method V in Table 1). Method VI is similar but with continuity correction.

Noting that, when $X=0$ or $X=n$, the Wald interval has zero length, something an applied statistician will be reluctant to present, it is wise to consider replacing it, just in these two cases, by the exact Clopper–Pearson expressions (2.3). This is denoted Method VII (without continuity correction). Method VIII is just the same but with continuity correction. Vollset (1993) also considered these two modifications of the Wald interval.

Another possibility for correcting the Wald interval (mentioned for instance by Brown, Cai and DasGupta, 2002) is to recenter it at the center of the score interval, which is given by $(X + c^2/2)/(n + c^2)$, see Table 1. This modification is considered here together with the previous one, both without and with continuity correction (Methods IX and X, respectively).

2.3. Bootstrap methods

It was considered interesting to include in this study some bootstrap methods because, in this particular situation, it is not necessary to use a Monte Carlo approximation, and also because this example is not usually mentioned in the bootstrap literature (the only reference found was Hjorth, 1994, pp. 110–111, and not with the options taken here).

The non-parametric bootstrap method introduced by Efron (1979), consists on making inferences about a population using solely the empirical distribution of the observed sample. In the present context, as the sample consists on X successes and $n-X$ failures, the empirical distribution function is given by

$$F_n(y) = \begin{cases} 0, & y<0 \\ 1-\dfrac{X}{n}, & 0\leq y<1 \\ 1, & y\geq 1 \end{cases} ,$$

that is, the distribution function of a Bernoulli$\left(\frac{X}{n}\right)$ random variable. Considering the estimator of p, $\hat{p}=X/n$, we obtain the bootstrap distribution of this estimator by noting that the bootstrap distribution of $n\hat{p}$ is the distribution of the number of successes on a random sample of size n from a Bernoulli$\left(\frac{X}{n}\right)$ population, that is, Binomial$\left(n, \frac{X}{n}\right)$.

Given this distribution several methods can be used to obtain two-sided confidence intervals for the parameter of interest, p. One of those methods is the Percentile Method, which in this case consists on taking the percentiles $\alpha/2$ and $1-\alpha/2$ of the bootstrap distribution of $\hat{p}$, leading to Method XI in Table 2.

Since the parameter p varies continuously in $[0,1]$ and the quantities given by $L_{XI}(X)$ and $U_{XI}(X)$ vary discontinuously by $1/n$ it makes sense to introduce here a kind of continuity correction. This is called Method XII (see Table 2).

The Percentile Method is usually considered to have some drawbacks, and several corrections have been proposed for it. One is the BCP (*Bias Corrected Percentile*, see e.g. Shao and Tu 1995) which consists on replacing the previous percentiles ($\alpha/2$ and $1-\alpha/2$) by other percentiles accounting for the asymmetry of the bootstrap distribution. Thus, $\alpha/2$ is replaced by

$$\alpha' \;=\; \Phi\left(z_{\alpha/2} + 2\times\Phi^{-1}\left(K_B\left(\frac{X}{n}\right)\right)\right) \tag{2.5}$$

and $1-\alpha/2$ by

$$\alpha'' \;=\; \Phi\left(z_{1-\alpha/2} + 2\times\Phi^{-1}\left(K_B\left(\frac{X}{n}\right)\right)\right) , \tag{2.6}$$

where K_B denotes the bootstrap distribution function. Due to the discrete nature of this distribution a further correction must be applied and the one used was

$$K_B\left(\frac{X}{n}\right) = \left[F_{\mathrm{Bin}\left(n,\frac{X}{n}\right)}(X) + F_{\mathrm{Bin}\left(n,\frac{X}{n}\right)}(X-1)\right]\Big/2 \; .$$

Note that if the bootstrap distribution is symmetric (which happens when $X/n \simeq 0.5$) then $K_B(X/n) \simeq 0.5$, $\alpha' \simeq \alpha/2$ and $\alpha'' \simeq 1-\alpha/2$, and there is practically no correction to the raw percentile method. Method XIII refers to the bootstrap BCP method and Method XIV is similar but with the continuity correction introduced above.

In Methods XI to XIV the zero length intervals for $X=0$ and $X=n$ have been replaced by the exact Clopper–Pearson expressions (as it was done for Methods VII to X).

2.4. Normal approximations after a transformation

The next methods considered are based on the approximate normal distribution after the variance stabilizing transformation, that is on

$$\arcsin\sqrt{\frac{X}{n}} \;\xrightarrow{d}\; \mathcal{N}\left(\arcsin\sqrt{p},\, \frac{1}{4n}\right) \; .$$

Solving for p and correcting for inconsistencies in the extremes one obtains Method XV in Table 2. Introducing a continuity correction leads to Method XVI. A refinement of this method using a correction due to Anscombe (1948) is also considered (Method XVII).

2.5. Other approximations

This subsection introduces the last three methods consisting on very recent suggestions.

Agresti and Coull (1998) noting that the score interval has very good properties (confirmed in the comparative studies of Vollset 1993, Newcombe 1998, and again in the next section of this paper), that it is centered around

$$\tilde{p}(c) = \frac{X + c^2/2}{n + c^2} \; , \tag{2.7}$$

and that for 95% confidence $c^2 \simeq 4$, suggested a simple, yet effective, method:

add 4 observations to the sample, two successes and two failures, and then use the Wald formula (Method V). This method will be referred as Add 4 or Method XVIII. They also propose the use of $\tilde{p}(2)$ as point estimator and call it the Wilson point estimator, since Wilson (1927) was the first statistician recommending it (as a curiosity note that $\tilde{p}(\sqrt{2})$ is also an old estimator, the Laplace estimator).

Pan (2002) proposes a further modification, both on the Wald interval and the Add 4 interval, which consists on using percentiles of a suitable t-distribution instead of the normal. The modification of the Wald interval is denoted Method XIX whereas the one for the Add 4 interval is referred as Method XX. Let $V(p,n) = p(1-p)/n$ be the variance of $\hat{p}$. By introducing a scaled chi-square distribution and matching its first two moments with those of $V(\hat{p},n)$, Pan (2002) concludes that for the Wald-t interval the appropriate degrees of freedom are given by

$$\nu = \frac{2\,V(\hat{p},n)^2}{\Omega(\hat{p},n)}\ , \tag{2.8}$$

and that for the Add 4-t those are given by

$$\nu = \frac{2\,V\big(\tilde{p}(2), n+4\big)^2}{\Omega\big(\tilde{p}(2), n+4\big)}\ , \tag{2.9}$$

where

$$\begin{aligned}\Omega(p,n) \;=\; &\frac{p-p^2}{n^3} + \frac{p+(6n-7)p^2+4(n-1)(n-3)p^3-2(n-1)(2n-3)p^4}{n^5}\\ &-\,2\,\frac{p+(2n-3)p^2-2(n-1)p^3}{n^4}\ .\end{aligned}$$

3. COMPARISON OF THE METHODS

3.1. Central intervals

In order to evaluate and compare the performance of the twenty methods presented in Section 2, the coverage probability and the corresponding expected length have been computed for 5000 values of p, equally spaced in $[0.0001, 0.5]$, for every $10 \le n \le 1000$ and for $\alpha = 0.05,\, 0.01$.

The coverage probability, function of p, n and the method, $i = \mathrm{I}, ..., \mathrm{XX}$, is given by

$$CP\big(p,n,i\big) \;=\; \sum_{j=0}^{n} \binom{n}{j}\, p^j (1-p)^{n-j}\, I_{[L_i(j),U_i(j)]}(p)\ , \tag{3.1}$$

where $I_{[a,b]}(p)$ denotes the indicator function of the interval $[a,b]$, i.e. $I_{[a,b]}(p)=1$, if $p\in[a,b]$, and $I_{[a,b]}(p)=0$, if $p\notin[a,b]$. The expected length is

$$EL\big(p,n,i\big) \;=\; \sum_{j=0}^{n} \binom{n}{j}\, p^j(1-p)^{n-j}\,\big(L_i(j)-U_i(j)\big)\;. \tag{3.2}$$

The above computations were performed in R (code available upon request). The results are exact, within machine accuracy, and shall therefore not be interpreted as Monte Carlo results (from the papers cited before the only one giving Monte Carlo results is García-Pérez, 2005).

Figures 1 and 2 show plots of the coverage probability for two non extreme cases, $n=50$ and $n=500$, with $\alpha=0.05$. The plots for other values of n and for $\alpha=0.01$ are qualitatively similar to these two. The non-smooth aspect is expected due to the presence of the indicator function in expression (3.1). In fact, the coverage probability, as a function of p, has as many discontinuity points as the number of distinct values of L_i and U_i, about $2n$ in $(0,1)$. Between the discontinuity points $CP(p,n,i)$ is a polynomial of degree n.

Since it is impossible to analyze the several thousands of plots that could be produced, the results for the coverage probability were summarized in terms of observed minimum and mean on p for each n, and then plotted as a function of n. These plots are shown in Figures 3, 4, 5 and 6.

Considering the criteria described in Section 1, a possible classification of the methods is the following:

1.st Group – Strictly conservative methods, i.e. methods for which the minimum coverage probability is, for all $n\geq 10$ and for all p, greater or equal to $1-\alpha-0.005$ (nominal coverage probability rounded to two decimal places):

$$i:\quad \min_p CP\big(p,n,i\big)\;\geq\;1-\alpha-0.005\,,\quad \forall_{n\geq 10}\;.$$

2.nd Group – Methods not strictly conservative but correct on average, i.e. with mean coverage probability, for all $n\geq 10$, greater or equal to $1-\alpha-0.005$:

$$i:\quad \int_0^1 CP\big(p,n,i\big)\,dp\;\geq\;1-\alpha-0.005,\quad \forall_{n\geq 10}\;.$$

3.rd Group – Methods which are neither strictly conservative nor correct on average.

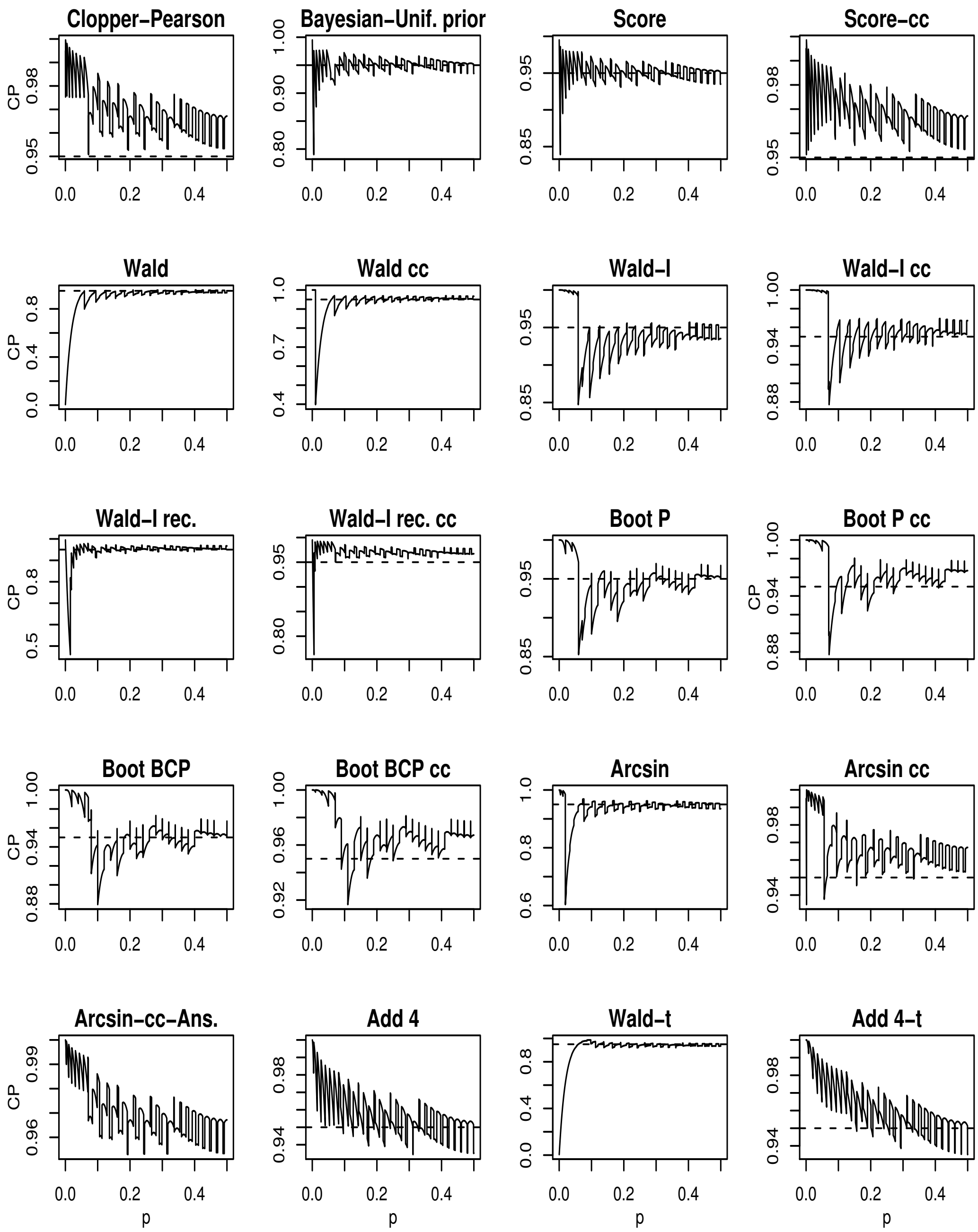

Figure 1: Coverage probability for each method as a function of p for $n = 50$ and $\alpha = 0.05$.

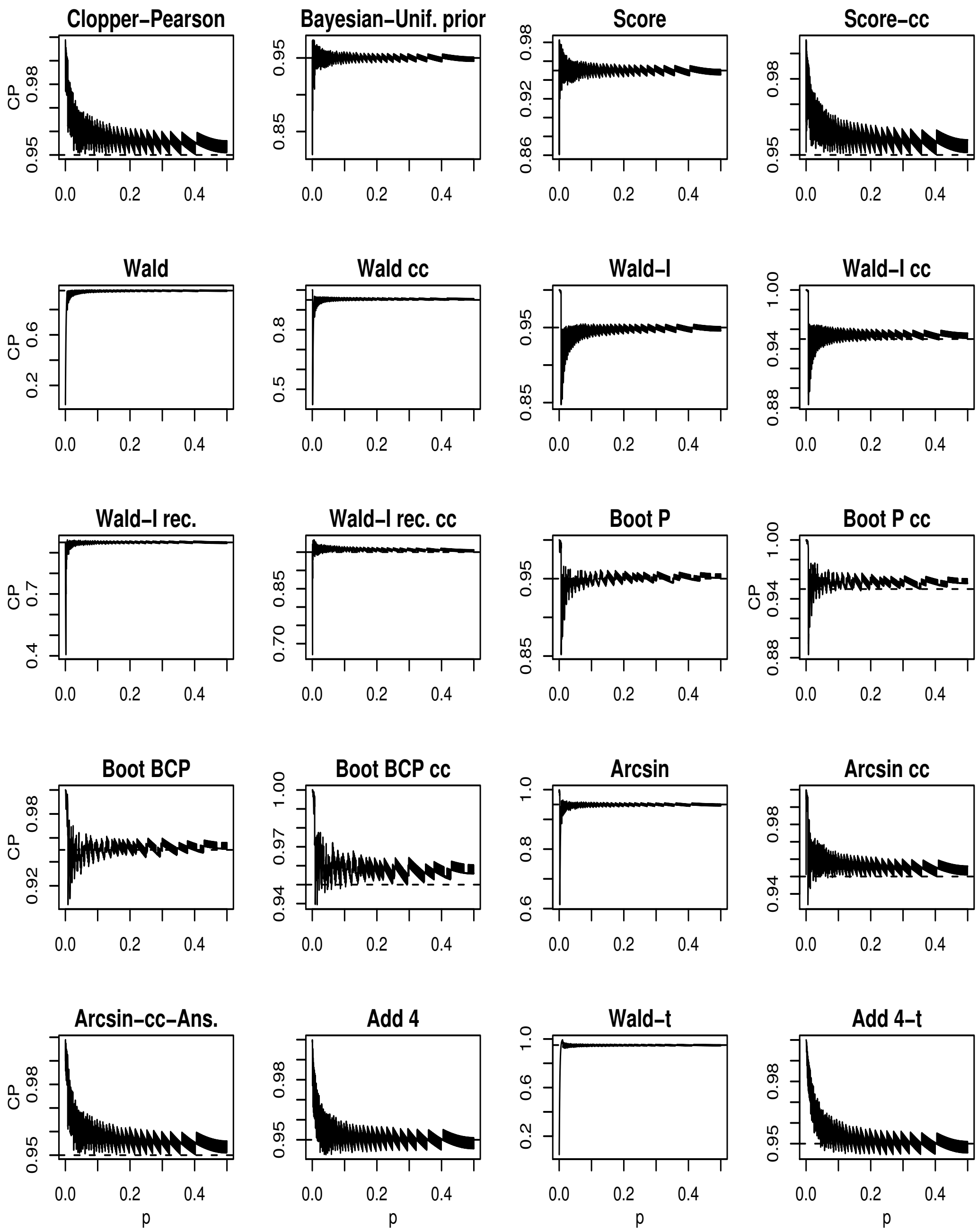

Figure 2: Coverage probability for each method as a function of p for $n = 500$ and $\alpha = 0.05$.

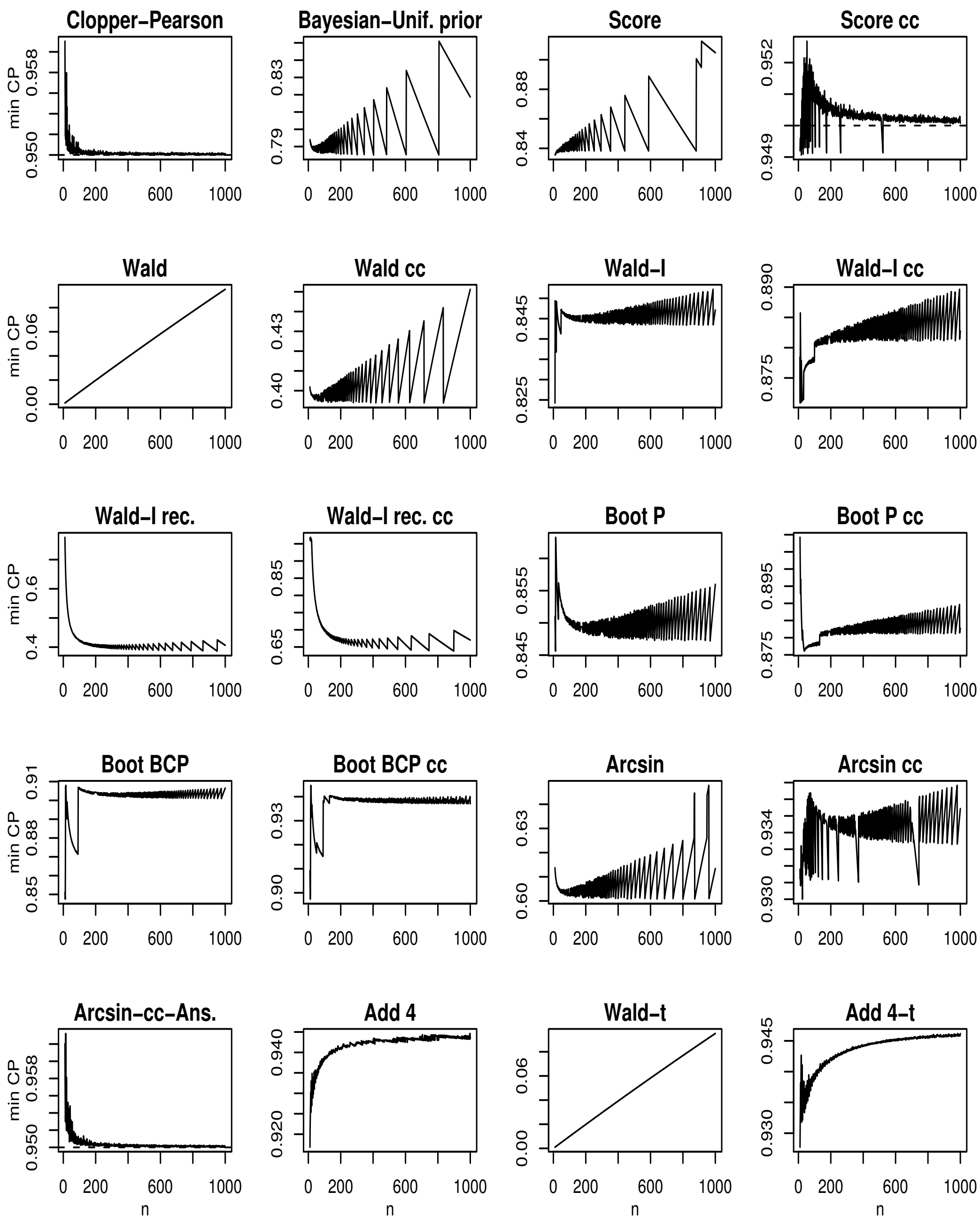

Figure 3: Minimum coverage probability for each method as a function of n $(10 \le n \le 1000)$ for 95% confidence.

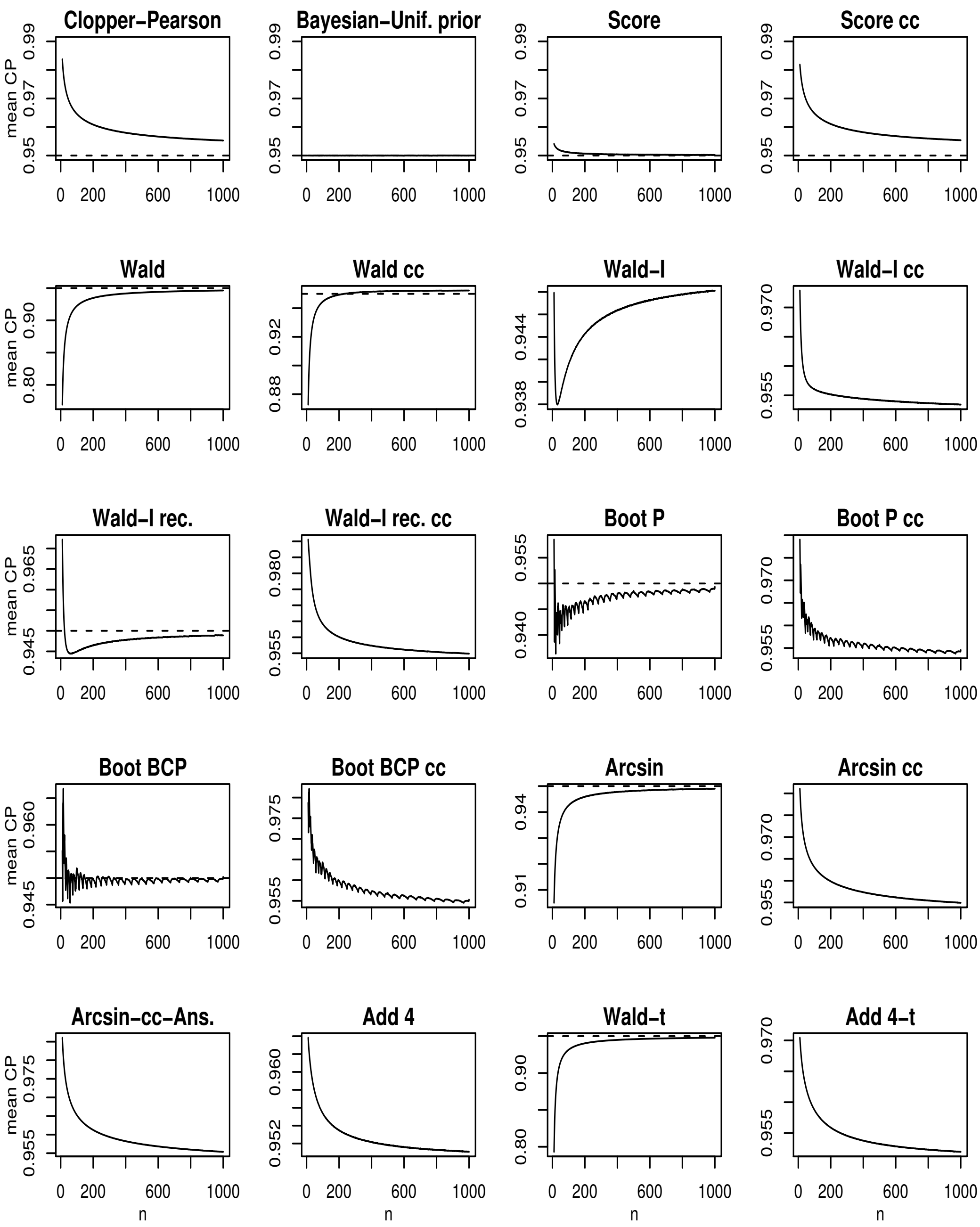

Figure 4: Mean coverage probability for each method as a function of n $(10 \le n \le 1000)$ for 95% confidence.

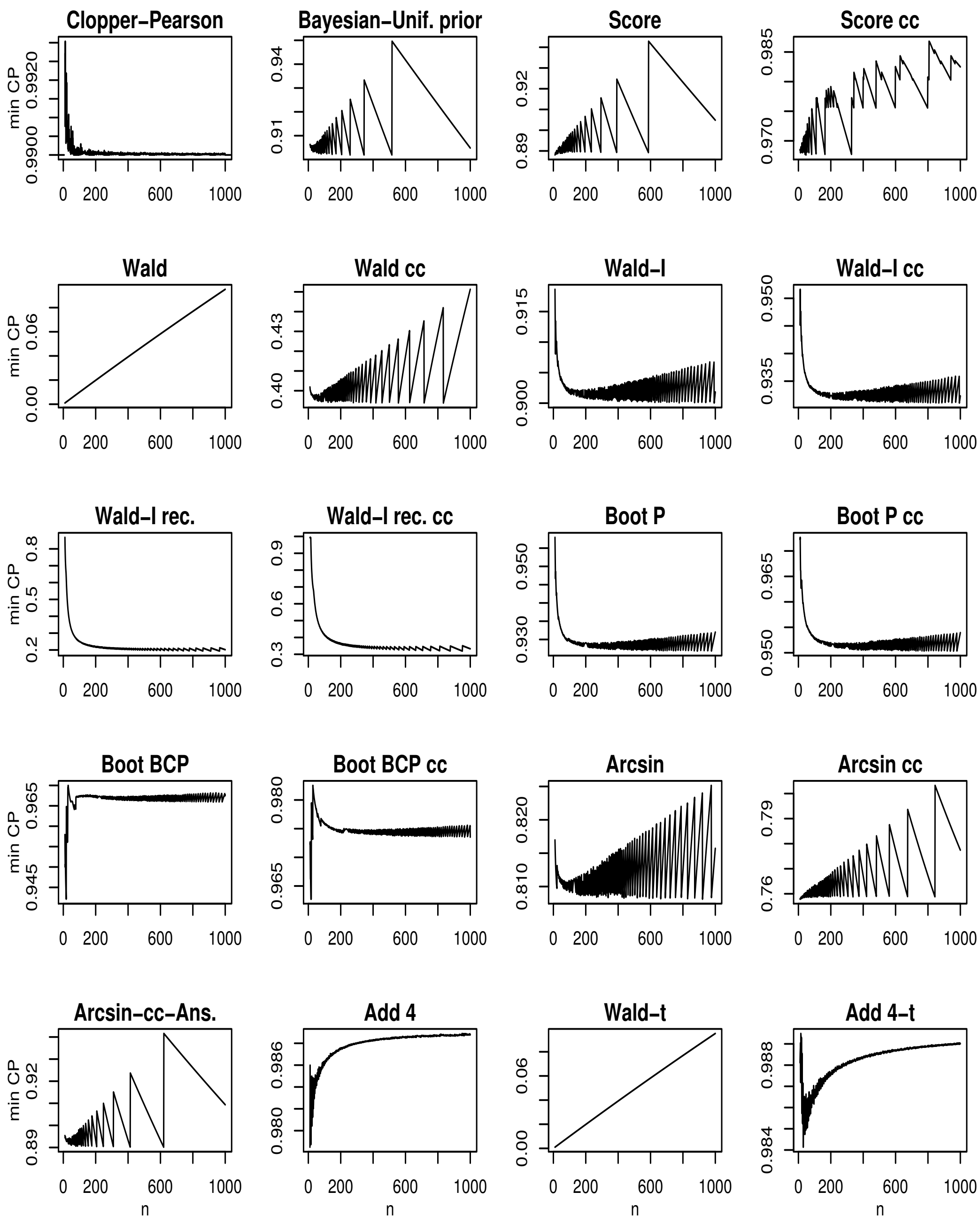

Figure 5: Minimum coverage probability for each method as a function of n $(10 \leq n \leq 1000)$ for 99% confidence.

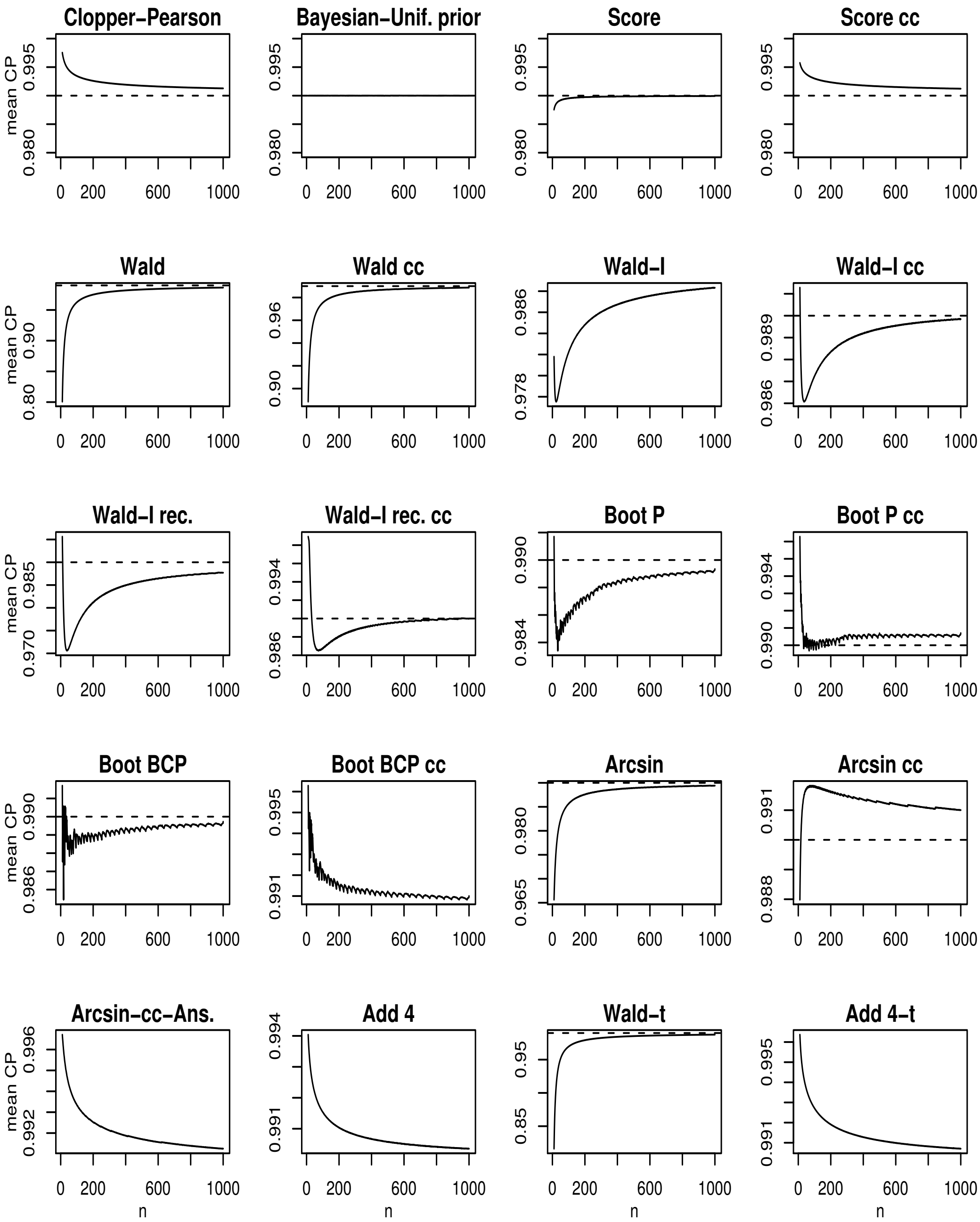

Figure 6: Mean coverage probability for each method as a function of n $(10 \leq n \leq 1000)$ for 99% confidence.

Table 3 summarizes the classification. 2% exceptions were allowed for classifying a method in the first group. The methods in the first group are shown by increasing order of mean coverage probability whereas those in the second group are shown by decreasing order of the minimum of the coverage probabilities overall n and p (shown into brackets). The composition of the groups depends slightly on the minimum value of n considered in this evaluation ($n \geq 10$). If the criterion was much less stringent, for instance $n \geq 200$ then: for 95% confidence Boot BCP cc and Arcsin cc would move to the first group and Wald cc and Boot P would enter the second group; for 99% confidence Add 4 would move to the first group and Boot P would enter the second group.

Table 3: Classification of the methods: strictly conservative (1st group); average correct (2nd group) (overall minimum of the coverage probabilities). Methods shown in **boldface** have not been considered in other comparative studies in the literature.

Group	95% confidence	99% confidence
1st	I – Clopper Pearson IV – Score cc **XVII – Arcsin cc Anscombe**	XX – Add 4-t **XVII – Arcsin cc Anscombe** I – Clopper Pearson
2nd	**XVI – Arcsin cc** (93%) XX – Add 4-t (93%) XVIII – Add 4 (92%) **XIV – Boot. BCP cc** (90%)	XVIII – Add 4 (98%) IV – Score cc (97%) **XIV – Boot. BCP cc** (96%) **XII – Boot. P cc** (95%)
	XII – Boot. P cc (88%) VIII – Wald-I cc (87%) **XIII – Boot. BCP** (85%) III – Score (84%) **II – Bayesian uniform prior** (79%) **X – Wald-I rec. cc** (64%)	**XIII – Boot. BCP** (94%) VIII – Wald-I cc (93%) **II – Bayesian uniform prior** (90%) III – Score (89%) **XVI – Arcsin cc** (76%) **X – Wald-I rec. cc** (32%)

It is worth noting that seven out of the thirteen methods classified in the first and second groups were not considered in the largest comparative studies in the literature (Vollset 1993, Newcombe 1998, Agresti and Coull 1998, Pan 2002).

Note that Method II is exact on average due to the coincidence between the prior distribution admitted and the distribution of p actually used to compute the mean coverage probability. The performance of this method is not so good due to the rather small value of the minimum coverage probability. If we had chosen the Jeffreys prior instead this aspect would have improved a little: 87% for 95% confidence (placing it between Methods VIII and XIV) and 96% for 99% confidence (between Methods IV and XIV).

It is also remarkable that some of the intervals maintain their good behavior for n as low as 10, in spite of being based on asymptotic results. It is the case namely of the score intervals (with and without continuity correction) and the ones based on the arcsine transformation.

Another interesting feature revealed by these results is that an apparently small modification may have a great impact in the performance of a method. For instance, the Wald method was completely disqualified in the vast majority of the papers mentioned in the introduction, however a simple modification at the boundary values and the simultaneous use of the continuity correction leads to an acceptable method, with better performance than the score or the Bayesian intervals. Figure 7 shows three plots illustrating this aspect for some values of n and α referred as “unlucky” by Brown, Cai and DasGupta (2001, 2002).

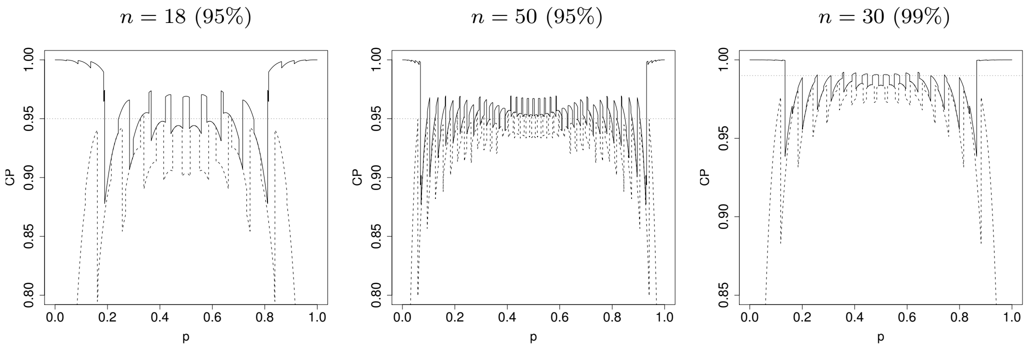

Figure 7: Coverage probability of the Wald interval (Method V, dashed) and the Wald interval with a modification at the boundary and continuity correction (Method VIII, solid) for some “unlucky” values of n and α.

Not surprisingly it is observed that the methods with continuity correction are always better in terms of coverage probability than the corresponding ones without that correction. It is also possible to verify that the bootstrap BCP method is slightly better than the percentile method and that both of these methods outperform the Wald methods. The results also show that the t correction of the Wald method proposed by Pan (2002) does not achieve its aim and it is in fact less effective than the usual continuity correction.

After this analysis of the coverage probabilities it is important to compare the expected lengths of the intervals. This comparison makes sense only within each group and only for the first and the second groups. Taking as reference Method I (Clopper–Pearson or exact) the ratio between the expected length of the intervals obtained using the other methods in the first group and the expected length by Method I was computed. The same was done with the first four methods in the second group (considering that the remaining, in spite of being correct on average, have an undesirable behavior in terms of the minimum coverage probability). Figure 8 shows the corresponding plots for $n = 100$. The plots for other values of n are qualitatively similar, but with the differences between the methods decreasing with n, especially at medium values of p.

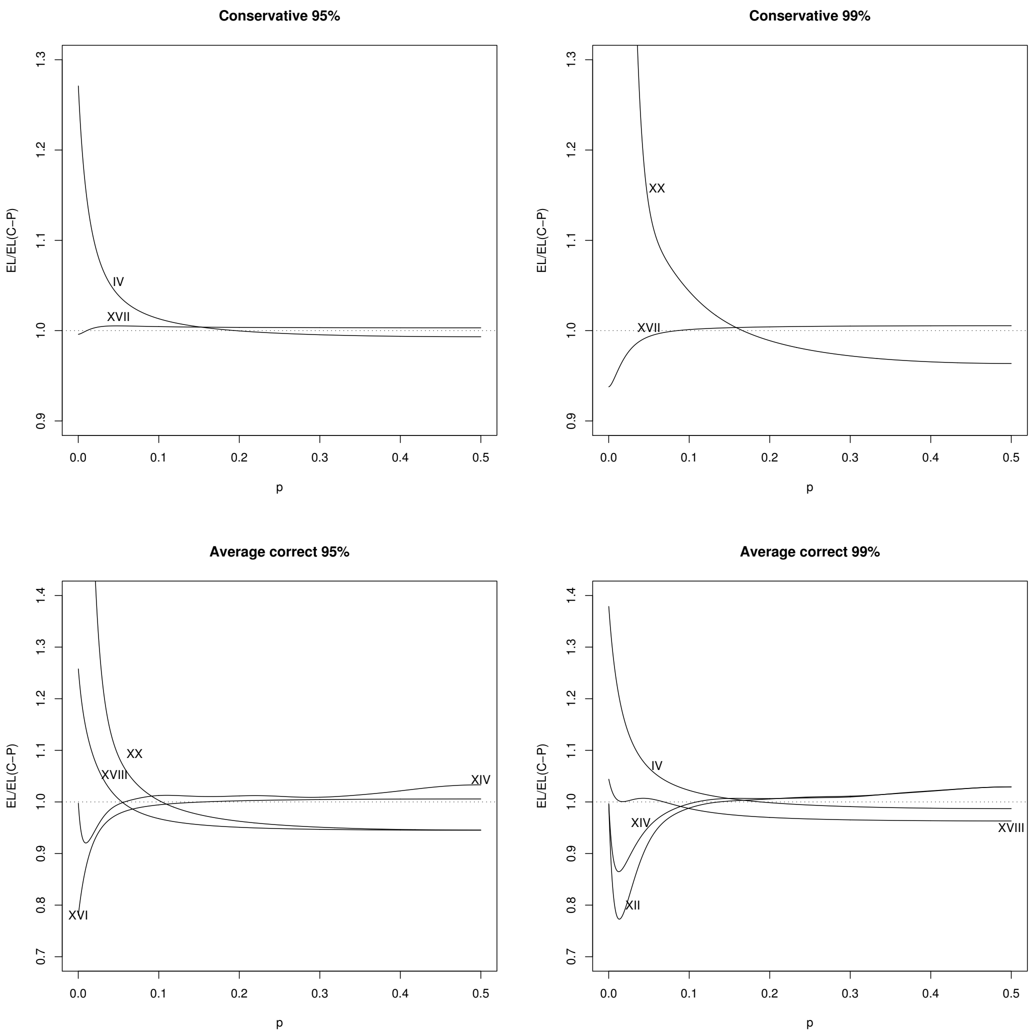

Figure 8: Ratio between the expected length of the conservative intervals and the expected length of the Clopper–Pearson interval (top) for $n = 100$. Ratio between the expected length of the first four top average correct intervals and the expected length of the Clopper–Pearson interval (bottom) for $n = 100$.

For the first group (top plots of Figure 8) the conclusion is that the Arcsine method with Anscombe's continuity correction is almost equivalent to the exact Clopper–Pearson interval in terms of length (and degree of conservativeness) but the Add 4-t is unnecessarily wide (or conservative).

In the second group (bottom plots of Figure 8) the conclusions are not so straightforward because there are more methods involved. However, for 95% confidence, it is possible to conclude that Method XVI (Arcsin cc) dominates Method XIV (Boot. BCP cc), having both better coverage and length. Method XVIII (Add 4) almost dominates Method XX (Add 4-t), it has better length but slightly smaller coverage probability. The choice between the dominating two methods is not so easy and depends on prior knowledge of the true p, if p is not small or large, Method XVIII is better because it leads to smaller length, if, on the contrary, p is either small or large, Method XVI is better. What is "small" or "large" depends on n and may not be easy to choose but after all it is not so important because any of these methods will produce a reasonable and safe interval whatever the value of p. For 99% confidence the conclusion is that Method XVIII (Add 4) dominates Method IV (Score cc) because of better length and coverage probability. The other two may lead to smaller lengths but at the cost of undesirably small minimum coverage probability. The best choice appears, therefore, to be the Add 4 method.

3.2. Non-central intervals

If one feels comfortable with the concept of a non-central interval then there are only two methods to choose from: the exact method according to the two criteria (minimum coverage probability of at least $1-\alpha$ or mean coverage probability equal to $1-\alpha$) and minimizing length.

To meet the first criterion and minimize the length of the interval one has to invert the test $H_0\colon p=p_0$ versus $H_1\colon p\neq p_0$ with size α, choosing for each p_0 the acceptance region, $A_n(p_0)\leq X\leq B_n(p_0)$, with smallest length. Then, given X, the confidence region is the set of those p_0 for which X is in the corresponding acceptance region. This is not an easy task and in fact many authors have addressed it (Sterne, 1954, Crow, 1956, Clunies-Ross, 1958, Blyth and Still, 1983, Casella, 1986, Reiczigel, 2003, see also the discussion in Santner and Duffy, 1989). The method we have implemented is based on Sterne's proposal (the acceptance interval for $p=p_0$ is made by including the most probable value of X, then the next most probable, ..., until the sum of their probabilities is greater than $1-\alpha$) with a slight modification. This modification is needed because although the acceptance region is always an interval the inverted region for p is not. We simply fill in the holes when they appear, which reduces to compute the interval by

$$L_{St}(X) \;=\; \min\Big\{p\colon\, A_n(p)\leq X\leq B_n(p)\Big\}$$

and

$$U_{St}(X) \;=\; \max\Big\{p\colon\, A_n(p)\leq X\leq B_n(p)\Big\}\;.$$

The slight modification has a very small practical effect: when compared with the intervals given in Table 2 of Blyth and Still (1983) for $n \leq 30$ there is only one different.

Figure 9 represents the results corresponding to Figures 1 and 2. It is clear that the Sterne interval is closer to the desired coverage than the Clopper–Pearson interval (it is less conservative but its coverage is still always over $1-\alpha$ as it should). This translates into smaller mean coverage probability (see Figure 10).

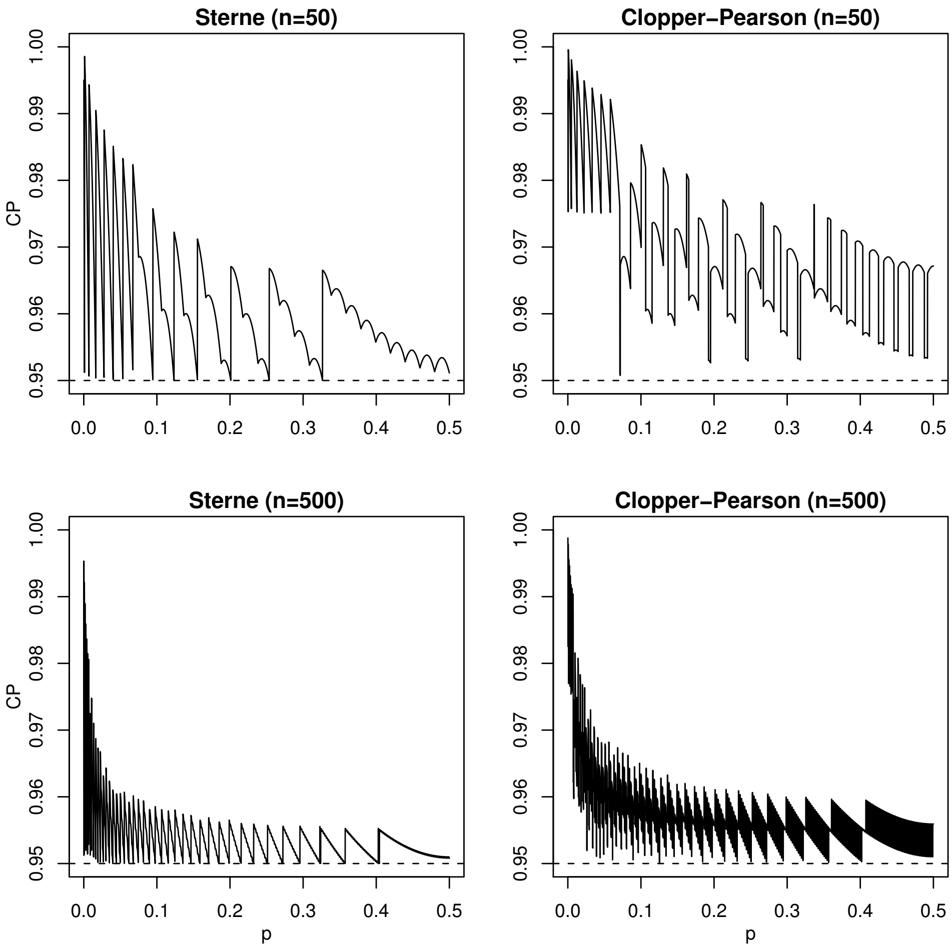

Figure 9: Coverage probability for the two exact frequentist intervals (non-central and central) as a function of p for $n = 50$ and $n = 500$ $(\alpha = 0.05)$.

Figure 11 reproduces the top two plots of Figure 8 with the curve corresponding to the Sterne interval. It does not come as a surprise that this interval has smaller expected length than the Clopper–Pearson interval except for values of p very close to the boundary.

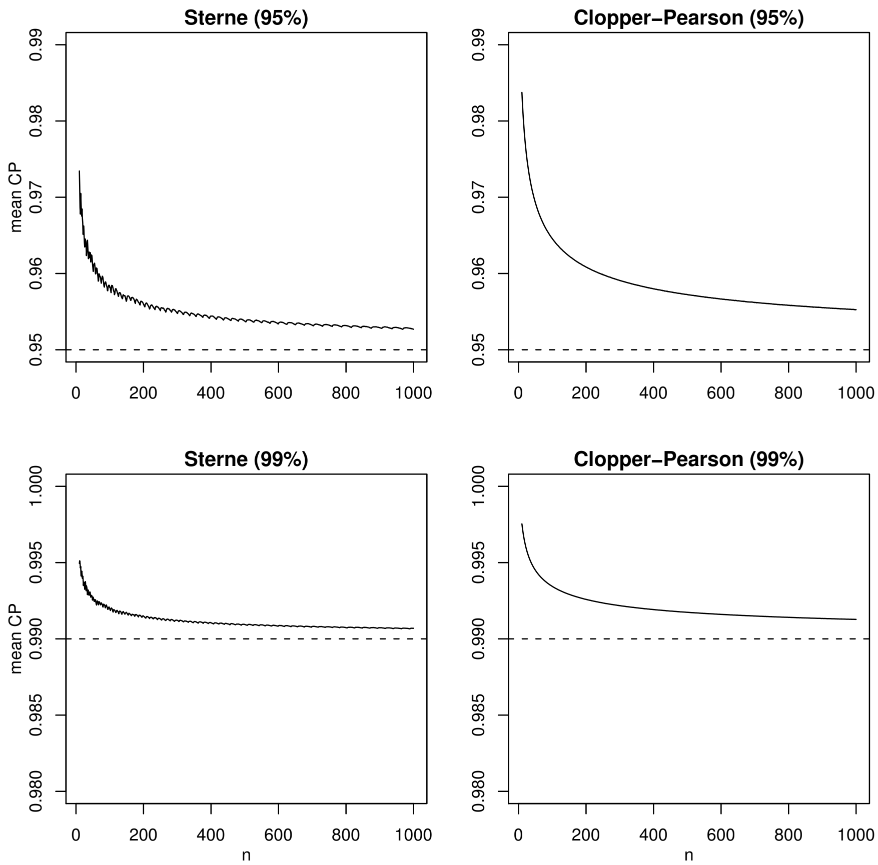

Figure 10: Mean coverage probability for the two exact frequentist intervals (non-central and central) as a function of n ($10 \leq n \leq 1000$) for 95% and 99% confidence.

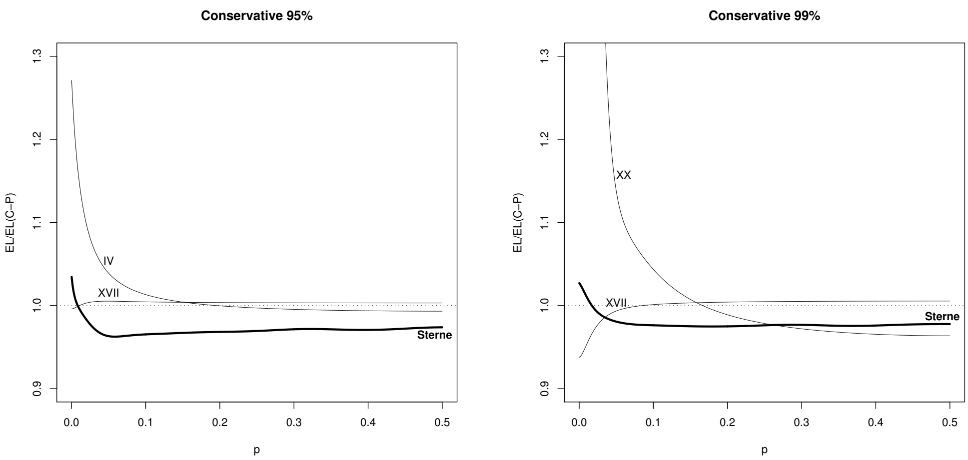

Figure 11: Ratio between the expected length of the conservative central intervals and the expected length of the Clopper–Pearson interval for $n = 100$ (solid thin lines) together with the same ratio for the non-central exact Sterne interval (solid thick lines).

When considering the second criterion (or mean coverage probability equal to $1-\alpha$) minimal length is achieved at ease by computing an HPD credibility interval. Given X, the HPD interval for the *a posteriori* Beta$(X+1, n-X+1)$ distribution can be determined in the following way: for every possible value of the left credibility tail, $0 \le \alpha' \le \alpha$, define

$$L(X,\alpha') = B_{X+1,n-X+1;\alpha'} \quad \text{and} \quad U(X,\alpha') = B_{X+1,n-X+1;1-(\alpha-\alpha')}$$

and determine $\alpha'(X)$ such that $U(X,\alpha') - L(X,\alpha')$ is minimum and denote it $\alpha'(X)$. The interval is given by $L_{HPD}(X) = L(X,\alpha'(X))$ and $U_{HPD}(X) = U(X,\alpha'(X))$. Note that, as mentioned in Subsection 2.1, we have $\alpha'(0) = 0$, $L_{HPD}(0) = 0$, $U_{HPD}(0) = 1 - \alpha^{1/(n+1)}$; and $\alpha'(n) = \alpha$, $L_{HPD}(n) = \alpha^{1/(n+1)}$, $U_{HPD}(n) = 1$.

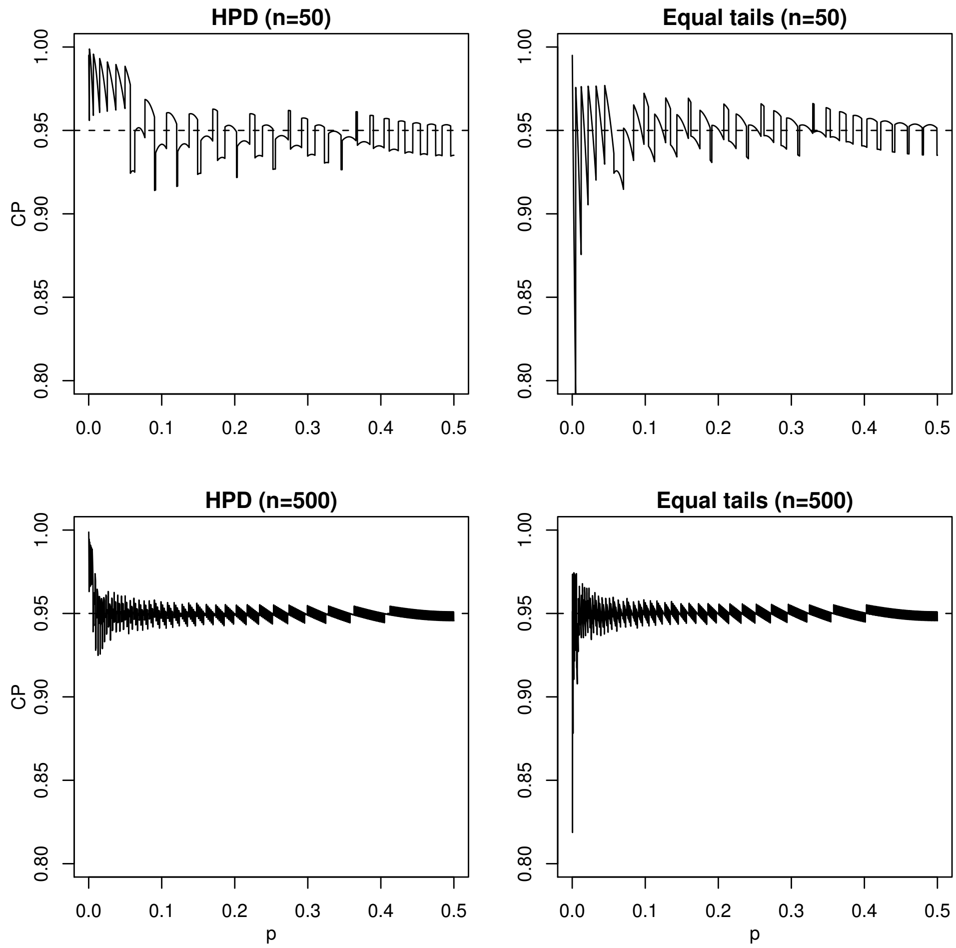

Figure 12: Coverage probability for the two Bayesian intervals with uniform prior (HPD and equal credibility tails) as a function of p for $n = 50$ and $n = 500$ $(\alpha = 0.05)$.

Figure 12 represents the results corresponding to Figures 1 and 2. For both intervals the coverage fluctuates around the target value but the HPD does not have the downward spikes lose to the boundaries. This is also evident from the minimum coverage probability plots (see Figure 13).

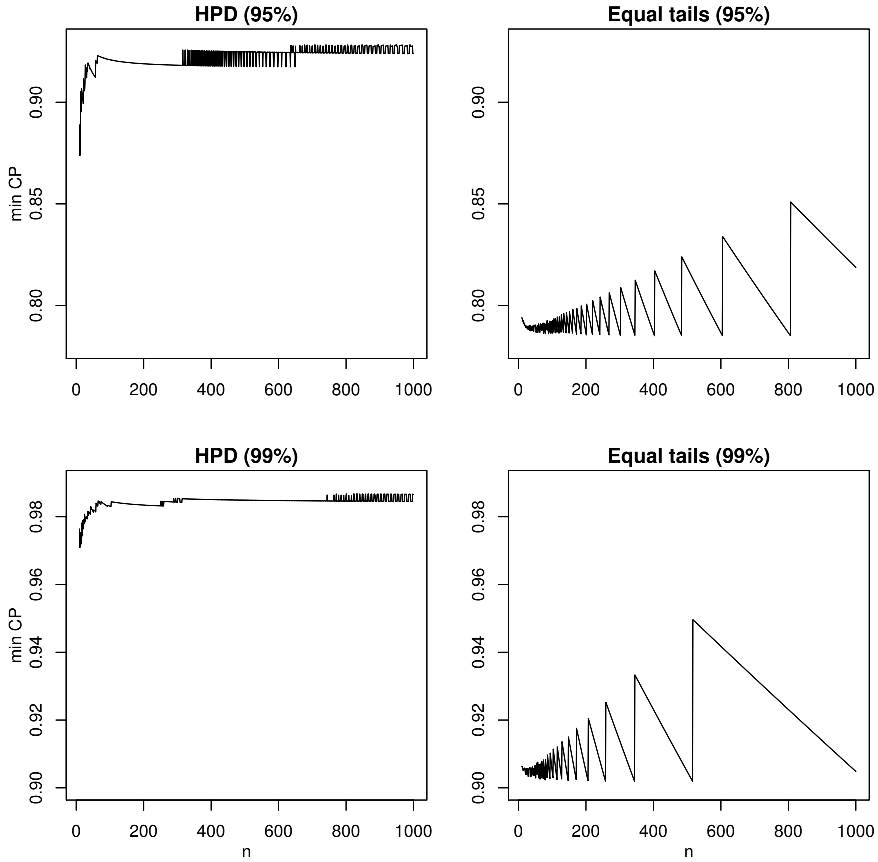

Figure 13: Minimum coverage probability for the two Bayesian intervals with uniform prior (HPD and equal credibility tails) as a function of n ($10 \leq n \leq 1000$) for 95% and 99% confidence.

Figure 14 reproduces the bottom two plots of Figure 8 with the curves corresponding to the two Bayesian intervals. As expected the HPD interval has smaller expected length than the other intervals except for values of p very close to the boundary (only for 99% confidence).

We have thus verified the optimality of the two exact procedures. What the statistician must decide is whether the reduction in expected length (of approximately 2% or 3% in the strictly conservative case and of approximately 5% or 7% in the average correct case) is worth the complications involved in the computations and the somehow different interpretation associated to non-central intervals.

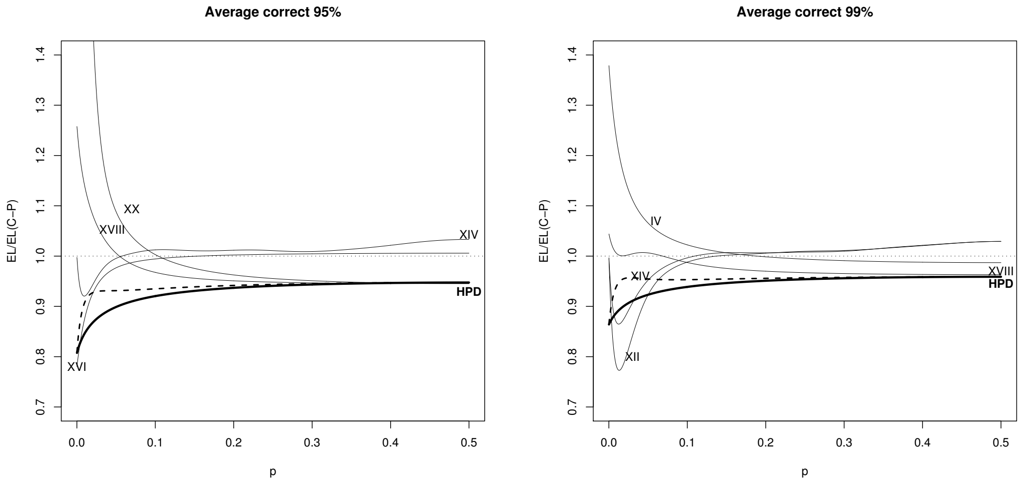

Figure 14: Ratio between the expected length of the first four top average correct central intervals and the expected length of the Clopper–Pearson interval for $n = 100$ (solid thin lines) together with the same ratio for the central Bayesian interval (dashed thick lines) and for the HPD interval (solid thick lines).

4. CONCLUDING REMARKS

The results reported in this paper have brought new insight into the apparently easy problem of determining an interval estimate for a binomial proportion. Although there is not a unique uniformly best choice, it is now easier to answer the two questions posed in the introduction and related, respectively, to applications and teaching.

4.1. Applications

When considering the computation of an interval estimate for a binomial proportion the first decision the applied statistician must take is related to the balance between degree of conservativeness and efficiency (equivalent in this case to the length of the intervals). Let us consider the two extreme options and only the class of central intervals:

(i) If strict conservativeness is mandatory than he or she must choose the Clopper–Pearson interval, or, almost equivalently, the arcsine interval with Anscombe's correction.

(ii) If strict conservativeness is not a major concern then length must be, subject to being at least correct on average. It is also wise to limit the "damage" measured by the overall minimum of the coverage probabilities and the recommendation is to choose, in the case of 95% confidence, either the Arcsin cc method or the Add 4 method. For 99% confidence the recommended method is the Add 4.

The score interval with continuity correction remains a valid choice, except that it may be too wide if the true p is close to 0 or to 1.

4.2. Teaching

In addition to the concerns of the applied statistician the teacher of statistics must also take into account the nature of the course and this may complicate the decision.

In a course for future statisticians the recommendations given in the previous subsection apply. The various methods should be taught and thoroughly discussed.

For elementary courses, typically less mathematically oriented and often unique, simplicity and lack of time for in depth discussions are a major concern. The Add 4 method of Agresti and Coull (1998) appears as a good choice, its properties are good and it is easy to compute. If, maybe for other reasons, one wants to stick to the Wald method then at least the continuity correction and the boundary modification should always be included.

4.3. Software

Four major statistical packages (SAS 9.1.3, S-Plus 8, SPSS 15 and R 2.6) were analyzed concerning the availability and correct implementation of interval estimates for the binomial parameter.

SAS provides, through PROC FREQ, the Wald interval, with and without continuity correction, and the exact Clopper–Pearson interval obtained with the percentiles of the F distribution.

In S-Plus there are two commands related to binomial proportions. The `prop.test` command gives the normal based hypotheses tests and the score intervals, with and without continuity correction. However, when using the continuity correction and when $X=0$ or $X=n$, the intervals given by this command are wrong, leading to $L(0) > 0$ or $L(n) < 1$. Another undesirable feature of this command is that it never applies the continuity correction when $X = n/2$, even when this option is set to TRUE. These apparently small details may have a strong visual impact and determine the classification of the method (see Figure 9). The `binom.test` command gives only the exact hypothesis test but could be easily modified in order to provide the exact Clopper–Pearson interval.

SPSS provides the asymptotic and the exact tests for binomial proportions but no confidence intervals (unlike other situations for which both the test and the confidence interval are provided, e.g. t test for the mean). Separately, there is a document describing how to compute the equal-tailed Jeffreys prior intervals (which are represented in bar charts).

The R software has commands with the same names as those of S-Plus (`prop.test` and `binom.test`), but the first makes the boundary correction and the second also gives the exact Clopper–Pearson interval. However the `prop.test` has the same problem with the continuity correction when $X = n/2$ (see Figure 15 and also Figure 1 of Geyer and Meeden, 2005).

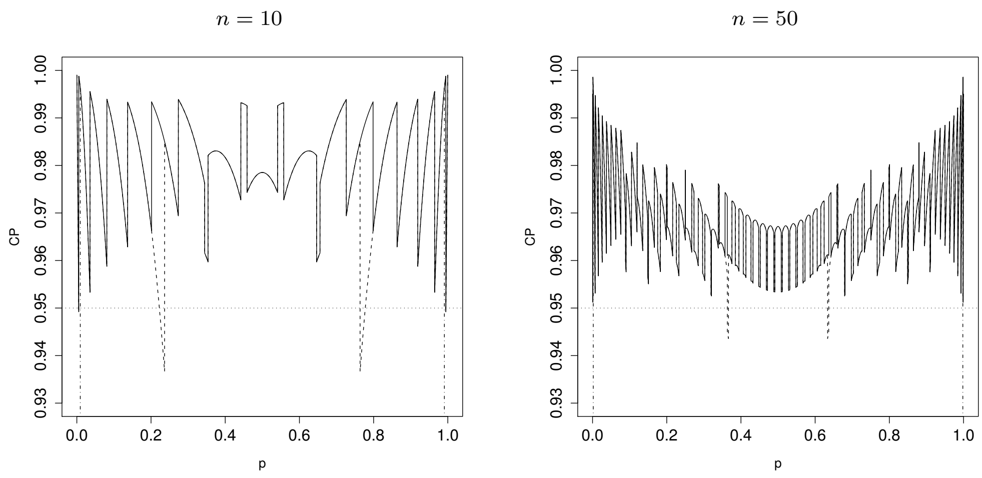

Figure 15: Effect of forcing no continuity correction for $X = n/2$ on the coverage probability of the Score cc interval (95% confidence). Correct implementation of the method (solid line) and implementation with the command `prop.test` of R and S-Plus (dashed line). The dashed-dotted line represents the results for `prop.test` of S-Plus at the extremes (no boundary correction when $X = 0$ or $X = n$).

In summary, the analyzed statistical packages do not treat the subject uniformly. This is perhaps a reflection of the recent spread of publications in the area. We hope that in the near future a consensus is reached and that it will be reflected in the software. This paper aims at contributing in that direction.

ACKNOWLEDGMENTS

We would like to thank an anonymous referee for helpful comments on the paper. This research was partially supported by the Center for Mathematics and its Applications, Lisbon, Portugal, through *Programa Operacional "Ciência, Tecnologia, Inovação"* (POCTI) of the *Fundação para a Ciência e a Tecnologia* (FCT), cofinanced by the European Community fund FEDER.

REFERENCES

[1] Agresti, A. and Caffo, B. (2000). Simple and effective confidence intervals for proportions and differences of proportions result from adding two successes and two failures, *The American Statistician*, **54**, 280–288.

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

[3] Angus, J.E. and Schafer, R.E. (1983). Improved confidence statements for the binomial parameter, *The American Statistician*, **38**, 189–191.

[4] Anscombe, F.J. (1948). Transformations of Poisson, binomial and negative-binomial data, *Biometrika*, **35**, 246–254.

[5] Armitage, P. and Berry, G. (1987). *Statistical Methods in Medical Research, Second Edition*, Blackwell, Oxford.

[6] Blyth, C.R. (1986). Approximate binomial confidence limits, *Journal of the American Statistical Association*, **81**, 843–855.

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

[8] Brown, L.D.; Cai, T.T. and DasGupta, A. (2001). Interval estimation for a binomial proportion (with discussion), *Statistical Science*, **16**, 101–133.

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

[10] Cai, Y. and Krishnamoorthy, K. (2005). A simple improved inferential method for some discrete distributions, *Computational Statistics and Data Analysis*, **48**, 605–621.

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

[12] Chen, H. (1990). The accuracy of approximate intervals for a binomial proportion, *Journal of the American Statistical Association*, **85**, 514–518.

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

[14] Clunies-Ross, C.W. (1958). Interval estimation for the parameter of a binomial proportions, *Biometrika*, **45**, 275–279.

[15] Cohen, G.R. and Yang, S.Y. (1994). Mid-p confidence intervals for the Poisson expectation, *Statistics in Medicine*, **13**, 2189–2203.

[16] Copas, J.B. (1992). Exact confidence limits for binomial proportions – Brenner and Quan revisited, *The Statistician*, **41**, 569–572.

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

[18] Edwardes, M.D. (1998). The evaluation of confidence sets with application to binomial intervals, *Statistica Sinica*, **8**, 393–409.

[19] Efron, B. (1979). Bootstrap methods: another look at the jackknife, *The Annals of Statistics*, **7**, 1–26.

[20] Fleiss, J.L.; Levin, B. and Paik, M.C. (2003). *Statistical Methods for Rates and Proportions, Third Edition*, Wiley, New York.

[21] Fujino, Y. (1980). Approximate binomial confidence limits, *Biometrika*, **67**, 677–681.

[22] García-Pérez, M.A. (2005). On the confidence interval for the binomial parameter, *Quality & Quantity*, **39**, 467–481.

[23] Geyer, C.J. and Meeden, G.D. (2005). Fuzzy and randomized confidence intervals and $p-$values (with discussion), *Statistical Science*, **20**, 358–387.

[24] Ghosh, B.K. (1979). A comparison of some approximate confidence intervals for the binomial parameter, *Journal of the American Statistical Association*, **74**, 884–900.

[25] Hjorth, J.S.U. (1994). *Computer Intensive Statistical Methods: Validation, Model Selection and Bootstrap*, Chapman and Hall, London.

[26] Johnson, N.L. and Kotz, S. (1969). *Discrete Distributions*, Wiley, New York.

[27] Leemis, L.M. and Trivedi, K.S. (1996). A comparison of approximate interval estimators for the Bernoulli parameter, *The American Statistician*, **50**, 63–68.

[28] Lee, S.-C (2006). Interval estimation of binomial proportions based on weighted Polya posterior, *Computational Statistics and Data Analysis*, **51**, 1012–1021.

[29] Lui, K.-J. (1995). Confidence limits for the prevalence ratio based on the negative binomial distribution, *Statistics in Medicine*, **14**, 1471–1477.

[30] Newcombe, R.G. (1994). Confidence intervals for a binomial proportion, *Statistics in Medicine*, **13**, 1283–1285.

[31] Newcombe, R.G. (1998). Two-sided confidence intervals for the single proportion: comparison of seven methods, *Statistics in Medicine*, **17**, 857–872.

[32] Pan, W. (2002). Approximate confidence intervals for one proportion and difference of two proportions, *Computational Statistics and Data Analysis*, **40**, 143–157.

[33] Puza, B. and O'neill, T. (2006). Generalised Clopper–Pearson confidence intervals for the binomial proportion, *Journal of Statistical Computation and Simulation*, **76**, 489–508.

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

[35] Santner, T.J. and Duffy, D.E. (1989). *The Statistical Analysis of Discrete Data*, Springer, New York.

[36] Shao, J. and Tu, D. (1995). *The Jackknife and Bootstrap*, Springer, New York.

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

[38] Stevens, W.L. (1950). Fiducial limits of the parameter of a discontinuous distribution, *Biometrika*, **37**, 117–129.

[39] Vollset, S.E. (1993). Confidence intervals for a binomial proportion, *Statistics in Medicine*, **12**, 809–827.

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