Bachelor's thesis

NTNU
Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Mathematical Sciences

Ylva Sofie Tollefsen

Different Approaches for Calculating the Confidence Intervals for the Binomial Proportion

Bachelor's thesis in Bachelor in Mathematical Sciences
Supervisor: Øyvind Bakke
June 2023

NTNU
Norwegian University of
Science and Technology

Ylva Sofie Tollefsen

Different Approaches for Calculating the Confidence Intervals for the Binomial Proportion

Bachelor's thesis in Bachelor in Mathematical Sciences
Supervisor: Øyvind Bakke
June 2023

Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Mathematical Sciences

NTNU
Norwegian University of
Science and Technology

Abstract

There are plenty of different methods for creating confidence intervals for the binomial proportion, and these can be divided into two groups: exact methods with coverage probability above the confidence level $1 - \alpha$, and asymptotic methods, which may have lower coverage probability (Thulin, 2014, p. 818). In this thesis we derive and compare four different methods, two asymptotic and two exact, mainly based on the actual coverage probability and the expected length of the intervals. The four methods are the Wald method (Laplace, 1812) (with and without continuity correction), the Wilson score method (Wilson, 1927), the Clopper–Pearson method (Clopper and Pearson, 1934) and the Blaker method (Blaker, 2000). The coverage probability of the original Wald method is in most cases unsatisfactory. In addition, it struggles with both zero-width intervals and overshooting. Although the coverage probability is better when using the Wald method with continuity correction, the interval becomes unnecessary large and the problem with overshooting worsen. The Wilson score method performs a lot better, with average coverage above the confidence level for all tested n, and narrower intervals. For this reason it proves to be a better choice when going for an asymptotic method. The Clopper–Pearson method is quite conservative, with wide intervals and coverage probability frequently a lot larger than the confidence level, unless when n is very large. The Blaker method produces intervals which are subsets of the Clopper–Pearson intervals (Klaschka and Reiczigel, 2021, pp. 1779–1780), thus they are less conservative. They are narrower with coverage probability closer to the confidence level for every tested n. However, this method has some unwanted properties, that the Clopper–Pearson method does not (Vos and Hudson, 2008, p. 81).

Sammendrag

Det finnes mange forskjellige metoder for å konstruere konfidensintervall for p i den binomiske fordelingen, der p er sannsynligheten for suksess i hvert forsøk. De kan bli delt inn i to grupper: eksakte og asymptotiske metoder. For eksakte metoder er sannsynligheten for at parameteren p blir dekket av konfidensintervallet høyere eller lik konfidensnivået $1 - \alpha$, mens for asymptotiske metoder kan sannsynligheten være mindre (Thulin, 2014, p. 818). I denne avhandlingen skal vi utlede og sammenlikne fire forskjellige metoder, to asymptotiske og to eksakte, hovedsakelig basert på den faktiske dekningsgraden og den forventede lengden til intervallene. De fire diskuterte metodene er Waldmetoden (Laplace, 1812) (med og uten kontinuitetskorreksjon), Wilson-score-metoden (Wilson, 1927), Clopper–Pearson-metoden (Clopper and Pearson, 1934) og Blakermetoden (Blaker, 2000). Dekningsgraden for den orginale Waldmetoden er som oftest utilstrekkelig. I tillegg sliter den med intervaller med lengde lik 0, og at de overstiger 1 eller går under 0. Dekningsgraden er bedre med Waldmetoden med kontinuitetskorreksjon, men intervallene blir unødvendig store og går oftere utenfor intervallet $[0, 1]$. Wilson-score-metoden presterer mye bedre, med gjennomsnittlig dekningsgrad over konfidensnivået for alle n, og med smalere intervaller. Av den grunn konkluderes det med at Wilson-score-metoden er den beste asymptotiske metoden. Clopper–Pearson-metoden er ganske konservativ, med store intervaller og med dekningsgrad som ofte er mye større enn konfidensnivået, med mindre n er veldig stor. Blakermetoden er mindre konservativ enn Clopper–Pearson-metoden, ettersom intervallene er delmengder av Clopper–Pearson-intervallene (Klaschka and Reiczigel, 2021, pp. 1779–1780). De er smalere med dekningsgrad nærmere konfidensnivået for alle n. Den har imidlertid flere uønskede egenskaper som vi ikke ser hos Clopper–Pearson-metoden (Vos and Hudson, 2008, p. 81).

Table of Contents

List of Figures

List of Tables

1 Introduction

Estimation of unknown parameters is a common statistical problem. For example, if a group of patients is receiving a type of medicine, medical practitioners are often interested in knowing the amount of patients expected to experience certain negative side effects. By using data about previous receivers of this medication, statisticians can provide an estimate for the expected proportion to experience said side effects. In this example, if we assume certain conditions, which will be explained later, we can say that the parameter we want to estimate is the binomial proportion, and that the number of patients with side effects is described by a binomial distribution.

The calculated estimate is either a point estimate or an interval estimate. In the medical example, the point estimate can be the number of patients with side effects divided by the total number of patients, while the interval estimate would be an interval in which the parameter is expected to lie. It is often an advantage to provide interval estimations rather than a point estimation, partially due to the extra information an interval estimation provides, as well as making the user more attentive to the uncertainty of the estimation (Larsen and Marx, 2012, pp. 297–299).

The most common type of interval estimates are confidence intervals. A confidence interval is created at a confidence level, which is the long-run proportion of intervals containing the true value of the unknown parameter (Larsen and Marx, 2012, pp. 297–299). There are plenty of different methods to choose between when wanting to compute a confidence interval for the binomial proportion. Where each method has its own advantages and disadvantages, some are easy to calculate, while others are more precise and reliable. They can be categorized into two groups: exact and asymptotic methods. The exact method guarantee the actual coverage to be larger than the confidence level, whereas for the asymptotic method, the actual coverage can go below, but the intervals are usually narrower and easier to compute (Thulin, 2014, p. 818).

Since there are so many different methods for computing confidence intervals for the binomial parameter, it is quite interesting to compare them, and discuss which should be used. This thesis present and compare four different methods for computing confidence intervals for the binomial distribution, beginning with two asymptotic methods: the Wald method, with and without continuity correction (Laplace, 1812), and the Wilson Score method (Wilson, 1927). This is followed by two exact methods, the most famous being the Clopper–Pearson method (Clopper and Pearson, 1934), and lastly the Blaker method (Blaker, 2000). The comparisons will be based on the complexity of the calculation, issues with the method, actual coverage and the length of the intervals.

2 Background Theory

This section introduces the relevant theory needed in order to understand the discussed topics, such as the binomial probability distribution, point and interval estimation, and most importantly, confidence intervals. This is followed up by introducing the Wald method, with and without continuity correction.

2.1 Binomial Probability Distribution

The definition of the binomial probability distribution can be found in any introductory book in the field of statistics (Larsen and Marx, 2012, p. 104; Agresti, 2002, pp. 5–6). The following is a short summary of the binomial distribution and its properties.

Assume you have n Bernoulli trials, meaning n independent and identical trials, each trial resulting in either "success" or "failure", with the probability of success in each trial is p. Furthermore, let the random variable X represent the number of successes of the n trials, which is to say that X is a binomial random variable; this can be denoted by $X \sim \text{binomial}(n, p)$. Then

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

is the probability mass function of the binomial distribution, with mean $\text{E}[X] = np$, and variance $\text{Var}[X] = np(1-p)$.

If n is large, one can use the normal distribution to approximate binomial probabilities. This follows from the De Moivre–Laplace theorem, which is a special case of the central limit theorem. Larsen and Marx (2012, p. 239) presents the theorem as follows:

Theorem 1 (De Moivre–Laplace Theorem)**.** *Let* $X \sim \text{binomial}(n, p)$*. Define* $Z_n = \frac{X-np}{\sqrt{np(1-p)}}$*. For any numbers* a *and* b*,*

$$\lim_{n\to\infty} P(a \leq Z_n \leq b) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-\frac{z^2}{2}}\, dz.$$

Theorem 1 states that Z_n converges to the standard normal distribution as $n \to \infty$. Hence we can deduce that if n is sufficiently large, the normal distribution is a good approximation, as illustrated in Figure 1.

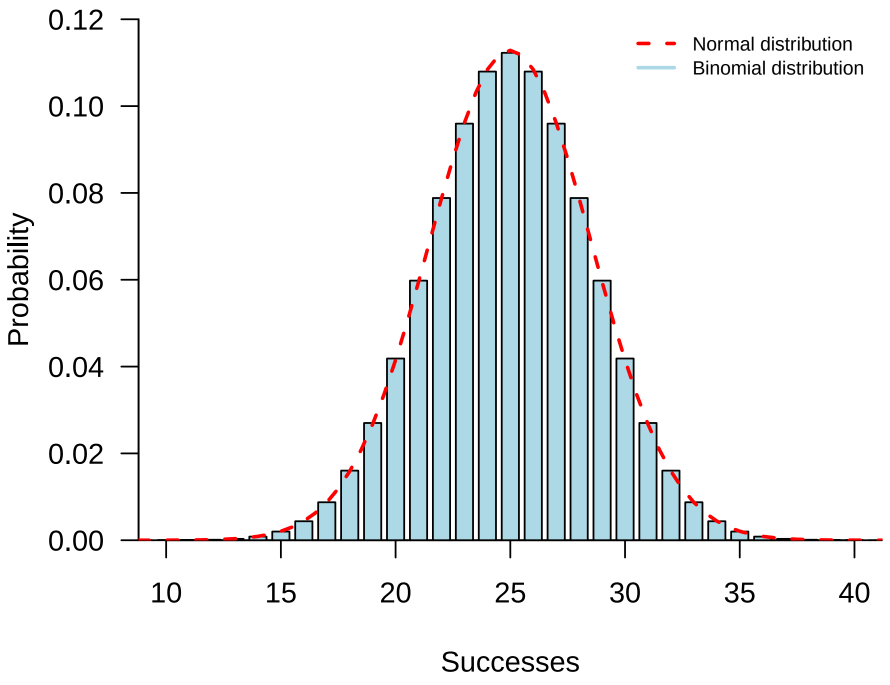


Figure 1: The binomial distribution with the normal approximation. The parameters for the binomial distribution are $p = 0.5$, $n = 50$. Both the normal and the binomial distribution has expectation and variance $\mu = 25$ and $\sigma^2 = 12.5$, respectively.

2.2 Estimation

Consider this scenario: You are interested in certain properties of a large population, for instance the average height of the women in Norway. Since the female population is quite large, it would be very impractical to measure every female. Under those circumstances, you could choose a smaller part of the population, called a sample, and only measure those women. Then set the average height of the sample, as an estimate for the average height of female population in Norway. This is called a point estimate.

Due to the fact that estimates are calculated from a sample of a population, it will rarely coincide with the true value. Thus, it is often of interest knowing the precision of the estimate. For example, the estimate from a sample consisting of 20 females is likely less accurate than the estimate from a sample of 1000 females. However, the point estimator alone gives no indication for its precision. As Larsen and Marx (2012) explains, this is where interval estimation comes in; an interval estimate is an interval that has a high probability of containing the unknown parameter. The width of the interval is a good indication for the estimates precision, a narrower interval indicates better precision (Larsen and Marx, 2012, pp. 297–298). Hence, this is a good alternative to the point estimate due to the extra information the interval gives.

One of the most prevalent types of interval estimates are confidence intervals. Following is a short summary from Larsen and Marx (2012) and Keener (2010). A confidence interval is computed using a confidence level $1 - \alpha$, usually 95%. The confidence level represents the percentage of the computed confidence intervals containing the true value in the long run. Formally, let X be a random variable having a probability distribution with parameter θ. We have that $(l(X), u(X))$ is a $1 - \alpha$-confidence interval for θ if

$$P(l(X) \leq \theta \leq u(X)) \geq 1 - \alpha. \tag{2.1}$$

Examples and more explanation are given by Larsen and Marx (2012, pp. 297–299) and Keener (2010, p. 161).

2.3 The Wald Method

One application of confidence intervals involves the binomial proportion parameter p. In introductory statistics books and courses the presented confidence interval for the binomial proportion is the Wald interval or the normal approximation interval, which is based on Theorem 1 (Brown et al., 2001, p. 103). It was first introduced by Pierre-Simon Laplace (1812), and is named after Abraham Wald, since it results from inverting a Wald test (Agresti and Coull, 1998, p. 119).

We will now derive the interval. Define z_α such that $P(Z \geq z_\alpha) = \alpha$, where Z is a standard normal random variable. Furthermore, let X, n and p be as described in Theorem 1, and assume n is sufficiently large, then we have that

$$P\left(-z_{\alpha/2} \leq \frac{X - np}{\sqrt{np(1-p)}} \leq z_{\alpha/2}\right) = P\left(-z_{\alpha/2} \leq \frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \leq z_{\alpha/2}\right) \approx 1 - \alpha, \tag{2.2}$$

where $\hat{p} = X/n$, which is the estimator of the population proportion. To find the confidence interval, we must get Equation (2.2) to the same form as Equation (2.1). First, we simplify the inequality

$$P\left(-z_{\alpha/2} \leq \frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \leq z_{\alpha/2}\right) \approx P\left(-z_{\alpha/2} \leq \frac{\hat{p} - p}{\sqrt{\hat{p}(1-\hat{p})/n}} \leq z_{\alpha/2}\right) \approx 1 - \alpha, \tag{2.3}$$

then solve it with respect to p, which results in

$$P\left(\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq p \leq \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) \approx 1 - \alpha.$$

Thus, the bounds of the Wald interval are

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}. \tag{2.4}$$

Thanks to the normal approximation and the simplification observed in Equation (2.3), this method only require simple calculations. However, the approximations have consequences for its performance; there is no guarantee that the coverage probability is above the confidence level. As a matter of fact, this is usually not the case (Agresti and Coull, 1998, p. 120).

For the Wald method to be somewhat useful, we depend on n being sufficiently large, and p not being close to 0 or 1. This is why this method is often accompanied with some conditions, which should hold if it is applicable in the current situation (Brown et al., 2001, p. 106). These conditions vary from text to text, luckily Brown et al. (2001) has collected a sample of the most popular; only use if either

1. $np, n(1-p)$ are ≥ 5 (or 10).
2. $np(1-p)$ are ≥ 5 (or 10).
3. $n\hat{p}, n(1-\hat{p})$ are ≥ 5 (or 10).
4. The interval with bounds $\hat{p} \pm 3\sqrt{\hat{p}(1-\hat{p})}/n$ does not contain 0 or 1.
5. n is quite large.
6. $n \geq 50$ unless p is very small.

These conditions are just guidelines, and even when one or more of these conditions are met, the performance is still unreliable. For this reason, many discourage using this method (Brown et al., 2001, pp. 103–107; Newcombe, 2013, pp. 56–58; Agresti and Coull, 1998, p. 122).

Another common problem with this method, discussed by Reed III (2007), is overshooting or zero-width intervals. Zero-width intervals occurs if $\hat{p} = 1$ or $\hat{p} = 0$, see Equation (2.4). Overshooting, on the other hand, can occur if $p \approx 0$ or $p \approx 1$; the lower bound may drop below zero, or the upper bound exceed one, respectively. Both of these problems is shown in Figure 2a; when $x = 0$ and $x = 10$ you see that the width of the interval is zero, and when $x = 1, 2, 8, 9$ you can see that the interval either drops below zero, or goes above one. One way to fix this is to intersect the interval with $[0, 1]$, but there should be better solutions (Reed III, 2007, p. 154).

For small samples it may be beneficial to use the continuity corrected version of the Wald interval. The bounds are given by

$$\frac{X - 1/2 - np}{\sqrt{np(1-p)}} \leq z_{\alpha/2} \quad \text{and} \quad \frac{X + 1/2 - np}{\sqrt{np(1-p)}} \geq -z_{\alpha/2}.$$

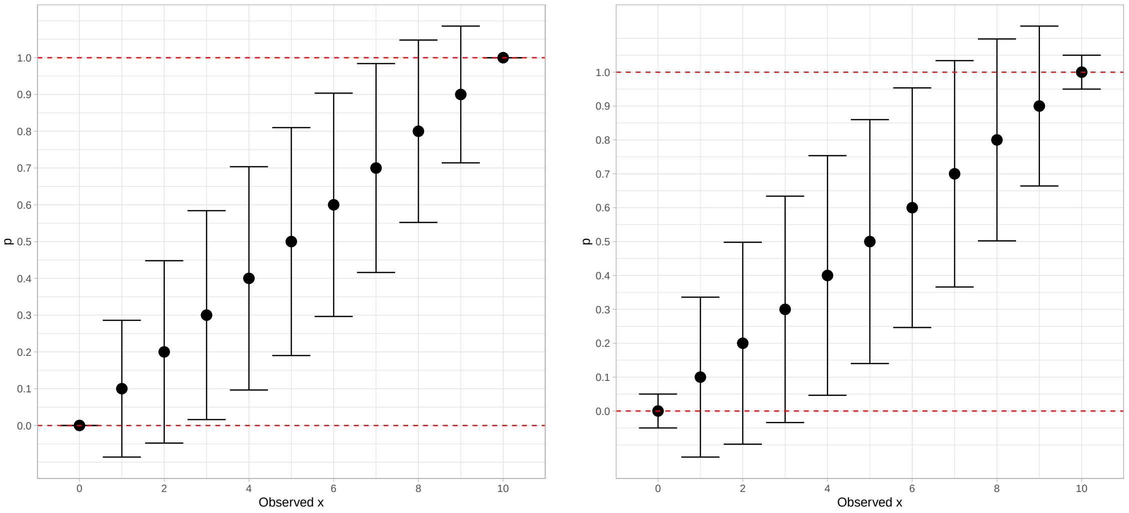
We solve these equations using the same notation and simplification as in Equations (2.2) and (2.3). Thus, the equations we want to solve are

$$\frac{\hat{p} - 1/(2n) - p}{\sqrt{\hat{p}(1-\hat{p})/n}} \leq z_{\alpha/2} \quad \text{and} \quad \frac{\hat{p} + 1/(2n) - p}{\sqrt{\hat{p}(1-\hat{p})/n}} \geq -z_{\alpha/2},$$

with respect to p. The solution is

$$\hat{p} - \frac{1}{2n} - z_{\alpha/2}\frac{\sqrt{\hat{p}(1-\hat{p})}}{n} \leq p \leq \hat{p} + \frac{1}{2n} + z_{\alpha/2}\frac{\sqrt{\hat{p}(1-\hat{p})}}{n}.$$

The continuity corrected Wald (CCW) interval simply subtract $1/(2n)$ from the lower bound, and add $1/(2n)$ to the upper bound. As a consequence, the actual coverage percentage is at least as good as the original method, and later we will see by how much. Figure 2b shows the confidence intervals at $n = 10$. As expected, overshooting occurs more often with this method, compared to the Wald method. However, observe that when $\hat{p} = 0$ or $\hat{p} = 1$, the intervals do not collapse to zero-width. We see that zero-width intervals do not occur when $x = 0$ or $x = 10$, as it does in Figure 2a, and in addition to the values of x overshooting occurs in Figure 2a, it overshoots when $x = 0, 3, 7, 10$.



(a) Wald without continuity correction

(b) Wald with continuity correction

Figure 2: Confidence intervals when $n = 10$, for the different observed values of x. $p = 0$ and $p = 1$ is marked with red dotted lines.

3 The Wilson Score Method

The next asymptotic method is the Wilson score method, or just the Wilson method, named after Edwin B. Wilson (Wilson, 1927). The Wilson intervals are asymmetric intervals, in contrast to Wald and CCW intervals, and thus, it works better with smaller sample sizes and skewed observations, and does not struggle with zero-width intervals and overshooting (Agresti and Coull, 1998, p. 120).

This method relies on the asymptotic normality of the binomial distribution as well, and thus, the calculation starts similarly, see Equation (2.2). Rather than using the approximation as in Equation (2.3), the bounds are the solution to the original quadratic equation (Agresti, 2002, pp. 15–16). Hence, solving

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

$$(\hat{p}-p)^2 = z_{\alpha/2}^2\left(\frac{1}{n}p(1-p)\right)$$

$$\hat{p}^2 - 2\hat{p}p + p^2 = \frac{z_{\alpha/2}^2 p}{n} - \frac{z_{\alpha/2}^2 p^2}{n}$$

$$0 = p^2\left(1+\frac{z_{\alpha/2}^2}{n}\right) - p\left(2\hat{p}+\frac{z_{\alpha/2}^2}{n}\right) + \hat{p}^2$$

with respect to p, results in the bounds of the Wilson interval. Solving this using the quadratic formula leads to

$$\begin{aligned} p &= \frac{\left(2\hat{p}+z_{\alpha/2}^2/n\right) \pm \sqrt{\left(2\hat{p}+z_{\alpha/2}^2/n\right)^2 - 4\left(1+z_{\alpha/2}^2/n\right)\hat{p}^2}}{2\left(1+z_{\alpha/2}^2/n\right)} \\ &= \frac{\left(2\hat{p}+z_{\alpha/2}^2/n\right) \pm z_{\alpha/2}/n\sqrt{z_{\alpha/2}^2+4n\hat{p}-4n\hat{p}^2}}{2\left(1+z_{\alpha/2}^2/n\right)} \\ &= \frac{2\hat{p}}{2\left(1+z_{\alpha/2}^2/n\right)} + \frac{z_{\alpha/2}^2/n}{2\left(1+z_{\alpha/2}^2/n\right)} \pm \frac{z_{\alpha/2}\sqrt{z_{\alpha/2}^2+4n\hat{p}-4n\hat{p}^2}}{2\left(n+z_{\alpha/2}^2\right)} \\ &= \frac{\hat{p}n}{n+z_{\alpha/2}^2} + \frac{z_{\alpha/2}^2}{2\left(n+z_{\alpha/2}^2\right)} \pm \frac{z_{\alpha/2}\sqrt{z_{\alpha/2}^2/4+n\hat{p}\left(1-\hat{p}\right)}}{n+z_{\alpha/2}^2}. \end{aligned}$$

The solutions to the above equation,

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

are the bounds of the Wilson interval.

4 The Clopper–Pearson Method

Since the previous intervals mentioned used the normal approximation to the binomial distribution, they cannot guarantee actual coverage above the confidence level. The next method, introduced by C.J. Clopper and E.S. Pearson (1934), does not depend on any approximation, and has actual coverage above the confidence level for every p and n. Intervals that guarantee actual coverage above the confidence level is sometimes referred to as exact intervals (Reed III, 2007, p. 154; Agresti and Coull, 1998, p. 119, Thulin, 2014, p. 818).

This method eliminates both the issue with overshooting, and zero-width intervals (Reed III, 2007, p. 154). Some of the disadvantages of this method are that the calculations are more complicated than with the asymptotic methods, and that it is considered too conservative; it regularly produces intervals that are unnecessary wide, with the coverage being far greater than the confidence level (Thulin, 2014, p. 818; Agresti and Coull, 1998, p. 119; Brown et al., 2001, p. 113).

4.1 Relevant Theorems

The Clopper–Pearson interval can be written as

$$\frac{1}{1+\frac{n-x+1}{x}F_{2(n-x+1),2x,\alpha/2}} \leq p \leq \frac{\frac{x+1}{n-x}F_{2(x+1),2(n-x),\alpha/2}}{1+\frac{x+1}{n-x}F_{2(x+1),2(n-x),\alpha/2}}, \tag{4.1}$$

where $F_{\nu_1,\nu_2,\alpha}$ is the upper quantile of the F-distribution, with ν_1 and ν_2 degrees of freedom (Casella and Berger, 2002, p. 454). Before proving that this is actually an exact confidence interval for the binomial parameter p, we need to present and prove some important results. The first is an identity that describes the relationship between the binomial and beta distribution, as formulated by Casella and Berger (2002, p. 454):

Theorem 2. *If $X \sim \text{binomial}(n,\theta)$, then $P_\theta(X \geq x) = P(Y \leq \theta)$, where $Y \sim \text{beta}(x, n-x+1)$.*

The notation $Y \sim \text{beta}(\alpha, \beta)$ denotes that Y follows a beta distribution with shape parameters α and β.

Proof. Let $X \sim \text{binomial}(n,\theta)$, and $Y \sim \text{beta}(x, n-x+1)$. We want to prove that $P(X \geq x) = P(Y \leq \theta)$. From the definitions of the distribution we have the following equations:

$$P(X \geq x) = \sum_{k=x}^{n} \binom{n}{k} \theta^k (1-\theta)^{n-k} \tag{4.2}$$

$$P(Y \leq \theta) = \int_0^\theta \frac{\Gamma(n+1)}{\Gamma(x)\Gamma(n-x+1)} t^{x-1}(1-t)^{n-x} dt \tag{4.3}$$

Equation (4.3) can be expressed as

$$P(Y \leq \theta) = x\binom{n}{x}\int_0^\theta t^{x-1}(1-t)^{n-x} dt.$$

Thus, checking whether $P(X \geq x) = P(Y \leq \theta)$, is the same as checking if

$$\sum_{k=x}^{n} \binom{n}{k} \theta^k (1-\theta)^{n-k} = x\binom{n}{x}\int_0^\theta t^{x-1}(1-t)^{n-x} dt.$$

Since it looks like both sides are easy to differentiate, we will prove that they are equal using their derivatives. We will start differentiating the left side with respect to θ.

$$\begin{aligned}
&\frac{\partial}{\partial\theta}\sum_{k=x}^{n}\binom{n}{k}\theta^{k}(1-\theta)^{n-k}\\
&\qquad\qquad= \sum_{k=x}^{n}\binom{n}{k}(k\theta^{k-1}(1-\theta)^{n-k}-(n-k)\theta^{k}(1-\theta)^{n-k-1}\\
&\qquad\qquad= \binom{n}{x}x\theta^{x-1}(1-\theta)^{n-x}-\binom{n}{x}(n-x)\theta^{x}(1-\theta)^{n-x-1}\\
&\qquad\qquad+\binom{n}{x+1}(x+1)\theta^{x}(1-\theta)^{n-x-1}-\binom{n}{x+1}(n-x-1)\theta^{x+1}(1-\theta)^{n-x-2}\\
&\qquad\qquad+\binom{n}{x+2}(x+2)\theta^{x+1}(1-\theta)^{n-x-2}-\binom{n}{x+2}(n-x-2)\theta^{x+2}(1-\theta)^{n-x-3}+\cdots
\end{aligned}$$

Using that

$$\binom{n}{x+1}=\binom{n}{x}\frac{n-x}{x+1}$$

simplifies our equation a lot, since we can rewrite the equation to

$$\begin{aligned}
&\frac{\partial}{\partial\theta}\sum_{k=x}^{n}\binom{n}{k}\theta^{k}(1-\theta)^{n-k}\\
&\qquad= \binom{n}{x}x\theta^{x-1}(1-\theta)^{n-x}-\binom{n}{x}(n-x)\theta^{x}(1-\theta)^{n-x-1}\\
&\qquad+\binom{n}{x}(n-x)\theta^{x}(1-\theta)^{n-x-1}-\binom{n}{x+1}(n-x-1)\theta^{x+1}(1-\theta)^{n-x-2}\\
&\qquad+\binom{n}{x+1}(n-x-1)\theta^{x+1}(1-\theta)^{n-x-2}-\binom{n}{x+2}(n-x-2)\theta^{x+2}(1-\theta)^{n-x-3}+\cdots,
\end{aligned}$$

and now it is easy to recognize it as a telescopic series. After the first term, every consecutive term cancel each other out, thus

$$\frac{\partial}{\partial\theta}\ \sum_{k=x}^{n}\binom{n}{k}\theta^{k}(1-\theta)^{n-k}=x\binom{n}{x}\theta^{x-1}(1-\theta)^{n-x}.$$

The other side is a lot easier to differentiate, using the fundamental theorem of calculus we get

$$\frac{\partial}{\partial\theta}\ x\binom{n}{x}\int_{0}^{\theta}t^{x-1}(1-t)^{n-x}dt=x\binom{n}{x}\theta^{x-1}(1-\theta)^{n-x}.$$

Now we observe that

$$\frac{\partial}{\partial\theta}P(X\geq x)=\frac{\partial}{\partial\theta}P(Y\leq\theta).$$

The only thing left to prove is that the constant term appearing after integrating both sides are equal. Looking at Equations (4.3) and (4.2), when $\theta=0$ it is easy to see that both equals 0, which concludes our proof. □

From Casella and Berger (2002, p. 225), the next result follows

Theorem 3. *If $X \sim F_{p,q}$, then* $\frac{p}{q}\frac{X}{1+(p/q)X} \sim \text{beta}\left(\frac{p}{2}, \frac{q}{2}\right)$.

Here $X \sim F_{p,q}$ denotes that X follows a F-distribution with p and q degrees of freedom.

Proof. Let $X \sim F_{p,q}$ with probability density function (PDF) $f_X(x)$, and

$$Y = \frac{p}{q}\frac{X}{1+(p/q)\,X}.$$

We want to find the PDF of Y in order to confirm its distribution. To do so we start with finding the cumulative distribution function (CDF).

$$\begin{aligned} P(Y \leq y) &= P\left(\frac{p}{q}\frac{X}{1+(p/q)\,X} \leq y\right) = P\left(\frac{1}{q/(pX)+1} \leq y\right) \\ &= P\left(\frac{q}{pX} + 1 \geq \frac{1}{y}\right) = P\left(\frac{q}{pX} \geq \frac{1}{y} - 1\right) = P\left(X \leq \frac{q}{p/y - p}\right) \end{aligned}$$

Differentiating the CDF using the chain rule leaves us with

$$\frac{d}{dy}P(Y \leq y) = f_X\left(\frac{q}{p/y-p}\right) \cdot \frac{d}{dy}\left(\frac{q}{p/y-p}\right) = f_X\left(\frac{q}{p/y-p}\right) \cdot \frac{q}{p(y-1)^2}.$$

The last step is expanding the equation and see if it coincide with the PDF of the beta distributing with the wanted parameters. We obtain that

$$\begin{aligned} \frac{d}{dy}P(Y \leq y) &= \frac{\Gamma\left(\frac{p+q}{2}\right)}{\Gamma\left(\frac{p}{2}\right)\Gamma\left(\frac{q}{2}\right)}\left(\frac{p}{q}\right)^{p/2} \frac{\left(\frac{q}{p/y-p}\right)^{p/2-1}}{\left(1+\frac{1}{1/y-1}\right)^{(p+q)/2}} \frac{q}{p(y-1)^2} \\ &= \frac{\Gamma\left(\frac{p+q}{2}\right)}{\Gamma\left(\frac{p}{2}\right)\Gamma\left(\frac{q}{2}\right)}\left(\frac{p}{q}\right)^{p/2-1} \frac{\left(\frac{q}{p/y-p}\right)^{p/2-1}}{\left(1+\frac{1}{1/y-1}\right)^{(p+q)/2}} \frac{1}{(y-1)^2} \\ &= \frac{\Gamma\left(\frac{p+q}{2}\right)}{\Gamma\left(\frac{p}{2}\right)\Gamma\left(\frac{q}{2}\right)}\left(\frac{y}{1-y}\right)^{p/2-1} \frac{1}{\left(1+\frac{1}{1/y-1}\right)^{(p+q)/2}} \frac{1}{(y-1)^2} \\ &= \frac{\Gamma\left(\frac{p+q}{2}\right)}{\Gamma\left(\frac{p}{2}\right)\Gamma\left(\frac{q}{2}\right)} y^{p/2-1}\left(\frac{1}{1-y}\right)^{p/2-1} (1-y)^{(p+q)/2} \frac{1}{(y-1)^2} \\ &= \frac{\Gamma\left(\frac{p+q}{2}\right)}{\Gamma\left(\frac{p}{2}\right)\Gamma\left(\frac{q}{2}\right)} y^{p/2-1}\left(\frac{1}{1-y}\right)^{p/2-1} (1-y)^{p/2-1}(1-y)^{q/2-1} \\ &= \frac{\Gamma\left(\frac{p+q}{2}\right)}{\Gamma\left(\frac{p}{2}\right)\Gamma\left(\frac{q}{2}\right)} y^{p/2-1}(1-y)^{q/2-1}. \end{aligned}$$

We recognize this as the PDF of the beta distribution with parameters $p/2$ and $q/2$. □

4.2 Proof of the Clopper–Pearson Method

Having the necessary theorems presented, it is time to start the proof by inverting an equal-tailed binomial test. Consider the hypothesis test for the binomial proportion

$$\text{H}_0 : \pi = \pi_0 \;\text{ and }\; \text{H}_1 : \pi \neq \pi_0,$$

with significance level α. Let $X \sim \text{binomial}(n, \pi_0)$ be the test statistic, with the present value x. We have the p-value

$$p(x, \pi_0) = \min\{2\min\{P_{\pi_0}(X \leq x), P_{\pi_0}(X \geq x)\}, 1\}.$$

We prove that the set consisting of all π_0 such that H_0 is not rejected,

$$C(x) = \{\pi_0 \mid p(x, \pi_0) > \alpha\},$$

is a $1-\alpha$-confidence set. The probability of committing a Type I error, that is, rejecting H_0 when H_0 is true, is less than or equal to the significance level α. In our case, assuming H_0 to be true and therefore $\pi = \pi_0$, means that

$$P_{\pi_0}(\pi_0 \notin C(X)) \leq \alpha.$$

Equivalently,

$$P_{\pi_0}(\pi_0 \in C(X)) \geq 1-\alpha,$$

which concludes the proof that $C(x)$ is a $1-\alpha$-confidence set, it is in fact the Clopper–Pearson interval. Finally, we want to calculate the lower and upper bounds, p_L and p_U, for $C(x)$. From $p(x, \pi_0)$ we get the following equations to solve

$$P_{p_L}(X \geq x) = \sum_{k=x}^{n} \binom{n}{k} p_L^k (1-p_L)^{n-k} = \frac{\alpha}{2} \quad \text{and} \quad P_{p_U}(X \leq x) = \sum_{k=0}^{x} \binom{n}{k} p_U^k (1-p_U)^{n-k} = \frac{\alpha}{2},$$

with respect to p_L and p_U, with $p_L = 0$ when $x = 0$, and $p_U = 1$ when for $x = n$ (Thulin, 2014, p. 819; Agresti and Coull, 1998, p. 119). The goal is to show that the interval derived from these equations is Equation (4.1).

First we are starting with the lower bound. From from Theorem 2 we know that

$$P(X \geq x) = P(Y_1 \leq p_L),$$

where $X \sim \text{binomial}(n, p_L)$ and $Y_1 \sim \text{beta}(x, n-x+1)$. Thus, the equation to solve is

$$P(Y_1 \leq p_L) = \frac{\alpha}{2}. \tag{4.4}$$

Now we wish to describe this equation using the F-distribution rather than the beta distribution. This can be done with the help of Theorem 3. Define variables w, v and random variable S such that

$$Y_1 = \frac{v}{w}\frac{S}{1+(v/w)S} \sim \text{beta}\left(\frac{v}{2}, \frac{w}{2}\right).$$

From Theorem 3 we know that $S \sim F_{v,w}$. When solving $P(Y_1 \leq p_L)$ with respect to S we get

$$P(Y_1 \leq p_L) = P\left(\frac{v}{w}\frac{S}{1+(v/w)S} \leq p_L\right) = P\left(S \leq \frac{w}{v/p_L - v}\right),$$

and now we can see how it coincides with Equation (4.4). We can calculate the parameters by using the definition of Y_1,

$$\frac{v}{2} = x \text{ and } \frac{w}{2} = n-x+1,$$

hence, the parameters are

$$v = 2x \text{ and } w = 2(n-x+1).$$

When inserting the parameters we get

$$P\left(S \leq \frac{w}{v/p_L - v}\right) = P\left(S \leq \frac{2(n-x+1)}{2x/p_L - 2x}\right).$$

With the new information we can formulate Equation (4.4) as

$$P(Y_1 \leq p_L) = P\left(S \leq \frac{2(n-x+1)}{2x/p_L - 2x}\right) = \frac{\alpha}{2}$$

$$P\left(S > \frac{2(n-x+1)}{2x/p_L - 2x}\right) = 1 - \frac{\alpha}{2}.$$

In order to solve the equation, we have to look up the appropriate F-value.

$$\begin{aligned}
\frac{2(n-x+1)}{2x/p_L - 2x} &= F_{2x,2(n-x+1),(1-\alpha/2)} \\
\frac{2x}{p_L} - 2x &= \frac{2(n-x+1)}{F_{2x,2(n-x+1),(1-\alpha/2)}} \\
\frac{1}{p_L} &= \frac{n-x+1}{xF_{2x,2(n-x+1),(1-\alpha/2)}} + 1 \\
\frac{1}{p_L} &= \frac{n-x+1+xF_{2x,2(n-x+1),(1-\alpha/2)}}{xF_{2x,2(n-x+1),(1-\alpha/2)}} \\
p_L &= \frac{xF_{2x,2(n-x+1),(1-\alpha/2)}}{n-x+1+xF_{2x,2(n-x+1),(1-\alpha/2)}} \\
p_L &= \frac{1}{\frac{n-x+1}{xF_{2x,2(n-x+1),(1-\alpha/2)}} + 1}
\end{aligned}$$

Since

$$F_{v,w,\alpha} = \frac{1}{F_{w,v,(1-\alpha)}}, \tag{4.5}$$

we can deduce that

$$p_L = \frac{1}{\frac{n-x+1}{x}F_{2(n-x+1),2x,\alpha/2} + 1}.$$

This coincides with the lower bound of the Clopper–Pearson interval, which concludes the first part of our proof.

Now we do the same with the upper bound of the interval. From Theorem 2 we have

$$P(X \leq x) = (n-x)\binom{n}{n-x}\int_0^{1-p_U} t^{n-x-1}(1-t)^x dt = P(Y_2 \leq 1 - p_U),$$

where $X \sim \text{binomial}(n, p_U)$ and $Y_2 \sim \text{beta}(n-x, x+1)$. Thus, the equation to solve is

$$P(Y_2 \leq 1 - p_U) = \alpha/2.$$

In the same manner as before, we define new variables w, v and random variable S such that

$$Y_2 = \frac{v}{w}\frac{S}{1+(v/w)S} \sim \text{beta}\left(\frac{v}{2}, \frac{w}{2}\right).$$

Consequently, we have

$$P(Y_2 \leq 1 - p_U) = P\left(S \leq \frac{w}{v/(1-p_U) - v}\right) = P\left(S \leq \frac{w(1-p_U)}{vp_U}\right),$$

together with the parameters

$$v = 2(n-x) \text{ and } w = 2(x+1).$$

Our equation is

$$P\left(S \leq \frac{w(1-p_U)}{vp_U}\right) = P\left(S \leq \frac{(x+1)(1-p_U)}{(n-x)p_U}\right) = \frac{\alpha}{2}$$

$$P\left(S > \frac{(x+1)(1-p_U)}{(n-x)p_U}\right) = 1 - \frac{\alpha}{2}.$$

In terms of a quantile of the F-distribution,

$$\frac{(x+1)(1-p_U)}{(n-x)p_U} = F_{2(n-x),2(x+1),1-\alpha/2}$$

$$x+1 = xp_U + p_U + F_{2(n-x),2(x+1),1-\alpha/2}(n-x)p_U$$

$$p_U = \frac{x+1}{x+1+F_{2(n-x),2(x+1),1-\alpha/2}(n-x)}$$

$$p_U = \frac{\frac{x+1}{(n-x)F_{2(n-x),2(x+1),1-\alpha/2}}}{\frac{x+1}{(n-x)F_{2(n-x),2(x+1),1-\alpha/2}}+1}$$

Once more, with Equation (4.5), we get the desired upper bound

$$p_U = \frac{\frac{x+1}{n-x}F_{2(x+1),2(n-x),\alpha/2}}{\frac{x+1}{n-x}F_{2(x+1),2(n-x),\alpha/2}+1}.$$

This concludes the proof, that that the interval produced by the Clopper–Pearson method has bounds as shown in Equation (4.1).

5 The Blaker Method

We have now discussed two asymptotic and one exact method for computing confidence intervals. All methods have different issues, the asymptotic methods does not guarantee the wanted coverage, while the exact method introduced by Clopper and Pearson is quite conservative. The Blaker method is an exact method that creates intervals which are subsets of Clopper–Pearson intervals (Klaschka and Reiczigel, 2021, pp. 1779–1780). Thus, the intervals are narrower with lower actual coverage, while still being above the confidence level.

Remember that when solving

$$P_{p_L}(X \geq x) = \frac{\alpha}{2} \quad \text{and} \quad P_{p_U}(X \leq x) = \frac{\alpha}{2},$$

with respect to p_L and p_U, we get the lower and upper bounds of the Clopper–Pearson interval, respectively. This condition is stronger than necessary when the aim is to get an exact confidence interval for the binomial proportion. By replacing this condition with

$$P_{p_L}(X \geq x) + P_{p_U}(X \leq x) = \alpha, \tag{5.1}$$

Blaker (2000) found a method that generally results in narrower exact intervals. In addition, it satisfies a nesting condition: $\alpha < \alpha'$ implies that the $1-\alpha'$-confidence interval is included in the $1-\alpha$-confidence interval (Blaker, 2000, p. 783). Just to clarify, Clopper–Pearson intervals are also nested, but there are exact methods using Condition (5.1) which are not (Thulin, 2014, p. 821).

Let's take a look at how the interval is defined, and prove that the actual coverage is indeed above the confidence level. In this proof we will consider a confidence set $C(x)$, which can easily be made into a confidence interval. Similarly to what we did with the Clopper–Pearson interval, the proof will be done by inverting a hypothesis test. However, instead of a one-tail test applied to both tails, this is a both-tails test. This means that the p-value involves both the tails at the same time, instead of treating them separately, which often results in less conservative exact intervals (Vos and Hudson, 2008, p. 81). We want to prove that $C(x)$ satisfies

$$P(\pi \in C(X)) \geq 1-\alpha.$$

Assume X is as described, and let x be our observation. Consider the hypothesis test

$$\mathrm{H}_0 : \pi = \pi_0 \ \text{ against } \ \mathrm{H}_1 : \pi \neq \pi_0,$$

with significance level α, and the test statistic T having value

$$t = \min\{P_{\pi_0}(X \leq x), P_{\pi_0}(X \geq x)\}$$

when the present value of X is x. The p-value is then

$$p(x, \pi_0) = P(T \leq t),$$

and H_0 is rejected if $p(x,\pi_0) < \alpha$.

Note that

$$p(x,\pi_0) = \begin{cases} P_{\pi_0}(X \geq x) + P_{\pi_0}(X \leq x^*), & \text{if } P_{\pi_0}(X \geq x) \leq P_{\pi_0}(X \leq x) \\ P_{\pi_0}(X \leq x) + P_{\pi_0}(X \geq x^{**}), & \text{if } P_{\pi_0}(X \geq x) > P_{\pi_0}(X \leq x) \end{cases}$$

where x^* is the largest u such that $P_{\pi_0}(X \leq u) \leq P_{\pi_0}(X \geq x)$, and x^{**} is the smallest v such that $P_{\pi_0}(X \geq v) \leq P_{\pi_0}(X \leq x)$ (Blaker, 2000, p. 785).

From the proof of the Clopper–Pearson method, Section 4.2, we know that

$$C(x) = \{\pi_0 \mid p(x,\pi_0) > \alpha\},$$

is a $1-\alpha$-confidence set, which concludes our proof that this $C(x)$ is a $1-\alpha$-confidence set. The Blaker interval is the smallest interval containing $C(x)$ (Blaker, 2000, p. 786). The bounds for this interval will not be calculated, since the p-value involves both tails simultaneously, which makes the calculation a lot harder. However, it is possible to determine the interval by directly checking whether the values of π_0 falls within it.

Vos and Hudson (2008) points out some problems that may arise when using methods that comes from inverting both-tails tests, instead of equal-tailed tests. These problems arise because the function for the p-value are neither continuous or bimonotonic in the parameter value, which results in undesirable behavior in the confidence intervals when the sample size increases. They present different examples for problematic behavior when it comes to the Blaker method. For instance, there are occurrences where two samples have the same estimated binomial proportion, but the larger sample has a higher p-value (Vos and Hudson, 2008, pp. 84–88). The behavior of the p-value function leads to the confidence sets not being intervals (Klaschka, 2010; Vos and Hudson, 2008, pp. 85–86), and caused problems with the algorithm presented by Blaker (2000). The algorithm was later fixed by Klaschka (2010).

6 Comparison

Up until this point, we have briefly discussed the coverage and length of the intervals, but we will now study them quantitatively and qualitatively. The goal is to compare the performance of the suggested methods in order to understand which one would be preferable in practise. The ideal interval is narrow, since this indicates higher precision, with actual coverage above and close to the confidence level (Agresti and Coull, 1998, pp. 120–122).

6.1 Actual Coverage

The most essential difference highlighted between the asymptotic and exact methods are their coverage probabilities; asymptotic methods can have actual coverage below the wanted level, in contrast to exact methods. Define $I_n(k,p)$ such that $I_n(k,p) = 1$ if p is contained in the interval when $X = k$, and $I_n(k,p) = 0$ when p is not contained in the interval. Then

$$C_n(p) = \sum_{k=0}^{n} I_n(k,p) P_p(X = k) \tag{6.1}$$

is the actual coverage at value p (Agresti and Coull, 1998, p. 120).

Figures 3–7 displays the actual coverage $C_n(p)$ for every method except the continuity corrected Wald (CCW) method, using four different values for n. In this way it is possible to see if the methods in question performs better with certain values of n. Figure 8 shows the actual coverage of the Wald methods, with and without continuity correction, together. The reason for omitting the CCW method in the first plots is to keep the plots cleaner and easier to read, and making it easier to compare the CCW and the Wald method. In all the plots, the confidence level is at 95%, which is shown as a black dotted line. It might be beneficial to look at Table 1, while examining the plots, to see the mean coverage for each method.

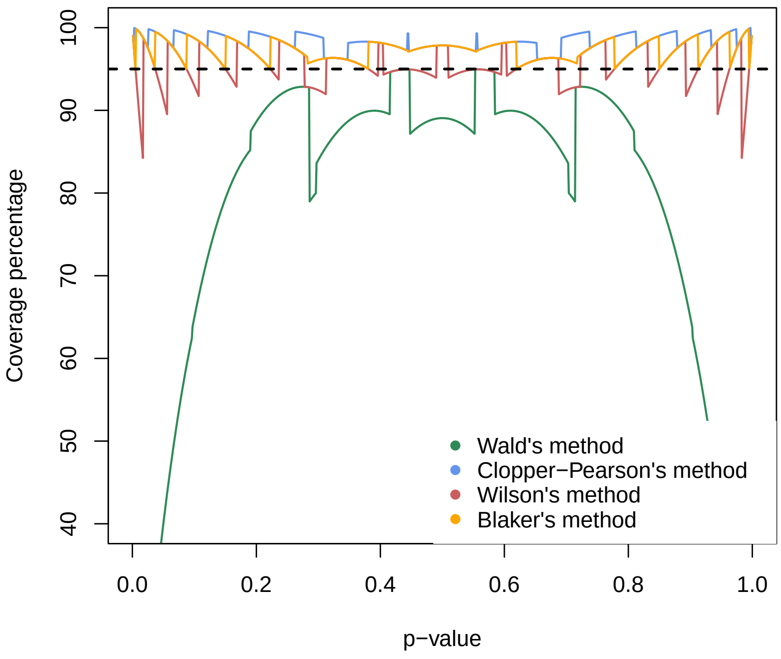


Figure 3: The actual coverage provided by the different methods, with $n = 10$.

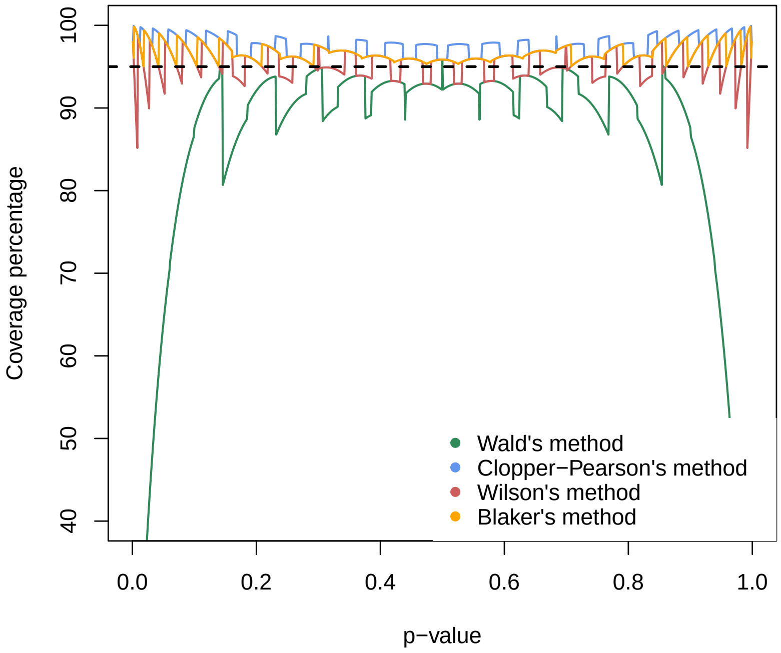


Figure 4: The actual coverage provided by the different methods, with $n = 20$.

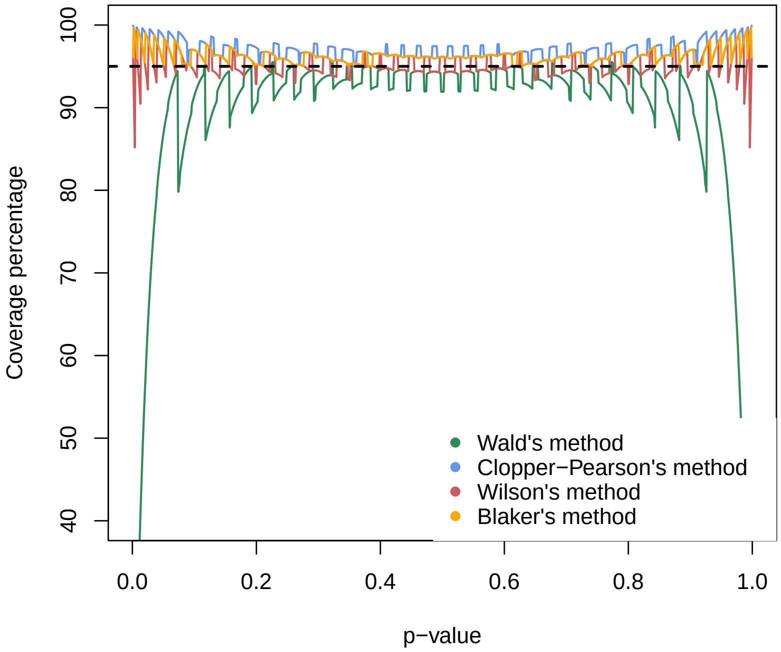


Figure 5: The actual coverage provided by the different methods, with $n = 40$.

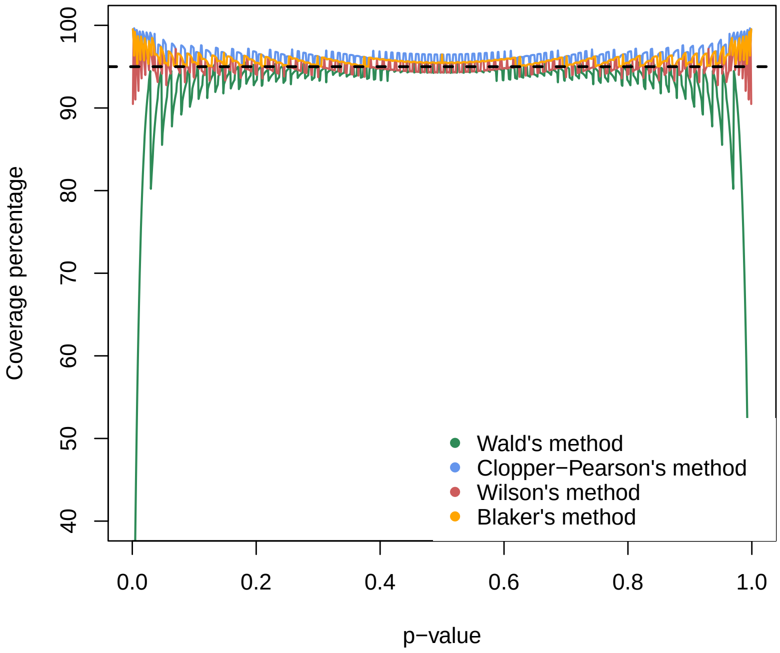
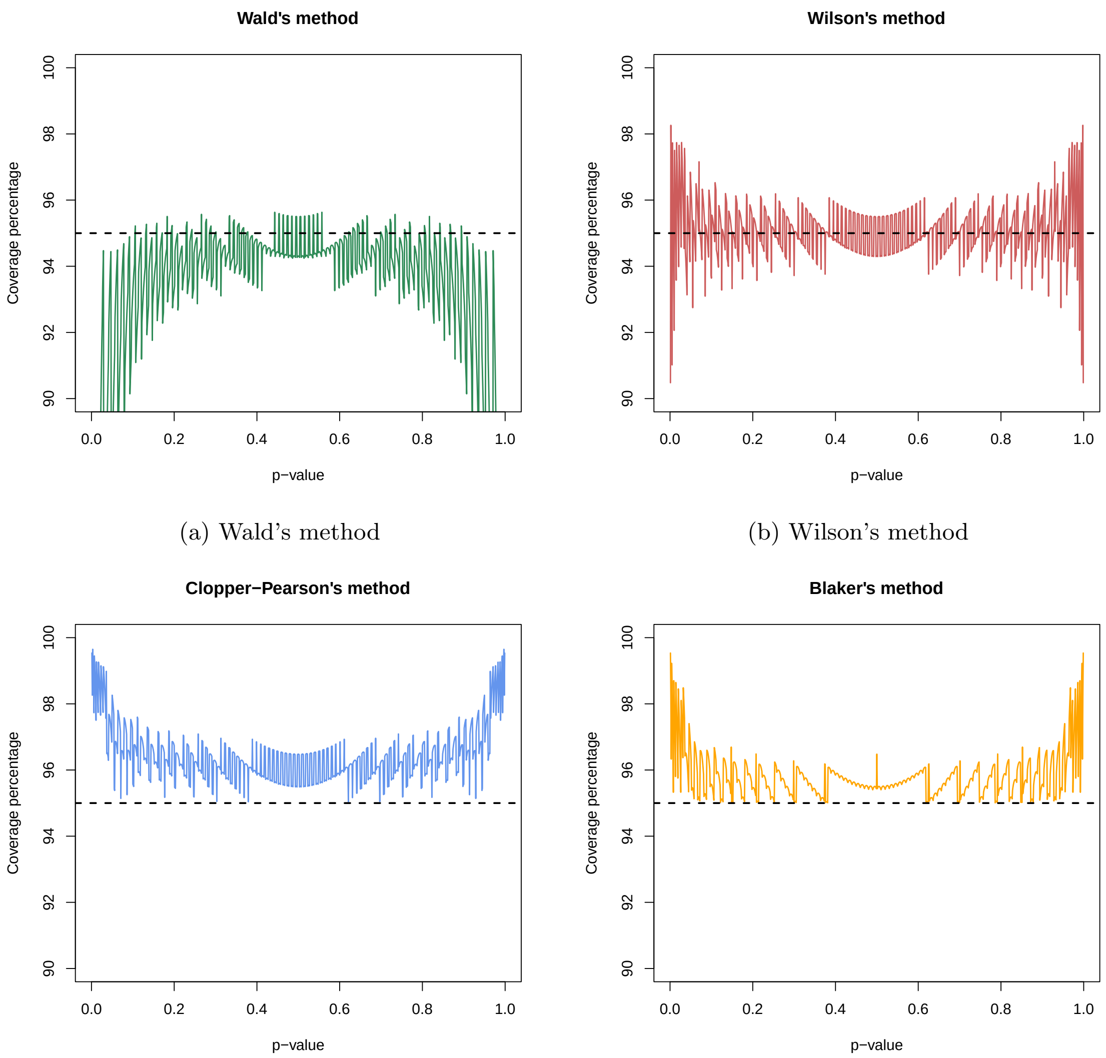


Figure 6: The actual coverage provided by the different methods, with $n = 100$.

For clarity, the four graphs of Figure 6 are separated in Figure 7.



(a) Wald’s method

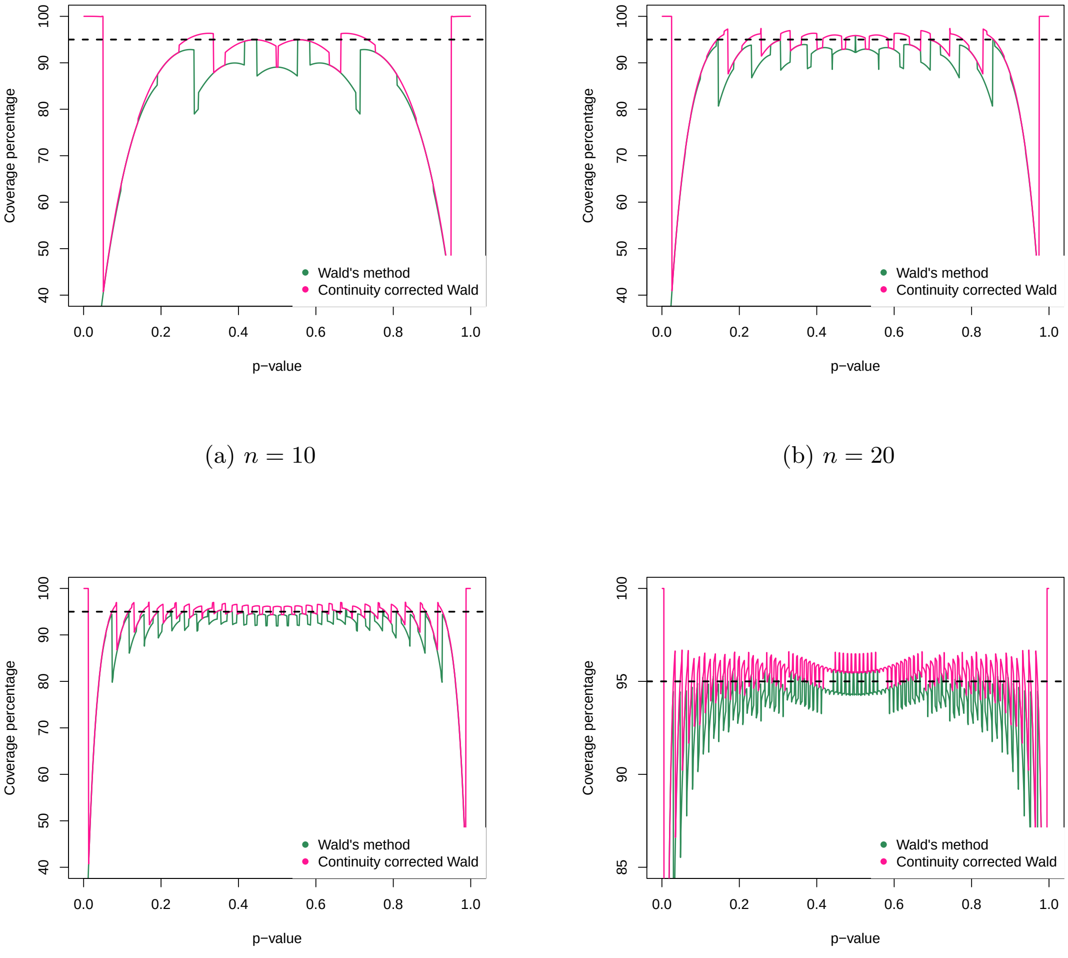
(b) Wilson’s method

(c) Clopper–Pearson’s method

(d) Blaker’s method

Figure 7: The actual coverage for each method in separate graphs, $n = 100$.

Figure 8 shows the actual coverage of the Wald and CCW method, be aware of the y-axis in Figure 8d which is scaled differently.



(a) $n = 10$

(b) $n = 20$

(c) $n = 40$

(d) $n = 100$

Figure 8: The actual coverage for the Wald methods, with and without continuity correction, for different values of n.

Each method will be discussed in the same order as before, starting with the Wald methods, with and without continuity correction, then following with Wilson, Clopper–Pearson and lastly the Blaker method. In addition to the presentation of the actual coverage for each method in Figures 3–8, the average coverage is displayed in Table 1.

From Figures 3 and 4, it is obvious the Wald method is not the one to choose for the considered n values. The plots shows that the actual coverage is almost method constantly lower than the confidence level at 95%, frequently under 90%. It is not expected that the actual coverage is above the confidence level at all times using an asymptotic method, but it is expected that it is frequently above, and does not fall too far below. To be precise, an asymptotic should have average coverage above the confidence level, which is not the case for this method, as can be seen in Table 1. Having said that, keep in mind that this is including the values of p such that none of the conditions in Section 2.3 are satisfied. When considering this list of conditions, the situation where $n = 10$, close to none of the conditions are satisfied, regardless of p. However, if $n = 20$ we can observe that some conditions hold for certain p values, for instance, when $p \geq 0.25$ and $p \leq 0.75$, condition 1 holds. In these cases it is still obvious from Figure 4 that the actual coverage provided can be far below the confidence level, and that the average coverage is not above the confidence level.

As expected, the coverage for the Wald method improves a lot when increasing n, looking both at the Figures 5 and 7a. For $n = 100$ and p close to 0.5 it seems to have an average coverage around the confidence level. Nevertheless, even in those situations, it does show an erratic behavior and it depends on a "lucky" p for the method to produce an interval with the wanted coverage. In fact, the actual coverage is rarely above the confidence level. Much of the same can be observed in Table 1, that the average coverage does improve with larger n, but it is below the confidence level for all n. Thus, according to these plots and the table, it is quite understandable that many authors do not consider this method reliable, even for large values of n.

From Figure 8 we see that the CCW method shows improvements compared to Wald when considering the coverage probability, which is expected considering the CCW intervals contains the Wald intervals. The graphs behaves similarly for all n, and even with the slight improvements, this method is still not applicable when $n \leq 10$. This method can considered be somewhat useful when $n \geq 20$ and $0.3 < p < 0.7$, since the actual coverage seems to be above 0.9, but this is obviously not what we expect from an 95% confidence interval. It is more reliable for larger n, and when only considering the actual coverage it is obviously superior over the Wald method.

Moving on to the Wilson method, the last asymptotic method. Figures 3–7 show vast improvements with this method compared to both the CCW and the Wald method. In Table 1 we see that the average coverage is above and close to the confidence level for all n, which is what we want to see. However, when looking at the plots we observe that the coverage seems to oscillate around the confidence level; consequently, knowing if it will provide actual coverage above the confidence level in a given situation is not easy. Increasing n seems to improve this model, and if p is closer to 0.5 it is more reliable. When p is close to 0 or 1 the provided coverage can be quite far from the confidence level, and consequently, it might not be the best method in these situations, but compared to the other asymptotic methods, it will likely perform better either way.

The Clopper–Pearson method, on the other hand, has coverage above the confidence level for all n and p, which is what we would expect. From Table 1 and Figures 3–7, we observe that the average coverage gets closer to the confidence level when increasing n, and that it is frequently far above the confidence level. The best intervals seems to be when $0.2 < p < 0.8$, and like the previous methods, it is clear that for p closer to 0 or 1, it is less useful. However, rather than the coverage dropping far below the confidence level like the asymptotic methods seems to do, the actual coverage increase and approach 100%.

The actual coverage for the Blaker method is likewise above the confidence level for all n and p. When comparing the two exact methods in Table 1 and Figures 3–6 we see that the Blaker method has actual coverage closer to the confidence level, which is preferable. Other than that, the behavior to the actual coverage is quite similar to the Clopper–Pearson method, with the coverage approaching 100% when p gets closer to 0 or 1, and it seems to provide good coverage when $0.2 < p < 0.8$. From these plots alone it seems that the Blaker method is the superior exact method, compared to the Clopper–Pearson method.

Figure 9 shows simulated coverage probabilities. This is used in order to see if the average coverage coincides with the actual coverage we have calculated. In this simulation $n = 20$, and for each p, 5000 samples are drawn from the population, and then the average of the coverage provided is computed. We see that the plot looks quite similar to Figure 4, as it should.

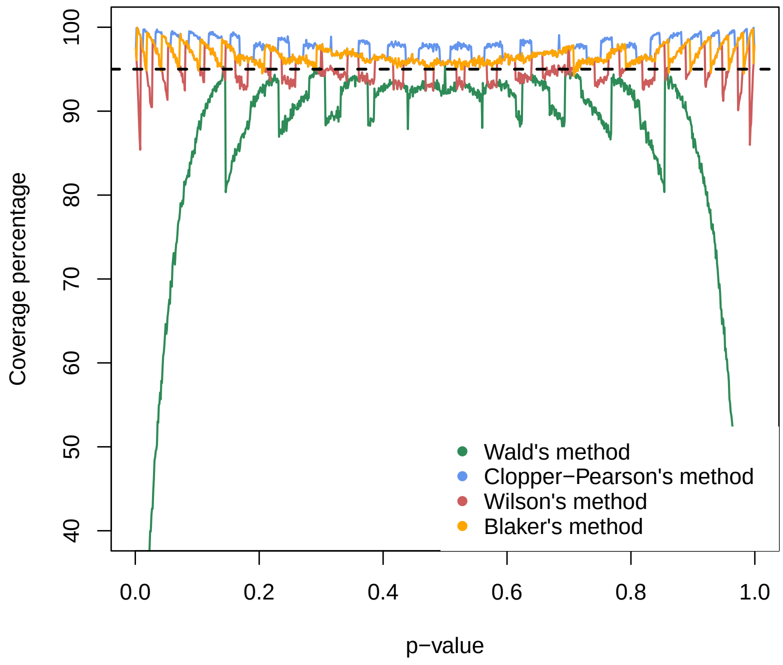


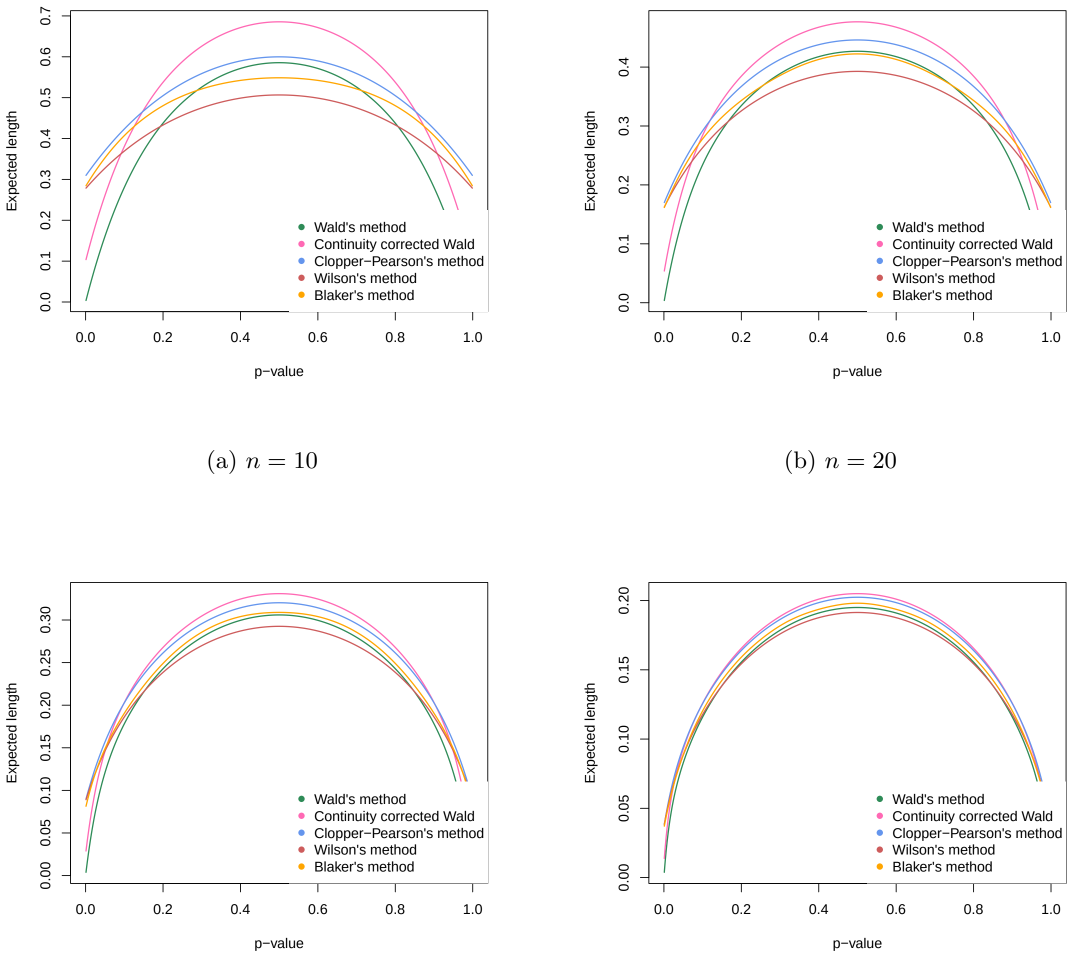
Figure 9: The simulated coverage provided by the different methods, for the given value $n = 20$.

6.2 Expected Length

When comparing the methods, another place of interest is the expected length of the interval. Let $l_n(k)$ and $u_n(k)$ denote the upper and lower confidence limits, respectively. Then

$$L_n(p) = \sum_{k=0}^{n} [u_n(k) - l_n(k)] P_p(X = k), \tag{6.2}$$

is the expected length of the interval (Vos and Hudson, 2005, p. 138).



(a) $n = 10$

(b) $n = 20$

(c) $n = 40$

(d) $n = 100$

Figure 10: The expected length for different values of n.

Table 1 summarizes all the plots, using the means.

	n			
Method	10	20	40	100
Wald	0.7699 (0.4282)	0.8467 (0.3243)	0.8915 (0.2367)	0.9231 (0.1524)
CCW	0.8730 (0.5282)	0.9095 (0.3743)	0.9302 (0.2617)	0.9451 (0.1624)
Wilson	0.9542 (0.4356)	0.9532 (0.3255)	0.9521 (0.2369)	0.9510 (0.1524)
Clopper–Pearson	0.9838 (0.5087)	0.9760 (0.3663)	0.9710 (0.2586)	0.9644 (0.1616)
Blaker	0.9733 (0.4762)	0.9664 (0.3450)	0.9626 (0.2479)	0.9581 (0.1568)

Table 1: Mean values of the actual coverage, and the expected length in parenthesis.

Starting with the Wald method, it looks like the intervals are shorter for p closer to 0 or 1, regardless of n. Recall that these situations are where this method does not produce reliable results, and where the actual coverage is usually far below the confidence level, see Figure 7a. Since we are not interested in producing such unreliable intervals, the length of these intervals is not very interesting. Therefore the interesting part of these plots is where the Wald method can be somewhat useful, when p is closer to 0.5 and $n \geq 20$. Here we observe that the Wald interval has expected lengths similar to Blaker, narrower than Clopper–Pearson, and wider than the Wilson intervals.

Remember that the length of the intervals by the CCW method is always $1/n$ wider than with Wald, as can be seen in Figure 10. Although this method creates intervals that have better coverage, and are more reliable for lower values of n, it is still not useful for $n = 10$, as can be observed in Figure 8a. Just like the Wald method, this method is more reliable for p close to 0.5 and $n \geq 20$, and in these situations, this method has the greatest expected length, which indicates low precision. As a consequence it might be preferable to choose another method.

The Wilson method seems to have narrower intervals than the other methods. Although from Table 1 it seems like Wald may have the shortest intervals, we observe in Figure 10 that for the values of p such that the Wald intervals are somewhat useful, Wilson intervals are shorter. Even though this is true for all the shown n, increasing n is lessening the difference for all the methods. Nevertheless, this does again indicate that the Wilson method is superior in comparison to the Wald method, both with and without continuity correction. In addition, we see that the Wilson intervals are shorter than the intervals produced by both of the exact methods, for all n and p.

Considering the excessive coverage of Clopper–Pearson intervals, it is not surprising that it has wide intervals, for all n and p, see Figure 10 and Table 1; the only method producing wider interval is the CCW intervals. It might be more interesting comparing this method with the Blaker method, since they are both exact methods, and as we know, Blaker intervals are contained in Clopper–Pearson intervals (Klaschka and Reiczigel, 2021, pp. 1779–1780). We do observe that Clopper–Pearson intervals are on average wider, with the difference decreasing when n is increasing. Hence, when considering the length of the intervals, the Blaker method would be the preferred exact method.

7 Discussion

The first method introduced, the original Wald method is the method that has the lowest coverage probability of the studied methods. This method is almost always introduced in introductory statistical courses, and while it is known that the actual coverage can be below the confidence level (Brown et al., 2001, p. 103), it was quite surprising how poor the performance actually was. The coverage is often way below the confidence level, and at the same time it struggles with zero-width intervals and overshooting. It is easy to calculate, and for that reason it is often presented in different statistical courses, but it is not a method anyone should rely on in any real life situations.

As expected, the coverage probability provided by the CCW method is better for all p and n, and in addition, it does not struggle with zero-width intervals, however, the problem with overshooting occurs far more often. In the situations when CCW is applicable, when $n \geq 20$ and $0.3 < p < 0.7$, the expected length is greater than all the other methods, which is not very favourable. All things considered, both the Wald and the CCW method did not seem like reliable methods, even for large values of n.

The last asymptotic method, the Wilson method, does not struggle with zero-width intervals or overshooting. It may not be as easy to calculate by hand, but it is still only a quadratic equation. The coverage probability is better than the other asymptotic methods, with the average being above the confidence level for all tested values of n. In addition, it has the narrowest intervals of all the methods, if we exclude the intervals where the Wald methods, with and without continuity correction, should not be used. Since the Wilson method is asymptotic, it should not be used if it is absolutely necessary for the coverage to be above the confidence level. However, in situations where an asymptotic method is considered good enough, this is a far better choice than the other asymptotic methods considered in this thesis.

For the situations where the actual coverage percentage needs to be above the confidence level, it is necessary to use an exact method. In this thesis I have discussed two methods, the Clopper–Pearson method, and the Blaker method. The Clopper–Pearson method is known for being conservative (Brown et al., 2001, p. 113), as we see in both the actual coverage, which is often far greater than the confidence level, and in the lengths of the interval, which is wider than most of the other proposed intervals.

When it comes to the Blaker method, we see that the intervals produced are narrower, and that the actual coverage is closer to the confidence level than with the Clopper–Pearson method; thus, the Blaker method is less conservative. There are, however, problems regarding this method as well; for example, the confidence set created by inverting the hypothesis test is not actually an interval, and some of the properties of the p-value function results in unexpected behavior from the created confidence intervals (Vos and Hudson, 2008, p. 81). Thus, this method is not objectively preferable compared to the Clopper–Pearson method. The Clopper–Pearson method leads to well behaved conservative intervals, while the Blaker method gives less conservative intervals, but can have undesirable behavior.

Bibliography

Agresti, Alan (2002). *Categorical Data Analysis.* 2nd ed. Wiley Series in Probability and Statistics. Wiley-Interscience. ISBN: 0-471-36093-7.

Agresti, Alan and Brent A. Coull (1998). 'Approximate is Better than "Exact" for Interval Estimation of Binomial Proportions'. In: *The American Statistician* 52.2, pp. 119–126. DOI: 10.1080/00031305.1998.10480550. URL: https://doi.org/10.1080/00031305.1998.10480550.

Blaker, Helge (2000). 'Confidence curves and improved exact confidence intervals for discrete distributions'. In: *Canadian Journal of Statistics* 28.4, pp. 783–798. DOI: 10.2307/3315916. URL: https://onlinelibrary.wiley.com/doi/abs/10.2307/3315916.

Brown, Lawrence D., T. Tony Cai and Anirban DasGupta (2001). 'Interval Estimation for a Binomial Proportion'. In: *Statistical Science* 16.2, pp. 101–133. DOI: 10.1214/ss/1009213286. URL: https://doi.org/10.1214/ss/1009213286.

Casella, G. and R.L. Berger (2002). *Statistical Inference.* Duxbury advanced series in statistics and decision sciences. Thomson Learning. ISBN: 9780534243128.

Clopper, C. J. and E. S. Pearson (1934). 'The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial.' In: *Biometrika* 26.4, pp. 404–413. DOI: 10.2307/2331986. URL: https://doi.org/10.2307/2331986.

Keener, R.W. (2010). *Theoretical Statistics: Topics for a Core Course.* Springer Texts in Statistics. Springer New York. ISBN: 9780387938394. URL: https://books.google.co.in/books?id=aVJmcega44cC.

Klaschka, Jan (2010). *On calculation of Blaker's binomial confidence limits.* http://www2.cs.cas.cz/~klaschka/c10/417_ext.pdf.

Klaschka, Jan and Jenő Reiczigel (2021). 'On matching confidence intervals and tests for some discrete distributions: methodological and computational aspects'. In: *Computational Statistics* 36.3, pp. 1775–1790. DOI: 10.1007/s00180-020-00986-0. URL: https://doi.org/10.1007/s00180-020-00986-0.

Laplace, Pierre Simon Marquis de (1812). *Théorie analytique des probabilités.* V. Courcier. ISBN: 0-471-36093-7. URL: https://archive.org/details/thorieanalytiqu00laplgoog/page/n15/mode/2up.

Larsen, R.J. and M.L. Marx (2012). *An Introduction to Mathematical Statistics and Its Applications.* Prentice Hall. ISBN: 9780321693945. URL: https://books.google.no/books?id=tZdbRAAACAAJ.

Newcombe, Robert G. (2013). *Confidence Intervals for Proportions and Related Measures of Effect Size.* Chapman & Hall/CRC Biostatistics Series. CRC Press. ISBN: 9781439812792.

Reed III, James F. (2007). 'Better Binomial Confidence Intervals'. In: *Journal of Modern Applied Statistical Methods* 6.15 (1), pp. 153–161. DOI: 10.22237/jmasm/1177992840. URL: https://doi.org/10.22237/jmasm/1177992840.

Thulin, Måns (2014). 'The cost of using exact confidence intervals for a binomial proportion'. In: *Electronic Journal of Statistics* 8.1, pp. 817–840. DOI: 10.1214/14-EJS909. URL: https://doi.org/10.1214/14-EJS909.

Vos, Paul W. and Suzanne Hudson (2005). 'Evaluation Criteria for Discrete Confidence Intervals: Beyond Coverage and Length'. In: *The American Statistician* 59.2, pp. 137–142. ISSN: 00031305. URL: http://www.jstor.org/stable/27643646.

— (2008). 'Problems with binomial two-sided tests and the associated confidence intervals'. In: *Australian & New Zealand Journal of Statistics* 50 (1), pp. 81–89. DOI: 10.1111/j.1467-842X.2007.00501.x. URL: https://onlinelibrary.wiley.com/doi/epdf/10.1111/j.1467-842X.2007.00501.x.

Wilson, Edwin B. (1927). 'Probable Inference, the Law of Succession, and Statistical Inference'. In: *Journal of the American Statistical Association* 22.158, pp. 209–212. DOI: 10.2307/2276774. URL: https://doi.org/10.2307/2276774.

NTNU

Norwegian University of
Science and Technology